



UNIVERSIDAD DE GRANADA



Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación



Departamento de Arquitectura y Tecnología de Computadores

TESIS DOCTORAL

**PROGRAMA DE DOCTORADO EN TECNOLOGÍAS  
DE LA INFORMACIÓN Y LA COMUNICACIÓN**

***“Mecanismos de seguridad para Big Data basados en  
circuitos criptográficos”***

**Autor:**

Ilia Blokhin

**Directores:**

Prof. Dr. Antonio Francisco Díaz García

Prof. Dr. Julio Ortega Lopera

Granada, 17 de julio de 2020

Editor: Universidad de Granada. Tesis Doctorales  
Autor: Ilia Blokhin  
ISBN: 978-84-1306-606-6  
URI: <http://hdl.handle.net/10481/63614>



## ***Agradecimientos***

A mis directores de tesis, D. Antonio F. Díaz García, D. Julio O. Lopera y también coordinador del doctorado D. Héctor E. Pomares Cintas, por su apoyo e inestimable ayuda durante el tiempo empleado.

También deseo expresar mi más profunda gratitud a todos los que han estado próximos a mí durante la realización de esta memoria, como mi familia, mis amigos y, a todos los que me han ayudado.

***Gracias a todos!***





# Índice

<i>Índice</i> .....	7
<i>Resumen</i> .....	19
<b>1. Introducción, objetivos y estructura de la Tesis</b> .....	<b>21</b>
1.1 ¿Qué es Big Data? .....	22
1.2 ¿Por qué los datos se han vuelto tan grandes? .....	22
1.3 Fuentes de información. ....	22
1.4 Datos de actualización rápida.....	24
1.5 Usando la base de datos.....	24
1.6 Big Data para al estado. ....	24
1.7 Big Data en la salud, la ciencia, el transporte. ....	25
1.8 Big Data en el comercio.....	27
1.9 Big Data en los servicios públicos. ....	29
1.10 Big Data en las compañías de seguros.....	29
1.11 Big data y medios de comunicación. ....	30
1.12 Big Data en diferentes países. ....	30
<b>2. El problema de la seguridad en Big data</b> .....	<b>37</b>
2.1 Los métodos de análisis de Big Data. ....	37
2.2 Sistemas de análisis del Big Data. ....	39
2.3 Programas de IBM para el análisis de Big Data. ....	40
2.4 Adaptación de las tecnologías.....	41
2.5 Lo que afecta a la ejecución de Big Data.....	41
2.6 Cómo procesar de Big Data.....	42
2.7 El mercado mundial de tecnologías en Big Data. ....	43
2.8 Los jugadores líderes en Big Data. ....	43
2.9 Protección de datos, privacidad y seguridad.....	44
2.10 El éxito de Big Data. ....	45
2.11 El retraso tecnológico. ....	46
<b>3. Descripción de algoritmos utilizados</b> . ....	<b>49</b>
3.1 Modernización de los equipos para el procesamiento de Big Data. ....	49

3.2 Descripción general de los métodos de protección de datos. ....	50
3.3 Cifrado.....	51
3.4 Hash.....	52
3.5 MAC. ....	53
3.6 Las firmas digitales. ....	54
3.7 Generación de números aleatorios.....	55
3.8 Fuentes de Fuentes de números aleatorios.....	55
3.9 Los estándares IEEE ....	56
3.10 GOST P 34.10-2001. ....	57
3.11 GOST P 34.11-2012. ....	59
3.12 Mecanismos de autenticación ....	60
<b>4. Almacenamiento en Big Data, descripción de Hadoop. ....</b>	<b>65</b>
4.1 Hadoop.....	65
4.2 Hadoop y seguridad en Big Data. ....	67
4.3 Solución de problemas de seguridad en Hadoop.....	69
4.4 Soluciones de aplicación Sentry en un entorno con HDFS, Hive y Impala. ....	69
4.5 Otros aspectos de la seguridad en el entorno Hadoop.....	71
4.6 Proyecto Rhino.....	71
4.7 Apache Knox Gateway.....	72
4.8 Tokens de Autenticación. ....	73
4.9 Formatos de tokens. ....	75
4.9.1 SAML estándar. ....	76
4.9.2 Estándar WS-Trust y WS-Federation.....	77
4.9.3 Estándar OAuth y OpenID Connect. ....	78
4.10 Autenticación del cliente en Hadoop.....	79
4.11 Tecnología de tokens delegado.....	81
4.12 Ventajas de tokens.....	82
4.13 Las características exclusivas de tokens.....	83
<b>5. Experimentos.....</b>	<b>85</b>
5.1 La estructura de Hadoop.....	85
5.2 Distribuciones de Hadoop.....	86
5.3 Cloudera Hadoop. ....	86
5.3.1 Componentes de Cloudera Hadoop.....	86

5.3.2 Requisitos de hardware. ....	86
<b>5.4 Instalación y configuración.....</b>	<b>87</b>
5.4.1 La instalación del Administrador de Cloudera. ....	87
5.4.2 Agregar espejo Cloudera y instalar los paquetes necesarios: .....	88
<b>5.5 Instalación de Cluster Cloudera Hadoop.....</b>	<b>88</b>
<b>5.6 La recopilación de datos a través Flume.....</b>	<b>95</b>
5.6.1 Sobre el proyecto Flume. ....	96
5.6.2 Arquitectura de Flume. ....	96
5.6.3 La estructura del flujo. ....	96
5.6.4 Fiabilidad y Control de errores. ....	98
<b>5.7 Instalación de Flume a través de Cloudera Administrador.....</b>	<b>98</b>
5.7.1 Configuración del agente de Flume. ....	102
5.7.2 La estructura del archivo de configuración.....	103
5.7.3 Configuración de fuente de Syslog UDP. ....	103
5.7.4 Configuración del cliente.....	104
<b>5.8 Pig.....</b>	<b>105</b>
<b>5.9 Experimento.....</b>	<b>107</b>
<b>5.10 CryptoARM Linux (Trusted eSign) ver. 0.1.0.....</b>	<b>113</b>
5.10.1 Introducción.....	113
<b>5.11 CriptoPro CSP. ....</b>	<b>113</b>
<b>5.12 El mecanismo para Instalacion Cripto Pro.....</b>	<b>115</b>
<b>5.13 El mecanismo para instalacion CryptoARM Linux (Trusted eSign sin GUI).....</b>	<b>117</b>
<b>5.14 Los datos obtenidos durante el experimento.....</b>	<b>117</b>
<b>5.15 Apache Ranger.....</b>	<b>122</b>
5.15.1 Cómo funciona Apache Ranger? .....	123
5.15.2 Ventajas de Apache Ranger. ....	124
5.15.3 Desventajas de Apache Ranger.....	124
<b>5.16 Desarrollo de software para cifrado de datos con algoritmo GOST-28147.....</b>	<b>124</b>
5.16.1 Sobre la programación desarrollada .....	124
5.16.2 Guía como compilar el programacion.....	125
<b>5.17 Los datos obtenidos durante el experimento:.....</b>	<b>126</b>
<b>5.18 Conclusiones: .....</b>	<b>131</b>
<b>5.19 Los experimentos con Azure.....</b>	<b>132</b>
<b>5.20 Los experimentos con Excel:.....</b>	<b>135</b>
<b>5.21 Comparación entre Azure y ordenador. ....</b>	<b>139</b>
<b>5.22 Conclusiones: .....</b>	<b>140</b>
<b>5.23 El experimento con maquina virtual de AWS Amazon.....</b>	<b>140</b>

5.24 Los experimentos con Amazon AWS.....	146
5.25 Conclusiones.....	151
<b>6. Sistemas de autenticación para Hadoop. ....</b>	<b>153</b>
6.1 eToken .....	153
6.2 ECDSA.....	154
6.3 ECDH.....	155
6.4 Como evitar el uso del eToken de forma remota.....	155
6.5 API eHTSecurity.....	155
6.6 GOST 28147-89.....	158
6.7 El método de autenticación del usuario en Cloudera Hadoop con GOST. ....	163
6.8 El esquema de autenticación usuarios en Hadoop con certificado GOST. ....	164
<b>7. Sistema de autenticación multiprotocolo.....</b>	<b>169</b>
7.1 Introduccion.....	169
7.2 Los sistemas de seguridad .....	171
7.3 Sobre el sistema propuesto. ....	171
7.4 eToken y ESP 32.....	173
7.5 Arquitectura del sistema. ....	176
7.6 Algoritmos Usados .....	178
7.6.1 Funciones hash .....	178
7.6.2 Criptografía de curva elíptica .....	178
7.6.3 Esquema de cifrado integrado de curva elíptica .....	179
7.6.4 Generador de números aleatorios verdaderos .....	180
7.6.5 Combinando Algoritmos .....	180
7.7. Registro de recursos .....	181
7.7.1 Registro del servidor de configuración .....	181
7.7.2 Registro del broker MQTT .....	181
7.7.3 Registro de elementos funcionales .....	182
7.8 Petición de servicio.....	182
7.8.1 El usuario solicita eToken para acceder al servicio .....	182
7.8.2 Etoken solicita acceso al servidor de autenticación .....	183
7.8.3 El servidor de autenticación solicita acceso a la puerta de enlace.....	184
7.8.4 El servidor de puerta de enlace habilita el servicio.....	184
7.8.5 Comunicación a través del portal.....	184
7.9 Autorización Multiservicio .....	185
7.10 Formato de mensaje .....	186
7.11 Definición de reglas de acceso.....	187

<b>7.12</b>	<b>Análisis de seguridad</b>	<b>187</b>
7.12.1	Resistencia en el broker MQTT	187
7.12.2	Resistencia a la manipulación del EToken	188
7.12.3	Resistencia al robo de EToken	188
7.12.4	Resistencia de los ataques MitM	189
7.12.5	Resistencia de los ataques al nodo del cliente	189
7.12.6	Resistencia de los ataques en el servidor de autenticación	190
7.12.7	Resistencia de ataques en el servidor de configuración	190
7.12.8	Resistencia de los ataques que acceden a las puertas de enlace	190
7.12.9	Resistencia de los ataques de denegación de servicio	191
7.12.10	Resistencia de cortes de red	191
7.12.11	Comunicación de bloque selectivo en tiempo real	191
<b>7.13</b>	<b>Análisis de rendimiento</b>	<b>192</b>
7.13.1	Tiempos de ejecución de ECC en EToken	192
7.13.2	Tiempos de funciones ECDSA	192
7.13.3	Tiempos del algoritmo ECIES en diversos clientes y servidores	194
7.13.4	Tiempos de comunicación y de acceso a los servicios	194
7.13.5	Sobrecarga en el ancho de banda y latencia	195
7.13.6	Impacto en las comunicaciones SSH	195
7.13.7	Impacto en el acceso a datos en Hadoop HDFS	196
7.14	Conclusiones	197
<b>8.</b>	<b>Conclusiones</b>	<b>198</b>
8.1	Principales aportaciones y conclusiones	198
8.2	Trabajo futuro	200
	<i>Bibliografía</i>	<b>202</b>
	<i>Acrónimos</i>	<b>214</b>

# Índice de figuras

<b>Figura 1 El universo digital: 50 veces más desde principios de 2010 hasta finales de 2020.</b>	21
<b>Figura 2 El aumento de los volúmenes de datos (a la izquierda) contra el desplazamiento de los medios de almacenamiento analógicos (a la derecha).</b>	23
<b>Figura 3 El tráfico de Big Data.</b>	27
<b>Figura 4 Big Data en Rusia.</b>	31
<b>Figura 5 Uso e interés (sobre 502 empresas entrevistadas).</b>	34
<b>Figura 6 Paisaje de Big Data.</b>	40
<b>Figura 7 Ecosistema de Big Data.</b>	44
<b>Figura 8 Aumentar el interés de unas empresas de Big Data.</b>	47
<b>Figura 9 El mecanismo de cifrado de datos.</b>	52
<b>Figura 10 Función hash.</b>	53
<b>Figura 11 El mecanismo de los algoritmos MAC.</b>	54
<b>Figura 12 El mecanismo de firma digital.</b>	54
<b>Figura 13 Un ejemplo del generador de números aleatorios.</b>	55
<b>Figura 14 Diagrama de bloques de XTS.</b>	56
<b>Figura 15 Demostración modelo MapReduce.</b>	66
<b>Figura 16 Modelo MapReduce.</b>	67
<b>Figura 17 La arquitectura Sentry.</b>	70
<b>Figura 18 La protección del perímetro de Apache Knox.</b>	72
<b>Figura 19 La autenticación del cliente con token "activa".</b>	73
<b>Figura 20 Autenticación "pasiva" mediante la reorientación de las solicitudes del cliente.</b>	74
<b>Figura 21 Respuesta a SP.</b>	76
<b>Figura 22 Respuesta a IP.</b>	77
<b>Figura 23 Recibir datos desde el servidor.</b>	78
<b>Figura 24 Acceso Hadoop RPC.</b>	80
<b>Figura 25 Hadoop acceso navegador web.</b>	81
<b>Figura 26 Una ventana que le pide que seleccione la versión de Cloudera Hadoop.</b>	88
<b>Figura 27 Indicar los hosts.</b>	89
<b>Figura 28 Especificar los hosts.</b>	90
<b>Figura 29 La ventana de selección de repositorio.</b>	90
<b>Figura 30 Los parámetros para el acceso a través de SSH.</b>	91
<b>Figura 31 Inicio del proceso de instalación.</b>	91
<b>Figura 32 La lista información.</b>	92
<b>Figura 33 Los componentes y servicios de Cloudera Hadoop para instalar.</b>	92
<b>Figura 34 Instalar todos los componentes.</b>	93
<b>Figura 35 La creación de una base de datos.</b>	93
<b>Figura 36 Configuración los elementos esta un cluster.</b>	94
<b>Figura 37 Configuración del clúster.</b>	94
<b>Figura 38 Dashboard</b>	95

Figura 39 Interfaz web. ....	95
Figura 40 La estructura del flujo.....	97
Figura 41 Esquema se muestra corrientes combinadas a continuación. ....	97
Figura 42 El esquema de manejo de error.....	98
Figura 43 Instalación de Flume. ....	98
Figura 44 Instalación de Flume. ....	99
Figura 45 Elija Zookeeper-servicio.....	99
Figura 46 Especificar el host del clúster. ....	100
Figura 47 Panel de control. ....	100
Figura 48 La página de servicio.....	101
Figura 49 La página configuración. ....	101
Figura 50 Interfaz web. ....	106
Figura 51 Gestor de archivos.....	110
Figura 52 Gráficos estadísticos.....	111
Figura 53 Gráficos estadísticos.....	112
Figura 54 Estructura CryptoPro.....	114
Figura 55 Cifrado/descifrado con GOST.....	118
Figura 56 Cifrado/descifrado con GOST.....	118
Figura 57 Cifrado/descifrado con GOST.....	118
Figura 58 Cifrado/descifrado con GOST.....	119
Figura 59 Cifrado/descifrado con XTS-AES.....	119
Figura 60 Cifrado/descifrado con XTS-AES.....	119
Figura 61 Cifrado/descifrado con XTS-AES.....	120
Figura 62 Cifrado/descifrado con XTS-AES.....	120
Figura 63 Los componentes de Apache Ranger.....	123
Figura 64 Cambiar la configuración del entorno.....	125
Figura 65 Los ficheros de experimento №1.....	126
Figura 66 Procesamiento de datos de experimento №1.....	126
Figura 67 Los ficheros de experimento №2.....	127
Figura 68 Procesamiento de datos de experimento №2.....	127
Figura 69 Los ficheros de experimento №3.....	128
Figura 70 Procesamiento de datos de experimento №3.....	128
Figura 71 Los ficheros de experimento №4.....	129
Figura 72 Procesamiento de datos de experimento №4.....	129
Figura 73 Configurado HDInsight en Azure.....	132
Figura 74 Configurado HDInsight en Azure.....	133
Figura 75 Configurado HDInsight en Azure.....	133
Figura 76 Configurado HDInsight en Azure.....	134
Figura 77 Configurado HDInsight en Azure.....	134
Figura 78 Iniciando el disco remoto.....	135
Figura 79 Buscar en el fichero en Azure una palabra.....	135
Figura 80 Buscar y cambiar una palabra el fichero en Azur.....	136
Figura 81 Buscar en el fichero en Azure una palabra (sin estructura de datos).....	136
Figura 82 Buscar y cambiar en el fichero un folio sin estructura de datos.....	137



<b>Figura 83</b>	<b>Buscar en el fichero en ordenador una palabra.</b>	137
<b>Figura 84</b>	<b>Buscar en el fichero en ordenador un palabro.</b>	138
<b>Figura 85</b>	<b>Buscar en el fichero un palabro en ordenador.</b>	138
<b>Figura 86</b>	<b>Buscar y cambiar en el fichero un folio sin estructura de datos.</b>	139
<b>Figura 87</b>	<b>Maquina virtual Virtual Box.</b>	140
<b>Figura 88</b>	<b>Pagina web de maquina virtual Amazon AWS.</b>	141
<b>Figura 89</b>	<b>Maquina virtual Amazon AWS.</b>	141
<b>Figura 90</b>	<b>Interface de ZeroTier.</b>	142
<b>Figura 91</b>	<b>Interface del cliente de Bitvise.</b>	142
<b>Figura 92</b>	<b>Interface del servidor de Bitvise.</b>	143
<b>Figura 93</b>	<b>Cifrado de datos.</b>	143
<b>Figura 94</b>	<b>Migración de datos.</b>	144
<b>Figura 95</b>	<b>Fichero cifrado.</b>	144
<b>Figura 96</b>	<b>Fichero descifrado.</b>	145
<b>Figura 97</b>	<b>Fichero descifrado.</b>	145
<b>Figura 98</b>	<b>Experimento №1.</b>	147
<b>Figura 99</b>	<b>Experimento №2.</b>	147
<b>Figura 100</b>	<b>Experimento №3.</b>	148
<b>Figura 101</b>	<b>Experimento №4.</b>	149
<b>Figura 102</b>	<b>Intercambio de datos entre el PC y el ATECC508A.</b>	154
<b>Figura 103</b>	<b>API Server.</b>	156
<b>Figura 104</b>	<b>Comunicacion entre API eSecurity y el ATECC508A.</b>	157
<b>Figura 105</b>	<b>Diagrama del sistema XTS-AES.</b>	158
<b>Figura 106</b>	<b>Diagrama de bloques del algoritmo de transformación criptográfica.</b>	159
<b>Figura 107</b>	<b>El algoritmo de cifrado en el modo de simple sustitución.</b>	161
<b>Figura 108</b>	<b>El modo de cifrado de sustitución simple, tiene la misma forma que en la codificación.</b>	162
<b>Figura 109</b>	<b>Esquema de autenticación usuarios en Hadoop con certificado GOST.</b>	164
<b>Figura 110</b>	<b>Cripto Login.</b>	165
<b>Figura 111</b>	<b>El método de autenticación.</b>	166
<b>Figura 112</b>	<b>Autorizacion en Cloudera Manager.</b>	166
<b>Figura 113</b>	<b>Cloudera Manager.</b>	167
<b>Figura 114</b>	<b>No hay autorización.</b>	167
<b>Figura 115</b>	<b>ESP32.</b>	173
<b>Figura 116</b>	<b>Diagrama de bloques del microcontrolador ESP32.</b>	174
<b>Figura 117</b>	<b>Contactos (PinOut) de ESP32.</b>	175
<b>Figura 118</b>	<b>Elementos principales del sistema de autenticación y comunicación.</b>	177
<b>Figura 119</b>	<b>El broker MQTT comunica todos los elementos.</b>	178
<b>Figura 120</b>	<b>Esquema utilizado para cifrar la clave de sesión.</b>	180
<b>Figura 118</b>	<b>Diagrama MSC entre los elementos del sistema.</b>	185
<b>Figura 122</b>	<b>Autorización multiservicio.</b>	186





## Índice de tablas

<b>Tabla 1</b>	<b>Los datos cifrado/descifrado con GOST y AES-XTS.</b>	120
<b>Tabla 2</b>	<b>Las ventajas /desventajas Criptologin y Apache Ranger.</b>	124
<b>Tabla 3</b>	<b>Cifrado/descifrado experimento №1.</b>	127
<b>Tabla 4</b>	<b>Cifrado/descifrado experimento №2.</b>	127
<b>Tabla 5</b>	<b>Cifrado/decifrado experimento №3.</b>	128
<b>Tabla 6</b>	<b>Cifrado/descifrado experimento №4.</b>	129
<b>Tabla 7</b>	<b>Cifrado/descifrado de experimentos.</b>	130
<b>Tabla 8</b>	<b>Comparación el tiempo de búsqueda entre Azure y Ordenador.</b>	139
<b>Tabla 9</b>	<b>Los datos del experimento №1.</b>	147
<b>Tabla 10</b>	<b>Los datos del experimento №2.</b>	148
<b>Tabla 11</b>	<b>Los datos del experimento №3.</b>	148
<b>Tabla 12</b>	<b>Los datos del experimento №4.</b>	149
<b>Tabla 13</b>	<b>Comparación el tiempo de búsqueda entre Amazon AWS y Ordenador portátil.</b>	149
<b>Tabla 14</b>	<b>El relleno constante C1 (constante) acumulador N6.</b>	160
<b>Tabla 15</b>	<b>El relleno constante C2 (constante) acumulador N5.</b>	160
<b>Tabla 16</b>	<b>Comparación del sistema propuesto con otros esquemas de autenticación.</b>	173
<b>Tabla 17</b>	<b>Comparación del tamaño de clave en RSA Y ECDSA.</b>	179
<b>Tabla 18</b>	<b>Notación de los elementos utilizados.</b>	181
<b>Tabla 19</b>	<b>Tiempos de generación, firma y verificación para diversas curvas elípticas en el ESP32.</b>	192
<b>Tabla 20</b>	<b>Tiempos de ejecución ECC en Intel i7-4980HQ @ 2.80GHz MacOS 10.15.4. .</b>	193
<b>Tabla 21</b>	<b>Tiempos de ejecución ECC en Intel Xeon Silver 4116 CPU @ 2.10GHz Centos 7.7 kernel 3.10.0.</b>	193
<b>Tabla 22</b>	<b>Tiempos de ejecución ECC en AMD EPYC 7571 @ 2.4 GHz RHEL 8.2.</b>	193
<b>Tabla 23</b>	<b>Tiempos de ejecución ECC en Raspberry PI4 ARMv7 Raspbian GNU/Linux 10.</b>	193
<b>Tabla 24</b>	<b>Tiempos de ejecución ECIES en diferentes plataformas clientes y servidores.</b>	194
<b>Tabla 25</b>	<b>Tiempos medios de ejecución de cada etapa.</b>	194
<b>Tabla 26</b>	<b>Tiempos de transferencia en Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, CentOS 7.7.</b>	195
<b>Tabla 27</b>	<b>Tiempos de transferencia en SSH para Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, CentOS 7.7.</b>	196
<b>Tabla 28</b>	<b>Tiempos de transferencia en Hadoop HDFS para lectura y escritura en Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, CentOS 7.7.</b>	197



## Resumen

El incremento en el volumen de datos a procesar ha desencadenado una línea específica de investigación que se engloba bajo el término de “Big Data”.

Conforme se abren nuevas soluciones computacionalmente más complejas, al mismo tiempo aparecen nuevos desafíos tecnológicos, y es lo que ha ocurrido con la seguridad en el procesamiento de dichos flujos masivos de datos.

Las leyes de protección de datos tratan de dar un marco jurídico de defensa de derechos, pero detrás tiene que haber un soporte tecnológico que garantice el acceso sólo a aquellas personas que estén realmente autorizadas.

Resulta vital contar con elementos que garanticen la protección y el almacenamiento seguro de dichos datos, por lo que los mecanismos de autenticación juegan un rol importante. Sistemas basados en usuario y contraseña no son suficientes en un mundo interconectado por Internet, donde la facilidad de acceso global también entraña riesgos por accesos no autorizados.

Continuamente vemos en las noticias como las empresas y organismos sufren ataques donde ponen al descubierto datos sensibles. Es por ello que resulta necesario tomar medidas que puedan hacer frente a posibles ciberataques.

La criptografía de clave pública es una poderosa herramienta matemática que permite crear escenarios seguros. En particular, la criptografía basada en algoritmos de curva elíptica permite

El objetivo principal de esta tesis es el estudio del problema de la seguridad en entornos Big Data y el desarrollo de soluciones basada en sistemas electrónicos que permitan mejorar la seguridad en el acceso a sistemas donde deben procesarse un elevado volumen de datos. Los sistemas propuestos ofrecen una solución eficiente y flexible para aumentar la seguridad en el acceso a servicios y sistemas que pueden procesar gran cantidad de información.



# 1. Introducción, objetivos y estructura de la Tesis.

El concepto de "Big Data" no es nuevo, se originó en los días de mainframe y computación científica relacionada. Como es bien sabido, cálculos complejos han sido siempre compleja y por lo general inextricablemente ligado con la necesidad de procesar grandes cantidades de información.

Sin embargo, sólo el término "Big Data" es relativamente reciente. Tiene una fecha bastante exacta de nacimiento - 3 de septiembre 2008, cuando un número especial de la revista "Nature" [CHE11], dedicada a la búsqueda de la respuesta a la pregunta "¿Cómo puede afectar a la ciencia, el procesamiento de Big Data?". Número especial de la revista resumió el papel de los datos de la ciencia.

Hay varias razones para el interés de Big Data. El volumen de información crece, y se aplica a la mayoría de los datos no estructurados. La reacción del mercado de IT fue inmediata - grandes empresas comenzaron a desarrollar herramientas para trabajar con Big Data, el número de nuevas ideas superó las expectativas de todo concebibles. Con el crecimiento de la potencia informática y el desarrollo de tecnologías de almacenamiento y análisis de Big Data convertido gradualmente a disposición de las pequeñas y medianas empresas. Esto contribuye al desarrollo del modelo de cloud computing.

Una mayor penetración de las IT en el entorno empresarial y la vida cotidiana, el flujo de información a procesar sigue creciendo continuamente. Y si Big Data de hoy - es petabytes, exabytes mañana. Obviamente, más herramientas el futuro para manejar tales cantidades masivas de información seguirá siendo difícil y costoso. El mundo de hoy es desbordante, literalmente, flujos de información.

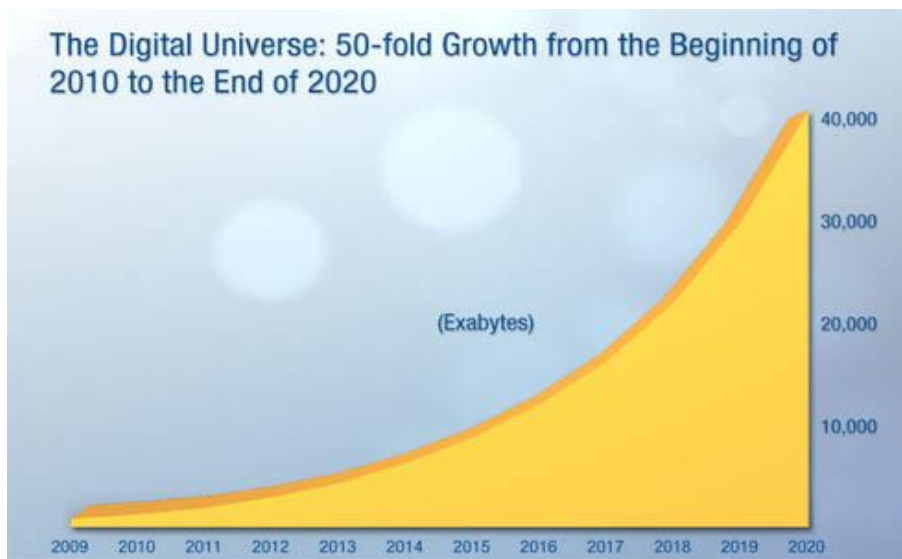


Figura 1 El universo digital: 50 veces más desde principios de 2010 hasta finales de 2020.



La Figura 1 nos muestra un crecimiento exponencial del universo digital desde 2010. Así pues, almacenamiento, manipulación y uso de este tipo de archivos enormes de información – resulta una tarea extremadamente difícil.

## **1.1 ¿Qué es Big Data?**

Desde el nombre se puede suponer que el término "Big Data" se refiere a la gestión y el análisis de Big Data. "Big data" se refiere a los conjuntos de datos cuyo tamaño es más allá de las capacidades de bases de datos típicos enumerados, almacenamiento, gestión y análisis de la información. Y almacén de datos mundial, que seguirá creciendo.

Sin embargo, los "grandes datos" sugieren que no es sólo el análisis de grandes cantidades de información. No es que las organizaciones puedan crear enormes cantidades de datos, pero el hecho de que la mayoría de ellos se presentan en un formato mal estructurado - recursos basadas en la web, vídeos, documentos de texto y etc. Es todo se almacena en diferentes repositorios, a veces incluso fuera organización. Como resultado, la empresa puede tener acceso a un gran volumen de sus datos, pero no puede procesarlos. En la actualidad, los datos se actualizan cada vez con más frecuencia, y las herramientas para analizar esta cantidad de datos que faltaba, y le dio el desarrollo de la tecnología de Big Data.

## **1.2 ¿Por qué los datos se han vuelto tan grandes?**

Existen gran cantidad de fuentes de datos en el mundo moderno. Estos datos de los dispositivos de medición, el flujo de los mensajes de las redes sociales, datos meteorológicos, corrientes de datos sobre la ubicación de los abonados de las redes celulares, audio y dispositivos de grabación de vídeo. La distribución masiva de las tecnologías anteriores y modelos innovadores de la utilización de diversos tipos de dispositivos y servicios de Internet fue el punto de partida para la penetración de grandes volúmenes de datos en casi todas las esferas de la actividad humana. En primer lugar, la investigación y el desarrollo, el sector comercial y la administración pública.

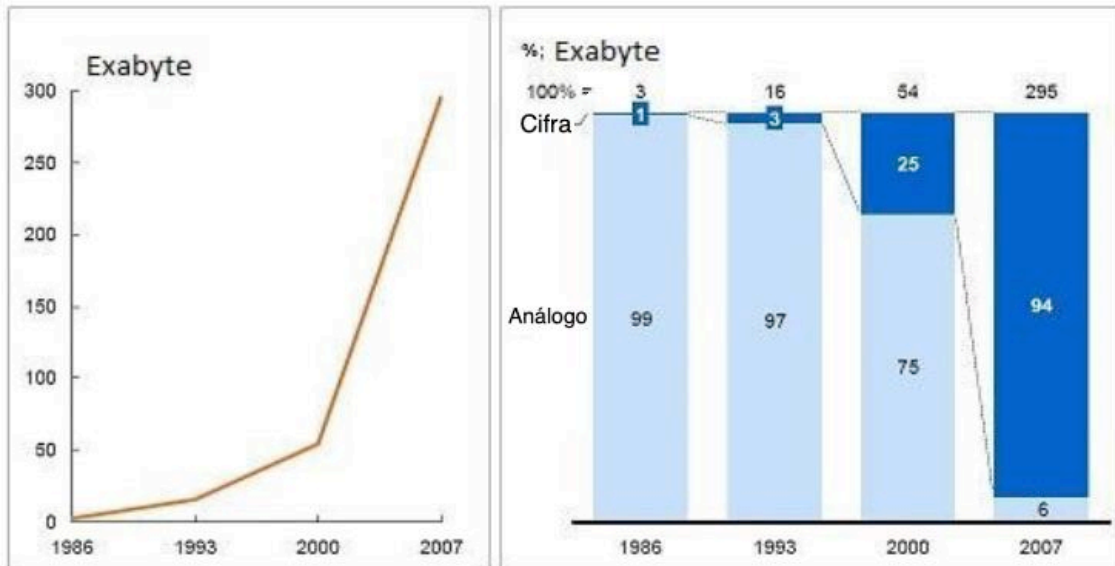
## **1.3 Fuentes de información.**

Las empresas recogen y utilizan muchos tipos diferentes de datos, tanto estructurados como no estructurados. Estas son las fuentes de las que se obtiene los datos encuestados Cisco Connected World Technology Report [PAL12]:

- 74% de la recogida de datos actual;
- 55% de los datos históricos recogidos;
- 48% de los datos se elimina de los monitores y sensores;
- 40 % son datos en tiempo real, y luego hay que borrarlos.

En la mayoría de los datos en tiempo real utilizado en la India (62 %), Estados Unidos (60 %) y Argentina (58 %).

32 % de los encuestados recogen datos no estructurados - como vídeo. En esta área, lo que lleva China, donde los datos no estructurados recogidos por el 56 % de los encuestados.



**Figura 2** El aumento de los volúmenes de datos (a la izquierda) contra el desplazamiento de los medios de almacenamiento analógicos (a la derecha).

Por ejemplo, los sensores montados en motores de aviones, generando 10 terabytes en media hora. Sobre la misma característica de flujo de equipos de perforación y refinerías de petróleo. Servicio de mensajes cortos Twitter, a pesar de la limitada extensión del mensaje en 140 caracteres, lo que genera un flujo de 8 TB día. Si todos estos datos se acumulan para su posterior procesamiento, su importe total se mide en decenas y cientos de petabytes. [SVI12]

El estudio denominado Cisco Connected World Technology Informe, realizado en 18 países de la firma de análisis independiente InsightExpress, se entrevistó a 1.800 estudiantes universitarios y el mismo número de profesionales jóvenes entre las edades de 18 a 30 años. La encuesta se llevó a cabo para determinar el nivel de preparación de los departamentos de TI para implementar proyectos de Big Data.

La mayoría de las empresas recogen, registrar y analizar los datos. Sin embargo, muchas empresas en relación con el Big Data se enfrentan a una serie de cuestiones de negocios y tecnología de información complejos. Por ejemplo, el 60 % de los encuestados admite que las soluciones Big Data pueden mejorar los procesos de toma de decisiones y aumentar la competitividad, pero sólo el 28 por ciento dijo que ya se benefician de la información acumulada.

## **1.4 Datos de actualización rápida.**

La necesidad de una rápida actualización de datos es una de las principales fuentes de tecnologías de Big Data. Los procesos en el almacén de datos son lentos, no son lo suficientemente flexibles y no están satisfechos con el negocio.

Los almacenes de datos son caros, y los administradores de bases de datos calificados tienen que resolver los problemas de síntesis y estructuración de datos. La necesidad de involucrar a los administradores de bases de datos provoca retrasos en el acceso a nuevas Fuentes de datos y la creación de estructuras rígidas que son difíciles de modificar. Los almacenes de datos no tienen la flexibilidad suficiente para satisfacer las necesidades de la mayoría de las organizaciones modernas. En lugar de aumentar la cantidad de datos que tienen a su disposición, las empresas comenzarán a evaluar más su relevancia y a aumentar la rapidez con que obtienen la información requerida.

## **1.5 Usando la base de datos.**

El uso a gran escala de la base de datos es capaz de afectar seriamente a algunos segmentos del mercado de los servicios de información. El volumen de estos servicios en todo el mundo hoy en día se estima por lo menos 100 millones de dólares. Sector financiero. Es necesario resolver los problemas más complejos del algoritmo para la determinación de la renta garantizada, pequeña distorsión de la información podría dar lugar a graves pérdidas económicas. Telecomunicaciones, la publicidad, el comercio minorista, el transporte, la construcción, la producción industrial. Hoy en día, sólo el 0,5% de la información recogida se procesa adecuadamente. La implementación de la base de datos es capaz de cambiar la situación.

Cerca del 65% de las empresas más grandes del mundo para invertir en el estado de la base de datos o planean grandes inversiones para este fin en un futuro próximo. Hace dos años, estas empresas fueron 7% más baja de lo que son ahora. El saldo de la información del costo de operación para cambiar significativamente en el futuro próximo. Se espera que en 2020 el costo de almacenamiento de información en las grandes empresas se reducirá en diez veces, pero el costo de un estudio sobre el posible uso de la base de datos se incrementará en un 40%.

## **1.6 Big Data para al estado.**

El conjunto de información obtenida por las agencias gubernamentales es enorme. Pero hoy en día, la mayoría de los datos (70%) es poco estructurado. Se trata de un gran número de informes, declaraciones, informes, notas sobre políticas, pronósticos, proyecciones, declaraciones, notas explicativas. La gente siempre recurre a varias

agencias del gobierno, y estos recursos son fijos. El uso eficiente de las autoridades gubernamentales Big Data puede asustar a los ciudadanos que ocultan sus ingresos.

Big Data ofrece oportunidades increíbles para las autoridades fiscales. Historia de la creación y liquidación de las empresas de detalles efímeros de la relación entre el gasto y los ingresos de las personas físicas y jurídicas, la organización detallada de los impuestos, todos los hechos cometidos por ciudadanos y personas jurídicas de una amplia variedad de delitos a la disciplinaria - toda esta información estaría disponible a cualquier inspector de Hacienda. Las instituciones gubernamentales sabrán exactamente cómo mover las personas en todo el país y las redes sociales en el extranjero, se conocen los funcionarios: que es con quien estudió, que a quien se casó divorciado, alguien con quien se sentó en la misma mesa. Cualquier mensaje en la red social será objeto de un análisis exhaustivo.

Ya se ha tomado la decisión fundamental en el uso de las agencias gubernamentales Big Data. El siguiente paso debe ser la creación de único espacio de información, que introducirá un acceso centralizado a las autoridades fiscales federales regionales basadas en la tecnología y el establecimiento de información mutua entre todos los departamentos.

El Servicio de Impuestos Federales se está embarcando en un proyecto para implementar el sistema de modelado y los contribuyentes de comportamiento. Las autoridades fiscales tendrán un sistema que permita identificar zonas de alto riesgo de cometer delitos tributarios. Es en se llevarán a cabo estos sectores perspectivas. En la capacidad financiera e institucional vez del estado es inconmensurablemente superior a los recursos similares de cualquier empresa privada, por lo que he indicado anteriormente las dificultades de aplicación Big Data será fácilmente superado por los organismos gubernamentales.

## **1.7 Big Data en la salud, la ciencia, el transporte.**

Cada año, el crecimiento es del 20-40% de las historias clínicas del Hospital [SER14]. Los pequeños hospitales en 2015 generarán 665 terabytes de datos médicos. Los ámbitos de aplicación del análisis de Big datan en la atención médica son numerosos, tanto en investigación como en la práctica. Por ejemplo, el uso del control remoto y monitoreo del paciente para los enfermos crónicos puede reducir el número de visitas al médico, visitas a la sala de emergencia y el número de días en el hospital, para que la ayuda sea más eficaz. El análisis de Big Data que contienen las características del paciente, los resultados del tratamiento y su costo, puede ayudar a determinar los tratamientos más eficaces. Además, el análisis de los patrones globales de la enfermedad con el fin de identificar las tendencias en una etapa temprana es una tarea importante, no sólo en la superación de las crisis en la salud pública, sino también para

proporcionar oportunidades para el sector médico farmacéutico y para simular la futura demanda de sus productos, como base para la toma de decisiones sobre las inversiones en investigación y desarrollo.

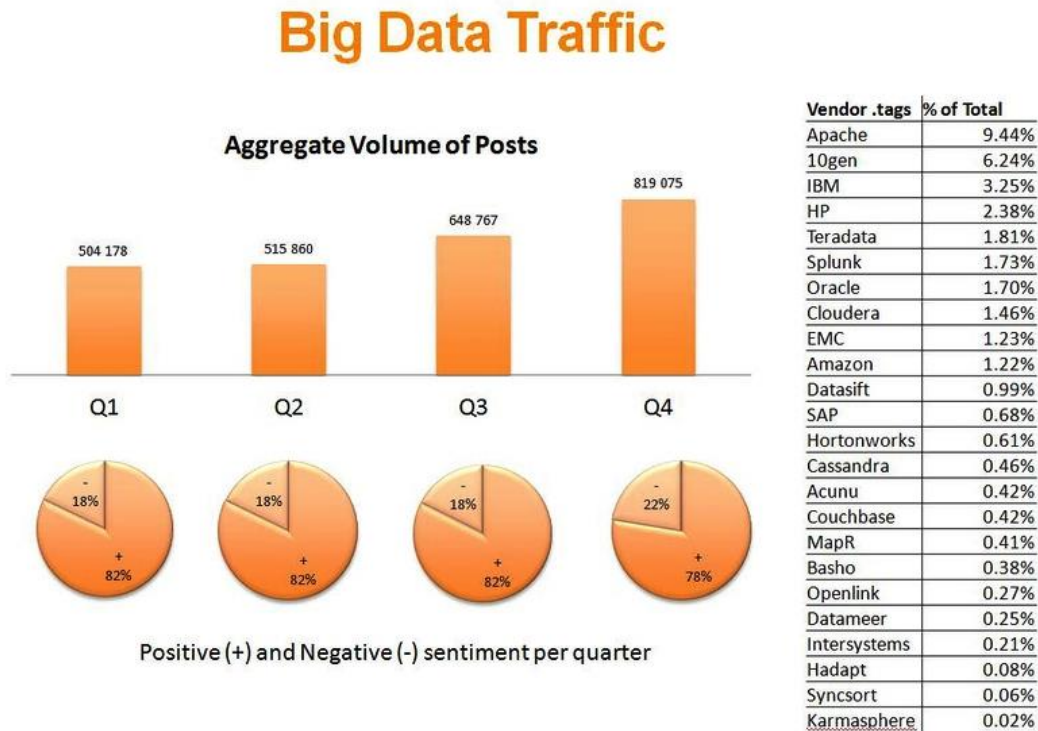
Además, Big Data se utiliza para descubrir los secretos del universo. Organización Europea para la Investigación Nuclear (OEIN) es la sede de uno de los mayores experimentos conocidos en el mundo. Más de 50 años de OEIN opera una corriente creciente de los datos obtenidos en el curso de experimentos para estudiar las partículas elementales y las fuerzas por las que interactúan. El Gran Colisionador de Hadrones tiene 27 kilómetros del anillo de imanes superconductores con un número de unidades para mejorar la energía de aceleración de las partículas que se mueven. El detector dispone de 150 millones de sensores y trabaja como cámara 3D, llevando a cabo el rodaje del momento de la colisión de partículas a una velocidad de 40 millones de marcos por segundo. Debido a la necesidad de almacenamiento, distribución y análisis de hasta 30 petabytes de datos producidos cada año en 2002, se creó la Red Mundial de Área, que se ha convertido en una red global distribuida de centros de datos. Muchos de los datos del OEIN no está estructurado, y que sólo muestran que algo pasó. Científicos de todo el mundo están trabajando juntos en la estructuración, la reconstrucción y el análisis de lo que pasó y por qué.

Las redes móviles pueden ser usados para simular el tráfico. Esto es de particular interés para los países que carecen de otros datos relacionados con el transporte. Por ejemplo, para la planificación de los flujos de tráfico con el fin de reducir la congestión de la empresa Orange, que ofrece acceso a los conjuntos de datos anónimos, que contiene 2500 millones cuentas de llamadas telefónicas y mensajes de texto intercambiados por 5 millones miembro durante 5 meses. Del mismo modo, la compañía ha ayudado a las autoridades de Seúl Korea Telecom para determinar las rutas óptimas de autobuses nocturnos. Como resultado de ello, se añadieron siete líneas de autobuses nocturnos adicional al plan original del tráfico de la ciudad.

En una escala geográfica más amplia de datos móviles contribuyen al análisis de los flujos migratorios y las tendencias son de gran valor para la gestión de las operaciones durante las crisis. La iniciativa "Global Pulse", presentado por la Oficina Ejecutiva del Secretario General de las Naciones Unidas, es una respuesta a la necesidad de más información a tiempo para el seguimiento y monitoreo de los efectos de las crisis socioeconómicas globales y locales.

En el campo de Network Analyst IT ayuda a los proveedores de servicios a optimizar sus medios de red de enrutamiento y predecir los errores y los cuellos de botella de rendimiento antes de que causen daños. Combinando las capacidades de red en tiempo real, y perfiles de clientes completos crea beneficios adicionales, lo que permite ofrecer paquetes de servicios personalizados que mejoren las oportunidades de ingresos, mientras que atraer y retener clientes.

Las discusiones sobre Big Data en la red son muy activas. Además, como se puede ver en los gráficos circulares anteriores, el punto álgido de la discusión solo está creciendo: si en el primer trimestre de 2012 hubo más de 504 mil referencias al término, en el cuarto trimestre ya había más de 800 mil.



*Figura 3 El trafico de Big Data.*

## 1.8 Big Data en el comercio.

Utilizar Big Data para analizar el comportamiento de los clientes, diseñar itinerarios en una sala de operaciones, el derecho de colocar la mercancía, las compras del plan y, en última instancia, para aumentar las ventas. El comercio por internet Big Data se construye un mecanismo de ventas: a los usuarios se les ofrecen productos teniendo en cuenta sus preferencias personales, información sobre la cual se recopila, por ejemplo, en las redes sociales. En ambos casos, el análisis de Big Data ayuda a reducir costos, aumentar la lealtad del cliente y llegar a una gran audiencia. Todas estas son solo características básicas que se pueden implementar con la ayuda de las tecnologías de Big Data.

A pesar de la crisis económica, se espera que aumente el número de proyectos para implementar grandes datos, incluso en el sector minorista. Aunque la introducción de las nuevas tecnologías no sólo amenaza a las ganancias, pero también de alto riesgo, las empresas ya están familiarizados con los éxitos de colegas de negocios más fuertes. En una difícil situación económica a la vanguardia en la necesidad de salvar y mejorar la

lealtad del cliente. Sólo con estos problemas y soluciones están diseñados para manejar de manera grandes de datos.

Para atraer a más clientes están recurriendo cada vez más a las tecnologías innovadoras, como el análisis de datos de Big Data, el comercio electrónico, etc. Por ejemplo, en Corea, que ha abierto recientemente su primera tienda virtual del mundo. Al escanear los códigos QR con los paneles pegados imágenes de diversos productos, los residentes de Seúl pusieron en su carrito de compras virtual de los elementos seleccionados, que a su vez los entrega a casa a una hora conveniente.

El mayor proveedor de Tesco del Reino Unido está experimentando con la realidad aumentada. Para los compradores que se ha desarrollado una aplicación que les permite recibir información sobre el contenido de calorías de ciertos productos y otra información que no es relevante para la etiqueta de precio, solo apuntar la placa de la cámara en la plataforma y tomar una foto.

Otro ejemplo: hace poco la venta online de ropa y calzado eran muy comunes debido a la incapacidad para llevar a cabo la instalación en el espacio virtual. El comprador fue un alto riesgo de cometer un error con el tamaño o el estilo. Pero la situación está cambiando. Pronto la tienda eBay probador virtual en línea esté disponible, lo que permite a los compradores a "probar" su ropa favorita desde el directorio compartido en el modelo tridimensional de su propio cuerpo. Un proyecto similar probador virtual fue presentado por SAP y ha sido elogiado por los expertos. Gracias a las tecnologías, persona puede con la ayuda de sus fotos y los parámetros de entrada para tratar de ordenar la nueva ropa, pagar por ello con un teléfono móvil.

El desarrollo de las tecnologías móviles se puede considerar una de las principales tendencias que afectan al desarrollo comercial. El móvil se ha convertido en la herramienta más importante para el comercio, su importancia no hará sino crecer. 69% de los consumidores ya creen que el móvil es una necesidad para ir de compras y aumenta en gran medida su disfrute del proceso. Nadie puede negar la conveniencia de ordenar desde cualquier lugar y forma de pago a través de Internet o por teléfono, en cualquier forma adecuada. Como resultado de desarrollar activamente el concepto de «Canal Omni» - cuando los canales de venta reales y virtuales combinan en un solo proceso de negocio. Incluso hoy en día, cualquier comprador quiere ser capaz, por ejemplo, para comenzar a comprar en Internet, por lo que es una orden y terminar el pago en la tienda, y viceversa.

Los proveedores están obligados a buscar herramientas que le permiten crear ofertas personalizadas y la dirección para promover bienes. Por ejemplo, Interfaz Amazon.com. Cada vez que visitar el sitio, el cliente recibe una variedad de propuestas sobre la base de un análisis de la historia de compras anteriores, páginas vistas, comentarios de izquierda, etc. Enormes cantidades de los procesos del sistema de información en una fracción de segundo, cada vez que convertirlos para ofrecer dirigidos, lo que lleva finalmente a un aumento en las ventas.

## **1.9 Big Data en los servicios públicos.**

La compañía IDC Energy Insights publicó un informe sobre la disponibilidad de los servicios públicos de Estados Unidos para trabajar con la tecnología Big Data. En el estudio [CNE14], IDC examina el trabajo de 760 empresas estadounidenses, incluyendo 59 empresas con ingresos de más de \$ 500 millones. El objetivo del informe - para ayudar a las empresas a evaluar su disposición a trabajar con las tecnologías de Big Data.

Expertos IDC identifican los criterios fundamentales por los que evaluar la capacidad de la empresa para trabajar con la tecnología Big Data. El informe también contiene recomendaciones para mejorar la situación de las tecnologías de Big Data.

Los autores del informe creen que la voluntad de utilizar la tecnología Big Data se compone de cinco componentes: el deseo, los datos acumulados, las tecnologías de adaptación, procesos y personal racionalizados. El éxito de la empresa en el área de Big Data depende igualmente de la voluntad de la empresa en todas estas áreas.

Según IDC, el sector municipal se encuentra en las primeras etapas de la adopción de tecnología Big Data. Por lo tanto, la voluntad de las dos terceras partes de las empresas a trabajar con Big Data, IDC estima como "promedio". Estimación "Bajo" recibió cuatro veces más empresas que el "alto".

## **1.10 Big Data en las compañías de seguros.**

Las compañías de seguros están interesadas en aplicar la tecnología Big Data, pero sólo unos pocos han comenzado a trabajar activamente en esta dirección. Estos datos llevan a un estudio conjunto de la empresa Bravura Solutions y Consejo de Servicios Financieros. Los investigadores entrevistaron a un número de compañías de seguros líderes en sus planes para la modernización y la introducción de Big Data.

Según la encuesta [SMI14], el 67% de las compañías de seguros creen que tienen un acceso limitado a los datos del usuario. Según los encuestados, estos datos son suficientes para personalizar las interacciones con los clientes, pero no lo suficiente como para predecir su comportamiento.

Sin embargo, más del 56% de los encuestados, es decir, la creación de campañas personalizadas es el objetivo principal.

Alrededor del 30% de esas compañías de seguros encuestadas hoy están utilizando la tecnología y analítica Big Data para anticipar las necesidades del cliente y crear



mensajes personalizados. El principal problema para las empresas que aún no lo hayan hecho, es la ausencia de los sistemas necesarios, según el estudio. Las compañías de seguros tienen matrices de datos, pero no hay posibilidad de sacar el máximo provecho de ellos. La mayoría de las compañías de seguros están interesadas en la modernización de sus sistemas de IT en los próximos cinco años. Sin embargo, el 23,7% de las organizaciones de la cuestión de la modernización no vale la pena.

### **1.11 Big data y medios de comunicación.**

Requisitos para el almacenamiento de grandes volúmenes de datos en la industria de medios y entretenimiento crece resolución aumenta vídeo muy rápidamente. Distribución de HD y el consumo de vídeo móvil para estimular el surgimiento de una fuerte demanda de los contenidos digitales relevantes. En este sentido, se está aumentando la demanda de soluciones de almacenamiento de datos y el disco duro para el archivo de vídeo.

Aumento significativo de uso de unidades flash. La memoria flash es muy importante en la difusión de los datos, los investigadores. En el período 2012-2019 los requisitos de capacidad de almacenamiento digital en la industria del entretenimiento crecerán en 5,6 veces, y los requisitos de almacenamiento de datos que participan de volumen en un año - en 4 veces [SEV15]. Las ganancias de las ventas de sistemas de almacenamiento en la industria del entretenimiento crecerán en más de 1,4 veces en el período 2012-2019 de \$ 5.6 mil millones a \$ 7800 millones. La solución de almacenamiento máximo en el año 2012 se utilizó para la conservación y el archivo de nuevos contenidos (98%).

Los ingresos totales de la venta de vehículos y equipos utilizados en la industria de medios y entretenimiento crecerán en el período 2012-2019 en 1,3 veces a partir de \$ 774 millones a \$ 974 millones.

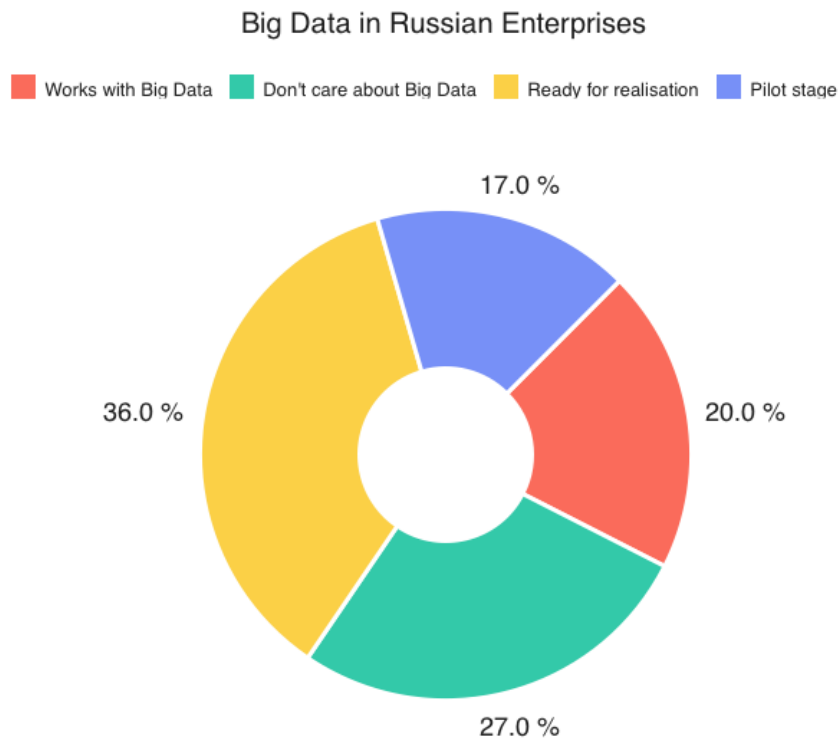
### **1.12 Big Data en diferentes países.**

#### **Rusia:**

En octubre de 2015, EMC publicó una encuesta [FED13] en la que 678 ejecutivos de IT de las empresas rusas compartieron sus puntos de vista sobre qué desafíos y oportunidades, e incluyendo nuevas competencias, que están vinculados con los datos de

Big

Data



***Figura 4 Big Data en Rusia.***

Los expertos rusos dicen que el uso de datos grande conduce a una mejora significativa en los procesos de decisión, el impacto positivo en la competitividad de las empresas y simplifica la gestión de riesgo.

70% de los encuestados en Rusia creen que el análisis de los datos de su empresa va a tomar decisiones más informadas, mientras que el 35% de los encuestados confirman que la alta dirección de la Compañía cree que los resultados del análisis de Big Data en la toma de decisiones fundamentales de negocio.

31% de los encuestados informaron que sus empresas tienen una ventaja competitiva como resultado de la introducción de la tecnología de la información general, mientras que el 51% de los encuestados cree que la industria, que utilizan este tipo de herramientas mostrará el crecimiento más alto.

Más de la mitad (51%) de los encuestados coinciden en que el análisis de las tecnologías de Big Data jugará un papel crucial en la detección y prevención de ataques cibernéticos; puede ser un factor decisivo, ya que sólo el 67% de los encuestados en Rusia confía en que podrán, en caso necesario, para restaurar completamente todos sus datos.

Al mismo tiempo, la encuesta reveló una serie de factores que influyen en la decisión de implementar Big Data analytics en empresas rusas:

25% de las empresas encuestadas actualmente no planea introducir la tecnología de Big Data.

Entre los encuestados que no planean introducir más datos, el 37% dijo que la razón principal de prevenir su introducción, no la relevancia de esta tecnología para el negocio.

Como empresas en Rusia continúan para ver la innovación en la fundación de IT de la ventaja competitiva en los mercados nacionales y extranjeros:

entre las prioridades más comunes para las empresas, estimular la transformación de IT incluye la eficiencia de los procesos de negocios / operaciones (68%), mejorar el servicio al cliente y la interacción con ellos (37%);

76% de los encuestados dice que la inversión en tecnología es un factor de importancia estratégica en la consecución de los objetivos de negocio de la empresa;

71% de los encuestados predicen que, en los próximos tres años, una tarea importante será mantener las habilidades de los profesionales en el nivel correspondiente a la tasa de desarrollo de las tecnologías de IT.

El ascenso y la caída de Hadoop ocurrieron en unos diez años. Hoy, los clientes corporativos en Rusia están más preocupados por las cuestiones - qué tecnología de análisis elegir, dónde ubicar los datos, localmente o en la nube. En muchos casos, esto depende de las leyes de repatriación de datos.

En particular, las distribuciones de Hadoop se volverán moralmente obsoletas, ya que, debido a su alto nivel de complejidad y los dudosos beneficios de las pilas completas de Hadoop, muchas organizaciones rusas las abandonan en favor de alternativas convenientes en la nube de pago por uso y optimización para tareas específicas.

Cloudera ha hecho un buen trabajo al organizar la configuración y la administración de clústeres, pero hoy en día la mayor parte del mundo de los datos está en la nube. Vale la pena considerar cuán atractivo es Hadoop para los negocios en Rusia.

La tendencia de influencia en la nube ha provocado la aparición de interfaces SQL para el almacenamiento de objetos. En el futuro, estas interfaces, que permiten interactuar con los datos en la nube con la sintaxis SQL habitual, se convertirán en la función central de las plataformas de bases de datos en la nube como un servicio y aumentarán la productividad en un plazo de dos a cinco años, ya que la mayoría de los proveedores y desarrolladores de la nube se dirigen ahora a ellas. Los analistas agregan que los repositorios de objetos son adecuados para acomodar grandes cantidades de información de estructura mixta.

Algunas compañías argumentan que los sistemas orientados a cadenas como MySQL y PostgreSQL pueden satisfacer las necesidades de las cargas de trabajo analíticas, así como sus cargas de trabajo transaccionales tradicionales. Ambas ofertas pueden realizar análisis, y si se usan menos de 20 GB de datos, es probable que la escala no valga la pena.

Las empresas que usan datos en miles de millones de filas no podrán aprovechar MySQL y PostgreSQL. No hay nada en ellos que pueda manejar tal carga. El gasto en infraestructura para almacenar conjuntos de datos, incluso durante días, en el almacenamiento orientado a filas, eclipsó los costos de personal.

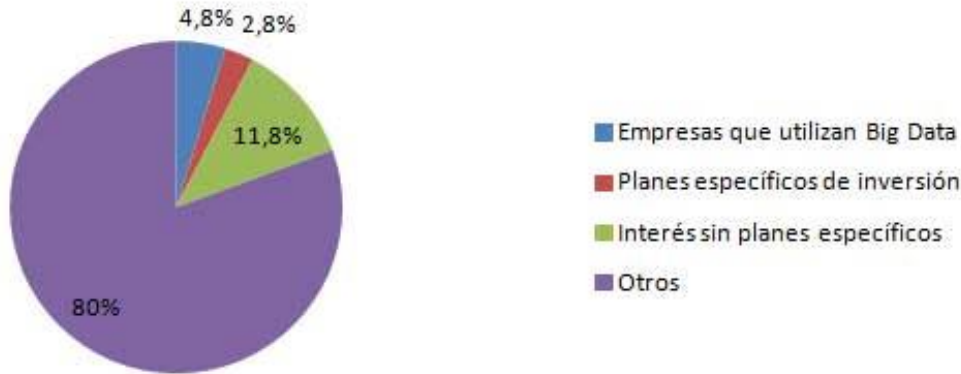
Puede que Cloudera y MapR estén pasando por tiempos difíciles en este momento, pero no hay nada que haga creer que haya algo en AWS EMR, DataBricks y Qubole que pueda competir. Incluso Oracle está lanzando una oferta impulsada por Spark. Sería bueno si la industria viera algo más en Hadoop que solo la oferta de Cloudera.

### **España:**

En julio de 2015 se supo que Madrid utilizará la tecnología para el gran manejo de datos de la infraestructura urbana. El coste del proyecto - € 14,7 millones, la fundación implementa soluciones hacen que la tecnología para analizar y gestionar datos de Big Data. Con su ayuda, la administración de la ciudad va a controlar el funcionamiento de cada proveedor de servicios y, en consecuencia, pagar por ello, dependiendo del nivel de los servicios.

Se trata de la administración de contratistas que siguen a la iluminación de las calles del estado, zonas verdes, llevan a cabo la limpieza y remoción del territorio, así como el reciclaje. El proyecto de los inspectores designados trabajó 300 indicadores clave de rendimiento de los servicios urbanos, con base en qué día será 1500 varias pruebas y mediciones. Además, la ciudad va a empezar a utilizar innovadora plataforma tecnológica denominada Madrid Inteligente - Smart Madrid.

## Uso e Interés (sobre 502 empresas entrevistadas)



*Figura 5 Uso e interés (sobre 502 empresas entrevistadas).*

### **India:**

Mercado de IT de la India está empezando gradualmente para reducir la tasa de desarrollo y la industria tienen que buscar nuevas formas de mantener el crecimiento. Los desarrolladores de software y aplicaciones están comenzando a ofrecer nuevas opciones para el uso de las últimas tecnologías. Así que algunas empresas de la India producen un análisis de la actividad de consumo basado en Big Data, y luego ofrecen los resultados de los estudios principales de compras y redes minoristas.

Analizan las cámaras de circuito cerrado de televisión, informa sobre las compras, consultas en Internet, informes sobre compras realizadas desde la web.

Djiray Rajaram señaló que la mayor parte de este tipo de análisis se hace en los EE.UU., pero ahora, cuando el desarrollo del mercado de IT de la India comenzó a debilitarse, la empresa prestará más atención a este segmento prometedor.

Al mismo tiempo, las empresas indias están trabajando con Big Data a menudo utilizan la tecnología de nube para almacenar y procesar datos y los resultados de sus actividades.

El volumen de los datos globales producidos en 2011 se estima, de acuerdo con Djiray Rajaram, en alrededor de 1.8 mil millones de terabytes, el equivalente a 200 mil millones películas de alta definición de longitud completa.

Además del análisis de la consulta y de procesamiento de resultados de imágenes de

cámaras de vigilancia, un gran espacio para el trabajo puede ser visto en la cantidad de información de los usuarios y los compradores aparece en las redes sociales.

La Asociación Nacional de la India de Software y Servicios de IT (India's National Association of Software and Services Companies) predice un aumento de seis clases de soluciones de segmento para trabajar con grandes volúmenes de datos a 1,2 mil millones de dólares.

Este crecimiento global del Big Data será de más de 2 veces a partir de 8250 millones de hoy a \$ 25 mil millones en los próximos años.



## 2. El problema de la seguridad en Big data.

### 2.1 Los métodos de análisis de Big Data.

Hay muchas metodologías diferentes para el análisis de conjuntos de datos, que se basan en algoritmos de estadística e informática. Esta no es una lista exhaustiva, pero refleja los enfoques más populares en diferentes sectores. Se debe entender que los investigadores continúan trabajando en la creación de nuevas técnicas y mejorar los existentes. Además, algunas de estas técnicas no son necesariamente aplicables únicamente a Big Data pueden ser utilizados con éxito para las matrices de menor volumen. Se analiza la matriz volumétrica, los datos más precisos y relevantes se pueden obtener en la salida.

**A/B testing.** La técnica, que a su vez muestra de control se compara con los demás. Esto hace que sea posible identificar la mejor combinación de rendimiento de lograr, por ejemplo, la mejor respuesta de los consumidores en una propuesta de marketing.

**Association rule learning.** Un conjunto de técnicas para la identificación de las relaciones, es decir, las normas entre las variables en grandes conjuntos de datos. **El uso en data mining.**

**Classification.** Un conjunto de técnicas que nos permite predecir el comportamiento de los consumidores en un segmento particular. **El uso en data mining.**

**Cluster analysis.** El método estadístico para clasificar objetos en grupos mediante la identificación de los signos comunes no es conocido. **El uso en data mining.**

**Crowdsourcing.** Métodos de recogida de datos de un gran número de fuentes.

**Data fusion and data integration.** Un conjunto de técnicas que permite analizar los comentarios de los usuarios de redes sociales, y se compara con los resultados de las ventas en tiempo real.

**Data mining.** Un conjunto de técnicas que le permite identificar los más susceptibles a los productos o servicios promovidos categorías de los consumidores, en particular la identificación de los empleados con más éxito para predecir modelo de comportamiento de los consumidores.

**Machine learning.** La dirección en la informática (inteligencia artificial), que tiene como objetivo crear algoritmos basados en la auto-análisis.



**Natural language processing.** Un conjunto de técnicas informáticas tomados de los derechos de reconocimiento de voz.

**Network analysis.** Un conjunto de técnicas para el análisis de los vínculos entre los nodos de una red. Esto se aplica a los medios sociales permite analizar la relación entre el individuo usuarios, empresas, comunidades, etc.

**Optimization.** Un conjunto de técnicas para el rediseño de los sistemas y procesos complejos para mejorar uno o más indicadores. Le ayuda en la toma de decisiones estratégicas.

**Pattern recognition.** Un conjunto de técnicas con elementos del modelo de auto-estudio para predecir el comportamiento de los consumidores.

**Predictive modeling.** Un conjunto de técnicas que permiten crear un modelo matemático dado el escenario probable. Por ejemplo, el análisis de la base de datos de posibles condiciones que estimulen a los abonados cambiar de proveedor de servicio.

**Regression.** Un conjunto de métodos estadísticos para identificar patrones entre el cambio en la variable independiente y varias variables independientes. A menudo se utiliza para el pronóstico y la predicción.

**Signal processing.** Tomado de el aparato de radio de técnicas que reconoce señal a ruido de fondo y su posterior análisis.

**Spatial analysis.** Análisis de Datos - área Topología, coordenadas geográficas, la geometría de los objetos. La fuente de los datos son las grandes bases de datos geográficos.

**Statistics.** La ciencia de la recolección y organización de datos, incluye el desarrollo de encuestas y experimentos.

**Supervised learning.** Un conjunto de métodos de aprendizaje automático de base tecnológica que pueden detectar las relaciones funcionales en los conjuntos de datos analizados.

**Simulation.** Modelado del comportamiento de los sistemas complejos se utiliza a menudo para predecir, prever y desarrollar diferentes escenarios para la planificación.

**Time series analysis.** Un conjunto de procesamiento digital de señales, análisis repite sobre los datos de series de tiempo. Se utiliza - seguimiento del mercado de valores o la morbilidad de los pacientes.

**Unsupervised learning.** El conjunto de métodos que pueden detectar las relaciones funcionales ocultos en los conjuntos de datos analizados.

**Visualization.** Los métodos de representación gráfica de los resultados del análisis de Big Data en forma de diagramas o imágenes animadas para facilitar la comprensión de los resultados.

## 2.2 Sistemas de análisis del Big Data.

Algunos de estos enfoques permiten realizar en algoritmos para manejar Big Data. Aquí hay un ejemplo de sistemas de análisis de presupuesto de Big Data:

-1010data;

-Apache Chukwa;

-Apache Hadoop;

-Apache Hive;

-Apache Pig;

-Jaspersoft;

-LexisNexis Risk Solutions HPCC Systems;

-MapReduce;

-Revolution Analytics

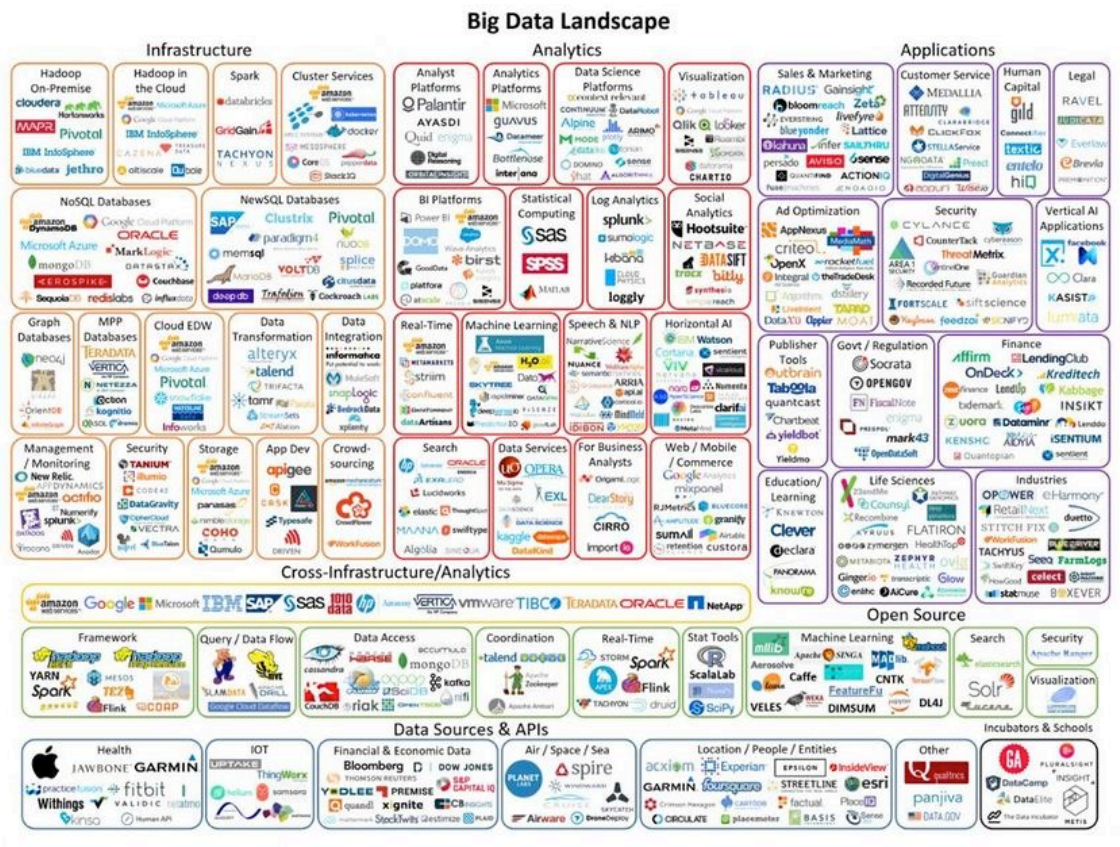


Figura 6 Paisaje de Big Data.

De particular interés en esta lista es Hadoop - software con código abierto que, en los últimos años, se utiliza como un analizador de datos la mayoría de las empresas. En la actualidad, casi todos los medios modernos de análisis de Big Data proporcionan un medio para integrarse con Hadoop. Sus desarrolladores actúan como empresas de crecimiento y las empresas mundiales conocidas.

### 2.3 Programas de IBM para el análisis de Big Data.

IBM ha introducido nuevos programas como Smarter Analytics Signature Soluciones para detectar el fraude, la evaluación de riesgos y análisis del comportamiento del consumidor. Según IDC, la compañía invertirá más de \$ 120 millones a través de 2015 en las soluciones de software orientadas a la identificación de patrones ocultos en los "Big Data".

El software Anti-fraud, Waste and Abuse está diseñado para detectar el fraude en tiempo real relacionada con la evasión de impuestos y pagos de seguros, evitando así que los pagos en efectivo no autorizadas. Además, los servicios competentes de las compañías de seguros y servicios públicos obtendrán las mejores recomendaciones para el despliegue adicional de incidentes. Para violaciones de menor importancia, pueden ser limitado a enviar una carta exigiendo la devolución del pago, con más graves - hay una propuesta para llevar a cabo una investigación completa.

Solución Siguiendo Mejor Acción debería ayudar a la empresa conocer mejor a sus clientes y construir una mejor relación entre ellos. Será capaz de analizar no sólo los datos acumulados en el sistema corporativo, sino también información de las redes sociales. Como su nombre indica, el sistema da recomendaciones para la acción aún más en función de las preferencias y el comportamiento de un único cliente.

## **2.4 Adaptación de las tecnologías.**

Después de varios años de experimentación con la tecnología Big data y la primera introducción en 2013, la adaptación de este tipo de decisiones se incrementará significativamente. Los investigadores encuestaron a los líderes de IT de todo el mundo y encontraron que el 42% de los encuestados ya han invertido en Big Data de la tecnología, o está pensando en hacer este tipo de inversiones en el próximo año.

Las empresas están obligadas a invertir en tecnología para manejar grandes volúmenes de datos debido a que el panorama de la información está cambiando rápidamente, exigen nuevos enfoques de tratamiento de la información. Muchas empresas se han dado cuenta de que las grandes cantidades de datos son críticos, trabajando con ellos pueden lograr beneficios no disponibles con las fuentes tradicionales de información y formas de tratarla. Además, las constantes referencias a "grandes datos" en los combustibles medios del interés en las tecnologías pertinentes. En previsión de nuevas características que traerá la tecnología de procesamiento de "grandes datos," muchas empresas organizan la recogida y almacenamiento de diversos tipos de información.

## **2.5 Lo que afecta a la ejecución de Big Data.**

¿Qué es lo que dificulta la introducción y el uso de la base de datos en muchas empresas? Es el factor humano. La eliminación de información innecesaria para la tarea de expertos cualificados. ¿Qué información se guardan (analizar, procesar, uso), y qué desechar - tales decisiones deben llevar a la gente? Si la empresa no emplea analistas y vendedores de alto nivel - el uso de Big Data no traerá mucho beneficio. Además, los procesos de negocio en compañía utilizando Big Data, también deben cumplir con las nuevas tareas de nivel.

La práctica demuestra que la mayor parte de los costes en el marco de la aplicación de Big Data es el proceso de recopilación de información. Por otra parte, el volumen de dicha información de antemano para determinar con gran precisión no es posible. Y el costo del fracaso en esta etapa es muy alta.

Por ejemplo, es necesario determinar la efectividad de la publicidad. Podemos poner al lado de cada uno de los observadores de diseño publicitario que observarán el número de la gente que pasa por unidad de tiempo. Este trabajo se puede realizar con una cámara. Podemos utilizar esta información a disposición de las empresas que trabajan en el campo de las telecomunicaciones. Tienen información de geolocalización de sus suscriptores: ubicación, edad, sexo, etc. Agencia de publicidad, Con esta información, sea capaz de determinar dónde y qué mensajes debe ser colocado. En un momento, los viajes a las agencias de anuncios en la otra - la llamada a visitar la sala de exposición, o para comprar boletos para el concierto de la cantante popular.

## **2.6 Cómo procesar de Big Data.**

Una amplia variedad de datos, como resultado de un gran número de interacciones proporciona una base excelente para los negocios mediante la evaluación de las perspectivas de desarrollo de productos y áreas enteras, un mejor control de costos, evaluaciones. Por otra parte, Big Data suponen un reto para cualquier departamento de TI. Los datos de Big

Data fundamentalmente nueva naturaleza, cuando la decisión es importante tener en cuenta las restricciones presupuestarias impuestas a los costos operativos y de capital.

IT-jefe que tiene la intención de beneficiarse de Big Data estructurados y no estructurados, deben ser guiadas por el principio siguiente técnica:

- Mover y la integración de datos son necesarios, pero ambos enfoques son más altos que el costo de las herramientas de extracción de información, transformación y cargarlo. Así que no descuide los programas estándar como Oracle, y las tiendas de análisis de datos, como Teradata.

- En función de la situación particular solicitudes de rango análisis varía ampliamente. A menudo, para obtener la información que necesita para obtener suficiente respuesta a una consulta SQL, pero hay consultas analíticas profundas que requieren el uso de herramientas de inteligencia de negocios adquiridos, y toda una gama de capacidades del tablero de mandos y visualización. Con el fin de evitar un fuerte aumento de los gastos de explotación, debe ser el enfoque cuidadosamente equilibrado a la elaboración de una lista de las tecnologías patentadas necesarias en conjunto con Apache Hadoop. - Facilidad de escala se ha convertido en una de las principales razones de la rápida propagación de Hadoop. Especialmente en el campo del procesamiento paralelo en clusters de servidores convencionales y así ahorrar la inversión en los recursos de IT.

## **2.7 El mercado mundial de tecnologías en Big Data.**

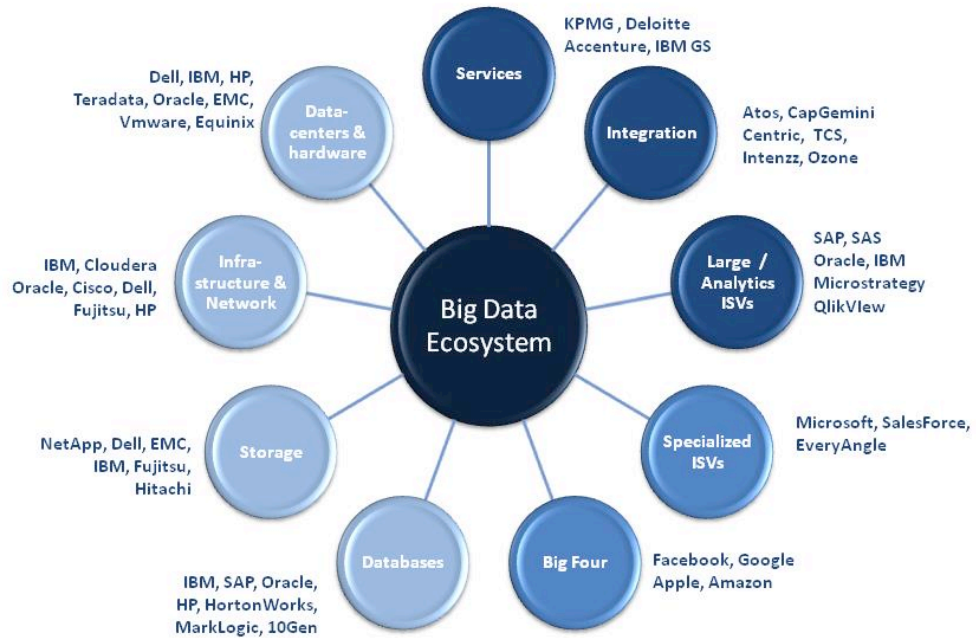
De acuerdo con el pronóstico de tecnología en el mundo Big Data y servicios de predicción, las tecnologías del mercado y los servicios para el procesamiento de Big Data crecerá a partir de \$ 3,2 mil millones en 2010 a \$ 16.9 mil millones en 2016. Esto corresponde a una tasa de crecimiento anual promedio de 40%, que es de aproximadamente 7 veces mayor que la tasa de crecimiento medio anual del mercado total de IT en general.

Un estudio realizado Iniciar Lógica, mostró que el 49% de los gerentes de IT creen disposición de sus empresas para hacer frente a los grandes datos. Al mismo tiempo, el 38% admitió que no tenían ninguna idea acerca de la esencia de este fenómeno.

## **2.8 Los jugadores líderes en Big Data.**

El interés en la colección de instrumentos, elaboración, gestión y análisis de Big Data muestran casi toda la compañía líder en tecnología, que es bastante natural. En primer lugar, se ven directamente afectadas por este fenómeno en su propio negocio, y en segundo lugar, los grandes volúmenes de datos abren grandes oportunidades para el desarrollo de nuevos nichos de mercado y atraer nuevos clientes.

- Amazon
- Dell
- eBay
- EMC
- Facebook
- Fujitsu
- Google
- Hitachi Data Systems Corporation
- HP
- IBM
- LinkedIn
- Microsoft
- NetApp
- Oracle
- SAP
- SGI
- Teradata
- VMware
- Yahoo



*Figura 7 Ecosistema de Big Data.*

## 2.9 Protección de datos, privacidad y seguridad.

Los dos principios básicos de la protección de datos - la prevención de la acumulación de Big Data y minimizar los datos - en marcado contraste con la capacidad de Big Data para facilitar el seguimiento del movimiento, el comportamiento y las preferencias de las personas que predicen el comportamiento de la persona con gran precisión, a menudo sin el consentimiento de esa persona.

Por ejemplo, los registros médicos electrónicos y retiros independientes registros médicos en tiempo real, puede ser un gran paso adelante en la simplificación del sistema de emisión de recetas para medicamentos o dieta y gimnasio. Sin embargo, muchos consumidores creen que estos datos son la naturaleza altamente sensible.

Conjuntos de Big Data, incluso al tiempo que garantiza su anonimato y eliminar cualquier información personal puede ser usada para crear una "huella digital", que, en combinación con otros datos, como fotos geolocalizadas o registros de visitas lugares diferentes, capaces de revelar la identidad de una persona.

A medida que la cantidad de datos personales y la información digital mundial está aumentando y el número de entidades que tienen acceso a esta información y la utilizan.

Las garantías necesarias para el correcto uso de los datos personales.

Un problema estrechamente relacionado de la ciberseguridad. Necesita reevaluación de las amenazas y riesgos en vista de Big Data y la adaptación correspondiente de las soluciones técnicas. Es necesario reconsiderar la política en el ámbito de las leyes de seguridad de la información, la confidencialidad y protección de datos.

Fuentes importantes de los nuevos datos, como la información de las redes celulares y, en particular para los servicios de redes sociales, podrían complementar las estadísticas oficiales. Sin embargo, con el Simposio sobre indicadores de telecomunicaciones mundiales señalado una serie de cuestiones de interés en materia de confidencialidad y privacidad asociados con el uso de Big Data. WTIS instó a considerar la elaboración de directrices sobre la creación, uso y almacenamiento de Big Data. Servicio Nacional de Estadística, en colaboración con otros organismos competentes deberían considerar las oportunidades que ofrecen los grandes datos, participar en la solución paralela de problemas actuales en términos de calidad y fiabilidad de Big Data y protección de la privacidad en el marco de los principios fundamentales de las estadísticas oficiales.

## **2.10 El éxito de Big Data.**

Según la investigación de Accenture, el 60% de las empresas han completado con éxito al menos un proyecto relacionado con Big Data. La mayoría (92%) de los representantes de estas empresas parecía satisfecho con los resultados, y el 89% dijo que el gran de datos se ha convertido en una parte fundamental de su transformación empresarial. Entre el 36% restante de los encuestados no cree que la introducción de esta tecnología, y el 4% aún no han terminado sus proyectos.

El estudio de Accenture participó más de 1.000 consejeros delegados de 19 países. El estudio se basa en datos de PwC Economist Intelligence encuesta Unidad entre 1.135 encuestados en todo el mundo.

Las principales ventajas de Big Data de los encuestados dijeron que "la búsqueda de nuevas fuentes de ingresos" (56%), "mejorar la experiencia del cliente" (51%), "nuevos productos y servicios" (50%) y la "afluencia de nuevos clientes y la preservación de la vieja" (47%). Con la introducción de las nuevas tecnologías, muchas empresas se enfrentan a problemas similares. Para el 51% de seguridad se ha convertido en una característica importante para el 47% - el presupuesto, con el 41% - la falta de personal necesario, y el 35% - dificultades en la integración con el sistema existente. Casi la



totalidad de las empresas encuestadas (91%) están planeando pronto resolver el problema de la escasez de personal y contratar a expertos en Big Data. Las empresas son optimistas sobre el futuro de las tecnologías de Big Data.

89% cree que van a cambiar el negocio tanto como Internet.

79% de los encuestados dijo que las empresas que no tienen que ver con Big Data, van a perder ventaja competitiva.

65% de los encuestados cree que es "archivos grandes de datos,"

60% seguro de que es "la analítica avanzada y análisis"

50% - que se trata de "herramientas de visualización de datos."

Si el rendimiento de los sistemas informáticos modernos ha crecido a lo largo de varias décadas en muchos órdenes y no ir a cualquier comparación con los primeros ordenadores personales de los años 80. del siglo pasado, desde el sistema de almacenamiento, las cosas son mucho peores. Por supuesto, muchas veces los volúmenes disponibles aumentaron drásticamente redujeron el costo de almacenamiento de información en base a, pero la tasa de recuperación y búsqueda de información relevante es pobre.

Sin teniendo en cuenta que todavía es demasiado caro y no muy fiable y duradera USB flash, tecnologías de almacenamiento no han ido muy lejos. Todavía tienen que lidiar con los discos duros, la velocidad de rotación de las placas se encuentra en los modelos más caros se limita al nivel de 15 vol. / min. Cuando se trata de grandes volúmenes de datos, obviamente, un número considerable de los colocó en las unidades a una velocidad de 7200. Vol. / Min. Esto es no muy bien.

## **2.11 El retraso tecnológico.**

Big Data puede llegar a ser un gran problema o una gran oportunidad para las agencias del gobierno, a menos que sean capaces de tomar ventaja de ellos. Esta es la conclusión alcanzada en el segundo trimestre de 2013, los autores del estudio Los resultados de la encuesta Big Data. En los próximos dos años, el volumen de los datos almacenados en las instituciones públicas mutila 1 petabytes. Al mismo tiempo, beneficiarse del creciente flujo de información es cada vez más difícil, que afecta la falta de espacio disponible en el sistema de almacenamiento de datos, el difícil acceso a los datos de la derecha no es suficiente poder de cómputo y personal cualificado.

A disposición de los administradores de IT y aplicaciones de tecnología de mostrar una acumulación significativa de los requisitos de problemas del mundo real, cuya solución puede aportar un gran valor añadido a los datos. 60% de las agencias civiles y de defensa de 42% solamente se dedica al estudio de Big Data y están buscando para su

posible uso en sus actividades. El principal debe ser la mejora de la eficiencia del trabajo - así decir el 59% de los encuestados. En segundo lugar, está el aumento de la velocidad y la precisión de las decisiones (51%), el tercero - la capacidad de prever (30%) [SUL15].



*Figura 8 Aumentar el interés de unas empresas de Big Data.*

La consulta para encontrar cientos de líneas de un millón, no puede hacer frente a una tabla de cien mil millones de filas. Si los datos cambian con frecuencia, es importante mantener un registro y auditoría. La aplicación de estas reglas simples tendrá una técnica importante para el desarrollo de almacenamiento y trabajo con la información de los datos sobre el volumen de datos, la velocidad y la frecuencia de cambio.

Sea como sea, los flujos de datos procesados continúan creciendo. El aumento del volumen de información almacenada en los dos últimos años indica 87% de los encuestados en la continuación de esta tendencia en la perspectiva de los próximos dos años se calcula con un 96% de los encuestados. Para poder disfrutar de todos los beneficios que son los grandes datos, teniendo instituciones encuestas necesitarán un promedio de tres años.



## **3.Descripción de algoritmos utilizados.**

### **3.1 Modernización de los equipos para el procesamiento de Big Data.**

Los resultados de Oracle Corporation sugiere que muchas empresas se ven atrapados por sorpresa el crecimiento del " Big Data ". "La introducción del «Big Data», será el mayor reto para las empresas de IT en los próximos dos años - dice Luigi Frege, vicepresidente de Oracle. Al final de este período, o bien tratar con él o significativamente a la zaga en los negocios y serán mucho tanto de las amenazas y las capacidades de los " Big Data ".

La tarea de "desarrollo" de Big Data es única, reconocida en Oracle. La respuesta de la empresa principal a los desafíos de Big Data debe ser la modernización de los centros de datos empresariales.

Para evaluar la disposición de las empresas a los cambios en el centro de datos, durante casi dos años con la firma de análisis de Oracle Quocirca recolectó datos para el estudio del índice de Oracle Índice Centro de datos de próxima generación. Esta evalúa el progreso de las empresas en los centros de datos de uso cuestión de mejorar el rendimiento de las IT infraestructura y optimización de procesos de negocio.

El estudio consistió en dos fases, fue visto cambios significativos de indicadores clave en el umbral de la segunda etapa. Promedio por Índice NGD Oracle, que ganó los encuestados de Europa y el Medio Oriente, fue 5,58. La puntuación máxima de -10.0 - refleja la estrategia más meditada de la utilización del centro de datos.

Promedio 5.58 convirtió mayor en comparación con el primer ciclo de los estudios realizados en febrero de 2011, - 5,22. Esto sugiere que las empresas en respuesta al auge de "Big Data" para aumentar la inversión en la estrategia de desarrollo del centro de datos. A nivel mundial, la industria y las tendencias dentro de las industrias cubiertas por el estudio, el aumento del índice NGD Oracle en los resultados de la segunda vuelta que en la primera. Escandinavia, Alemania, Suiza, es una empresa líder en el índice de desarrollo sostenible a 6,57. A continuación en el ranking debe Benelux 5,76, y luego el Reino Unido, con un índice de 5,4, lo que ya está por debajo de la media.

En Rusia, que ha sido incluido en la lista de países, sólo en los estudios de segundo ciclo y participó en la primera, existe un considerable potencial de crecimiento de 4,62, según los analistas.

Según el estudio, las organizaciones rusas están considerando apoyar el crecimiento de las empresas como una razón importante para la inversión en centros de datos. Más del 60% de las empresas ven la necesidad de tal inversión, ahora o en un futuro próximo, lo que sugiere que la organización pronto descubrirá que se hace muy difícil competir, si no para hacer las inversiones adecuadas.

En el mundo de la proporción de encuestados con los centros de datos corporativos privados ha disminuido del 60% en el primer ciclo del estudio al 44% en el segundo ciclo del estudio, por el contrario, el uso de los centros de datos externos se ha incrementado en 16 puntos a 56%.

Sólo el 8% de los encuestados dijo que ellos no necesitan una nueva capacidad del centro de datos en el futuro previsible. 38% de los encuestados ve la necesidad de nueva capacidad en el centro de datos dentro de los próximos dos let. Casi 6,4% de los encuestados informó que su organización no tiene un plan de desarrollo sostenible asociado con el uso de los centros de datos. La participación de los administradores de centros de datos que consideran copias de facturas en costos de electricidad aumentó de 43,2% a 52,2% durante el período de estudio.

## **3.2 Descripción general de los métodos de protección de datos.**

- Medios de almacenamiento de protección contra la corrupción de datos no autorizada o accidental. Por la distorsión entender eliminar, modificar o introducir la información de terceros en el bloque de datos protegida. Para protegerse de este tipo de influencias, el sistema debe ser capaz de detectarlos, es decir, el receptor recibe un bloque de datos debe garantizar que no se han cambiado.
- La confidencialidad significa que el acceso a los datos privados puede haber sólo los usuarios autorizados. El acceso de usuarios no autorizados debe ser excluidos.
- Identificación y autenticación permite a las partes implicadas en el intercambio de datos, para identificar a sí mismos. Si la autenticación es satisfactoria, el usuario accede a la información. Del mismo modo, hay una necesidad de identificar la fuente de datos.
- Fiabilidad. De acuerdo con AENOR R ISO 7498-2-99, este método puede tomar dos formas.

Confirmación de la fiabilidad del remitente asume que los datos facilitados por el destinatario comprobar los datos del remitente. Esto protege contra cualquier intento por parte del emisor de negar falsamente la transferencia de los datos o su contenido.

Confirmación fiabilidad de la entrega se asegura de que el transmisor de datos es proporcionado por la transferencia de datos. Esto protege contra cualquier intento posterior al destinatario falsamente negar la recepción de los datos o su contenido.

El problema de la protección de datos no es nuevo. Por el momento, no son juzgados y funciones, y prácticas de seguridad a prueba.

### **3.3 Cifrado.**

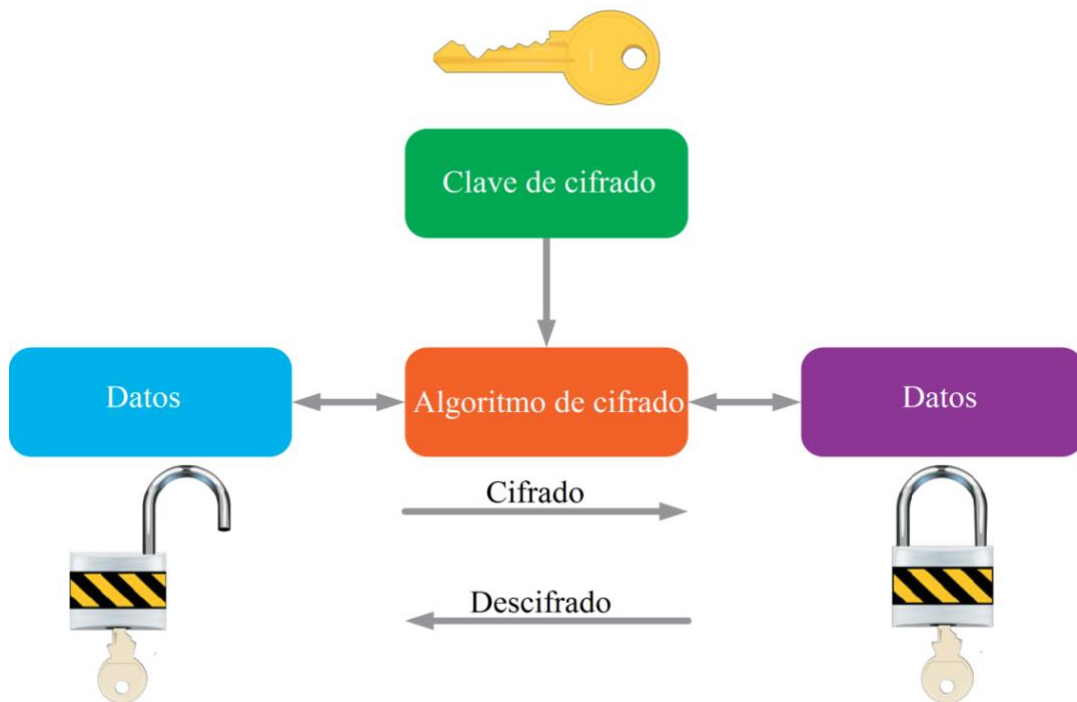
El cifrado se utiliza para almacenar información importante y transmitirla a través de canales de comunicación inseguros. Dicha transferencia de datos consta de dos procesos mutuamente inversos:

Antes de enviar los datos a través de la línea de comunicación o antes del almacenamiento, se cifra. Para restaurar los datos originales cifrados, se aplica el procedimiento de descifrado.

El cifrado se usó originalmente solo para transmitir información confidencial. Sin embargo, más tarde comenzaron a cifrar la información para almacenarla en fuentes poco confiables. El cifrado de información para su almacenamiento se usa ahora, esto evita la necesidad de un almacenamiento físicamente protegido.

Un cifrado es un par de algoritmos que implementan cada una de estas transformaciones. Estos algoritmos se aplican a los datos con una clave. Las claves para el cifrado y el descifrado pueden variar, pero pueden ser las mismas. El secreto del segundo (descifrado) hace que los datos sean inaccesibles para un conocimiento no autorizado, y el secreto del primero (cifrado) hace que sea imposible ingresar datos falsos. Los primeros métodos de cifrado usaban las mismas claves, clave privada o simétrica, pero en 1976 se abrieron algoritmos usando diferentes claves, clave pública o asimétrica. Mantener estas claves confidenciales y compartirlas adecuadamente entre los destinatarios es una tarea muy importante en términos de mantener la confidencialidad de la información transmitida.

Este método es para convertir los datos de forma cifrada (Figura 9). La codificación se realiza por medio de algoritmos especiales y la clave de cifrado. Proceso inverso también requiere clave de decodificación. Es posible dos opciones. Con el cifrado simétrico utilizando la misma clave para codificar y decodificar. Cuando se usa el cifrado asimétrico para codificar una clave para la decodificación y - otro.



*Figura 9 El mecanismo de cifrado de datos.*

### 3.4 Hash.

Una función hash criptográfica es una clase especial de función que tiene ciertas propiedades que lo hacen adecuado para su uso en criptografía. Es un algoritmo matemático que muestra datos de tamaño arbitrario con una cadena de bits de tamaño fijo (un hash) y está destinado a ser una función unidireccional, es decir, una función que es inviable para invertir. La única manera de recuperar los datos de entrada de la salida de una función hash criptográfica ideal es tratar de buscar a través de las posibles entradas para ver si están produciendo un partido, o utilizar una tabla de hashes coincidentes (Rainbow table). Bruce Schneier llamó a las funciones hash unidireccionales "los caballos de la criptografía moderna". La entrada de datos se suele nombrarse mensaje, y la salida (el valor de una función hash o hash) se refiere a menudo como un mensaje de resumen o simplemente hash.

Una función hash criptográfica ideal tiene cinco propiedades básicas:

- Con mismo mensaje determinista siempre resulta en el mismo hash
- Es rápido para calcular el valor hash para cualquier mensaje dado
- Es inviable crear un mensaje a partir de un valor hash, excepto tratando todos los mensajes posibles un pequeño cambio en el mensaje debe cambiar el valor hash tan extensamente que el nuevo valor hash aparece correlacionado con el antiguo valor hash
- Es inviable encontrar dos mensajes diferentes con el mismo valor hash

La función hash criptográfica tiene una gran cantidad de aplicaciones en aplicaciones de seguridad, en particular en la firma digital, códigos de autenticación de mensajes (MPC), y otras formas de autenticación. También se pueden utilizar como funciones hash normales, para indexar datos en tablas hash, para tomar huellas “dactilares”, para detectar datos duplicados o identificar archivos de forma única, y comprobación para detectar daños accidentales en los datos. De hecho, en términos de seguridad de la información, los valores hash criptográficos a veces se llaman huellas digitales (digitales), sumas de comprobación, o sólo valores hash, incluso si todas estas condiciones valen funciones más generales con propiedades y propósitos muy diferentes.



*Figura 10 Función hash.*

### 3.5 MAC.

#### **Código de autenticación MAC (Message Authentication Code).**

MAC, en inglés. message authentication code (código de autenticación de mensajes) es una herramienta de protección en protocolos de autenticación de mensajes con participantes de confianza, un conjunto especial de caracteres que se agrega al mensaje y está diseñado para garantizar su integridad y autenticación de origen de datos.

El MAC se utiliza generalmente para garantizar la integridad y la protección contra la manipulación de la información transmitida.

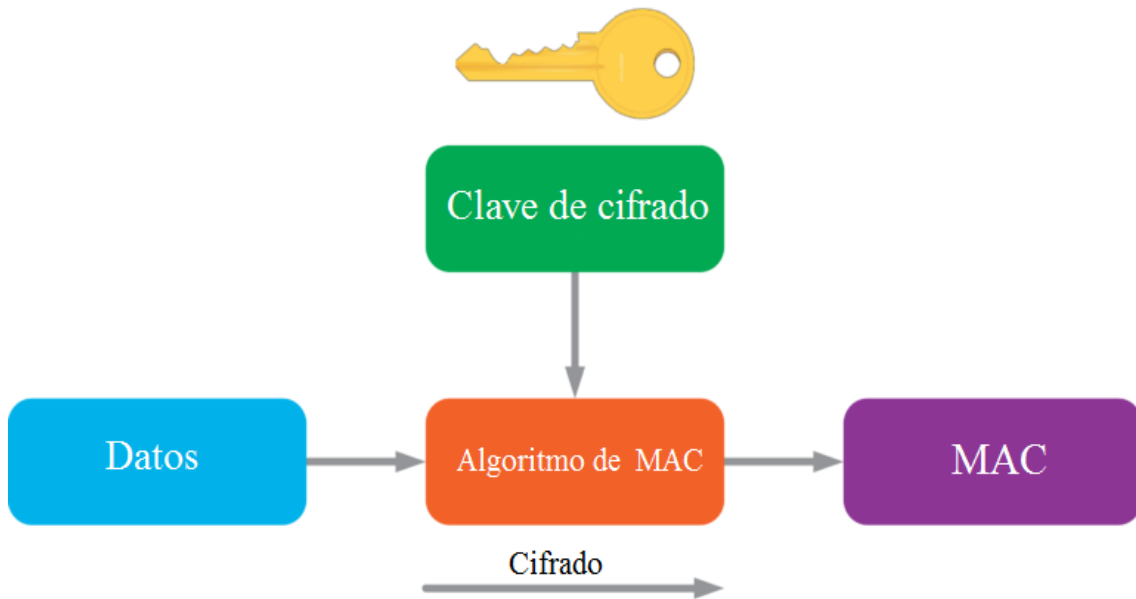
Para verificar la integridad (pero no la autenticidad) del mensaje en el lado de envío, se agrega un valor de función hash de ese mensaje al mensaje, también se genera un hash del mensaje recibido en el lado de recepción. Se comparan el hash generado en el lado receptor y el hash resultante. En caso de igualdad, se considerará que el mensaje recibido no ha cambiado.

Para proteger contra la falsificación del MAC del mensaje, se utiliza una imitación elaborada utilizando un elemento secreto (clave) conocido solo por el remitente y el receptor.

El mecanismo utilizado para generar el hash. Sin embargo, esto requiere no sólo el mensaje original, sino también una clave secreta que sólo conoce el emisor y el receptor (Figura 11). Esto permite que el destinatario para verificar la integridad de los



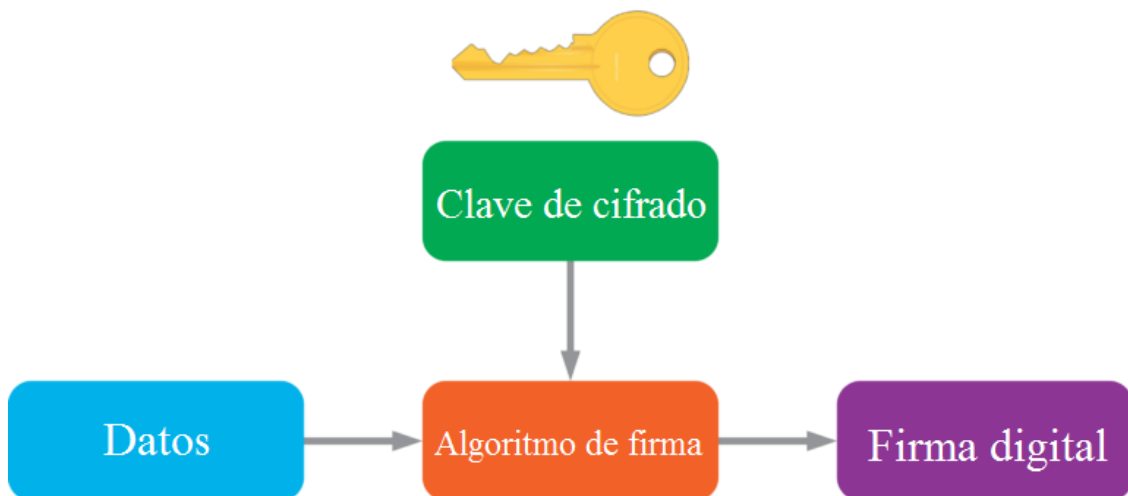
datos y la información para identificar al remitente. Si el remitente no tiene la clave secreta - código hash se generará de forma incorrecta, es fácil de detectar el destinatario.



*Figura 11 El mecanismo de los algoritmos MAC.*

### 3.6 Las firmas digitales.

Este mecanismo permite la autenticación de mensajes, es decir, para probar su autenticidad con una firma digital. La firma digital funciona de la misma manera como una firma normal - en que siempre es posible identificar al remitente. Cuando un cifrado de firma digital que se utiliza asimétrico (Figura 12). Clave privada y el descifrado se utiliza para cifrar un mensaje - abierta. La clave privada se conocida sólo por el emisor, mientras que el acceso a la clave pública puede tener el conjunto de destinatarios de los datos.



*Figura 12 El mecanismo de firma digital.*

### 3.7 Generación de números aleatorios.

**Generación de números aleatorios.** Codificación de protección se utiliza cuando es imposible resolver la clave de cifrado. Por esta razón, la clave debe ser aleatoria. Para ello, utilice un generador de números aleatorios (Figura 13). Se puede utilizar como métodos de software de formación de la secuencia inicial de números aleatorios, y una fuente natural de señal aleatoria, por ejemplo, la tensión de ruido a través del diodo.



*Figura 13 Un ejemplo del generador de números aleatorios.*

### 3.8 Fuentes de Fuentes de números aleatorios.

Las fuentes de números aleatorios reales son extremadamente difíciles de encontrar. Los ruidos físicos, como los detectores de eventos de radiación ionizante, el ruido de disparo en una resistencia o la radiación cósmica, pueden ser tales Fuentes. Sin embargo, estos dispositivos rara vez se utilizan en aplicaciones de seguridad de red. Las complejidades también causan ataques bruscos en dispositivos similares. Existen varias desventajas en las Fuentes físicas de números aleatorios:

- Tiempo y mano de obra durante la instalación y configuración en comparación con el software generador de números aleatorios;
- Carestía;
- La generación de números aleatorios es más lenta que la implementación programática de un generador de números aleatorios;
- No se puede reproducir una secuencia de números aleatorios generada previamente.

Al mismo tiempo, los números aleatorios derivados de una fuente física se pueden utilizar como un elemento de generación (en inglés. seed) para el software generador de números aleatorios. Estos generadores combinados se utilizan en criptografía, lotería, máquinas tragamonedas.

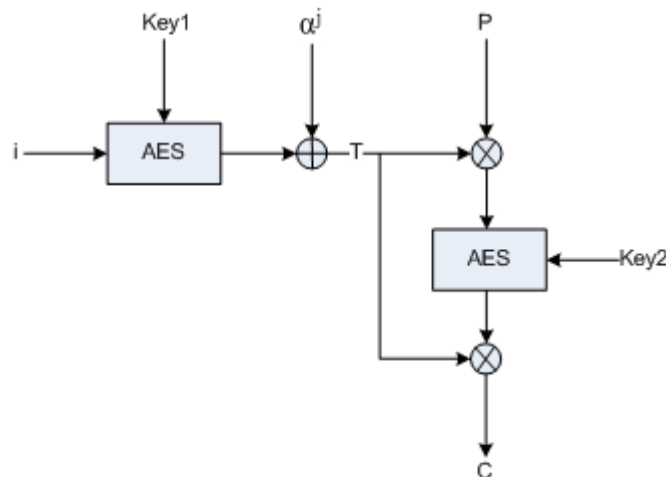
### 3.9 Los estándares IEEE

En muchas implementaciones de unidades de disco cifrado se utiliza el modo de cifrado XTS. El modo XTS está involucrado en el grupo de trabajo de estandarización SIS-WG IEEE, que ha desarrollado una serie de normas - IEEE P1619. Estos estándares, además de describir el modo XTS proporciona además otros modos de cifrado que utilizan es conveniente, dependiendo de la organización interna de los dispositivos de almacenamiento de datos, y los requisitos de seguridad.

**- IEEE P1619–2007 — IEEE Estándar para la protección criptográfica de datos en dispositivos de almacenamiento orientados a bloques.**

El modo de cifrado estándar es la descripción de la XTS-AES, que pertenece al grupo de los sistemas de cifrado, con se llama Tweakable Block Cyphers (enfoque de cifrado, que implica el uso en el cálculo de un parámetro adicional).

Figura 14 muestra un diagrama de bloques de los XTS.



*Figura 14 Diagrama de bloques de XTS.*

Este modo utiliza un par de claves, como lo hacía el número de sector,  $j$  - Número de bloque de 128 bits dentro del sector (no más de 220),  $\alpha$  - elemento primitivo de GF ( $2^{128}$ ). También se describe en el equipamiento de serie de texto cifrado Robar (CTS), que es el tratamiento especial de los dos últimos bloques de texto claro cuyo tamaño no es un múltiplo de 128 bits.

**- IEEE P1619.1–2007 — IEEE Estándar para cifrado autenticado con expansión de longitud para dispositivos de almacenamiento.**

Este estándar describe el modo recomendado para su uso en los casos de la necesidad de datos adicionales software. Esto conduce inevitablemente a un aumento adicional en

el tamaño del texto cifrado (mediante la adición de los valores del código de autenticación de mensaje). modos de cifrado recomiendan para uso estándar:

- CBC-MAC (CCM);
- Galois/counter mode (GCM);
- Cipher Block Chaining with HMAC-SHA (CBC-HMAC-SHA);
- Tweakable block-cipher with HMAC (XTS-AES-256-HMAC-SHA-512).

**- IEEE P1619.2–2010 — IEEE Estándar para cifrado de bloque ancho para medios de almacenamiento compartidos.**

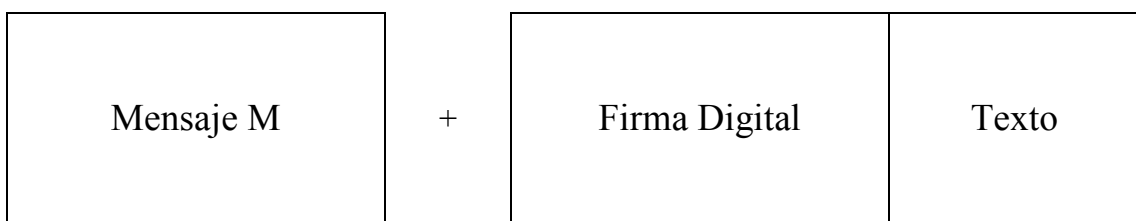
El estándar describe los modos de cifrado de bloques de datos que son más grandes que 512 bytes. Se supone que junto con los datos que requieren modos de procesamiento de confidencialidad y integridad se pueden proporcionar datos asociados (por ejemplo, un número de bloque lógico) para los que sólo se requiere un software de autenticación. Se recomiendan los modos de cifrado para su uso:

- EME2-AES;
- XCB-AES.

Modos de aplicación convenientes en situaciones en que el atacante tiene acceso directo a los datos cifrados, o se puede llevar a cabo la interceptación en el canal. La siguiente tabla muestra los costos de cifrado algorítmico N bloques de datos de 16 bytes. La selección de un modo particular depende de los requisitos de rendimiento y el tamaño de la implementación de hardware.

**3.10 GOST P 34.10-2001.**

**GOST P 34.10-2001** se base de curvas elípticas. Su resistencia se basa en el cálculo del logaritmo discreto en un grupo de puntos sobre una curva elíptica, así como la resistencia de la función hash. Después de firmar el mensaje M al mismo tamaño de firma digital adjunta de 512 bits y un cuadro de texto. En el campo de texto puede contener, por ejemplo, la fecha y hora de envío, o una variedad de datos sobre el remitente:



Este algoritmo no describe un mecanismo para la generación de parámetros necesarios para la formación de las firmas, pero sólo determina la forma sobre la base de estos parámetros para obtener una firma digital. El mecanismo de parámetros de generación determina en el lugar, dependiendo del sistema que está siendo desarrollado.

### Algoritmo

#### Parámetros del esquema de firma digital:

*Número simple* - módulo de curva elíptica de tal manera que  $p > 2^{255}$

Curva elíptica E se define por su invariante o coeficientes  $a, b \in F_p$ ,  $F_p$  - campo de primer finita.

J (E) asociado con los coeficientes a y b de la siguiente manera:

$$J(E) = 1728 \frac{4a^3}{4a^3 + 27b^2} \pmod{p}$$

*mientras que*  $4a^3 + 27b^2 \not\equiv 0 \pmod{p}$

*Número entero m* - el orden de la curva elíptica (es decir, el número de miembros del grupo). m no será la misma a la p.

*Número primo q*, el orden de un subgrupo cíclico de puntos sobre una curva elíptica, es decir  $m = nq$ , para algunas  $n \in \mathbb{N}$ . q se encuentra dentro de  $2^{254} < q < 2^{256}$ .

*Punto*  $P(x_P, y_P) \neq 0$  curva elíptica E, que es un generador del subgrupo de orden q, o en otras palabras  $qP = 0$ . Aquí  $qP = P + P + \dots$  (veses q). A 0 - elemento cero de la curva elíptica.

*h (M) - función hash* (AENOR 34.11-94), que por supuesto muestra un mensaje M en el vector binario de longitud de 256 bits.

#### Cada usuario tiene unas claves de firma digitales personales:

*clave de cifrado d* - número entero en el rango  $0 < d < q$ .

*clave de descifrado*  $Q(x_Q, y_Q)$  - un punto de la curva elíptica,  $dP = Q$ .

*Esto debe llevarse a cabo:*

- $p^t \not\equiv 1 \pmod{q}, \forall t = 1..B, \text{ где } B \geq 31$
- $J(E) \not\equiv 0 \pmod{p}, J(E) \neq 1728$

#### Formación de una firma digital:

1. El cálculo de un hash del mensaje M:  $\bar{h} = h(M)$
2. Calcular  $e = z \pmod{q}$ , y si  $e = 0$ ,  $e = 1$ . Donde z - entero correspondiente
3. Generar un número aleatorio k, tal que  $0 < k < q$
4. Cálculo del punto de la curva elíptica  $C = kP$ , y en la búsqueda de que  $r = xc \pmod{q}$  ( $xc$  - x C punto de coordenadas). Si  $r = 0$ , volvemos al paso anterior.

5. Finding  $s = rd + ke \pmod{q}$ . Si  $s = 0$ , vuelva al paso 3.
6. Formación de la firma digital  $\xi = (\bar{r}|\bar{s})$ , en el que  $r$  y  $s$  - los vectores correspondientes a la  $r$  y  $s$ .

### Verificación de la firma digital:

1. El cálculo de la firma digital de los números  $r$  y  $s$ , dado que  $\xi = (\bar{r}|\bar{s})$ , en el que  $r$  y  $s$  - el número correspondiente a los vectores  $r$  y  $s$ . Si al menos una de las desigualdades  $r < q$  y  $s < q$  es falsa, entonces la firma es incorrecta.
2. El cálculo de un hash del mensaje  $M$ :  $\bar{h} = h(M)$
3. Cálculo  $e = z \pmod{q}$ , y si  $e = 0$ ,  $e = 1$  donde  $z$  -  $z$  - entero correspondiente  $\forall t = 1..B$
4. Cálculo de  $v = e - 1 \pmod{q}$ .
5. Cálculo  $Z1 = sv \pmod{q}$  y  $Z2 = -rv \pmod{q}$ 
  6. Cálculo de elíptico punto de la curva  $C = Z1P + z2Q$ . Y la definición  $R = xc \pmod{q}$ , donde  $xc$  - coordenada  $x$  de la curva  $C$ . En el caso de igualdad de  $R = r$  firma es correcta, de lo contrario - no es correcta.

## 3.11 GOST P 34.11-2012.

**GOST P 34.11-2012** La norma especifica el algoritmo y el procedimiento para el cálculo de la función hash para una secuencia de caracteres.

La base de la función hash se basa en un proceso iterativo de diseño Merkla Damgarda utilizando MD-amplificación. Además, MD amplificación es de bloques parcial en el cálculo de una función hash a la completa añadiendo el vector (0 ... 01) de tal longitud como para obtener un bloque completo. De los elementos adicionales necesarios a tener en cuenta lo siguiente:

- completa la conversión, que es que la función de compresión se utiliza para controlar la suma de todos los bloques de mensajes del módulo de 2512;
- al calcular el código hash de cada iteración se aplica diferentes funciones de compresión. Podemos decir que la función de compresión depende del número de iteración.

Breve descripción de la función hash GOST P 34.11-2012

La entrada a la función de hash recibe un mensaje de tamaño arbitrario. A continuación, el mensaje se divide en bloques de 512 bits si el tamaño de mensaje no es un múltiplo de 512, entonces se complementa con el número necesario de bits. A continuación, de forma iterativa utiliza una función de compresión como resultado de lo cual la acción se actualiza el estado interno de la función hash. También calculados bloques de suma de control y el número de bits procesados. Cuando se procesan todos los bloques del mensaje original, produjo dos cálculos más, que completa el cálculo de la función hash:

- función de compresión de bloque de procesamiento con la longitud total del mensaje.
- función de compresión de bloque de procesamiento con una suma de comprobación.

### 3.12 Mecanismos de autenticación

Una solución estándar para identificar usuarios en entornos HPC y entornos Cloud es autorizar el acceso con un par de claves pública / privada, desde un nodo de cliente confiable. Una preocupación es el uso incontrolado de claves públicas / privadas. Los usuarios pueden olvidar dónde los han instalado, por lo que el acceso a esta computadora personal puede comprenderlos a todos [YLO19]. Si alguien accede a la computadora original, el acceso a todos los nodos se ve comprometido. Para resolver esto, Cloudfare [RHE20] utiliza certificados de corta duración y autenticación basada en credenciales de inicio de sesión único (SSO).

Además, Kerberos [MIL88] se propuso en 1988 como un sistema de autenticación y se actualizó a la versión 5 en 2005 para resolver algunas limitaciones en RFC4120 [RFC20]. Se utiliza como mecanismo de autenticación inicial para acceder a los sistemas Windows y Linux. También se puede utilizar en otro entorno [PER13], pero requiere un proceso de "kerberización", que limita su uso en aplicaciones que no están previstas. Estos sistemas fueron diseñados cuando había limitaciones en las redes de comunicación, los procesadores eran más lentos y la criptografía de clave pública era menos utilizada. Kerberos separa el proceso de autenticación del servidor de servicios creando un control centralizado para diferentes servicios. Además, Kerberos se basa en el uso de criptografía de clave privada, por lo que es esencial seleccionar contraseñas robustas para evitar ataques de diccionario. Algunos análisis de seguridad [TSA08] [ELE11] muestran esta debilidad en la fase de comunicación y representan un verdadero desafío para este protocolo. También requiere sincronización horaria entre computadoras, ya que los tickets se basan en marcas de tiempo.

Se ha propuesto alguna extensión a Kerberos, como la descrita por Tbatou et al. [TBA17]. Definen un protocolo de autenticación para sistemas distribuidos basado en los modelos Kerberos V5 y Diffie Hellman, pero no es un esquema de autenticación multifactor. Para hacer que Kerberos sea más seguro, Quoc et al. [HOA15] proponen

modificar el intercambio inicial en Kerberos 5 mediante el uso de datos biométricos y criptografía asimétrica. Usan un SDK de MCC de biblioteca DLL .Net para la autenticación biométrica de huellas digitales, por lo que se limita a las plataformas Windows. Además, estas soluciones no resuelven cómo adaptar el esquema de autenticación a otros protocolos.

Además, podemos encontrar algunos recursos de autenticación multifactor para mejorar la seguridad que confía en un solo componente. Algunos sistemas de Contraseña de un solo uso (OTP) basados en el tiempo, como Google Authenticator [GOO20], que implican una sincronización aproximada para generar el código de acceso. Además, el usuario necesita acceder a otro programa para obtener el código, escribirlo o esperar porque el vencimiento del tiempo está cerca, por lo que generalmente es un inconveniente usarlo. Aunque Google ha desarrollado este esquema OTP, prefiere que sus empleados usen claves de seguridad basadas en tokens de hardware [KRE20].

Se han propuesto algunos modelos de autenticación, como los sistemas basados en la latencia de comunicación [DOU18], pero estos sistemas no son aplicables en clústeres, ya que los tiempos de acceso entre nodos son casi constantes e independientes de la conexión a Internet, ni en entornos con tiempos de acceso.

Los sistemas basados en tarjetas inteligentes [KAN19] se han utilizado durante mucho tiempo para identificar de forma segura. Estas tarjetas requieren un lector que pueda integrarse en un teclado, computadora portátil o lectores USB. El problema es que, con el tiempo, algunos controladores han dejado de funcionar porque las versiones del sistema operativo los han vuelto obsoletos. Las nuevas tarjetas inteligentes también son compatibles con Near Field Communications (NFC), lo que facilita el acceso, pero implica el uso de una computadora como elemento intermedio para la autenticación. Si no tenemos este elemento, no es posible llevar a cabo la autenticación. Además, el usuario no puede validar o cancelar una solicitud si la tarjeta está insertada.

El uso de sistemas basados en tokens ha crecido porque son elementos simples de usar. Es un recurso que posee el usuario, por lo que no es suficiente tener una contraseña o un par de claves pública / privada. Además, el usuario valida mediante un clic el acceso solicitado.

La Fast Identity Online Alliance (FIDO) especificó dos marcos y protocolos de autenticación: el Marco de autenticación universal (UAF) para la autenticación sin contraseña desde dispositivos inteligentes y el protocolo Universal Second Factor (U2F) para la autenticación de dos factores utilizando un pequeño token de hardware para acompañar un dispositivo inteligente no FIDO que tenga un navegador web compatible con FIDO. Ambos operaban con el mismo principio subyacente de usar encriptación asimétrica para la autenticación, y ambos ahora se han combinado en la Recomendación de Autenticación Web (FIDO2) del Consorcio World Wide Web (W3C) [W3C20].



Yubiko [YUB20] vende diferentes dispositivos de seguridad U2F. Según el fabricante, para evitar ataques a YubiKey, que podrían comprometer su seguridad, YubiKey no permite el acceso o la modificación de su firmware. Pero, un error en la aleatoriedad de algunas de estas claves (Serie FIPS) hace que se recuperen 80 bits de un valor pseudoaleatorio. Si se usa para firmas con algoritmo de firma digital de curva elíptica (ECDSA), podría permitir que un atacante que obtenga acceso a varias firmas reconstruya la clave privada [YUB20b]. Por lo tanto, si usa un dispositivo que no puede reprogramarse y se detecta un fallo crítico, finalmente hay que desecharlo.

Chadwick y col. [CHA19] proponen el uso de teléfonos móviles con lectores de huellas digitales para autenticar a los usuarios con UAF, en lugar de usar usuarios y códigos de acceso. Ciolino y col. [CIO19] han estudiado el impacto en la usabilidad de las claves de seguridad y la importancia de usarlas en un contexto más amplio de servicios web. Existen esquemas de aprovisionamiento para dispositivos IoT basados en U2F como U2Fi [KAN19], pero estos no ofrecen una solución directa para comunicarse entre aplicaciones independientes del cliente y el servidor.

OAuth es una solución que intenta unificar la autorización de diferentes aplicaciones. La primera versión de OAuth data de 2007, y en 2012 se definió OAuth 2.0 [HAR20]. Este protocolo permite autorizar a terceros a acceder a los recursos de su servidor sin compartir sus credenciales, y se utiliza principalmente en aplicaciones web o móviles. De acuerdo con Chae et al. [CHA99], tiene muchas vulnerabilidades de seguridad en el procedimiento de certificación de aplicaciones de terceros. OAuth 2.0 no es un protocolo de autenticación, sin embargo, se puede usar como una capa base para otro protocolo de autenticación como OpenID Connect [OPE20]. OpenID Connect 1.0 permite a los clientes verificar la identidad del usuario en función de la autenticación realizada por el servidor de autorización. De acuerdo con Li et al. [LI18], las implementaciones del mundo real de ambos esquemas a menudo son vulnerables a los ataques, y en particular a los ataques de falsificación de solicitudes entre sitios (CSRF). OpenID Connect y OAuth 2.0 no incluyen una autenticación multifactor, pero existe una solución externa para proporcionarla, como SAASSPASS [SAA20], con una aplicación móvil. Además, SecSign [SEC20b] proporciona una extensión para la autenticación de dos factores utilizando un teléfono móvil. Pero, como se comentó anteriormente, los usuarios son reacios a instalar aplicaciones en sus teléfonos para acceder a los recursos de trabajo.

Otra solución es combinar diferentes elementos como Single Sign-On (SSO) con autenticación de dos factores y lenguaje de marcado de aserción de seguridad (SAML). Podemos usar SAML para intercambiar datos de autenticación de usuario como lenguaje de marcado extensible (XML) entre proveedores de identidad y proveedores de servicios. SSO es un esquema que utiliza una autenticación única para permitir el acceso a múltiples aplicaciones al pasar un token de autenticación sin problemas a las aplicaciones configuradas.

Pero al final, si se necesita implementar una solución, se debe usar diferentes capas proporcionadas por diversos desarrolladores, por lo que podría ser fundamental la compatibilidad de las diferentes versiones de cada componente para obtener una solución estable. Debemos evitar usar un mazo para romper una nuez.

En un entorno HPC, podemos encontrar un ecosistema muy diverso de aplicaciones. No solo tenemos acceso web a una plataforma, sino que también necesitamos acceder a recursos y aplicaciones ubicados internamente en el clúster, por lo que generalmente no hay acceso directo para facilitar la autenticación en computadoras y usuarios que están en la red de intranet privada.

Estos entornos informáticos están en continuo crecimiento, nuevos sistemas de almacenamiento de datos masivos, modelos informáticos como map / reduce, o nuevos paradigmas que surgen al tratar de abordar problemas informáticos complejos que pueden resolverse más fácilmente con nuevos modelos de comunicación. Muchos de estos sistemas se basan en esquemas de usuario y contraseña como primer nivel de autenticación.

Existen soluciones específicas para Hadoop, como la propuesta por Khalil et al. [KHA15], basado en TPM, pero esta solución no se puede utilizar en otras aplicaciones y depende del Módulo de plataforma confiable (TPM) instalado en cada nodo de cliente, por lo que bloquea al usuario en un nodo de cliente específico.

Además, otro problema general en múltiples protocolos es que no incluyen un sistema centralizado que pueda controlar fácilmente la revocación de las comunicaciones en tiempo real. Normalmente, cuando la comunicación cliente-servidor ha sido autorizada y establecida, no hay forma directa de cancelarla. En estos casos, se pueden establecer diferentes reglas en el nivel de firewall, pero requieren permisos de superusuario.



## 4. Almacenamiento en Big Data, descripción de Hadoop.

### 4.1 Hadoop.

Big data, un zumbido reciente en el mundo de Internet, está creciendo más fuerte con cada día que pasa. Facebook tiene casi 21 PB [BEA10] mientras que Jaguar ORNL tiene más de 5 PB. Los datos almacenados están creciendo tan rápidamente que es probable que los sistemas de almacenamiento a escala EB se utilicen próximamente. En ese momento debería haber más de mil 10 trabajos de PB [YAN16].

El siglo XXI se ha convertido principalmente una era de la información. Búsqueda eficiente y el procesamiento del flujo de datos se convierten en un requisito clave del nuevo siglo. El problema principal es limitar el funcionamiento de cualquier equipo, y se encontró una solución a los problemas de escala y distribuirlos entre varios ordenadores. El advenimiento de la computación distribuida ha dado lugar a la aparición de nuevos programas que pueden resolver problemas complejos. Entre estas herramientas uno de los más populares fue la plataforma de software Apache Hadoop.

El sistema de Hadoop está escrito en Java y está instalado en varios servidores que funcionan en un clúster en la tecnología de shared-nothing. Se pueden añadir o eliminar servidores de un cluster Hadoop, sin interrumpir el servicio. Si disponemos de más servidores podremos aportar más potencia de cálculo.

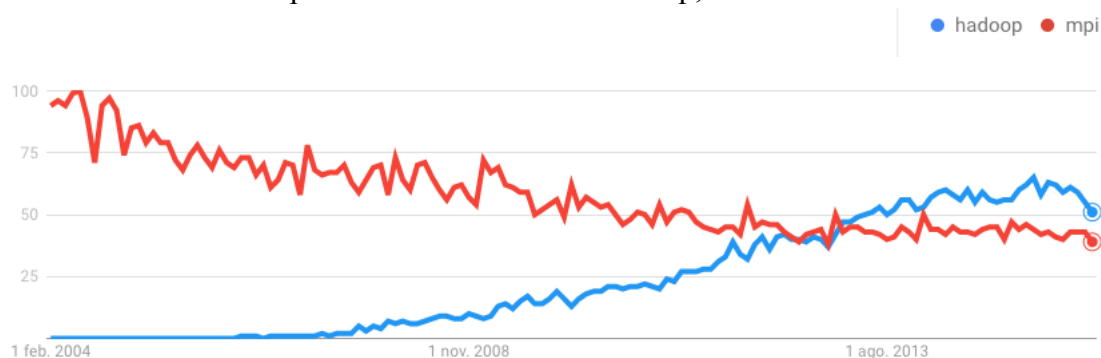
Este sistema es muy efectivo, confiable y altamente escalable, que funciona en todos los servidores, está disponible para los usuarios de software. Entre las principales aplicaciones de la plataforma Web búsqueda para la que la línea de base y desarrollado (con la compañía de Yahoo) [BIF12].

Hadoop- sistema es un conjunto gratuito de herramientas de software para el desarrollo y ejecución de programas distribuidos que se ejecutan en grupos de varios cientos de miles de nodos. Hadoop proporciona redundancia en caso de fallo de nodos, el sistema es compatible con múltiples copias de los datos de producción. El procesamiento de datos duplicados puede ser reasignados a los nodos de trabajo. Hadoop se basa en el principio de procesamiento en paralelo - esto le permite aumentar la velocidad. El volumen de información procesada petabytes medidos.

Se han realizado estudios comparativos como [REY15], donde se ha evaluado el uso de Hadoop frente a modelos más tradicionales basados en MPI. Aunque Hadoop pueda mantener datos en memoria, aún están lejos de lograr el rendimiento de las tecnologías HPC de última generación. Sin embargo, Hadoop puede preferirse porque también:

- ofrece un sistema de archivos distribuido con administración de fallos y replicación de datos.
- permite añadir de nuevos nodos en tiempo de ejecución.
- proporciona un conjunto de herramientas para el análisis y la gestión de datos que es fácil de usar, implementar y mantener.

En la siguiente figura de Google Trends se muestra desde 2004 hasta 2015 la creciente tendencia en el interés por herramientas como Hadoop, frente a MPI.



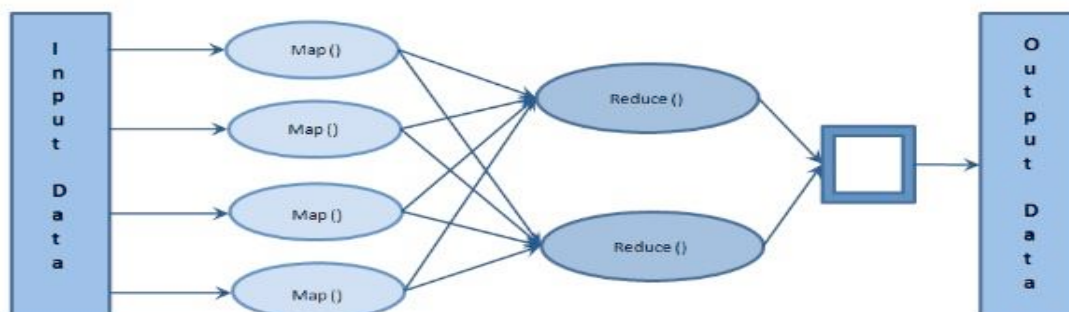
**Figura 15 Demostración modelo MapReduce.**

La estructura de Hadoop se basa en dos elementos:

1. Un sistema de archivos distribuido de Hadoop (HDFS): es un sistema de archivos desarrollado para un proyecto Hadoop que adoptó la arquitectura maestra / esclavo. HDFS consta de un NameNode y numerosos DataNodes. HDFS proporciona a los usuarios el espacio de nombres de archivo correspondiente para almacenar datos en un formato de archivo. En general, HDFS divide estos archivos en varios bloques de archivos, que se almacenan en un grupo de servicios de datos. Luego, NameNode proporciona funciones fundamentales como abrir, cerrar y renombrar los archivos y directorios, a la vez que es responsable de asignar los bloques de archivos a los DataNodes. Luego, DataNode es responsable de responder a las operaciones de lectura y escritura de archivos concretos en el lado del cliente, mientras se ocupa de los requisitos para establecer, eliminar y realizar copias de seguridad de los bloques de datos lanzados por NameNode. En la estructura típica de topología de clúster de Hadoop, hay es otro servidor residente en el NameNode secundario para evitar que el residente del NameNode del servidor se desplome del clúster como resultado de una falla. NameNode y JobTracker se pueden organizar en el mismo nodo maestro en grupos pequeños con menor presión para el procesamiento de datos, mientras que, en

los grupos grandes, es mejor organizar los datos en dos servidores diferentes debido a los intercambios de datos más frecuentes. DataNode y TaskTracker se organizaron en cada nodo informático como nodo esclavo responsable del almacenamiento y la computación.

2. Plataforma de programación y ejecutar un trabajo de cómputo distribuido, diseñado para trabajar con grandes volúmenes de datos MapReduce. El principio de MapReduce consiste en dos elementos. El primer elemento (Map) es la primera transformación de datos: el nodo maestro recibe la entrada, los dividen en partes y las transferencias a otros equipos, nodo trabajador, de pretratamiento. El segundo elemento (Reduce) es la agregación de los datos procesados: el nodo maestro recibe respuestas unidades operativas y produce un resultado.



*Figura 16 Modelo MapReduce.*

Además, el sistema también incluye módulos comunes Hadoop (conjunto de programas de infraestructura y componentes, software middleware) y YARN (sistema de planificación de tareas y administración del clúster).

Hadoop se utiliza a menudo para el tratamiento de una variedad de «Big Data». La eficacia en el tratamiento de manifiesto Hadoop «Big Data», también el sistema se utiliza para agregar los masives de datos.

## 4.2 Hadoop y seguridad en Big Data.

Información - es poder, y cada vez más especialistas IT usar el poder de Big Data para una mejor comprensión de las fuerzas impulsoras detrás de sus empresas. Mientras tanto, la información sigue acumulando, la cantidad de datos se duplica cada año.

Aproximadamente el 80% de los datos recogidos se refieren a los datos no estructurados y se debe formatear utilizando una plataforma de procesamiento por lotes, tales como Hadoop, que se puede recuperar la información. Las empresas quieren aprovechar las importantes conclusiones que pueden extraerse de los datos recogidos por ellos. Sin

embargo, su gran peligro: sistema de Hadoop se creó sin tener en cuenta los requisitos de seguridad. La cuestión de la protección que se presentó posteriormente. Debido a la creciente popularidad de Hadoop a la empresa de sus deficiencias en materia de seguridad son cada vez más evidentes.

Hadoop fue creado originalmente sin tomar en cuenta las necesidades de las empresas, por no hablar de la seguridad de la empresa. El objetivo de Hadoop fue la información de gestión, tales como enlaces en Internet, y que estaba destinado a dar formato a grandes cantidades de datos no estructurados en un entorno de computación distribuida, en concreto - en el entorno de Google. No fue creado para apoyar una mayor seguridad, controles de cumplimiento, cifrado, políticas de uso y gestión del riesgo.

Para la autenticación, Hadoop utiliza Kerberos. Sin embargo, este protocolo puede ser difícil de implementar. Además, no cumple una serie de requisitos de seguridad de la empresa, como la autenticación basada en roles, es compatible con LDAP y Active Directory para la participación política. Más Hadoop no es compatible con el cifrado de los datos en los nodos o en la transmisión entre los nodos.

Tecnologías de seguridad de datos tradicionales se basan en el concepto de la protección de una unidad física independiente (o servidor de base de datos), y no se distribuyen de forma única los entornos informáticos para Big Data representativos de las agrupaciones de Hadoop. Técnicas de seguridad convencionales son ineficaces en un entorno distribuido a gran escala.

Distribuida estructura de grupos Hadoop hace que muchos de los métodos tradicionales y políticas ineficaces de copia de seguridad y restauración de datos. El uso de Hadoop empresas debe producir la replicación, copia de seguridad y almacenamiento de datos en un entorno seguro independiente.

Para aprovechar las ventajas de Big Data, Hadoop se utiliza en combinación con otras tecnologías como Hive, HBase y Pig. Estas herramientas proporcionan acceso a grandes volúmenes de datos y el uso, la mayoría de ellos también protege a la empresa. Fortalecimiento de la seguridad directamente Hadoop - sólo una parte de los problemas más grandes de protección de datos.

Para Big Data hay un conjunto especial de requisitos. Independientemente de IT, que se utiliza para el almacenamiento y gestión de datos, las empresas deben cumplir con los requisitos de las autoridades reguladoras en materia de privacidad y protección de datos.

Las grandes compañías que utilizan Hadoop, proporciona controles y restricciones de acceso adicionales sobre el número de empleados con acceso a Big Data. Para proteger los datos requeridos software adicional. Esto será necesario, siempre y cuando no habrá información, eliminando vulnerabilidades entorno Hadoop. Las organizaciones deben analizar regularmente IT-ambiente para determinar las vulnerabilidades.

### **4.3 Solución de problemas de seguridad en Hadoop.**

Solución Sentry es compatible con el modelo creado previamente de acceso basado en roles de llamadas (RBAC) (Role-based Access Control), que funciona "por encima" de la plantilla, que es característica de las bases de datos relacionales (base de datos, tablas y etc.).

Modelo RBAC soporta varias funciones necesarias para proteger el entorno corporativo de Big Data. La primera función - una autenticación segura, que proporciona datos de control de acceso obligatorio a los usuarios autenticados.

Los usuarios se asignan a los roles, y luego presentó a las competencias respectivas de acceso a los datos. Este enfoque permite el uso de plantillas de modelo de escala, dividir a los usuarios en categorías según sus funciones.

Esto ahorra administradores de tener una asignación detallada de competencias a cada usuario. Además, esta característica simplifica la gestión de permisos y reduce la carga sobre los administradores, así como para reducir al mínimo la posibilidad de errores y el acceso inadvertido.

La siguiente función le permite organizar la administración de credenciales de usuario con el fin de distribuir la tarea entre varios administradores a nivel de circuito o en el nivel de base de datos. El control de gestión sobre el acceso de datos se puede realizar en el nivel de base de datos. Por ejemplo, una función puede permitir al usuario utilizar los datos de las operaciones de "select", y la otra función es permitir la aplicación para "insert" los datos de la operación (a nivel de servidor, los datos y las tablas de base de datos).

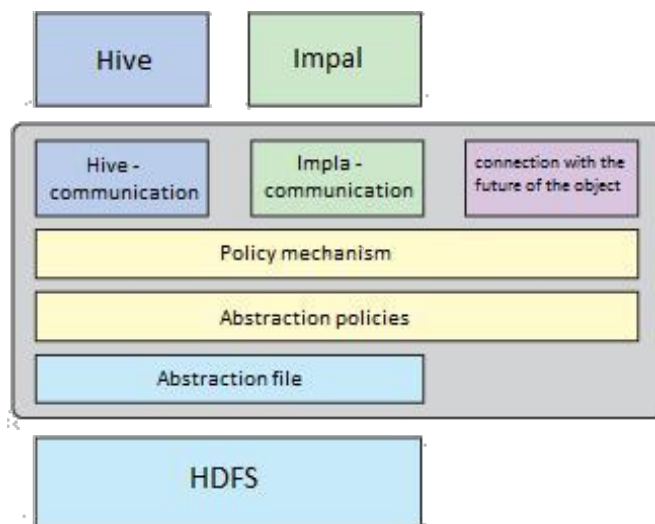
### **4.4 Soluciones de aplicación Sentry en un entorno con HDFS, Hive y Impala.**

Los servicios de Internet de rápido crecimiento, como Google, Yahoo, Amazon y Facebook, son un estilo representativo de aplicaciones de uso intensivo de datos. Utilizan infraestructuras orientadas a big data como plataformas de computación en la nube para servicios escalables. La calidad de servicio (QoS) de los sistemas de archivos distribuidos (DFS) subyacentes contribuye a la confiabilidad y el rendimiento de estas aplicaciones. Por lo tanto, existe una fuerte demanda de modelar el rendimiento de estos sistemas de archivos a escala de Internet como los Sistemas de archivos de Google (GFS), el Servicio de almacenamiento simple de Amazon (S3), el Sistema de archivos distribuidos de Hadoop (HDFS) y OpenStack Swift, para proporcionar un rendimiento predecible y una guía de configuración práctica para mejorar el



rendimiento. HDFS se ha convertido en un representante típico de los datos. DFS intensivo. En general, el rendimiento de HDFS está influenciado por varios factores, como la implementación física del clúster, la E / S de disco, el tráfico de red, las opciones de configuración y los patrones de acceso a datos. [TIA17]

Figura 16 muestra las soluciones básicas arquitectura de Sentry. Esta arquitectura se diseñó sobre la base de la capacidad de extensión (para soportar una amplia variedad de aplicaciones basadas en Hadoop) y la tolerancia (por diversas formas de los proveedores de datos).



**Figura 17 La arquitectura Sentry.**

Hasta la fecha, en el proyecto de Cloudera Impala compatible con muchos mecanismos que se utilizan comúnmente para el código abierto para realizar consultas SQL, incluyendo Apache Hive (por HiveServer 2 RPC-interfaz) y para Cloudera Impala. Cada aplicación está protegida por un conjunto de enlaces, implementado para esta aplicación particular. Estos interactúan con el mecanismo de fijación de políticas con el fin de evaluar y validar la política de seguridad predeterminada y, a continuación, después de acceder trabajo a través de la política de la abstracción con el fin de obtener acceso a los datos. La abstracción moderna basada en la integración de archivos HDFS apoyo o el acceso al sistema de archivos local de acuerdo con la política de seguridad.

Hive es compatible con el SQL, que además utiliza una gran cantidad de escritura habitual y, en algunos casos, del programa secreto que podríamos tener que verse obligados a implementar la programación de Map Reduce. Tenemos que usar Hive para interrumpir la información del conjunto de acciones y grandes conocimientos, en ese momento podríamos tener la pregunta primaria basada en el cálculo relativo avanzado y de la entidad para usar las habilidades SQL de Hive-QL y la información conectada se supervisa durante un mapa específico y mapeo de retroescala. depreciará el tiempo de avance y puede administrar uniones entre el conjunto de datos (por ejemplo, Información sobre acciones, datos industriales). Hive además tiene sus servidores principales, por lo que regalaremos nuestras consultas Hive desde cualquier lugar al servidor Hive, que se emplea para ejecutarlas. [SAS18]

Sentry ofrece una autorización detallada de la tarea de los controles de seguridad para el servidor, base de datos, tabla, y para la presentación, incluyendo características tales como especificar una referencia selectiva para las representaciones y mesas, mesas de inserción, y las potencias de la conversión de los servidores de la oficina. Cada base de datos y cada diagrama pueden tener la política de autorización por separado. Además, Sentry proporciona soporte para la arquitectura repositorio de meta Hive.

Con el fin de apoyar un alto grado de extensibilidad solución Sentry protege nuevas aplicaciones como el Apache Pig (a través de la configuración de Pig fijaciones), y el acceso a las nuevas abstracciones para su uso en las políticas de seguridad (por ejemplo, bases de datos). Todas estas características se implementan como interfaces de “plug-in”.

## **4.5 Otros aspectos de la seguridad en el entorno Hadoop.**

Sentry es una infraestructura de autorización basado en funciones, pero no es la única innovación en el campo de la seguridad, que apareció recientemente en Hadoop. Hay otro diseño moderno en el ámbito de la protección y el control de acceso a datos de Big Data.

## **4.6 Proyecto Rhino.**

Este es proyecto de código abierto. El proyecto Rhino desarrollado por Intel para mejorar la plataforma Hadoop con mecanismos de protección adicionales. El objetivo de este proyecto es eliminar los agujeros de seguridad en la pila de Hadoop. Para este propósito, Intel está desarrollando en varias direcciones y la seguridad se centra en la funcionalidad criptográfica. Entre la variedad de trabajos realizados en el marco del Proyecto de Rhino, las nuevas características más interesantes para el cifrado / descifrado múltiples modelos de uso. Por ejemplo, añadir una capa de abstracción común para codec criptográfica implementa la API de la interfaz para apoyar esta capacidad desarrollada entorno adecuado para la distribución y gestión de claves.

También en el desarrollo es un sistema de archivos de cifrado Hadoop (con se llama CFS Hadoop), que proporcionará los servicios de cifrado de bajo nivel para los archivos dentro de HDFS. En este nivel, cualquier usuario podrá utilizar sin problemas nuevas herramientas de seguridad de datos de Hadoop (desde aplicaciones MapReduce hasta Hive, Apache HBase y Pig).

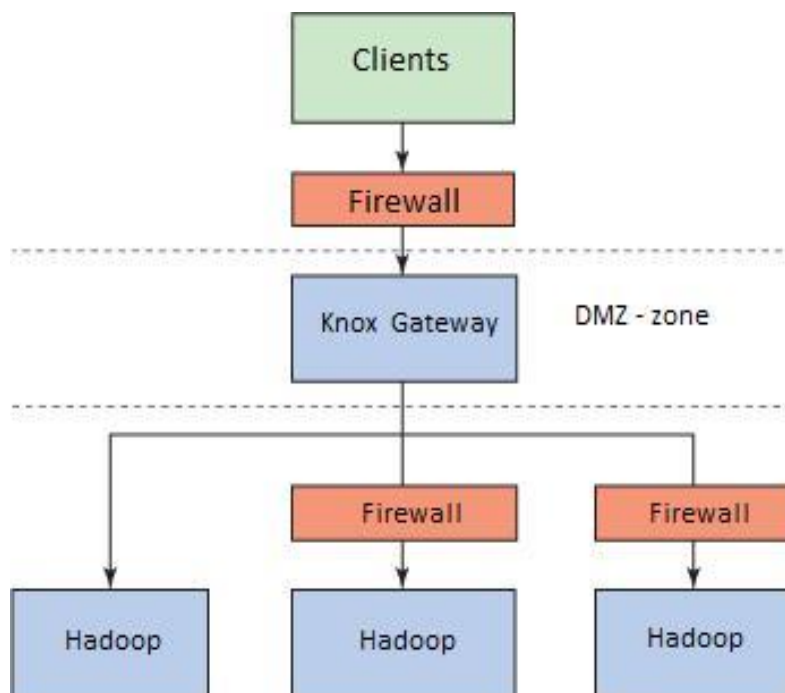
En desarrollo son servicios tales como instantáneas de cifrado y los registros de transacciones transparentes en el disco, así como nuevas oportunidades para las

funciones de apoyo de arranque del cerdo y de almacenamiento, incluyendo capacidades de cifrado.

## 4.7 Apache Knox Gateway.

Apache Knox Gateway - una solución que protege el perímetro de Hadoop. A diferencia de las soluciones de Sentry, que proporciona herramientas para la solución de control de acceso a datos detallados Knox Gateway proporciona control de acceso a los servicios de la plataforma de Hadoop. El propósito de Knox Gateway - para proporcionar un único punto de acceso seguro a Hadoop-clusters. Esta solución se implementa en una puerta de enlace (gateways o un pequeño racimo) que da acceso a los racimos con el reposo interfaz de estilo API de Hadoop (Representational State Transfer).

Esta pasarela es compatible con un servidor de seguridad entre grupos de Hadoop y usuarios (Figura17) y le permite controlar el acceso a las agrupaciones en las que diferentes versiones de Hadoop ejecutados.



*Figura 18 La protección del perímetro de Apache Knox.*

Knox Gateway Solution complementa Sentry, la aplicación de nivel de acceso externo de protección.

A medida que la puerta de entrada a la DMZ - zone de la disolución Knox Gateway proporciona acceso controlado a uno o más de Hadoop-grupos, separados por cortafuegos de red.

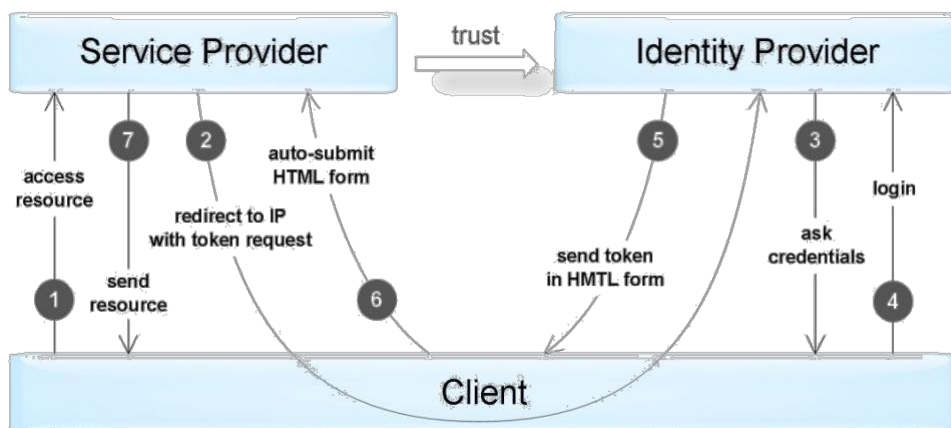
## 4.8 Tokens de Autenticación.

Este método de autenticación es la más utilizada en la construcción de sistemas distribuidos de sesión único Single Sign-On (SSO), donde una sola aplicación (service provider o relying party) autentica de usuario delegado a otra aplicación (identity provider o authentication service).

La implementación de este método reside en el hecho de que identity provider (IP) proporciona información fiable sobre el usuario en forma de un token, service provider (SP) aplicación utiliza esta señal para la identificación, autenticación y autorización de usuarios.

### ***Todo el proceso es el siguiente:***

1. El cliente se autentica en identity provider con uno de los métodos que le son propias (contraseña para acceder a la clave, el certificado, el Kerberos, etc ..).
2. Cliente solicita identity provider que le proporcionen una ficha para un SP-aplicación. Identity provider genera una señal y la envía al cliente.
3. El cliente se autentica en el SP-aplicación que utiliza este token.



**Figura 19** La autenticación del cliente con token "activa".

Este proceso refleja el mecanismo de autenticación del cliente activo, es decir, uno que puede realizar una secuencia programada de acciones (por ejemplo, el IOS / app Android). Browser - cliente pasivo en el sentido de que sólo puede mostrar la página de usuario solicitado.

En este caso, la autenticación se logra mediante la reorientación automáticamente el navegador entre el proveedor de aplicaciones web identity proveedor y service provider.



**Figura 20 Autenticación "pasiva" mediante la reorientación de las solicitudes del cliente.**

Existen varios estándares de definir con precisión la interacción entre los clientes de protocolo (activo y pasivo) y la aplicación IP/SP y fichas formato compatible. Entre las normas más populares OAuth, OpenID Connect, SAML, y WS-Federation.

El token es una estructura de datos que contiene la información que genera la señal que puede ser el receptor de la señal, la validez de un conjunto de información sobre el usuario (claims). Además, más token se utilizan para evitar cambios no autorizados, y ofrecer garantías de autenticación.

Cuando la autenticación utilizando token SP-aplicación debe realizar las siguientes comprobaciones:

1. El token fue emitido por una aplicación de identity provider de confianza (verificación de issuer).
2. El token significaba que el SP-aplicación actual (verificación de audiencia).
3. La validez de la ficha no ha expirado (verificación de expiration date).
4. El token es genuino y no se ha cambiado (verificación de la firma).

## 4.9 Formatos de tokens.

Hay varios formatos comunes de tokens:

1. Simple Web Token (SWT) — El formato más simple es un conjunto de nombre / valor en el formato de codificación HTML form. La norma define varios nombres reservados: Issuer, Audience, ExpiresOn y HMACSHA256. El token se ha firmado con una clave simétrica, por lo tanto, IP y la aplicación SP debe tener esa llave para permitir la comprobación de token.

Ejemplo SWT -token (después de la decodificación)

```
Issuer=http://auth.myservice.com&
Audience=http://myservice.com&
ExpiresOn=1435937883&
UserName=John Smith&
UserRole=Admin&
HMACSHA256=KOAQRRSpy64rvT4KnYyQKtKFXUIggnespE7ADA4o9w
```

2. JSON Web Token (JWT) - contiene tres bloques separados por puntos: un título, un conjunto de campos (activos) y firma. Los dos primeros bloques están representados en JSON-formateados y está codificado adicional en formato base64. Los campos contienen pares de nombre / valor arbitrario, estándar JWT define varios nombres reservados (iss, aud, exp y etc.). La firma puede ser generado con la ayuda y algoritmos de cifrado simétrico y asimétrico. Además, hay una norma separada, que describe el formato cifrado de JWT - token.

Ejemplo JWT-token (después de la decodificación bloques 1 y 2 )

```
{ «alg»: «HS256», «typ»: «JWT» }.
{ «iss»: «auth.myservice.com», «aud»: «myservice.com», «exp»: «1336938983»,
«userName»: «John Smith», «userRole»: «Admin» }.
S9Zs/8/rHEGTAVtLggFTizCVMtxOwneQjae2B3UAhdP
```

3. Security Assertion Markup Language (SAML) — especifica el token (SAML assertions) en formato XML, que incluye información sobre el emisor de la materia, las normas necesarias para la clave de verificación de token, un conjunto de demandas adicionales sobre el usuario. Firma del token hechas usando criptografía asimétrica SAML. Además, a diferencia de los formatos anteriores, SAML token contiene un mecanismo para verificar el contador del tiempo, lo que impide la interceptación de señales a través del ataque Man-in-the-Middle cuando se utilizan conexiones seguras.

### 4.9.1 SAML estándar.

Estándar de Security Assertion Markup Language (SAML) describe los métodos y protocolos de interacción entre el identity provider y service provider para el intercambio de datos por medio de autenticación y autorización token.

Esta norma fundamental - lo suficientemente difícil y es compatible con una gran cantidad de diferentes escenarios de integración de sistemas. Los básicos "bloques de construcción" de la norma:

1. Assertions - SAML formato de tokens a XML.
2. Protocols - se les hace una serie de comunicaciones soportado entre los participantes para crear un nuevo token, token de conseguir, cierre de sesión, gestión de identidades de usuario existente.
3. Bindings - mecanismos a través de varios protocolos de transporte de mensajería. Soportados HTTP Redirect, HTTP POST, HTTP Artifact, SAML SOAP, SAML URI.
4. Profiles - escenarios norma define un conjunto de assertions, protocols y bindings necesarios para su ejecución, lo que permite una mejor compatibilidad.

Además, la norma define el formato de intercambio de información entre los usuarios, que incluye una lista de funciones admitidas protocolo atributos claves de cifrado y etc.

Un ejemplo del uso escenario SAML para Single Sign-On. El usuario quiere acceder a un proveedor de recursos de seguro. Dado que el usuario no ha sido autenticado, SP envía al sitio "identity provider" para crear un token.

```
HTTP/1.1 302 Found
Location: https://idp.example.com/SAML/SSO/Browser?SAMLRequest=aQC5kAsTB...3QIX3
f7M&RelayState=kzdeP2qmdr576&SigAlg=http%3A%2F%2Fwww.w3.org%2F2000%2F09%2Fxmldsig
%23rsa-sha1&Signature=OPqCEvoc8Vz0NsQdctKbzLBNj74
```

- URI
- SAMLRequest — del servicio identity provider
- RelayState —
- SigAlg — solicitud de creación de un nuevo token cadena arbitraria que
- Signature — describe SP algoritmo de firma mensajes

firmar mensajes

*Figura 21 Respuesta a SP.*

En el caso de una solicitud, identity provider autentica al usuario.

```

...
<body onload="document.forms[0].submit()">
<form method="post" action="https://sp.example.com/SAML/SSO/POST">
  <input type="hidden" name="SAMLResponse" value="XHU2KeRbb... 94X/rX4" />
  <input type="hidden" name="RelayState" value="kzdeP2qmdr576" />
  <input type="submit" value="Submit" />
</form>
...

```

del servicio indentity provider

- URI
- **SAMLResponse** — solicitud de creación de un nuevo token cadena arbitraria que
- **RelayState** — describe SP

*Figura 22 Respuesta a IP.*

Después de que el navegador enviará automáticamente este formulario para el proveedor de servicio web, que decodifica la señal y autentica al usuario.

## 4.9.2 Estándar WS-Trust y WS-Federation.

WS-Trust y WS-Federation incluye un grupo estándares WS-\*, que describe servicios SOAP / XML.

Estos estándares son desarrollados por un grupo de empresas que incluye Microsoft, IBM, VeriSign y otros. Junto con el SAML, estas normas son bastante complejos, que se utiliza principalmente en escenarios empresariales.

Estándar WS-Trust describe el servidor de autenticación de la interfaz, con se llama Secure Token Service (STS). Este servicio funciona en el protocolo SOAP y apoya la creación, renovación y revocación de tokens. Este estándar permite el uso de fichas de diferentes formatos, pero en la práctica se utilizan principalmente SAML token.

Estándar WS-Federation describe el mecanismo de servicio de la interacción entre las empresas, en particular tokenov.WS-Federación de intercambio de protocolos amplía las funciones de servicio STS, tal como se describe en el estándar WS-Trust. Entre otras cosas, el estándar WS-Federation define:

- El formato y los métodos de intercambio de metadatos acerca de los servicios.
- La función de una sola salida de todos los sistemas (single sign-out).
- Atributos de servicio que proporciona información adicional sobre el usuario.
- Soporte para clientes pasivos (navegadores) redireccionando

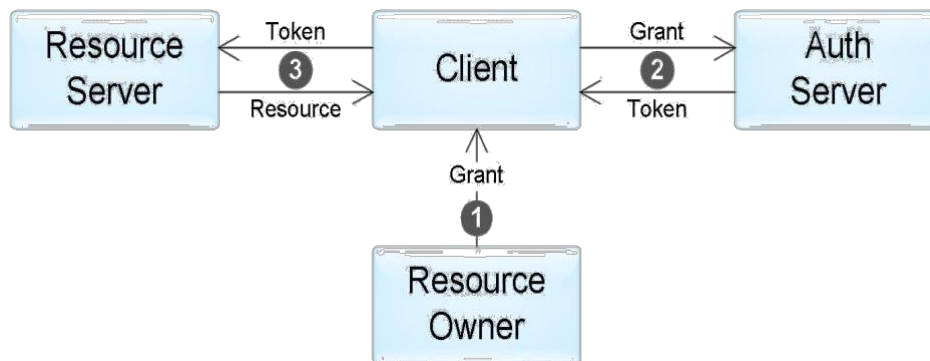


### 4.9.3 Estándar OAuth y OpenID Connect.

Por el contrario, SAML y WS-Federation, estándar OAuth (Open Authorization) no se describe el protocolo de autenticación de usuario. En lugar de ello, se define un mecanismo para acceder a una aplicación a otro nombre de usuario.

Por ejemplo, considere una aplicación web que ayuda a los usuarios planificar viajes. A medida que la funcionalidad, es capaz de analizar usuarios de correo electrónico a la presencia de las letras con una reserva confirmada y automáticamente incluirlos en la ruta planificada. Como una aplicación web se puede acceder de forma segura al correo electrónico del usuario Esto resuelve el estándar OAuth: él describe cómo la aplicación de viaje (client) puede acceder al buzón del usuario (resource server) con el permiso del usuario (resource owner). Todo el proceso consiste en varios pasos:

- El usuario (resource owner) autoriza la aplicación (client) para acceder a un recurso en particular en forma de una grant.
- La aplicación tiene acceso a la autenticación del servidor token y consigue el acceso al recurso, a cambio de una grant.
- La aplicación utiliza este token para obtener los datos requeridos desde el servidor de recursos.



**Figura 23** Recibir datos desde el servidor.

- Authorization Code - el usuario puede obtener una subvención del servidor de autorización después de la autenticación exitosa y la confirmación del consentimiento para el acceso. Este método se utiliza con mayor frecuencia en las aplicaciones Web. El proceso de obtención de una subvención es muy similar al mecanismo de autenticación de cliente SAML pasiva y WS-Federation.

- Implicit - se utiliza cuando la aplicación no es posible obtener con seguridad un símbolo desde el servidor de autenticación (por ejemplo, navegador compatible con JavaScript aplicación). En este caso el contador de subvención se recibe desde el

servidor de autenticación.

- Resource Owner Password Credentials - grant es un par de usuario nombre de usuario / contraseña. Se puede utilizar si la aplicación es la "interfaz" para los recursos del servidor

- Client Credentials - en este caso, no hay ningún usuario en la aplicación tiene acceso a sus recursos mediante el uso de su clave de acceso.

El estandar no define el formato de token, la cual recibe la solicitud: en escenarios direccionables aplicación estándar no necesita analizar el token, es decir, a la que sólo se utiliza para obtener acceso a los recursos ... Por lo tanto, ni token ni grant sí mismo no puede ser utilizado para la autenticación de usuario. Sin embargo, si una aplicación necesita obtener información fiable sobre el usuario, hay varias maneras de hacer esto:

- Servidores de la API incluye proporcionar información sobre el usuario (por ejemplo, Facebook API). Una aplicación puede realizar esta operación cada vez que después de recibir la señal para identificar al cliente.

- Utilice estándar OpenID Connect. De acuerdo con este estandar, Autorización Server proporciona un token adicional. Este token JWT en el formato contendrá un conjunto de campos específicos (claims) a la información del usuario.

## **4.10 Autenticación del cliente en Hadoop.**

Hadoop ofrece los siguientes comandos: init, renew, expire, validate y etc. ..

Token "init" del sistema de inicio de nombre de usuario y las credenciales para autenticar a TAS y obtiene token de identidad.

Token "result" se mantuvo en la memoria caché de token para un uso posterior, estando disponible para los componentes transversales, para apoyar de sesión único.

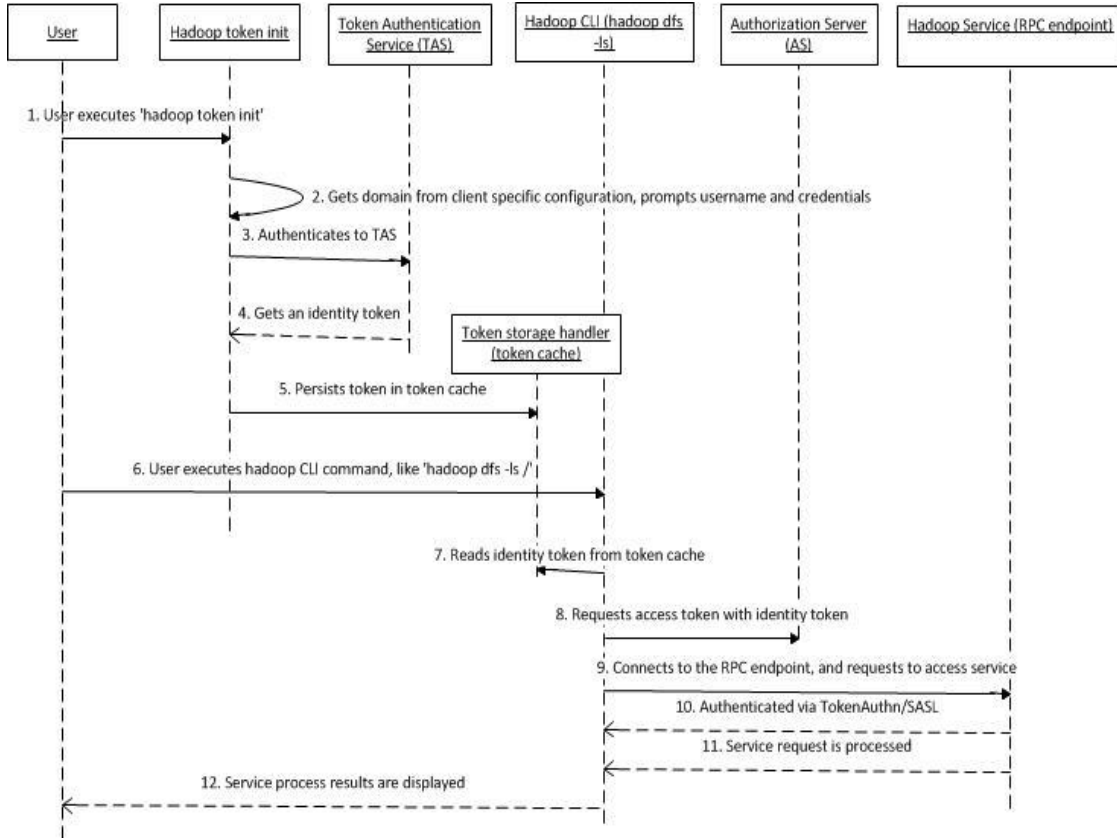
Token "renew" comando extiende el ciclo de vida del token.

Token "expire" comando expire señal caduca contador antes de que llegue hasta el final.

Token "validate" se comprueba el token es válido o no.

### - Acceso Hadoop RPC

La siguiente figura muestra cómo se intercambian en Hadoop el token de inicio y comandos CLI para acceder a un servicio a través de Hadoop extremo de RPC.



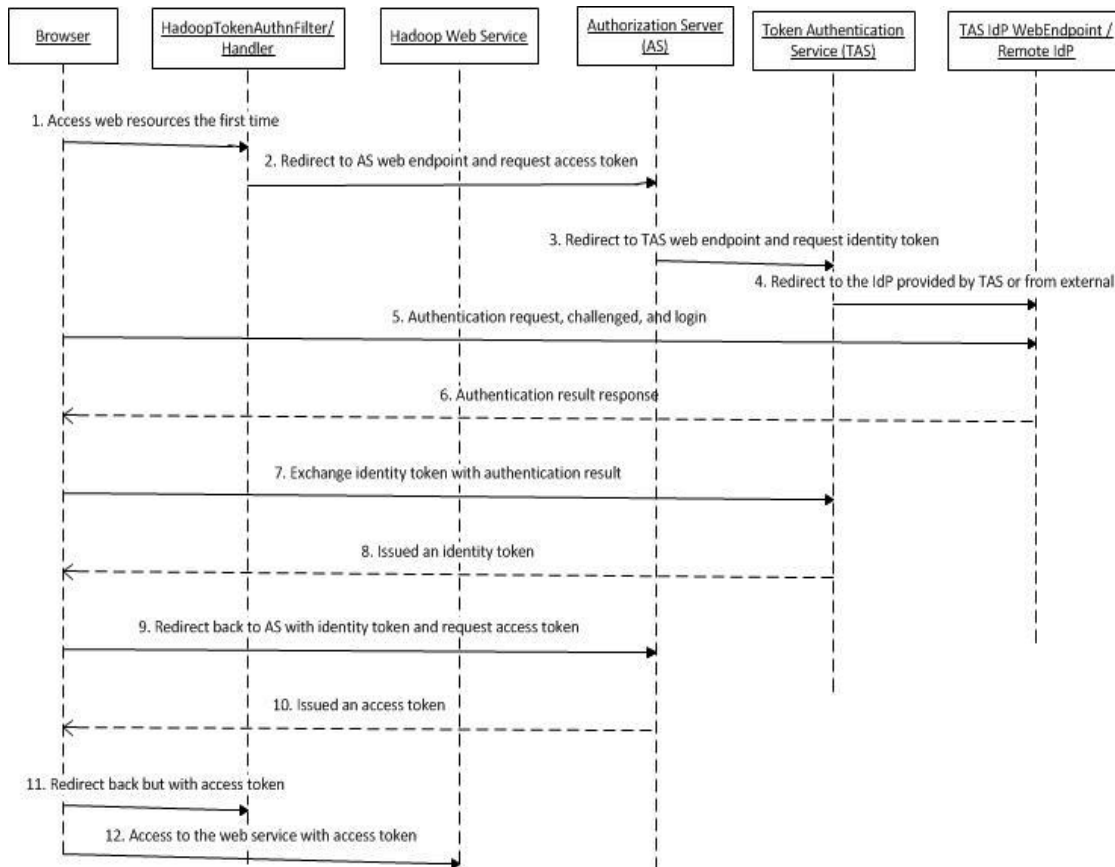
**Figura 24 Acceso Hadoop RPC.**

### - Acceso Hadoop REST

En Hadoop hay comandos del cliente que utilizan la clase HttpURLConnection Java para acceder a un servicio REST Hadoop. Comparando con el acceso RPC, las diferentes partes. Se conecta a través de servicio REST HTTP / HTTPS, en lugar de RPC; que la autenticación para descansar servicio a través de algún proceso como SPNEGO, en lugar de método simbólico authn / SASL.

### - Hadoop acceso navegador web

La siguiente figura muestra cómo el uso del navegador del usuario para autenticar el acceso y el servicio Hadoop en la interfaz web. Este es un caso típico usuario de SSO web, adaptado para Hadoop y el marco.



**Figura 25 Hadoop acceso navegador web.**

En caso de que el backend de identidad para el módulo de autenticación no admite página de registro y flujo de SSO web, por ejemplo, LDAP en cuestión, TAS ofrece punto final web como un IdP.

### 4.11 Tecnología de tokens delegado.

El desarrollo de la Internet de las Cosas (IoT) y el Sistema Ciberfísico (CPS) ha facilitado en gran medida muchos aspectos de las aplicaciones tecnológicas y el desarrollo. Esto puede conducir a un crecimiento significativo de los datos, especialmente para archivos pequeños. El análisis y procesamiento de un gran número de archivos pequeños se ha convertido en una parte crucial del desarrollo de IoT. Los sistemas de archivos distribuidos Hadoop se han convertido en poderosas plataformas para almacenar una mayor cantidad de big data. [WEI18]

Apache Knox Gateway y Sentry ofrece protección perimetral y la protección de acceso a datos. Además, es necesario proteger el acceso a HDFS-datos de las tareas de MapReduce. Una de estas decisiones se basa en el concepto de símbolo delegado (delegation token). Protocolo de autenticación de dos vías tecnología basada en token de delegado que permite que un usuario se autentifique ante nodo NameNode (utilizando Kerberos); después de recibir el token de delegación, el usuario puede proporcionar este símbolo en el nodo JobTracker, lo que resulta en Hadoop-asignación para este usuario será capaz de utilizar este token para el acceso seguro a los datos dentro de HDFS.

Tokens de delegado se basan en la autenticación de dos vías, que es más fácil y más eficiente que el de tres vías de autenticación se utiliza en Kerberos. Esto minimiza el tráfico de Kerberos diferencia y mejora la escalabilidad y reduce la carga sobre los activos de Kerberos.

### **4.12 Ventajas de tokens.**

- Si no se introduce la contraseña manualmente. Al utilizar eToken Single Sign-On usuario no hace la información de entrada desde el teclado. Todos los campos, incluidos los campos de contraseña se rellenan automáticamente. Esto reduce el riesgo de robo de contraseñas.
- Capacidad de utilizar contraseñas largas y complejas. Puesto que el usuario no introduce manualmente una contraseña, la contraseña en sí puede ser más largo y más complejo que el usuario puede recordar.
- El uso de la contraseña generada al azar, desconocido para el usuario. eToken Single Sign-On le permite generar contraseñas aleatorias de hasta 14 símbolos, guardarlos en la memoria eToken y el sustituto en la forma de cambiar la contraseña para que el usuario ni siquiera sabe su nueva contraseña, y por lo tanto no se puede escribir, y por lo tanto comprometerlo.
- Autenticación de dos factores. eToken Single Sign-On sustituye a la autenticación de dos factores en aplicaciones en las que el usuario sólo necesita saber los detalles, de dos factores, que requiere la presencia de un dispositivo de hardware (eToken) y el conocimiento del código PIN. Esto aumenta en gran medida la seguridad.
- La simplicidad y comodidad para los usuarios. Al utilizar eToken Single Sign-On usuarios no tienen que recordar los valores de varios campos que necesitan ser llenados en diferentes aplicaciones de Windows. En lugar de ello, sólo tenemos que recordar código PIN.

### **4.13 Las características exclusivas de tokens.**

- eToken Single Sign-On – es un producto universal que se puede utilizar con la gran mayoría de aplicaciones para Microsoft Windows.

- eToken Single Sign-On integración con eToken TMS (Token Management System), llave USB ciclo de vida de la gestión del sistema y las tarjetas inteligentes. Esto permite la gestión centralizada de eToken almacenados en los perfiles de memoria (para realizar su copia de seguridad, recuperación, replicación).



## 5. Experimentos.

El crecimiento de los datos continua y aparecen nuevos problemas de procesamiento y almacenamiento. Sin embargo, las publicaciones teóricas no siempre está claro cómo utilizar las tecnologías apropiadas para resolver problemas específicos de carácter práctico. Uno de los proyectos más conocidos en el campo de la computación distribuida es Hadoop. Desarrollo del Apache Software Foundation. El conjunto de libre disposición de herramientas, bibliotecas y el marco para el desarrollo y ejecución de programas de computación distribuida.

El autor de Hadoop Dug Katting ha desarrollado el sistema Nutch. El sistema buscador con código abierto. Nutch proyecto se inició en 2002, pero muy pronto se dio cuenta de que sus desarrolladores la arquitectura existente es poco probable que escalar a miles de millones de páginas web. En 2003, se publicó un artículo que describe el modelo GFS (Google File System) Sistema de archivos distribuido, utilizado en proyectos de Google. Tal sistema podría fácilmente hacer frente al problema de almacenar grandes archivos generados al rastrear e indexar los sitios. En 2004, el equipo de desarrollo Nutch tomó la implementación de un sistema de este tipo es de código abierto - NDFS (Nutch Distributed File System). En 2004, Google introdujo la tecnología MapReduce. Nutch ya desarrolladores en el inicio de 2005 crearon un Nutch basado en MapReduce; Pronto todos los algoritmos básicos Nutch han sido adaptados para su uso MapReduce y NDF. En 2006, Hadoop se ha asignado a un sub-proyecto independiente dentro del proyecto de Lucene. En 2008, Hadoop se ha convertido en uno de los principales proyectos de Apache. Por el momento ya ha sido utilizado con éxito en empresas como Yahoo, Facebook y Last.fm. Hoy en día, Hadoop es ampliamente utilizado en aplicaciones comerciales.

### 5.1 La estructura de Hadoop.

La composición de Hadoop incluye los siguientes subproyectos:

- Common - un conjunto de componentes y interfaces para sistemas de archivos distribuidos, y una entrada-salida;
- Map Reduce - modelo de computación distribuida, diseñado para la computación paralela en volúmenes muy grandes de datos;
- HDFS - El sistema de archivos distribuido Hadoop (HDFS) está instalado en los clústeres para almacenar, procesar y administrar archivos de Big Data. Los archivos en HDFS se dividen en múltiples bloques de datos y se almacenan en diferentes nodos de datos que pueden acceder a ellos en paralelo, lo que permite dar una velocidad más rápida para el procesamiento de datos [SHW18].



## 5.2 Distribuciones de Hadoop.

Hadoop es un sistema complejo que consiste en un gran número de componentes. Instalar y configurar un sistema es complejo. Por lo tanto, muchas empresas ofrecen ahora distribuciones funcionales de Hadoop, entre ellos el de las herramientas de implementación, administración y monitorización. Distribuciones de Hadoop distribuye como comerciales (productos de compañías como Intel, IBM, EMC, Oracle), y licencias libres (productos de la compañía Cloudera, Hortonworks y MAPR).

## 5.3 Cloudera Hadoop.

Cloudera Hadoop es un sistema con código abierto creado por la participación de Doug Cutting y Mike Kafarely. Se aplica tanto en libre como en la versión de pago, conocido como Cloudera Enterprise.

### 5.3.1 Componentes de Cloudera Hadoop.

Los componentes de Cloudera Hadoop distribuidos como paquetes binarios con se llama parsel. Tienen las siguientes ventajas:

1. Facilidad de carga: Cada parsel es un archivo único que combina todos los componentes necesarios;
2. Consistencia interna: todos los componentes son probados a fondo en parsel;
3. Distribución y activación: puede configurar primera parcela en todos los nodos gestionados, y luego activarlos con una sola acción, gracias a esta actualización del sistema se realiza de forma rápida;
4. Simple de cambios de vuelta: en caso de cualquier problema en el trabajo con la nueva versión que puede fácilmente volver a la anterior.

### 5.3.2 Requisitos de hardware.

Requisitos de hardware para instalar de Hadoop un tema muy difícil. Por los diferentes nodos del clúster tienen diferentes requisitos. Regla general: más memoria y discos. En un RAID-controladores no es necesario debido a Hadoop HDFS y la propia arquitectura, diseñado para funcionar en los servidores estándar simples.

Enumera las configuraciones de hardware para una variedad de opciones de arranque:

- Fácil configuración: 2 procesadores de seis núcleos, 24-64 GB de memoria, 8 unidades de disco duro 1-2 TB;

- Configuración racional: 2 procesador de seis núcleos, 48 a 128 GB de memoria, 12-16, unidades de disco duro (1 TB o 2) conectado directamente a la placa base a través del controlador;
- Configuración para el cálculo intensivo: 2 procesador de seis núcleos, 64 a 512 GB de memoria, unidades de disco duro 4-8, 1-2 TB.

## 5.4 Instalación y configuración.

Para todos los servidores utilizará CentOS 6.4 en la instalación mínima, pero se pueden utilizar otras distribuciones: Debian, Ubuntu, etc. Los paquetes requeridos están disponibles en acceso abierto en [archive.cloudera.com](http://archive.cloudera.com) y seleccionan el controlador de paquete estándar.

Al servidor de Cloudera recomendar el uso de software o de hardware RAID 1 y una partición raíz, se puede hacer en una partición separada / var / log /. En los servidores que se agregan a la hadoop-clúster, se recomienda crear dos particiones:

«/» tamaño de 50-100 GB para el sistema operativo y Cloudera Hadoop;

«/dfs» LVM en la parte superior de todas las unidades disponibles para el almacenamiento de datos HDFS;

«swap» tamaño de 500Mb

En todos los servidores, incluyendo servidor de Cloudera, debe deshabilitar SELinux y firewall. Por razones de seguridad, se recomienda aislar el clúster desde el mundo exterior a nivel de red, por ejemplo, usando un firewall de hardware o VLAN aislada (acceso a los espejos dispuestos a través de un proxy local).

```
# vi /etc/selinux/config # desactivar SELinux
SELINUX=disabled
# system-config-firewall-tui # desactivar firewall y guardar los ajustes
# reboot
```

### 5.4.1 La instalación del Administrador de Cloudera.

Instalación de Cloudera Manager, que se desplegará y configurarse para la instalación de Hadoop-cluster de servidores. Es necesario para asegurarse de que antes de la instalación que:

- Todos los miembros de las agrupaciones de servidores de están disponibles para ssh, y fijan la misma contraseña
- Todos los nodos deben tener acceso a los repositorios estándar
- Todos los servidores de un clúster tienen acceso a [archive.cloudera.com](http://archive.cloudera.com) o un repositorio local con los archivos de instalación necesarios
- Instalado en todos los servidores y configurar la sincronización de hora
- Todos los nodos de un clúster y el servidor Cloudera Manager configurado de DNS y PTR.

## 5.4.2 Agregar espejo Cloudera y instalar los paquetes necesarios:

### necesarios:

```
# wget -q -O /etc/yum.repos.d/cloudera-manager.repo
http://archive.cloudera.com/cm4/redhat/6/x86_64/cm/cloudera-manager.repo
# rpm --import http://archive.cloudera.com/cdh4/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera
# yum -y install jdk
# yum -y install cloudera-manager-daemons
# yum -y install cloudera-manager-server
# yum -y install cloudera-manager-server-db
```

### Al final de la instalación ejecutar una base de datos estándar y el servicio en sí Cloudera Manager:

```
# /etc/init.d/cloudera-scm-server-db start
# /etc/init.d/cloudera-scm-server start
```

## 5.5 Instalación de Cluster Cloudera Hadoop.

Después de instalar el Cloudera Manager toda la interacción aún más con el clúster se llevará a cabo utilizando la interfaz web de Cloudera Manager. Por defecto Cloudera Manager se utiliza el puerto 7180. Puede utilizar cualquiera de DNS o la dirección IP del servidor. Introducir esta dirección en su navegador. En la pantalla de inicio de sesión aparecerá en el sistema.

Nombre de usuario y contraseña (admin, admin). Se abrirá una ventana que le pide que seleccione la versión de Cloudera Hadoop: libre, ensayo de 60 días o licencia de pago.

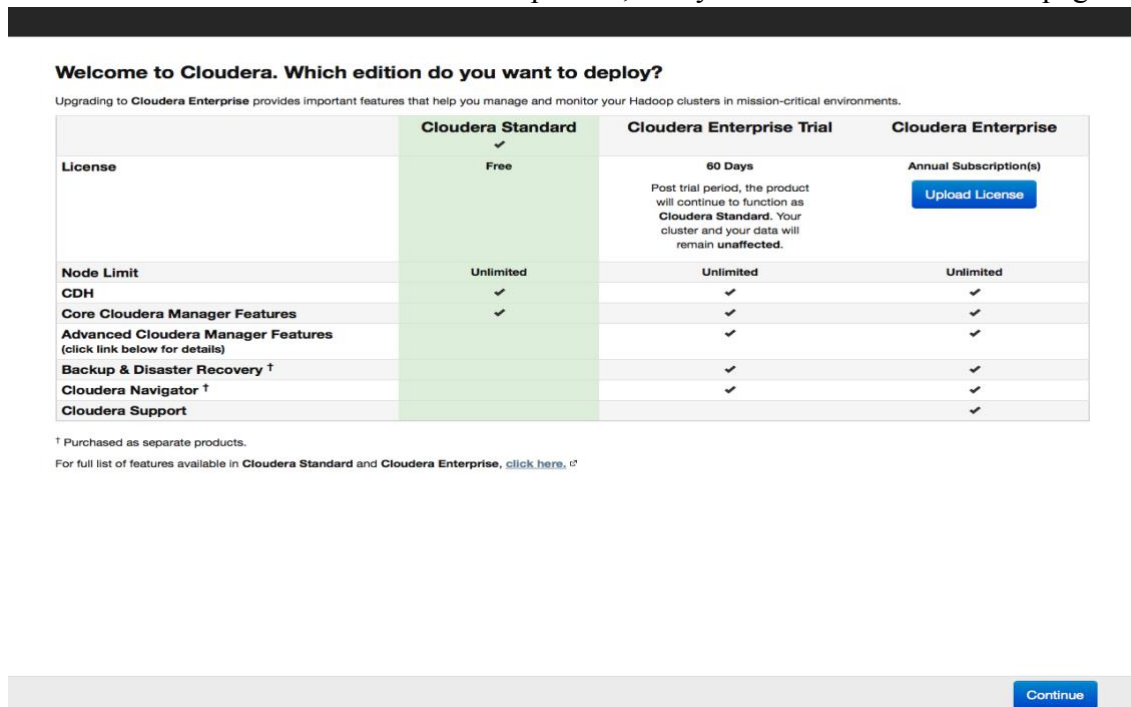


Figura 26 Una ventana que le pide que seleccione la versión de Cloudera Hadoop.

Seleccione una versión libre (Cloudera estándar). Durante la instalación del servicio Administrador de Cloudera conectará a través de SSH a los servidores dentro de la agrupación; todas las acciones en los servidores que realiza en nombre del usuario especificado en el menú, el valor predeterminado es root. Siguiendo Cloudera Manager le pide que especifique la dirección de host, el cual será instalado Cloudera Hadoop:

**Specify hosts for your CDH cluster installation.**

Cloudera recommends including Cloudera Manager server's host because it is often used for the Cloudera Management Service, and because this will enable health monitoring for that host.

Hint: Search for hostnames and/or IP addresses using [patterns](#) <sup>IP</sup>.

8.216/29

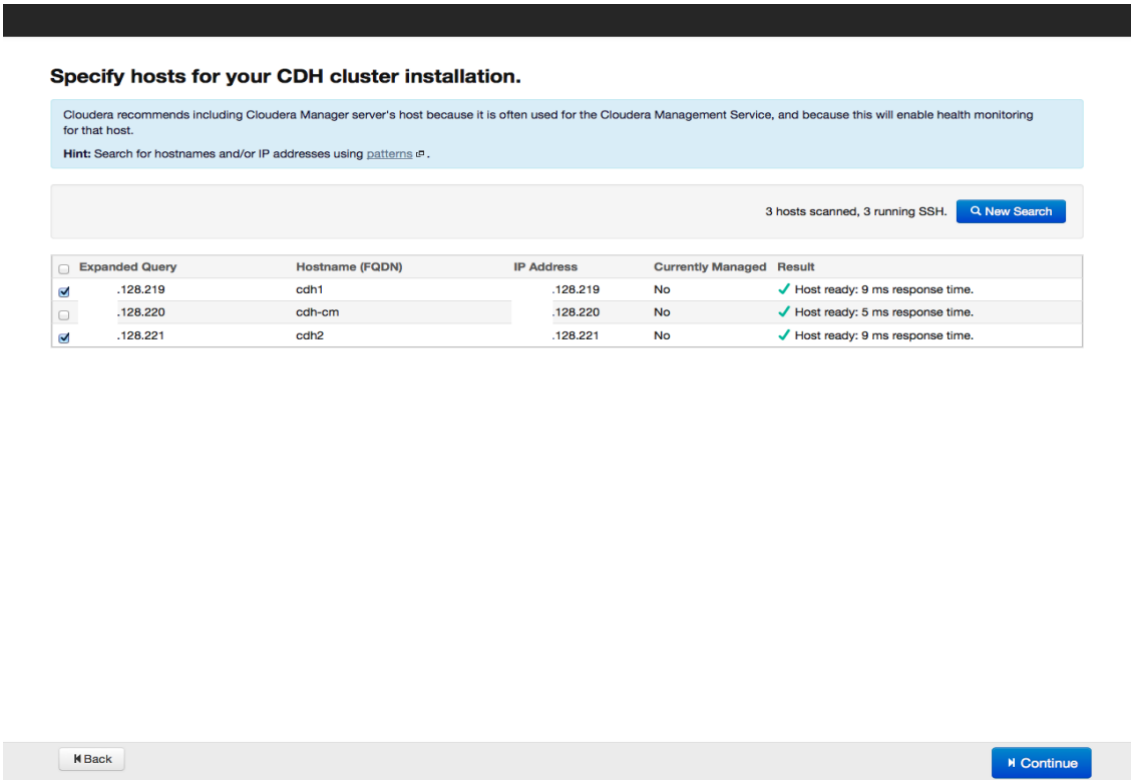
SSH Port: 22 Search

Back Continue

*Figura 27 Indicar los hosts.*

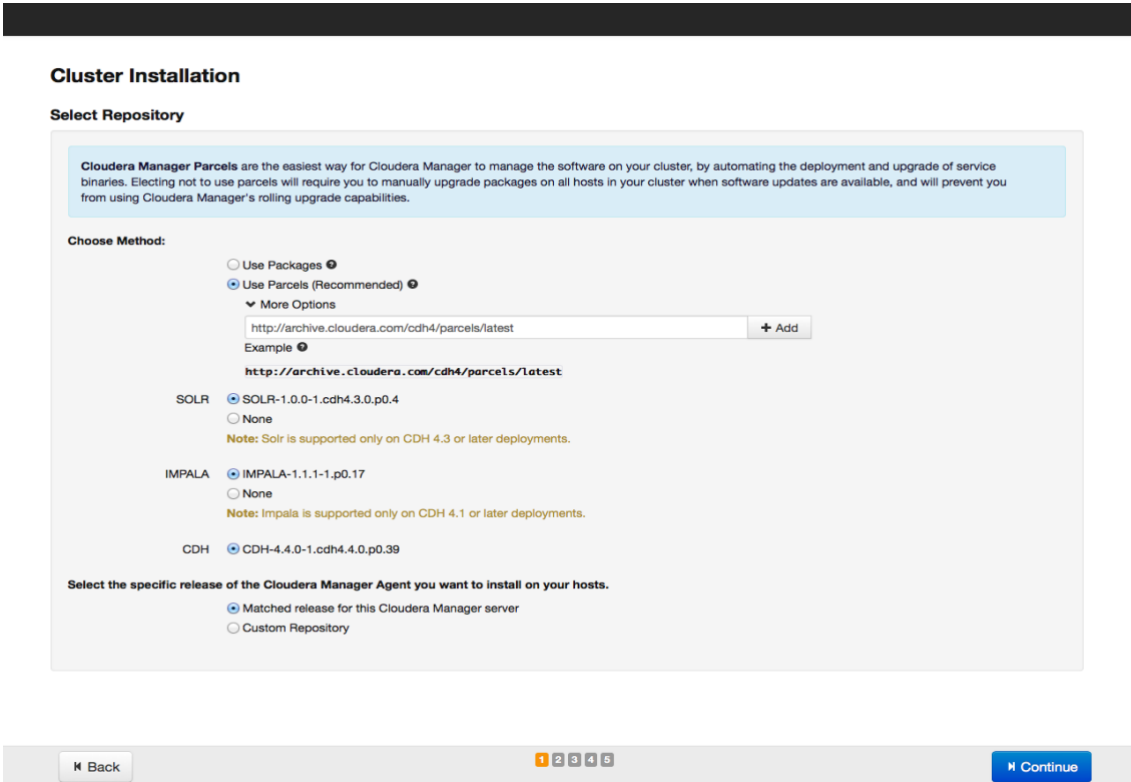
**Las direcciones pueden ser especificados por máscara y una lista como esta:**

10.1.1.[1-4] esto significa que los nodos del clúster se-direcciones IP 10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4. host[07-10].example.com — host07.example.com, host08.example.com, host09.example.com, host10.example.com. Luego hay que hacer clic en el botón de búsqueda. Cloudera Manager detecta los hosts especificados, y la pantalla mostrará una lista de ellos:



*Figura 28 Especificar los hosts.*

Una vez más comprobar si la figura en la lista de todos los hosts requeridos (añadir nuevos hosts haciendo clic en el botón Nuevo de búsqueda). A continuación, haga clic en el botón Continuar. Abra la ventana de selección de repositorio:



*Figura 29 La ventana de selección de repositorio.*

Como método de instalación, seleccione la instalación parcelami. Parsely conjunto de repositorio archive.cloudera.org. Además, Purcell CDH, del mismo repositorio se puede configurar la herramienta de búsqueda SOLR y base de datos basada en Hadoop Impl. La elección de la parcela para la instalación, haga clic en el botón Continuar. En la siguiente ventana, especifique los parámetros para el acceso a través de SSH (nombre de usuario, contraseña o clave privada, el número de puerto para la conexión):

**Cluster Installation**

**Provide SSH login credentials.**

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login to all hosts as:  root  Another User:

You may connect via password or public-key authentication for the user selected above.

Authentication Method:  All hosts accept same password  All hosts accept same private key

Enter Password:

Confirm Password:

SSH Port:

Number of simultaneous installations:  (Running a large number of installations at once can consume large amounts of network bandwidth and other system resources)

Navigation: Back | 1 2 3 4 5 | Continue

*Figura 30 Los parámetros para el acceso a través de SSH.*

A continuación, haga clic en el botón Continuar. Inicio del proceso de instalación:

**Cluster Installation**

**Installation in progress.**

0 of 2 host(s) completed successfully: Abort Installation

Hostname	IP Address	Progress	Status
cdh1	.128.219	<div style="width: 50%;"></div>	Refreshing package metadata... <a href="#">Details</a>
cdh2	.128.221	<div style="width: 50%;"></div>	Refreshing package metadata... <a href="#">Details</a>

Navigation: Back | 1 2 3 4 5 | Continue

*Figura 31 Inicio del proceso de instalación.*

Al finalizar, la pantalla mostrará una tabla con un resumen de la información sobre los componentes instalados y sus versiones:

✓ All checked Cloudera Management Agents versions are consistent with the server.

**Version Summary**

Group 1 (CDH4)		
Hosts		
cdh1, cdh2		
Component	Version	CDH Version
Impala	1.1.1	Not applicable
Lily HBase Indexer (CDH4 only)	1.2+2	Not applicable
Soir (CDH4 only)	4.4.0+69	Not applicable
Flume NG	1.4.0+23	CDH4
MapReduce 1 (CDH4 only)	2.0.0+1475	CDH4
HDFS (CDH4 only)	2.0.0+1475	CDH4
HttpFS (CDH4 only)	2.0.0+1475	CDH4
MapReduce 2 (CDH4 only)	2.0.0+1475	CDH4
Yarn (CDH4 only)	2.0.0+1475	CDH4
Hadoop	2.0.0+1475	CDH4
HBase	0.94.6+132	CDH4
HCatalog (CDH4 only)	0.5.0+13	CDH4
Hive	0.10.0+198	CDH4
Mahout	0.7+21	CDH4
Oozie	3.3.2+92	CDH4
Pig	0.11.0+33	CDH4
Sqoop	1.4.3+62	CDH4
Sqoop2 (CDH4 only)	1.99.2+85	CDH4
Whirr	0.8.2+15	CDH4
Zookeeper	3.4.5+23	CDH4
Hue	2.5.0+139	CDH4
Java	java version "1.6.0_31" Java(TM) SE Runtime Environment (build 1.6.0_31-b04) Java HotSpot(TM) 64-Bit Server VM (build 20.6-b01, mixed mode)	Not applicable
Cloudera Manager Agent	4.7.2	Not applicable

⏪ Back 1 2 3 4 5 Continue ⏩

Figura 32 La lista información.

Comprobamos que todo estaba en orden, y hacer clic en el botón Continuar. La pantalla le solicitará que seleccione los componentes y servicios de Cloudera Hadoop para instalar:

**Choose the CDH4 services that you want to install on your cluster.**

Choose a combination of services to install.

- Core Hadoop**  
HDFS, MapReduce, ZooKeeper, Oozie, Hive, and Hue
- Core with Real-Time Delivery**  
HDFS, MapReduce, ZooKeeper, HBase, Oozie, Hive, and Hue
- Core with Real-Time Query**  
HDFS, MapReduce, ZooKeeper, Impala, Oozie, Hive, and Hue
- All Services**  
HDFS, MapReduce, ZooKeeper, HBase, Impala, Oozie, Hive, Hue and Sqoop.
- Custom Services**  
Choose your own services. Services required by chosen services must also be selected. Note that Flume, Soir and KeyStore Indexer services can be added after your initial cluster has been set up.

This wizard will also install the **Cloudera Management Services**. These are a set of components that enable monitoring, reporting, events, and alerts; these components require databases to store information, which will be configured on the next page.

Inspect Role Assignments Continue

Figura 33 Los componentes y servicios de Cloudera Hadoop para instalar.

Por ejemplo, para instalar todos los componentes de la selección «Todos los servicios», más adelante será posible instalar o eliminar cualquiera de los servicios. Ahora tiene que especificar qué componentes de Cloudera Hadoop será instalado en un host en particular.

### Inspect role assignments

You can customize the role assignments for your new cluster here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, this can significantly impact the performance of your services. Cloudera does not recommend altering assignments unless you have specific requirements, such as having pre-selected a specific host for a specific role.

The host list presented here is prefiltered to remove hosts which are not valid candidates; these include hosts that are: unhealthy, members of other clusters, and/or which have an incompatible version of CDH installed on them.

Server	HDFS			HBase					MapReduce		
	All   None	DataNode All   None	NameNode	SecondaryNameNode	HttpFS All   None	Master All   None	RegionServer All   None	HBase REST Server All   None	HBase Thrift Server All   None	TaskTracker All   None	JobTracker All   None
cdh1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cdh2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

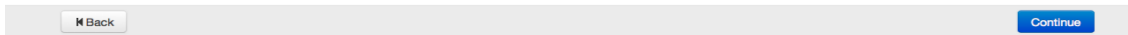


Figura 34 Instalar todos los componentes.

Haga clic en el botón Continuar y proceder al siguiente paso- la creación de una base de datos:

### Database Setup

On this page you configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#).

When using the Embedded Database, passwords are auto generated. Please copy them down.

Use Embedded Database  
 Use Custom Databases

---

**Hive** ✔ Skipped. Will create database in later step.

Database Host Name:  Database Type:  Database Name :  Username:  Password:

---

**Service Monitor** ✔ Successful

Currently assigned to run on cdh1.

Database Host Name:  Database Type:  Database Name :  Username:  Password:

---

**Activity Monitor** ✔ Successful

Currently assigned to run on cdh1.

Database Host Name:  Database Type:  Database Name :  Username:  Password:

---

**Host Monitor** ✔ Successful

Currently assigned to run on cdh1.

Database Host Name:  Database Type:  Database Name :  Username:  Password:

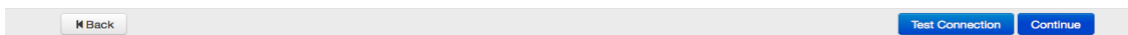


Figura 35 La creación de una base de datos.



Por defecto, toda la información relativa al sistema de monitorización y gestión, se almacena en una base de datos PostgreSQL, que se instala junto con el Administrador de Cloudera. Se puede utilizar otra base de datos en este caso, optar por utilizar el menú base de datos personalizada. Al establecer los parámetros necesarios, comprobar la conexión con el «Probar conexión» y, si tiene éxito, haga clic en el botón «Siguiente» para continuar con los elementos de la agrupación:

**Review configuration changes**

Set the following configuration values for your new role(s). Required values are marked with \*.

Group	Parameter	Recommended Value	Description
<b>Service hbase1</b>			
Service-Wide	HDFS Root Directory* hbase.rootdir	/hbase default value	The HDFS directory shared by HBase RegionServers.
<b>Service hdfs1</b>			
DataNode (Default) <a href="#">Show Members</a>	DataNode Data Directory* dfs.datanode.data.dir	/dfs/dn <a href="#">Reset to empty default value</a>	Comma-delimited list of directories on the local file system where the DataNode stores HDFS block data. Typical values are /data/N/dfs/dn for N = 1, 2, 3... These directories should be mounted using the noatime option and the disks should be configured using JBOD. RAID is not recommended.
DataNode (Default) <a href="#">Show Members</a>	DataNode Failed Volumes Tolerated dfs.datanode.failed.volumes.tolerated	0 default value	The number of volumes that are allowed to fail before a DataNode stops offering service. By default, any volume failure will cause a DataNode to shutdown.
DataNode (1) <a href="#">Show Members</a>	DataNode Data Directory* dfs.datanode.data.dir	/dfs/dn <a href="#">Reset to empty default value</a>	Comma-delimited list of directories on the local file system where the DataNode stores HDFS block data. Typical values are /data/N/dfs/dn for N = 1, 2, 3... These directories should be mounted using the noatime option and the disks should be configured using JBOD. RAID is not recommended.
DataNode (1) <a href="#">Show Members</a>	DataNode Failed Volumes Tolerated dfs.datanode.failed.volumes.tolerated	0 default value	The number of volumes that are allowed to fail before a DataNode stops offering service. By default, any volume failure will cause a DataNode to shutdown.
NameNode (Default) <a href="#">Show Members</a>	NameNode Data Directories* dfs.namenode.name.dir	/dfs/nn <a href="#">Reset to empty default value</a>	Determines where on the local file system the NameNode should store the name table (fsimage). For redundancy, enter a comma-delimited list of directories to replicate the name table.

[Back](#) [Continue](#)

*Figura 36 Configuración los elementos esta un clúster.*

Haga clic en el botón Continuar y ejecutar el proceso de configuración de este modo clúster. Configuración de trazo muestran en la pantalla:

**Starting your cluster services.**

Completed 3 of 22 steps.

- ✓ Waiting for ZooKeeper Service to initialize  
Finished waiting
- ✓ Starting ZooKeeper Service  
Service started successfully.
- ✓ Checking if the name directories of the NameNode are empty. Formatting HDFS only if empty.  
Successfully formatted NameNode.
- ⚙️ **Starting HDFS Service**
  - Creating HDFS /tmp directory
  - Creating HBase root directory
  - Starting HBase Service
  - Starting MapReduce Service
  - Creating Hive Metastore Database
  - Creating Hive Metastore Database Tables
  - Creating Hive user directory
  - Creating Hive warehouse directory
  - Starting Hive Service
  - Creating Oozie database
  - Installing Oozie ShareLib in HDFS
  - Starting Oozie Service
  - Creating Sqoop user directory
  - Starting Sqoop Service
  - Starting Impala Service

[Continue](#)

*Figura 37 Configuración del clúster.*

Cuando la sintonización de todos los componentes es completa, vamos a dashboard de cluster. Es el dashboard de cluster:

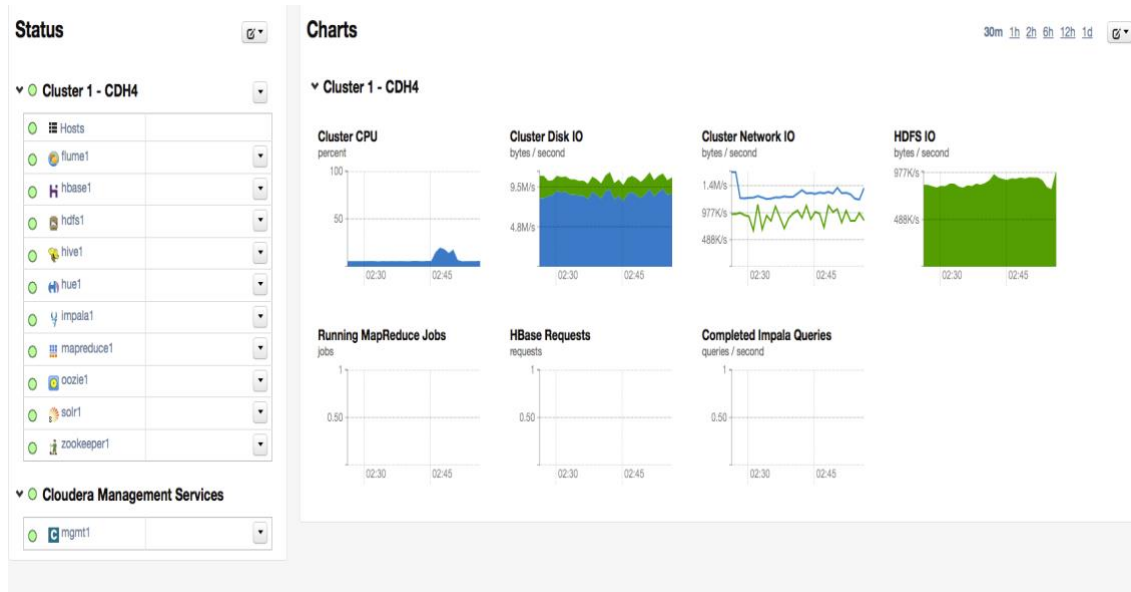


Figura 38 Dashboard

## 5.6 La recopilación de datos a través Flume.

Descargar los datos en Hadoop que puedas con una copia simple en HDFS, y con la ayuda de herramientas especiales. La forma más fácil de migrar los datos a un clúster - es copiar los archivos a través de una interfaz de administrador de archivos web en el panel de control de Hue. La Interfaz Web esta por dirección `http://[Hue_node]:8888/filebrowser/` (*dentro de* [Hue\_node] indica la dirección del nodo especificado, que desplegó a Hue). La interfaz web es bueno para los principiantes usuarios. Con que sea conveniente para explorar la estructura de directorio de HDFS. Al mismo tiempo, es inconveniente para descargar archivos grandes (varios GB).

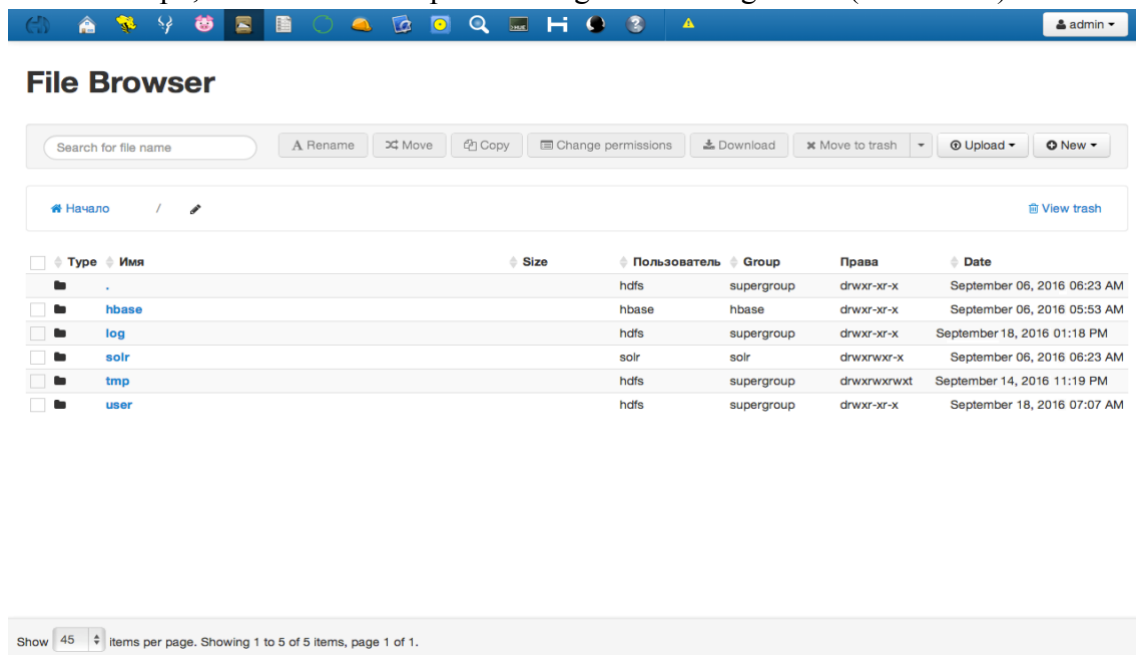


Figura 39 Interfaz web.

Para descargar un gran tamaño de archivos es preferible copiar archivos en HDFS con la utilidad de Hadoop. Esto se hace con el comando:

```
hadoop fs -put file_for_hadoop /path/to/put/file/in/HDFS/
```

En este caso siempre se puede pre-copiar los archivos al servidor. Estos métodos son muy adecuados para situaciones en las que desea transferir los datos existentes en HDFS. Sin embargo, una mejor manera de recoger datos de forma inmediata en Hadoop. Para esto utiliza herramientas especializadas. Una de estas herramientas se desarrolla dentro del proyecto Apache Hadoop. Este Flume- herramienta universal para la recogida de registros y otros datos.

### 5.6.1 Sobre el proyecto Flume.

Flume fue desarrollado originalmente por Cloudera antes de ser donado a la comunidad Apache. Flume funciona como un servicio distribuido para la recopilación de datos en tiempo real, el almacenamiento temporal y la entrega a un destino. Flume es una herramienta altamente confiable, distribuida y configurable. Está diseñado para recopilar datos de transmisión de varios servidores web a HDFS.

- La fuente: Flume tiene como objetivo recuperar mensajes de diferentes fuentes, especialmente archivos de registro, pero también como veremos en los datos de Twitter.
- El canal Flume: es un búfer que almacena mensajes antes de que se consuman.
- El objetivo Flume: el lote consume los mensajes que provienen del canal para escribirlos en un destino como HDFS, por ejemplo. [BIR17]

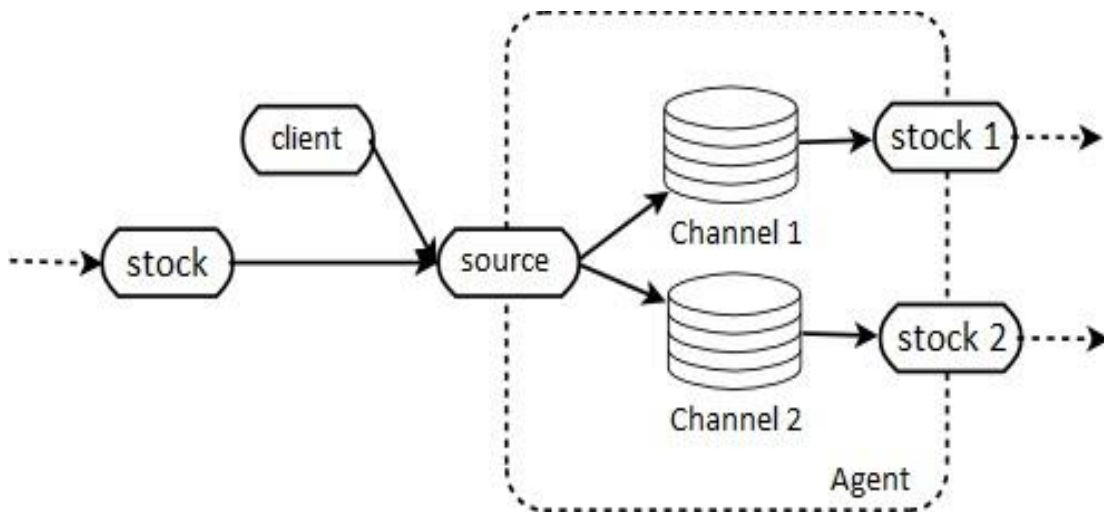
### 5.6.2 Arquitectura de Flume.

- *event* - los datos transmitidos para Flume desde el punto de origen al punto de destino;
- *flow* - eventos de movimiento ruta desde el punto de origen hasta el destino;
- *client* - cualquier aplicación que transfiere los datos al agente de Flume;
- *agent* - proporciona almacenamiento y transmisión de eventos al siguiente nodo;
- *source* - una interfaz para recibir mensajes a través de varios protocolos de comunicación. Estos eventos fuente transmite a los uno o más canales;
- *channel* - almacenamiento temporal para eventos;
- *sink* - toma el *event* del canal y transmite el siguiente agente

### 5.6.3 La estructura del flujo.

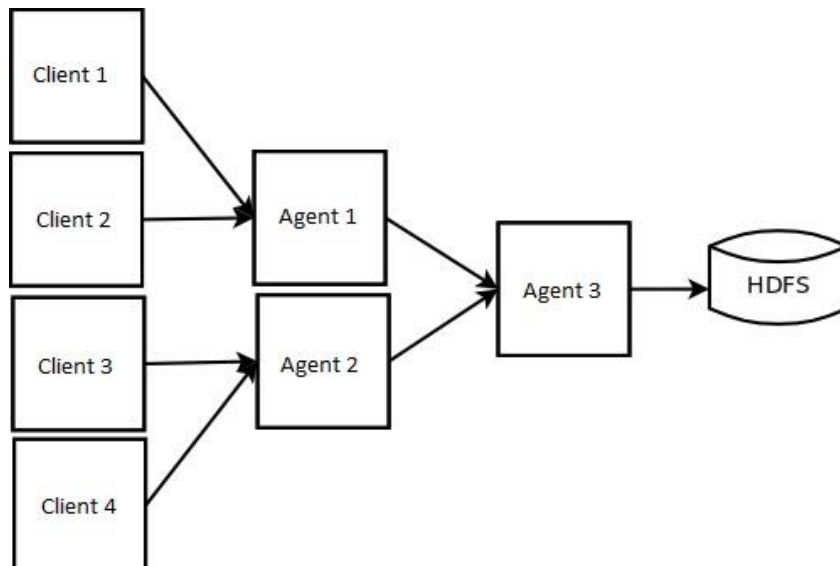
El flujo comienza con un cliente que envía un evento para el agente. Una fuente se pone un evento transmite a uno o más canales. Desde el canal se transmite a los receptores de sucesos que forman parte del mismo agente. Él puede darle a otro agente o para un nodo destino.

Puesto que la fuente puede transmitir eventos en múltiples canales, las corrientes se pueden dirigir a múltiples nodos de destino. Esto se ilustra en la siguiente figura: el agente lee el evento en los dos canales (canales 1 y 2) y, a continuación, y después transmite a stock.



**Figura 40** La estructura del flujo.

Varios subprocesos se pueden combinar en una sola. Para este propósito, las múltiples fuentes en la composición del agente también transmiten datos sobre el mismo canal. Componentes de interacción Esquema cuando se muestra corrientes combinadas a continuación (en este caso, cada uno de los tres agentes, que comprende varias fuentes, transmite datos en el mismo canal y después de flujo):



**Figura 41** Esquema se muestra corrientes combinadas a continuación.

### 5.6.4 Fiabilidad y Control de errores.

Procesamiento de datos entre fuentes y canales, así como entre los agentes se realiza utilizando transacciones, lo que asegura la integridad de datos. El manejo de errores se lleva a cabo sobre la base de un mecanismo transaccional. Cuando el flujo pasa a través de un número de diferentes agentes, y tienen problemas al pasar eventos de comunicación accesibles en la última tamponada en el agente de flujo. Más claramente se muestra a continuación el esquema de manejo de error:

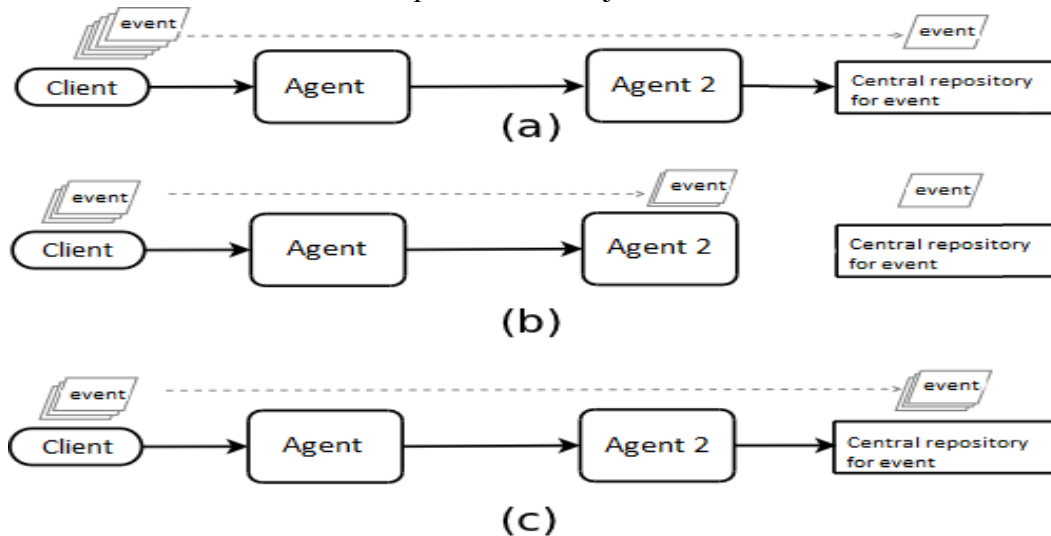


Figura 42 El esquema de manejo de error.

### 5.7 Instalación de Flume a través de Cloudera Administrador.

Instalación de Flume con ayuda de Cloudera Manager. En la página con una lista del servicio de clúster elegirá las «acciones» → «Añadir un servicio».

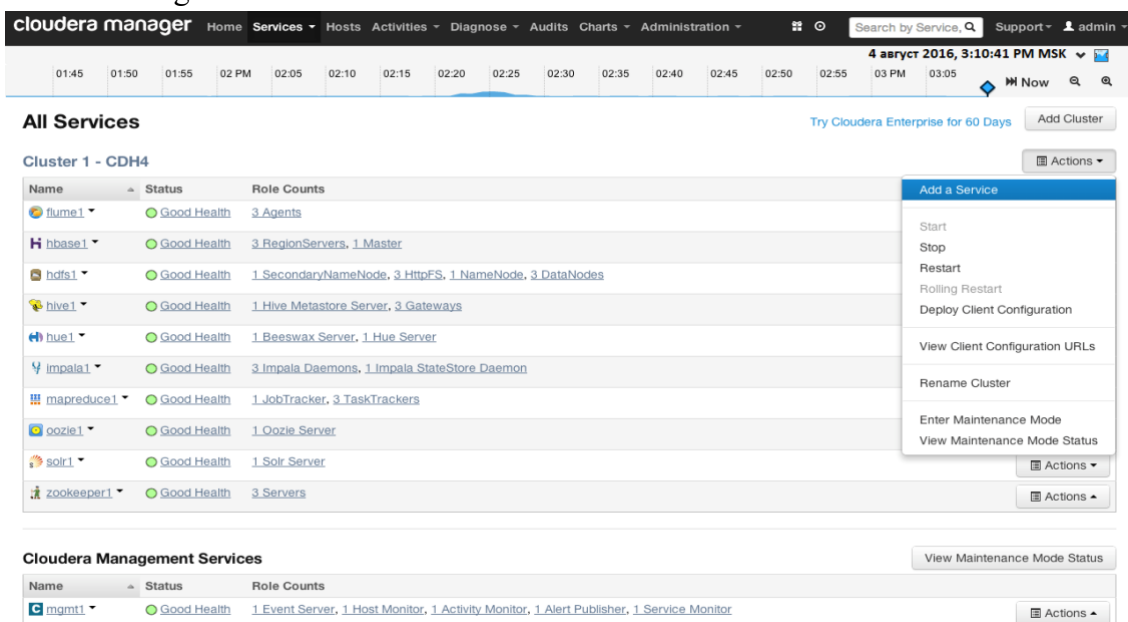
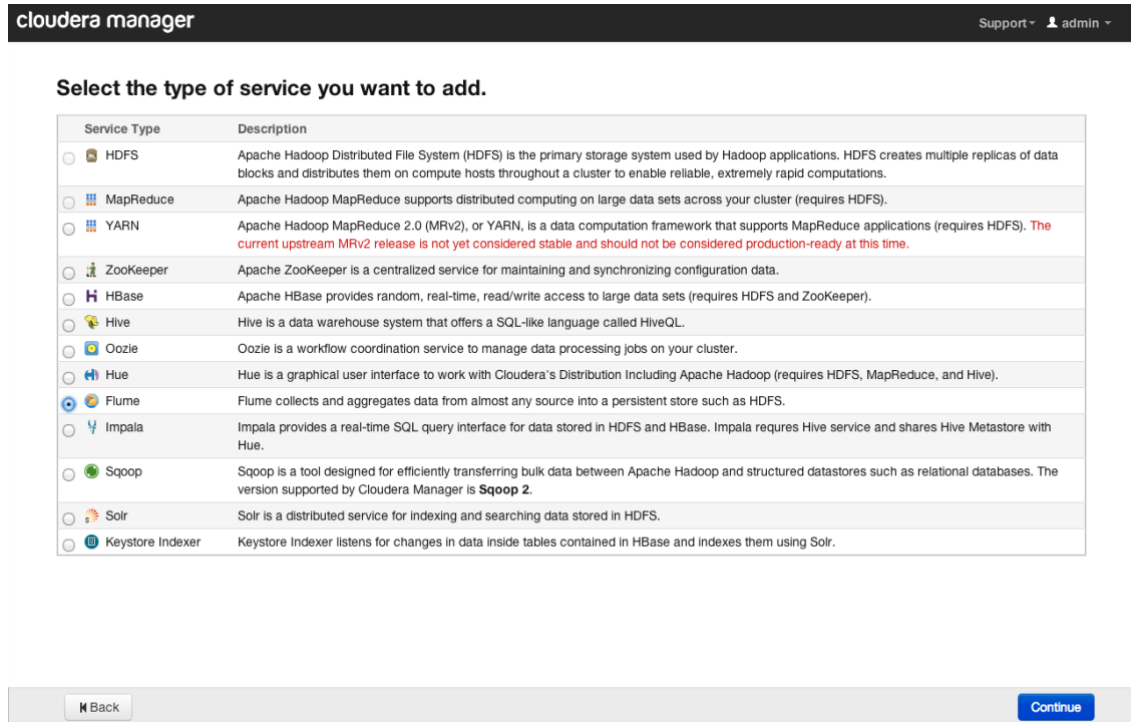


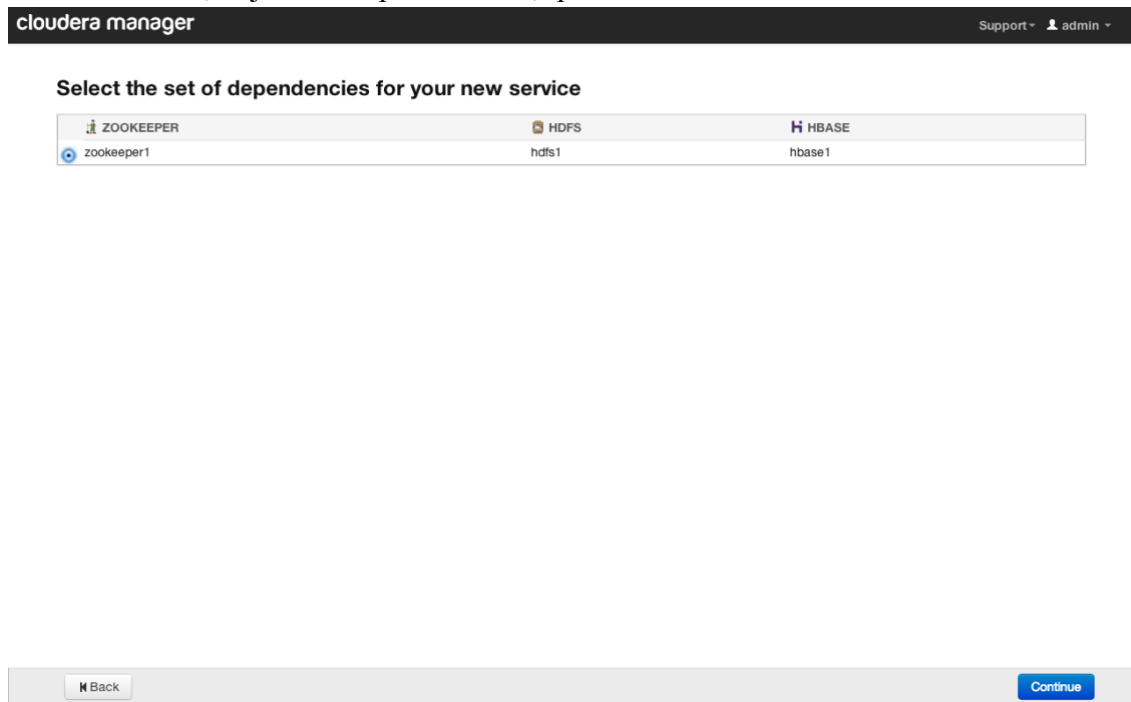
Figura 43 Instalación de Flume.

Elegir Flume y haga clic en el botón «Continué»:



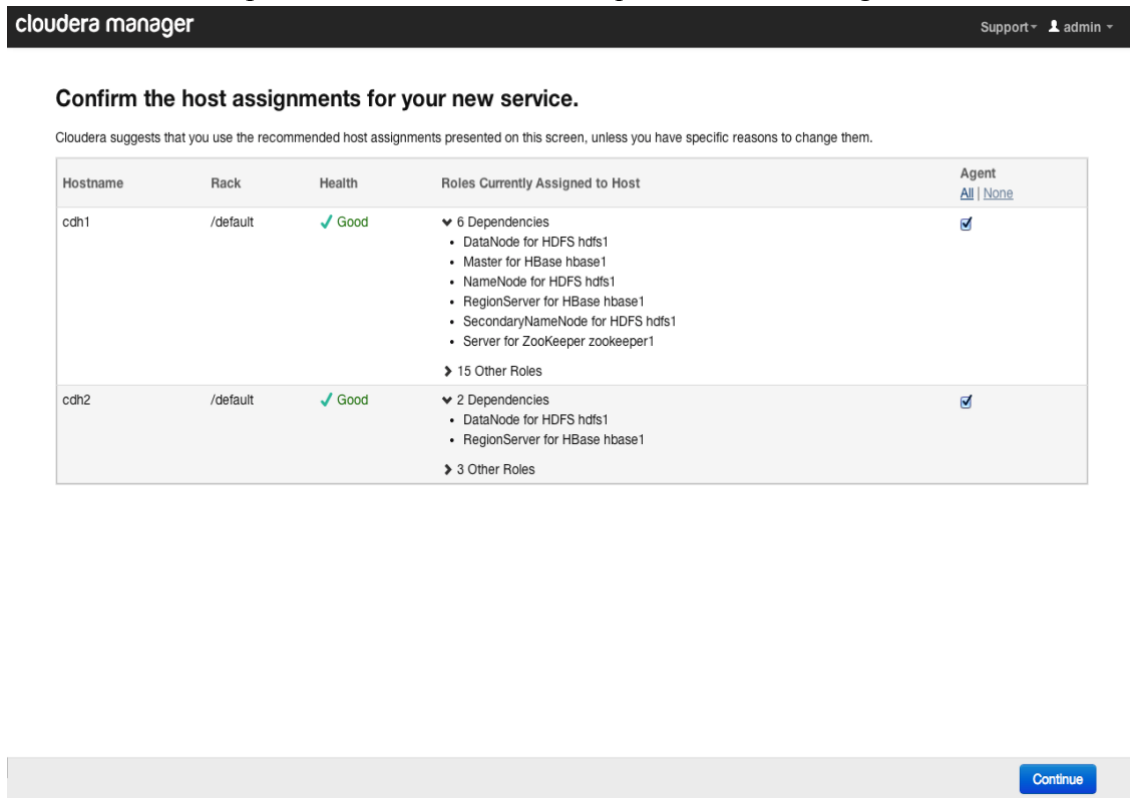
*Figura 44 Instalación de Flume.*

A continuación, elija Zookeeper-servicio, que estará vinculado al servicio de Flume.



*Figura 45 Elija Zookeeper-servicio.*

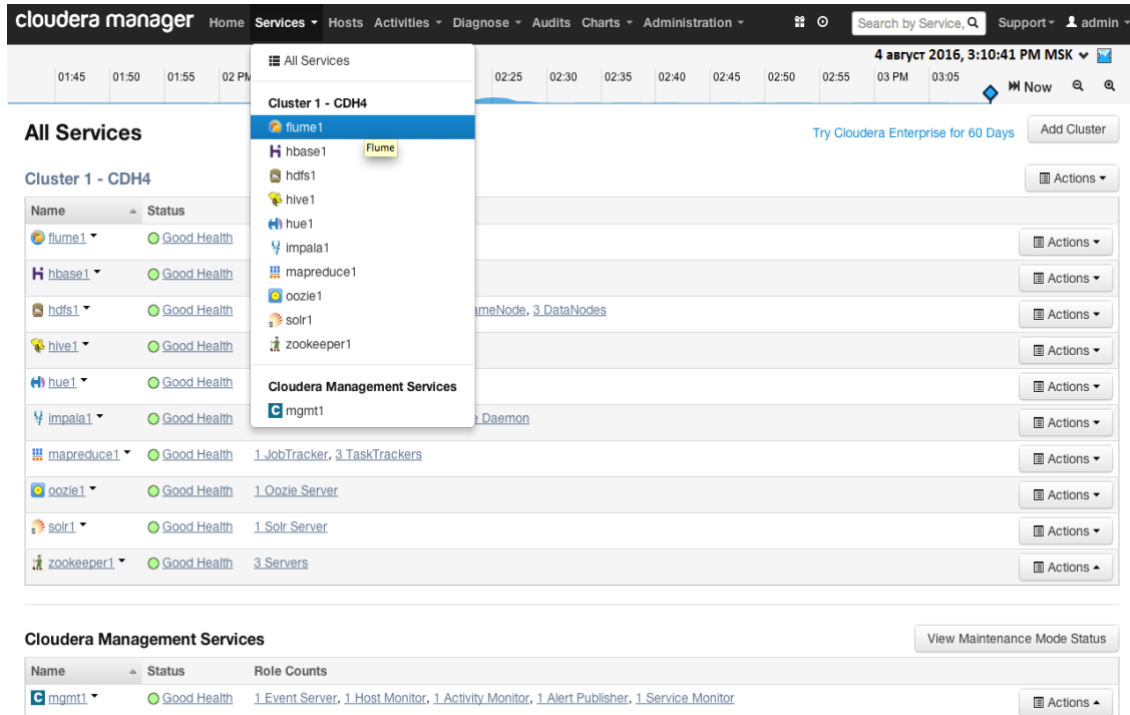
A continuación, especificar el host del clúster, que será instalado agentes Flume:



*Figura 46 Especificar el host del clúster.*

Haga clic en el botón «Continuar». Pronto verá un mensaje acerca del éxito del nuevo servicio.

Pasamos ahora al panel de control, elegido en Cloudera Manager «Services» → «flume1»:



*Figura 47 Panel de control.*

La página de servicio abre que contiene los siguientes enlaces: Estado general, servicio de instancias (agentes), comandos de gestión de servicios (encendido, apagado, reset), ajustes de servicio, configurar los permisos, estadísticas y gráficos de carga. Abra los ajustes de la ficha «Configuration» → «View and Edit»:

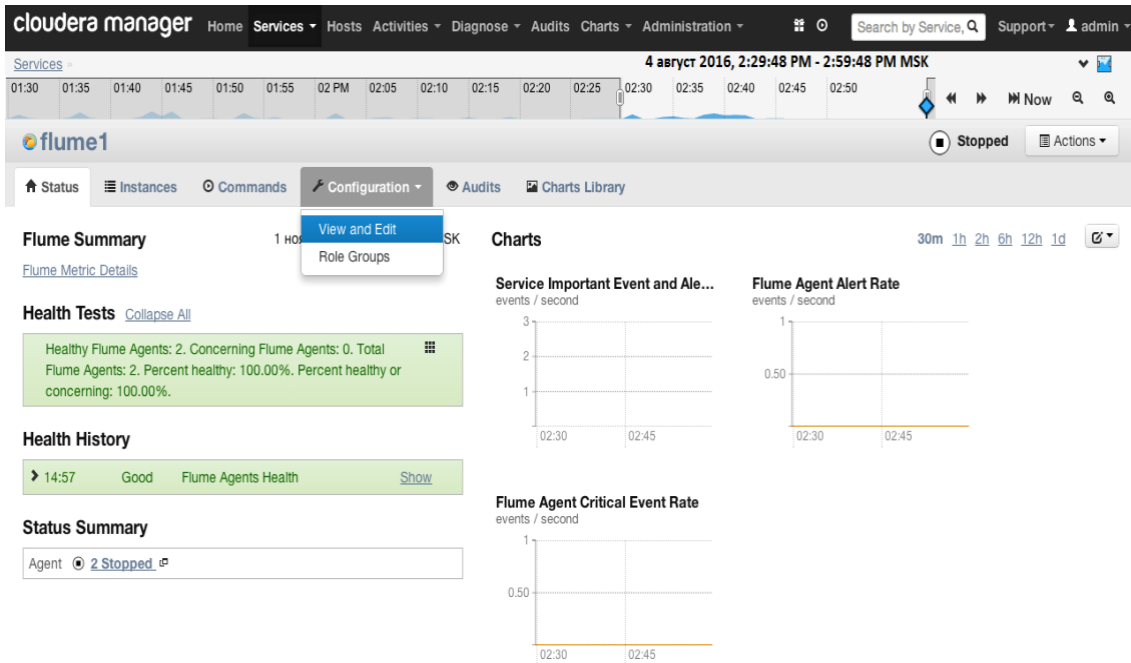


Figura 48 La página de servicio.

La configuración predeterminada para todos los agentes Flume se almacenan en un archivo de configuración (su contenido se muestra en el campo del archivo de configuración). Este archivo es común a todos los agentes y heredado por cada uno de ellos:

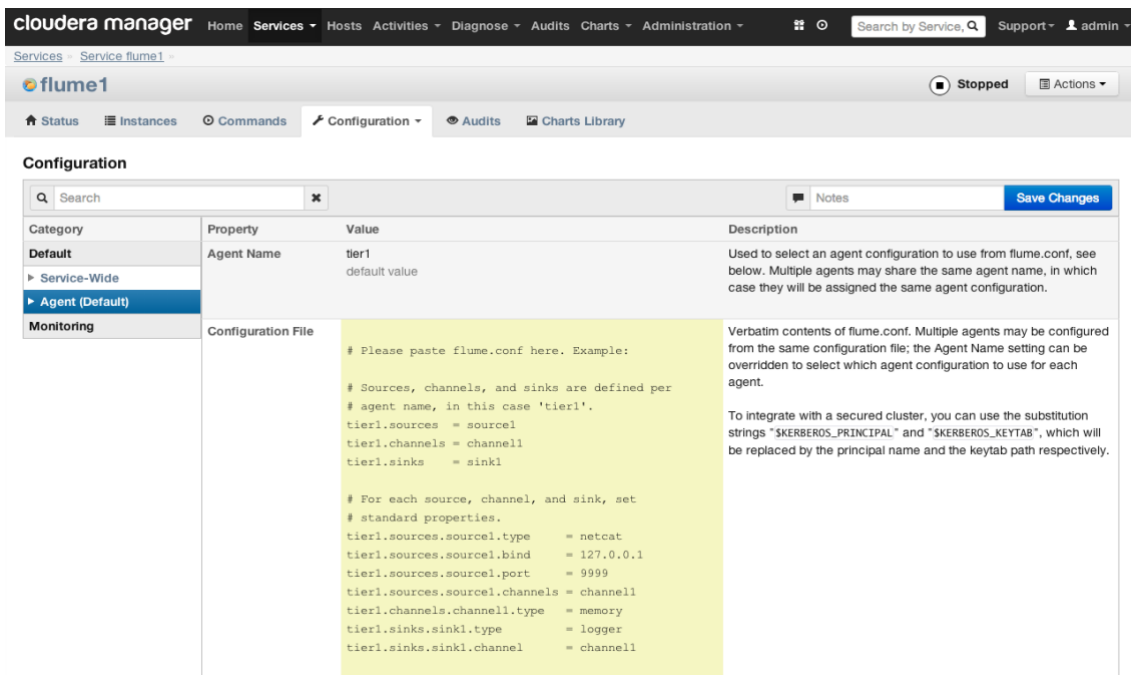


Figura 49 La página configuración.



## 5.7.1 Configuración del agente de Flume.

Producimos configuración del agente de Flume, que recogen registros de syslog a través de UDP y los almacena en el HDFS en el clúster:

```
### syslog cfg
a1.sources = r1
a1.channels = c1
a1.sinks = k1

# source
a1.sources.r1.type = syslogudp
a1.sources.r1.port = 5140
a1.sources.r1.host = cdh2.example.com
a1.sources.r1.channels = c1

# insert timestamp
a1.sources.r1.interceptors = i1
a1.sources.r1.interceptors.i1.type = timestamp

# sink
a1.sinks.k1.type = hdfs
a1.sinks.k1.channel = c1
a1.sinks.k1.hdfs.path = /log/flume/events/%y-%m-%d/%H-%M
a1.sinks.k1.hdfs.filePrefix = flume-
a1.sinks.k1.hdfs.round = true
a1.sinks.k1.hdfs.roundValue = 5
a1.sinks.k1.hdfs.roundUnit = minute
a1.sinks.k1.hdfs.fileType = DataStream
a1.sinks.k1.hdfs.rollCount = 100000
a1.sinks.k1.hdfs.rollSize = 0

# channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
```

Todos los registros en el archivo tienen una estructura jerárquica; orden de las filas no es importante. Antes de cada parámetro especifica el nombre del agente al que se refiere. A continuación, indicar el tipo de objeto (fuente, drenaje o canales) y su nombre, y luego - los tipos y subtipos de los parámetros y el valor en sí. Para la totalidad del agente predeterminado crea un único archivo de configuración. Debido al archivo de configuración general múltiples agentes pueden tener el mismo nombre y, en consecuencia, el mismo conjunto de opciones. Esto es útil para los agentes de tolerancia a fallos o para equilibrar la carga entre ellos. Para cambiar la función aegnta, basta con cambiar el nombre sin sobrescribir el nuevo archivo de configuración.

## 5.7.2 La estructura del archivo de configuración.

Especificamos los nombres de las principales instalaciones y el "atar" a un agente en particular. En nuestro caso, se indica que el agente «A1» fuente de «R1», el canal «C1» y stok «K1»:

```
a1.sources = r1
a1.channels = c1
a1.sinks = k1
```

Hemos creado el canal. A medida que el uso de canales de memoria de canal, el almacenamiento de la cola de eventos en la memoria. En el caso de ráfagas imprevistos de la actividad el tamaño máximo de cola se establece en 1000 los mensajes, aunque el número de mensajes en la cola normalmente no excede 10.

```
# channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
```

## 5.7.3 Configuración de fuente de Syslog UDP.

Como fuente vamos a utilizar UDP Syslog, incluidos a Flume:

```
a1.sources.r1.type = syslogudp
a1.sources.r1.port = 5140
a1.sources.r1.host = cdh2.example.com
a1.sources.r1.channels = c1
```

Canales de parámetros indican canales a los que se conectarán a la fuente. A continuación, especifique el objeto (interceptor). Interceptores no son elementos separados, y son parte de las fuentes.

```
# insert timestamp
a1.sources.r1.interceptors = i1
a1.sources.r1.interceptors.i1.type = timestamp
```

Las partes constituyentes del evento (events) en el canal de flujo son los datos y las cabeceras adicionales, la lista de los cuales pueden variar en función del tipo de fuente. Interceptor realiza pre-procesamiento de datos antes de ser transmitida al canal. Los interceptores pueden ser conectados en una cadena. En este caso, sólo un interceptor «i1» fuente de «r1». Registros del syslog del sistema de código en el evento sólo el propio mensaje. Registros del syslog del sistema Fuente de la marca de tiempo timestamp cabecera. Este título necesitará ajustes de stok.

Se procede a la configuración de stock, lo que permitirá ahorrar los datos en HDFS:

```
a1.sinks.k1.type = hdfs
a1.sinks.k1.channel = c1
a1.sinks.k1.hdfs.path = /log/flume/events/%y-%m-%d/%H-%M
a1.sinks.k1.hdfs.filePrefix = flume-
```

```

a1.sinks.k1.hdfs.fileType = DataStream
a1.sinks.k1.hdfs.round = true
a1.sinks.k1.hdfs.roundValue = 5
a1.sinks.k1.hdfs.roundUnit = minute
a1.sinks.k1.hdfs.rollCount = 100000
a1.sinks.k1.hdfs.rollSize = 0

```

El parámetro de *patch* especifica la ruta de los archivos en HDFS, en el que se guardarán los datos del evento. Al guardar los registros necesitan para distribuir archivos a las subcarpetas de acuerdo a la marca de tiempo - que simplifica el control y post-procesamiento. Es suficiente con especificar la carpeta con la fecha y una máscara para procesar los registros durante un período determinado.

**File Prefix** parámetro especifica el prefijo para el fichero del cual es útil cuando se recogen registros de diferentes agentes en la misma carpeta.

**FileType** parámetro especifica el formato de archivo que será guardado en el evento.

**DataStream** - un formato estándar en el que cada evento se almacena como una cadena en el archivo de texto sin formato.

**Round, RoundValue y RoundUnit** indica que el valor de marca de tiempo se redondea a un múltiplo de 5 minutos. Esto guardará los archivos de las subcarpetas en incrementos de 5 minutos.

**rollCount** = indica el número de mensajes en un solo archivo, más allá del cual el archivo actual es cerrada y un nuevo

**rollSize** = 0 indica que no limitamos el tamaño de cada archivo.

## 5.7.4 Configuración del cliente.

Un ejemplo de la configuración haproxy, responsable de almacenamiento en caché:

```

global
log [FQDN_Flume_agent]:5140 local0 info
maxconn 60000
nbproc 16
user haproxy
group haproxy
daemon
tune.maxaccept -1
tune.bufsize 65536

```

defaults

```

log global
mode http
# for hadoop
option httplog
#option logasap
log-format \%T\t%ci\t%cp\t%ft\t%b\t%s\t%ST\t%B\t%sq\t%bq\t%r
option dontlognull
#option dontlog-normal

```

Opción de **log** indica que la dirección del servidor y el puerto en el que el agente opera Flume, y los parámetros estándar para la local0 syslog y el nivel de registro se notifiquen.

**Mode http y option httplog** se indica que mantendrá registros sobre el acceso http.

Para guardar la máxima cantidad de información, desactivar la opción **logasap** y **dontlog-normal**. Cuando está desactivada **logasap haproxy** salvará a la petición HTTP ha sido completada con una indicación de los datos recibidos y transmitidos.

Para mantener **haproxy** todas las solicitudes, incluyendo exitosa, es necesario deshabilitar la opción **dontlog-normal**.

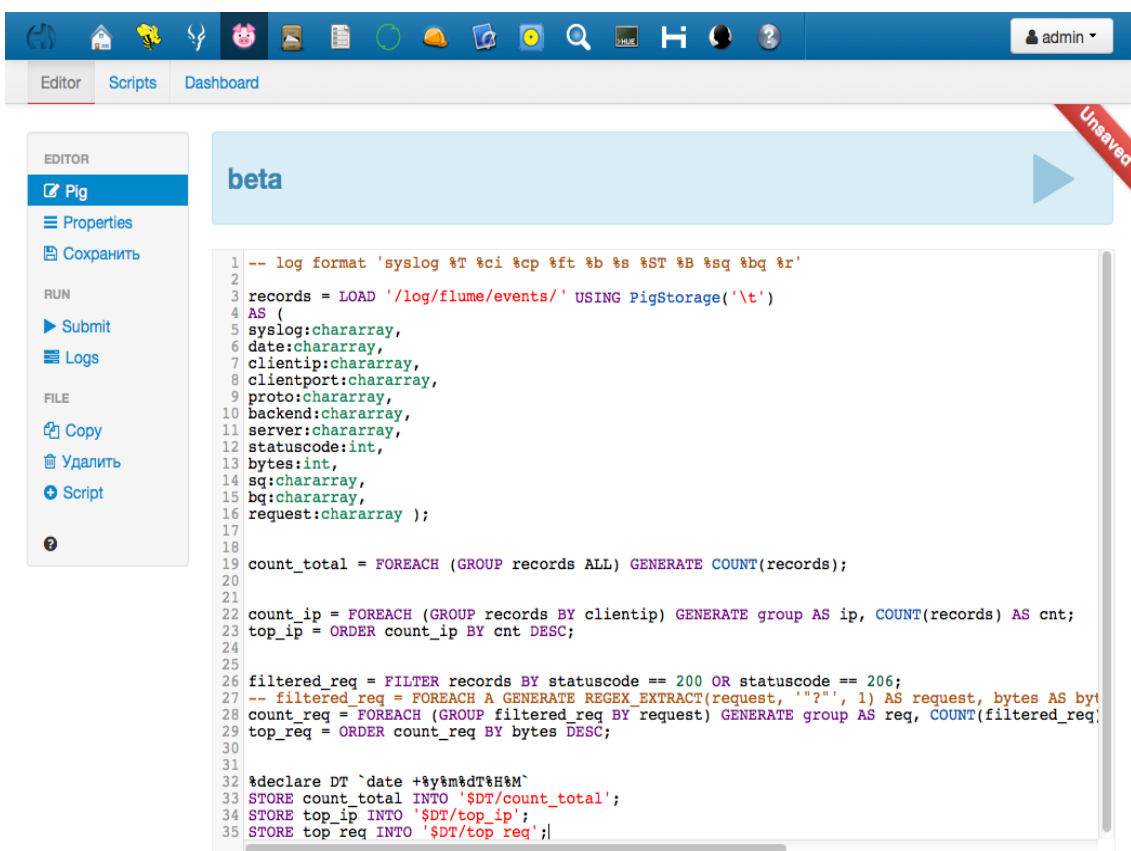
Para proporcionar datos para aproximar un formato legible por ordenador y para simplificar el procesamiento posterior de los datos, he cambiado el formato de los registros (log-formato). Por defecto, el separador utilizado en los registros del carácter de espacio, pero puede estar contenida en los datos en sí, lo que dificulta aún más el proceso. Así que lo reemplazó con un de tabulación. Además, apagué las cotizaciones de la URL y el método de consulta. ¡Un archivo de datos por un día casi 30 Gb!

## 5.8 Pig.

Apache Pig proporciona un lenguaje de flujo de datos similar a SQL sobre Apache Hadoop. Con Pig, los usuarios escriben scripts de flujo de datos en un lenguaje llamado Pig Latin. Luego, Pig ejecuta estos scripts de flujo de datos en Hadoop usando MapReduce. Proporcionar a los usuarios un lenguaje de secuencias de comandos, en lugar de exigirles que escriban programas MapReduce en Java, disminuye drásticamente su tiempo de desarrollo y permite a los desarrolladores que no son Java utilizar Hadoop. Pig también proporciona operadores para las operaciones de procesamiento de datos más comunes, como unir, ordenar y agregar. De lo contrario, se requeriría una gran cantidad de esfuerzo para que un programa Java MapReduce hecho a mano implemente estos operadores. Muchos tipos diferentes de procesamiento de datos se realizan en Hadoop. El cerdo no busca ser un propósito general solución para todos ellos. Pig se centra en casos de uso en los que los usuarios tienen un DAG de transformaciones para realizar en sus datos, que implican una combinación de operaciones relacionales estándar (unión, agregación, etc.) y procesamiento

personalizado que se puede incluir en Pig Latin a través de funciones definidas por el usuario, o UDF, que pueden escribirse en Java o en un lenguaje de script.<sup>1</sup> Pig también se centra en situaciones en las que los datos aún no se pueden limpiar y normalizar. Maneja con gracia situaciones en las que los esquemas de datos son desconocidos hasta el tiempo de ejecución o son inconsistentes. Estas características hacen de Pig una buena opción cuando se hacen transformaciones de datos en preparación para un análisis más tradicional. [PAT16]

Inicialmente Pig fue creado para la consola (ronco Shell Shell). En la ejecución de la obra con Cloudera Pig a través de un simple y fácil de usar interfaz web. Puede abrirlo a través de la interfaz Hue [http://\[.....\]\\_Hue\]:8888/pig/](http://[.....]_Hue]:8888/pig/)



*Figura 50 Interfaz web.*

Interfaz web incluye un editor de guión completo y gerente. Con ella, se puede guardar directamente en los guiones Hue, ejecutarlos, ver una lista de las tareas en ejecución, los resultados y los registros de aperturas.

## 5.9 Experimento.

A modo de experimentación, voy a realizar el procesamiento de los registros de acceso a la tienda para un día en particular (noche). Calculamos los siguientes parámetros:

- el número total de solicitudes;
- el número de solicitudes de cada IP única;
- el número de solicitudes de cada URL única;
- cantidad de datos transmitidos para cada URL.

A continuación, se muestra un script que permite resolver tareas. Este script (y todos los scripts en Pig) no se realiza por línea, en lenguajes interpretados. Cerdo compilador analiza según los arroyos y los conjuntos de datos. La compilación de una secuencia de comandos comienza con el final, es decir, con el comando STORE. Para los datos, después de lo cual no hay procesamiento de instrucciones tienda sin se crearán todos los problemas y no pueden leer los datos. Se le permite escribir un guión de una manera bastante arbitraria, todo el trabajo de optimización, para determinar el orden de ejecución y paralelización tomará Pig.

### Script:

```
records = LOAD '/log/flume/events/' USING PigStorage('\t')
AS (
date:chararray,
clientip:chararray,
clientport:chararray,
proto:chararray,
statuscode:int,
bytes:int,
sq:chararray,
bq:chararray,
request:chararray );

count_total = FOREACH (GROUP records ALL) GENERATE COUNT(records);

count_ip = FOREACH (GROUP records BY clientip) GENERATE group AS ip, COUNT(records) AS
cnt;
top_ip = ORDER count_ip BY cnt DESC;

filtered_req = FILTER records BY statuscode == 200 OR statuscode == 206;
count_req = FOREACH (GROUP filtered_req BY request) GENERATE group AS req,
COUNT(filtered_req) AS cnt, SUM(filtered_req.bytes) AS bytes;
top_req = ORDER count_req BY bytes DESC;

%declare DT `date +%y%m%dT%H%M`
STORE count_total INTO '$DT/count_total';
STORE top_ip INTO '$DT/top_ip';
STORE top_req INTO '$DT/top_req';
```

Se compone de tres partes: la carga de datos, procesamiento y guardar. Este procedimiento es común para la mayoría de tareas. En algunos casos, la resolución de

problemas puede incluir pasos adicionales. Por ejemplo, la generación o mantenimiento de los resultados de cálculo intermedios.

### Etapas del script

#### 1. Carga:

```
records = LOAD '/log/flume/events/' USING PigStorage('\t')
AS (
date:chararray,
clientip:chararray,
clientport:chararray,
proto:chararray,
statuscode:int,
bytes:int,
sq:chararray,
bq:chararray,
request:chararray );
```

Como una entrada utilizo registros del servidor web. Ejemplo de los datos entados:

```
08/Sep/2016:20:05:13 95.153.193.56 37877 http 200 1492030 0 0 GET /745dbda3-894e-43aa-9146-607f19fe4428.mp3 HTTP/1.1
```

```
08/Sep/2016:15:00:28 178.88.91.180 13600 http 200 4798 0 0 GET /public/cars/bmw71/down.png HTTP/1.1
```

```
08/Sep/2016:15:00:29 193.110.115.45 64318 http 200 1594 0 0 GET /K1/img/top-nav-bg-default.jpg HTTP/1.1
```

El objeto principal de Pig es "actitud". Esa relación de trabajo con todos los operadores de la lengua. En la forma de la relación son los datos de entrada y salida.

Cada relación es un conjunto de objetos similares - "tuples". Análogos en la base de datos: "tuples" - es una cadena, la actitud - esta tabla.

En Pig es el resultado de cualquier relación de operador, que es un conjunto de tuples.

#### ***Transformación:***

Calculamos el número total de entradas en los registros utilizando la sentencia COUNT. Antes de eso, es necesario combinar todos los registros en los registros de los estados y el grupo FOREACH mismo grupo.

```
count_total = FOREACH (GROUP records ALL) GENERATE COUNT(records);
count_ip = FOREACH (GROUP records BY clientip) GENERATE group AS ip, COUNT(records) AS cnt;
top_ip = ORDER count_ip BY cnt DESC;
```

Ahora contamos el número de solicitudes con direcciones únicas. En tuplas en relación con los registros de campo IPCliente contiene la dirección IP desde la cual la solicitud. Grupo de las tuplas en records de la clientip campo y definir una nueva relación que consiste en dos campos:

- ip campo, que se toma a partir del nombre del grupo en relación con los registros;

- el número de entradas en el grupo - cnt, calculado por la cuenta del operador COUNT, es decir, el número de registros que corresponden a una determinada dirección IP en el campo IP.

A continuación, determinamos top\_ip actitud, que consiste en los mismos datos que el count\_ip, pero ordenados por el campo ORDEN cnt operador. Por lo tanto, en la lista de clientes se top\_ip direcciones IP desde la que se producen más a menudo solicitudes.

```
filtered_req = FILTER records BY statuscode == 200 OR statuscode == 206;
```

```
count_req = FOREACH (GROUP filtered_req BY request) GENERATE group AS req,
COUNT(filtered_req) AS cnt, SUM(filtered_req.bytes) AS bytes;
```

```
top_req = ORDER count_req BY bytes DESC;
```

Calculamos el número de solicitudes correctas para cada URL, y el volumen total de datos descargados para cada URL. Para ello, utilice la primera FILTRO operador de filtro, la selección de las solicitudes sólo tiene éxito con un HTTP 200 OK y 206 códigos de "Partial Content". Esta declaración define un nuevo filtered\_req relación de registros de vinculación, filtrarlo por el campo "statuscode".

A continuación, de forma similar al cálculo del número de cuenta dirección IP del URL única, que registra agrupación respecto de las solicitudes para el campo "request". Calculamos la cantidad de datos para cada URL usando el operador de "SUM", por el campo en bytes agrupados relaciones "filtered\_req" registros.

Ordenamos de los bytes de campo, definiendo un nuevo top\_req relación.

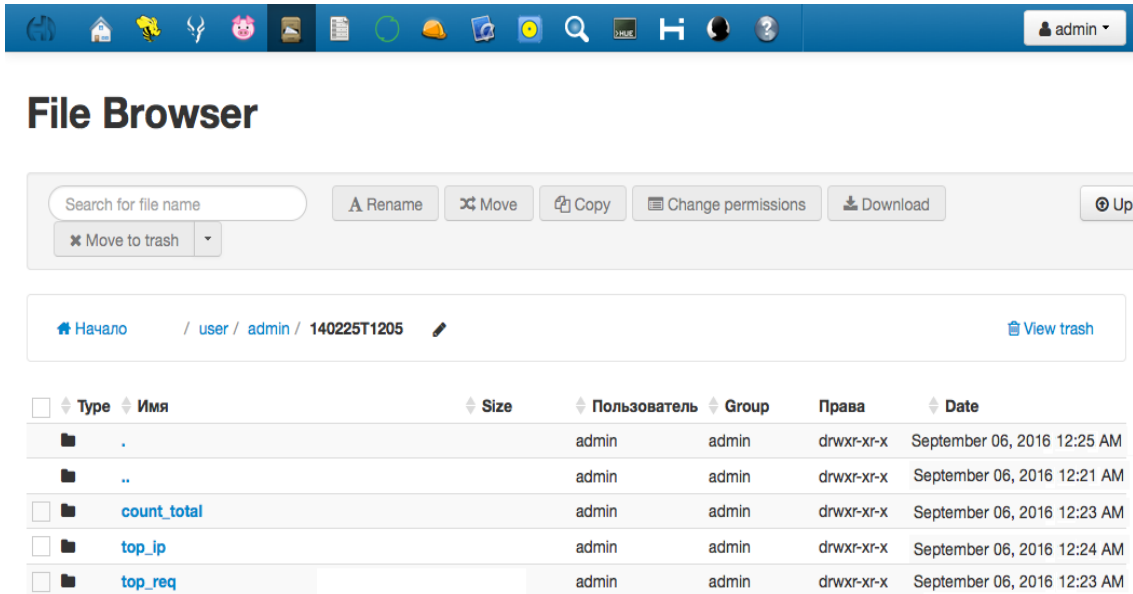
#### Guardar los resultados y conclusiones:

```
%declare DT `date +%y%m%dT%H%M`
STORE count_total INTO '$DT/count_total';
STORE top_ip INTO '$DT/top_ip';
STORE top_req INTO '$DT/top_req';
```

Preferiblemente almacenar los resultados de cada ejecución del script en un directorio independiente cuyo nombre incluye la fecha y la hora de la ejecución. Como resultado de la experimentación "date" comando se introduce en la variable DT, que luego se sustituye dentro de los datos de ruta de almacenamiento. Mandato para guardar los resultados STORE: cada relación en su catálogo.

Ver los datos salidos es posible a través del administrador de archivos en la ruta predeterminada Hue. De forma predeterminada, la ruta es relativa al HDFS en el directorio principal del usuario que ejecuta script.





*Figura 51 Gestor de archivos.*

### **Información sobre los resultados de la ejecución de tareas:**

<http://cdh3:8888/pig/#logs/1100715>

Input(s):

Successfully read 184442722 records (32427523128 bytes) from: "/log/flume/events/"

Output(s):

Successfully stored 1 records (10 bytes) in: "hdfs://cdh3:8020/user/admin/140225T1205/count\_total"

Successfully stored 8168550 records (1406880284 bytes) in:  
"hdfs://cdh3:8020/user/admin/140225T1205/top\_req"

Successfully stored 2944212 records (49039769 bytes) in: "hdfs://cdh3:8020/user/admin/140225T1205/top\_ip"

Counters:

Total records written : 11112763

Total bytes written : 1455920063

Informe de Oozie:

Last Modified Tue, 06 Sep 2016 00:22:00

Start Time Tue, 06 Sep 2016 00:05:16

Created Time Tue, 06 Sep 2016 00:05:15

End Time Tue, 06 Sep 2016 00:22:00

Después de que el final del experimento se manejó más de 180 millones de registros, por un total de más de 32 GB. El régimen de todo nos llevó unos 15 minutos.

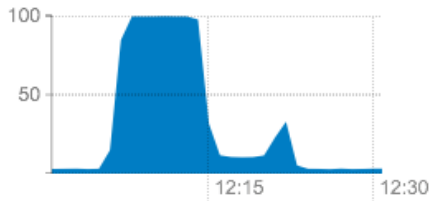
"Pig durante la ejecución del script crea un "tarear" MapReduce y los envía a realizar en el cluster "MR". Este proceso se ilustra claramente en las gráficas de las estadísticas en el panel de control "Cloudera Manager":

## Charts

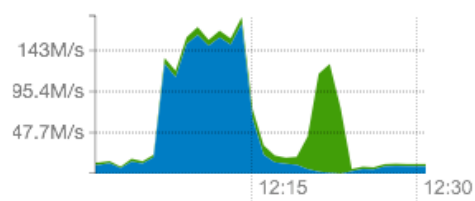
30m 1h 2h 6h 12h 1d

### Cluster 1 - CDH4

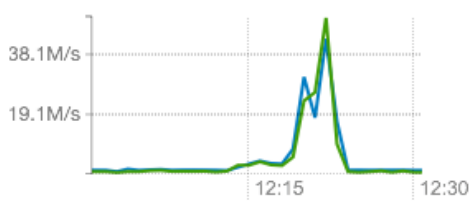
**Cluster CPU**  
percent



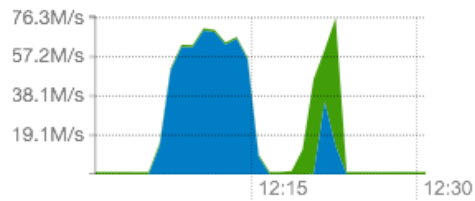
**Cluster Disk IO**  
bytes / second



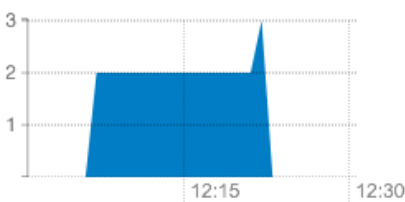
**Cluster Network IO**  
bytes / second



**HDFS IO**  
bytes / second



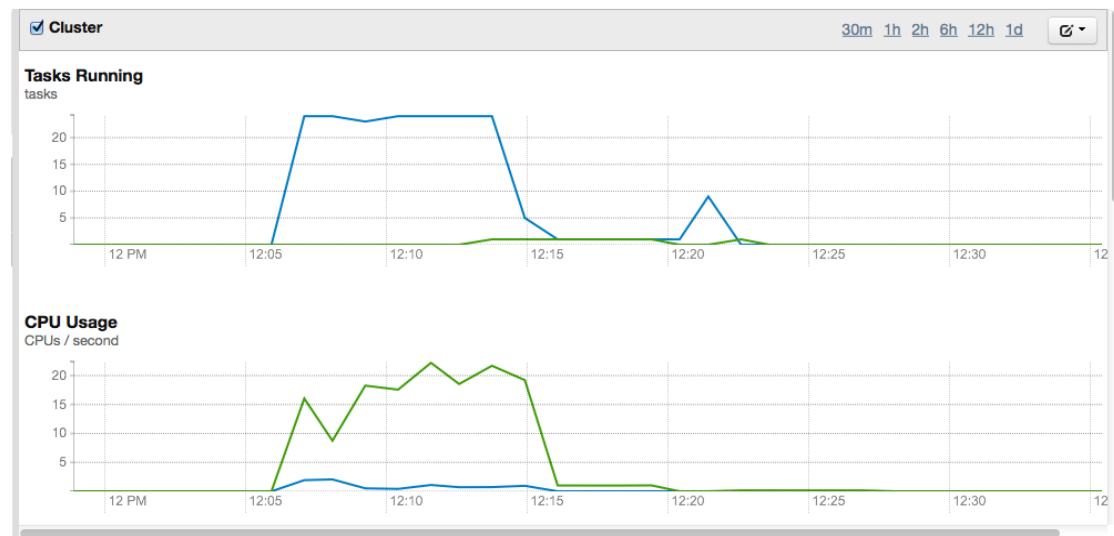
**Running MapReduce Jobs**  
jobs



**HBase Requests**  
requests



*Figura 52 Gráficos estadísticos.*

**Activities : mapreduce1***Figura 53 Gráficos estadísticos.*

1. Etapa "Map" : procesadores y discos en cada nodo ocupada procesando datos de sus partes.
2. Etapa "Reduce": resultados obtenidos en la primera etapa, y se transmiten por la red juntos.
3. **La tercera etapa** los resultados se almacenan en el sistema de archivos (se puede ver en la grabación de un salto en HDFS gráfico).

En los gráficos se puede ver que el trabajo hecho experimental incluye dos carreras de MapReduce. Durante la primera pasada cuentan registros únicos y se llevó a cabo en la segunda la clasificación. Estos procedimientos no pueden ser paralelizados y llevan a cabo en un solo paso, como un segundo procedimiento de los primeros resultados de trabajo.

## 5.10 CryptoARM Linux (Trusted eSign) ver. 0.1.0

### 5.10.1 Introducción.

12 de mayo de 2017 comenzó a atacar el virus Wanna Cry. Para unas horas, decenas de miles de ordenadores han sido infectados en todo el mundo. Hasta la fecha, más de 45.000 ordenadores infectados. En Rusia, fue atacada por las instituciones gubernamentales, los hospitales, los operadores móviles. En empresas rusas, que utilizan Cripto Pro para cifrar los datos el virus Wanna Cry no podría infectar los datos. En este trabajo desarrollará un mecanismo para la instalación del programa piloto con se llama Cripto ARM 0.1.0, también desarrollará un mecanismo funcionar Cripto ARM +Cripto Pro conjuntos, que mejorará la seguridad y el uso de claves de cifrado de datos y eToken. Se desarrollará el software para probar Cripto ARM que muestran el cifrado de datos en tiempo con GOST 28147-89.

En este trabajo he utilizado nuevos programas experimentales con beta testado:

CryptoARM Linux (Trusted eSign) es una aplicación personalizada para la firma electrónica y el cifrado de archivos en sistemas operativos Linux.

La aplicación se basa en un moderno " núcleo" NW.js (WebKit nodo), llame para las operaciones criptográficas utilizadas librería OpenSSL. En la versión 0.1.0 certificados y las claves se almacenan en un archivo de aplicaciones de almacenamiento.

CryptoARM Linux (Trusted eSign) incluye tres áreas de funcionar: funcionar con una firma digital, cifrado de datos y la gestión de certificados.

**CryptoARM Linux (Trusted eSign) puede soportar los algoritmos criptográficos:**

- **GOST P 34.10-2001** Los algoritmos para la generación y verificación de firma electrónica implementados
- **GOST P 34.11-2012** El algoritmo de generación de un valor hash
- **GOST 28147-89** Algoritmo de cifrado / descifrado de datos

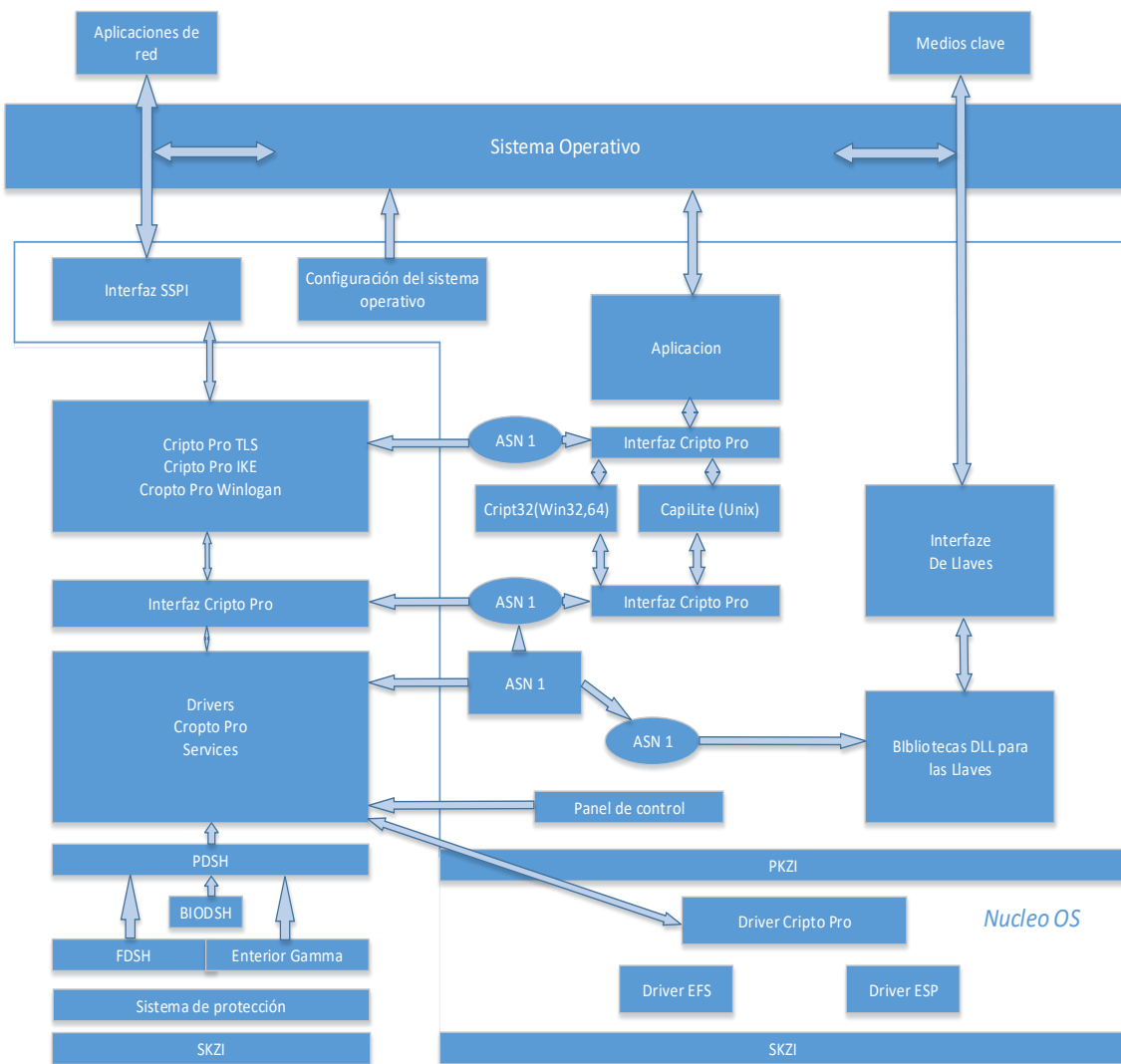
Para mejorar las seguridades desarrollar el mecanismo funcionar **CryptoARM Linux (Trusted eSign)** y **CriptoPro CSP**. Esto mecanismo permitirá eTokenes, pleno uso de certificados electrónicos.

## 5.11 CriptoPro CSP.

**Cryptopro CSP** - Utilidad de cifrado (criptografía). Se utiliza para crear las claves de cifrado y las claves de firma electrónica para el cifrado, la integridad y autenticidad de la información.

**Cryptopro CSP** le permite proteger la información confidencial en el intercambio de datos a través de Internet y garantizar la validez legal de los documentos electrónicos.

**Estructura CryptoPro:**



*Figura 54 Estructura CryptoPro.*

**Crypto Pro funciona en dos niveles:**

- Nivel de aplicación
- Nivel de nucleo OS

**El software incluye:**

- librerías DLL, servicios, drivers de Cripto Pro.
- Módulo de autenticación de red Cripto Pro TLS.
- Módulo de IKE para usar en protocolo Cripto Pro IPSec.
- Módulo de Cripto Pro Winlogon.
- Interfaz criptográfico Cripto Pro.

- El generador de números aleatorios de software (PDSH) con la instalación del generador de números aleatorios física (FDSH) hardware integrado y protección de software (PAA) del acceso no autorizado, BioDSH, escala exterior.
- La integridad del software de control.
- La ingeniería de sistemas y protección criptográfica.

### Abreviaturas

**BioDSH** - Bio sensor de números aleatorios

**SKZI** - Protección criptográfica de la información

**PKZI** - El paquete de software de herramientas de protección

**PDSH** - Sensor de software de números aleatorios

### Instalación los certificados en Linux

La principal utilidad para instalar certificados es certmgr (esta en /opt/cproccsp/bin/<la arquitectura>).

Manual: man 8 certmgr

#### Instalar un certificado raíz:

certmgr -inst -store root -file < el camino al archivo con el certificado >

#### Instalar un certificado personal:

certmgr -inst -file <путь к файлу с сертификатом> -cont < nombre del contenedor >

#### Instalación de un certificado otro:

certmgr -inst -file < el camino al archivo con el certificado >

## 5.12 El mecanismo para Instalacion Cripto Pro.

Los pacetes para instalar:

- lsb-cproccsp-base
- lsb-cproccsp-rdr
- lsb-cproccsp-capilite
- lsb-cproccsp-kc1
- lsb-cproccsp-kc2

El ejemplo solo para Ubuntu 14.04 x64 bit (los pacetes necesito instalar en este orden)

#### El codigo:

```
alien -kci lsb-cproccsp-base-4.0.0-4.noarch.rpm lsb-cproccsp-rdr-64-4.0.0-4.x86_64.rpm
lsb-cproccsp-capilite-64-4.0.0-4.x86_64.rpm lsb-cproccsp-kc1-64-4.0.0-4.x86_64.rpm
lsb-cproccsp-kc2-64-4.0.0-4.x86_64.rpm cproccsp-curl-64-4.0.0-4.x86_64.rpm
```

Despues tengo que instalar cripto-linux-amd64.tar

Con ayuda el programma "**cpconfig**" tengo que indicar el camino para CSP en libcurl (el camino para distintas arquitecturas diferentes)

#### El codigo:

```
cpconfig -ini \config\apppath -add string libcurl.so /usr/local/lib/64/libcurl.so
```

### Instalacion el certificado

El certificado tengo que pedir a los desarrolladores de Cripto Pro. Despues necesito instalar el certificado:

**El codigo:**

```
/opt/cproscsp/bin/amd64/certmgr -inst -store uMy -cont "\\.\ FLASH
\your_container_name' -provtype 75
```

Para probar la clave privada con el certificado se puede obtener en el centro de certificación de prueba

**El codigo:**

```
/opt/cproscsp/bin/amd64/cryptcp -creatcert -provtype 75 -provname "Crypto-Pro GOST
R 34.10-2001 KC1 CSP" -rdn 'CN.aaa.' -cont "\\.\HDIMAGE\test_container' -certusage
1.3.6.1.5.5.7.3.1 -ku -du -ex -ca http://cryptopro.ru/certsrv
```

Es necesario adjuntar un certificado expedido por este para instalar el certificado del contenedor, especificando KC2-proveedor:

**El codigo:**

```
/opt/cproscsp/bin/amd64/certmgr -inst -store uMy -cont "\\.\HDIMAGE\test_container' -
provtype 75 -provname "Crypto-Pro GOST R 34.10-2001 KC2 CSP"
```

Compruebe que no haya un manajo de llaves con un certificado

**El codigo:**

```
certmgr -list -store uMy
```

### Instalación y configuración gost\_capi

Gost Capi prestación de apoyo para las claves y algoritmos GOST. Distribución necesito pedir de los desarrolladores de CryptoPro.

**El codigo:**

```
alien -kci cproscsp-cpop-gost-64-4.0.0-4.x86_64.rpm
```

### Instalacion OpenSSL

Configuración de OpenSSL se realiza modificar el fichero de configuración openssl.cnf los parametros después de old\_section=new\_oids y los comentarios:

**El codigo:**

```
openssl_conf = openssl_def
```

```
[openssl_def]
```

```
engines = engine_section
```

```
[engine_section]
```

```
gost_capi = gost_section
```

```
[gost_section]
```

```
engine_id = gost_capi
```

```
dynamic_path = /opt/cprosp/cp-openssl/lib/amd64/engines/libgost_capi.so
```

```
default_algorithms = CIPHERS, DIGESTS, PKEY, PKEY_CRYPT, PKEY_ASN1
```

Después necesito aprobar que OpenSSL utilize gost\_capi :

**El código:**

```
openssl engine
```

Ejemplo después de usar el código " openssl engine"

**El código:**

```
(rsax) RSAX engine support
```

```
(dynamic) Dynamic engine loading support
```

```
(gost_capi) CryptoPro ENGINE GOST CAPI ($Revision: 116890 $)
```

### 5.13 El mecanismo para instalación CryptoARM Linux (Trusted eSign sin GUI).

1. Crear una carpeta en la mano con todos los derechos /etc/opt/Trusted/Trusted API/license
2. Copiar la licencia temporal: /etc/opt/Trusted/Trusted API/license.lic
3. Copiar los datos de cifrar/descifrar por camino /home/osboxes/Downloads
3. En los scripts cambian las rutas de archivos para cifrar archivos y certificados.
4. Antes de ejecutar el plug-in de script "trusted-crypto" con ayudas: var trusted = require('trusted-crypto');

Después de los pasos CryptoARM Linux (Trusted eSign) funcionar solo CriptoPro CSP.

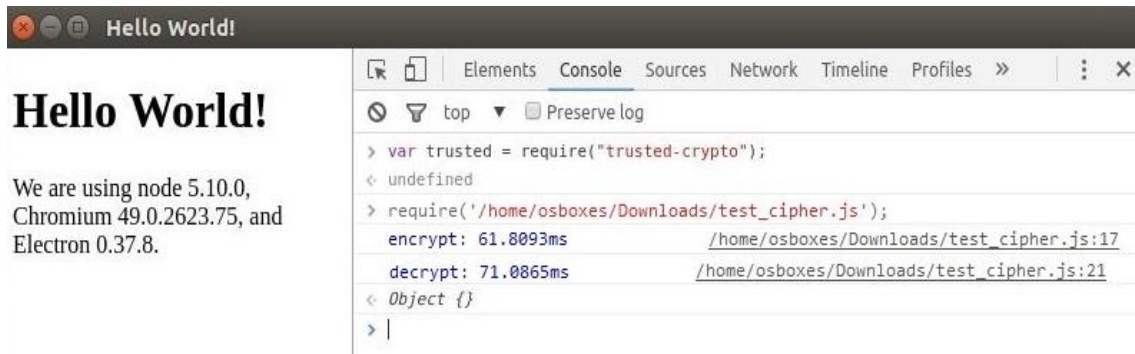
### 5.14 Los datos obtenidos durante el experimento.

Cifrado/descifrado los datos con tamanos :1Mb,10Mb,100Mb,200Mb



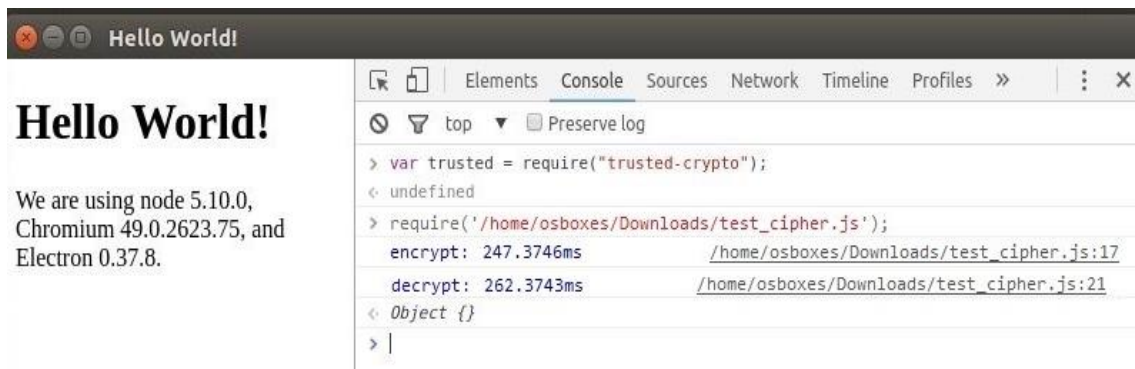
## ***Cifrado/descifrado con GOST***

*Experimento Nº1- 1 Mb*



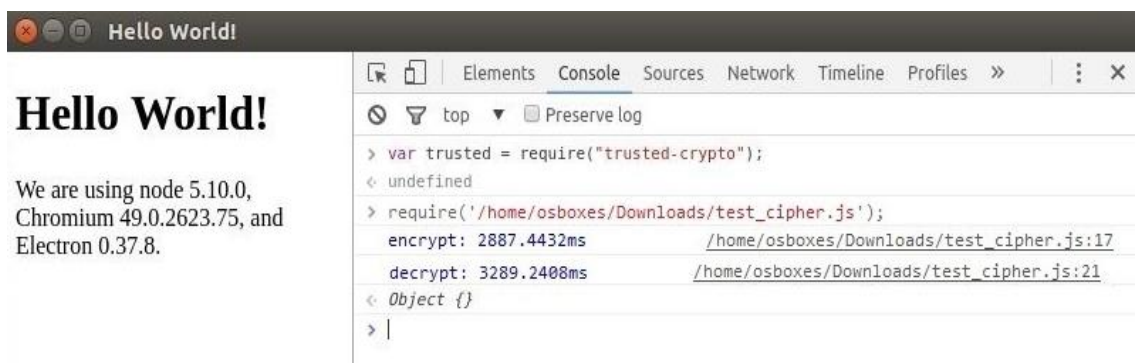
**Figura 55** *Cifrado/descifrado con GOST.*

*Experimento Nº2- 10Mb*



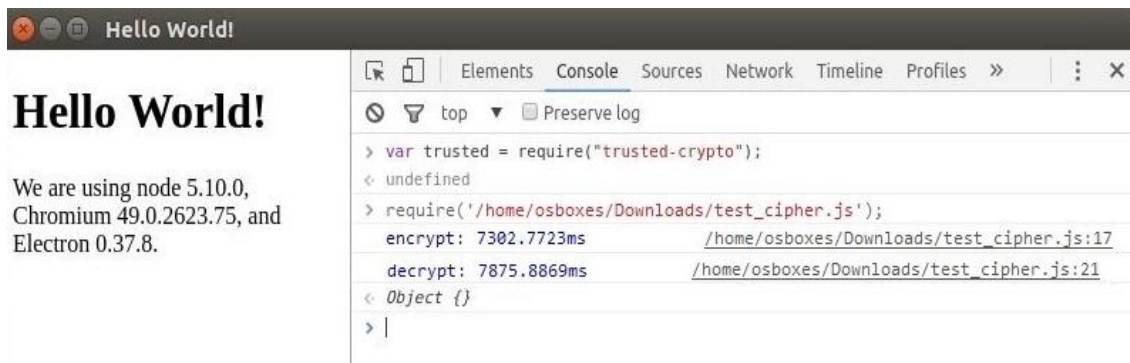
**Figura 56** *Cifrado/descifrado con GOST.*

*Experimento Nº3-100Mb*



**Figura 57** *Cifrado/descifrado con GOST.*

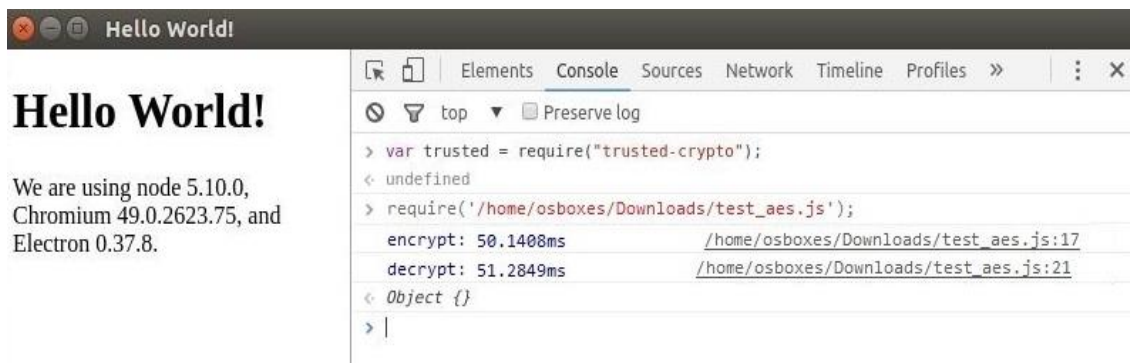
*Experimento Nº4-200Mb*



**Figura 58** Cifrado/descifrado con GOST.

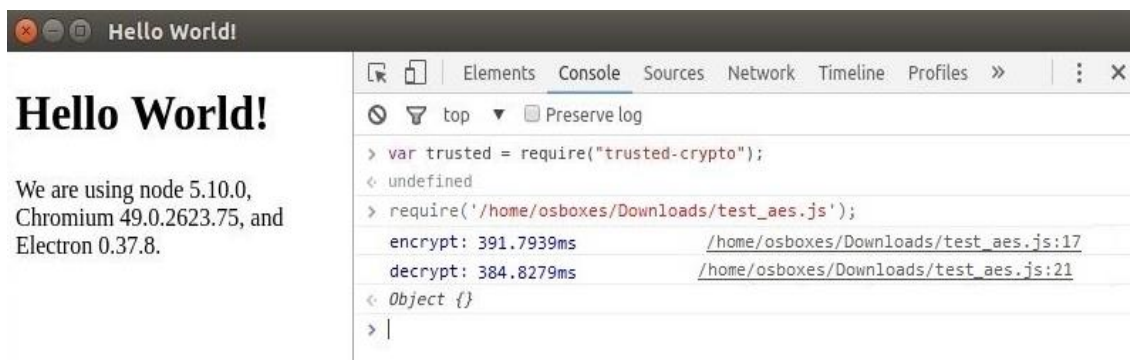
**Cifrado/descifrado con XTS-AES**

*Experimento Nº1- 1 Mb*



**Figura 59** Cifrado/descifrado con XTS-AES.

*Experimento Nº2- 10 Mb*



**Figura 60** Cifrado/descifrado con XTS-AES.

Experimento №3- 100 Mb

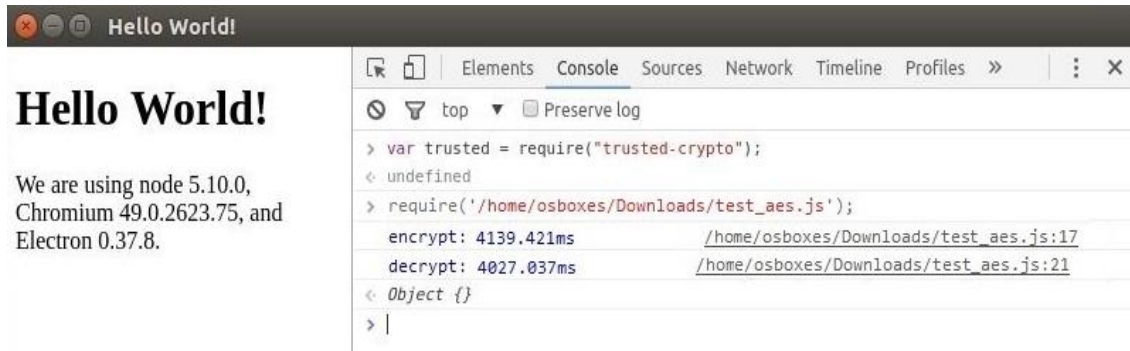


Figura 61 Cifrado/descifrado con XTS-AES.

Experimento №4- 200 Mb

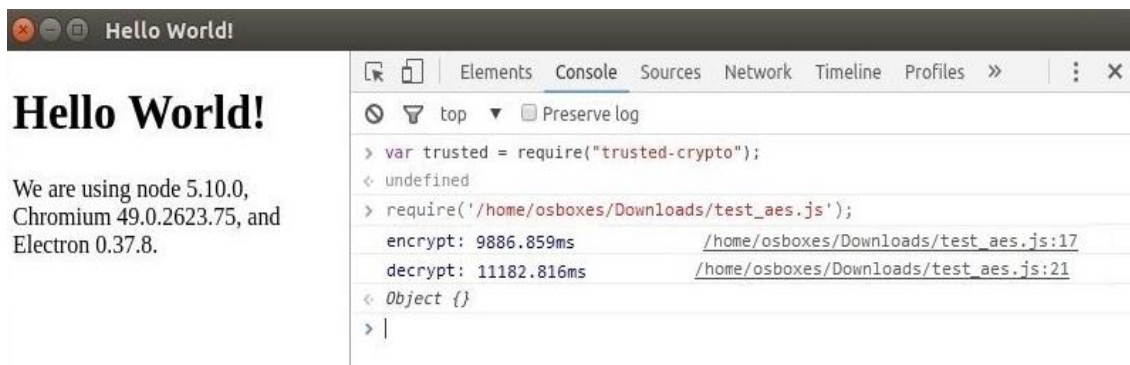


Figura 62 Cifrado/descifrado con XTS-AES.

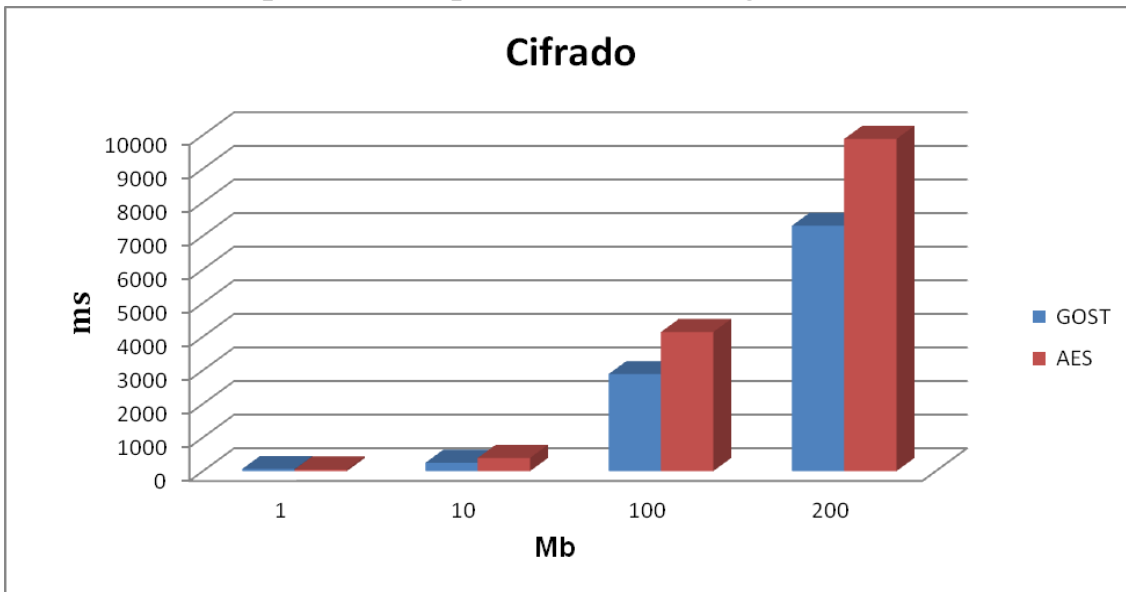
Los datos obtenidos durante el experimento:

El tiempo - Ms (milisegundo)

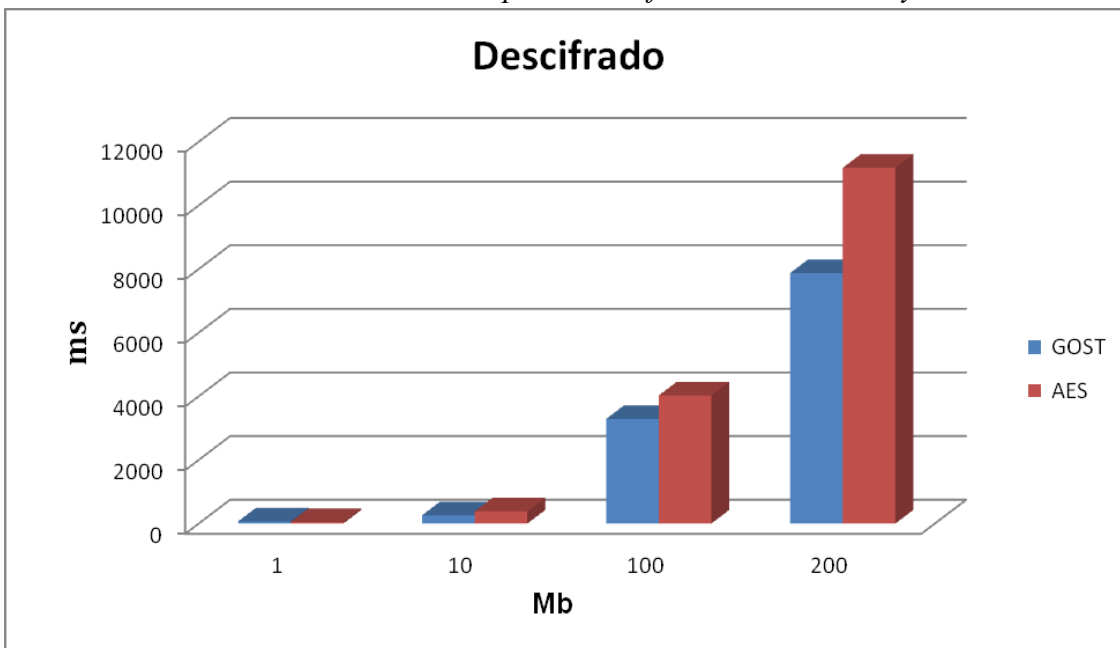
GOST 28147-89		
	Cifrado	Descifrado
1 M6	61,8093 Ms	71,0865 Ms
10 M6	247,3746 Ms	262,3743 Ms
100 M6	2887,4432 Ms	3289,2408 Ms
200 M6	7302,7723 Ms	7875,8869 Ms
XTS -AES		
	Cifrado	Descifrado
1 M6	50,1408 Ms	51,2849 Ms
10 M6	391,7939 Ms	384,8279 Ms
100 M6	4139,421 Ms	4027,037 Ms
200 M6	9886,859 Ms	11182,816 Ms

Tabla 1 Los datos cifrado/descifrado con GOST y AES-XTS.

**Gráficos para la comparación de dos algoritmos de cifrado:**



**Gráfico 1.** Comparación cifrados entre GOST y XTS-AES



**Gráfico 2.** Comparación descifrado entre GOST y XTS-AES

## 5.15 Apache Ranger.

Apache Ranger ofrece un enfoque integrado de seguridad para el clúster de Hadoop. Proporciona una plataforma centralizada para administrar y administrar las políticas de seguridad a través de los componentes de Hadoop.

Apache Ranger ofrece un sistema de seguridad para administrar el control de acceso:

Apache Hadoop HDFS

Apache Hive

Apache HBase

Apache Storm

Apache Knox

Apache Solr

Apache Kafka

Apache NiFi

YARN

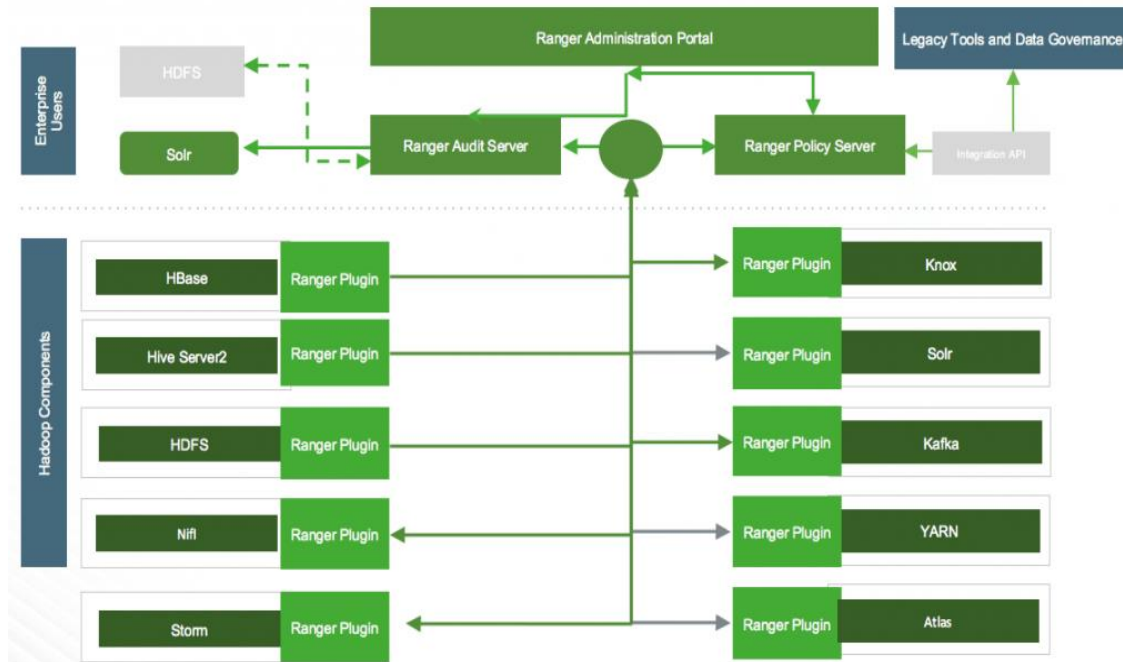
Con la consola de Apache Ranger, los administradores pueden administrar fácilmente las políticas para acceder a archivos, carpetas, bases de datos, tablas o columnas. Estas políticas se pueden establecer para usuarios individuales o grupos. Solo los administradores pueden acceder al clúster a través del protocolo SSH.

Key Management Service Ranger KMS proporciona un servicio escalable de administración de claves criptográficas para cifrar datos "en reposo" HDFS. Ranger KMS se basa en Hadoop KMS, originalmente desarrollado por la comunidad Apache, y amplía la funcionalidad de Hadoop KMS, lo que permite a los administradores del sistema almacenar claves en una base de datos segura.

Ranger también ofrece a los administradores de seguridad una gran visibilidad en su entorno Hadoop a través de una ubicación de auditoría centralizada que supervisa todas las solicitudes de acceso en tiempo real y admite múltiples destinos, incluidos HDFS y Solr.

### 5.15.1 Cómo funciona Apache Ranger?

Apache Ranger tiene una arquitectura descentralizada con los siguientes componentes internos:



*Figura 63 Los componentes de Apache Ranger.*

**Ranger admin portal:** el portal de administración de Ranger es la interfaz central para la administración de seguridad. Los usuarios pueden crear y actualizar políticas que luego se almacenan en la base de datos de políticas. Los complementos dentro de cada componente verifican regularmente estas políticas. El portal también consiste en un servidor de auditoría que envía datos de auditoría recopilados de los complementos para su almacenamiento en HDFS o en una base de datos relacional.

**Ranger plugins - Plugins** estos son programas ligeros de Java que están integrados en los procesos de cada componente del clúster. Por ejemplo, el complemento Apache Ranger para Apache Hive está integrado en Hiveserver2. Estos complementos toman políticas del servidor central y las almacenan localmente en un archivo. Cuando una solicitud del usuario llega a través del componente, estos complementos interceptan la solicitud y la evalúan de acuerdo con la política de seguridad. Los complementos también recopilan datos de la solicitud del usuario y siguen un hilo separado para enviar estos datos nuevamente al servidor de auditoría.

**User group sync - Apache Ranger** proporciona una utilidad de sincronización para que los usuarios extraigan usuarios y grupos de Unix o de LDAP o Active Directory. La información del usuario o grupo se almacena en el portal Ranger y se usa para definir la política.

### 5.15.2 Ventajas de Apache Ranger.

- Gestionar políticas para acceder a archivos, carpetas, bases de datos, tablas o columnas.
- Acceso SSH con clave RSA
- Licencia gratis

### 5.15.3 Desventajas de Apache Ranger.

- Sin soporte para el certificado GOST
- Un sistema complejo en instalacion y configuraciones.

	<b>CriptoLogin</b>	<b>Apache Ranger</b>
<b>Ventajas</b>	Soporte de certificado GOST	Gestionar políticas para acceder a archivos, carpetas, bases de datos, tablas o columnas.
	Soporte de autenticación de acceder en distintos web interfaces con certificado GOST	Acceso SSH con clave RSA
	Sistema fácil en instalacion y configuraciones	Licencia gratis
<b>Desventajas</b>	Sin soporte políticas para acceder a archivos, carpetas, bases de datos, tablas o columnas	Sin soporte para el certificado GOST
	Licencia GOST es pagado	Un sistema complejo en instalacion y configuraciones.

*Tabla 2 Las ventajas /desventajas Criptologin y Apache Ranger.*

## 5.16 Desarrollo de software para cifrado de datos con algoritmo GOST-28147.

### 5.16.1 Sobre la programación desarrollada

- Lengua de programación C#
- Cifra los contenidos de un archivo con algoritmo GOST-28147 con una llave diferente y una sincronización y almacena el resultado.

- Descifra un archivo cifrado con llave y sincronización

Para ejecutar este programa, debe crear el archivo test.txt en el directorio actual.

El ensamblado se realiza ejecutando el compilador, pasándolo como un parámetro para el nombre de archivo del programa, así como parámetros adicionales del compilador.

Para compilar el programa puede usar Microsoft Visual Studio 2010 o SharpDevelop o dentro de csc.exe

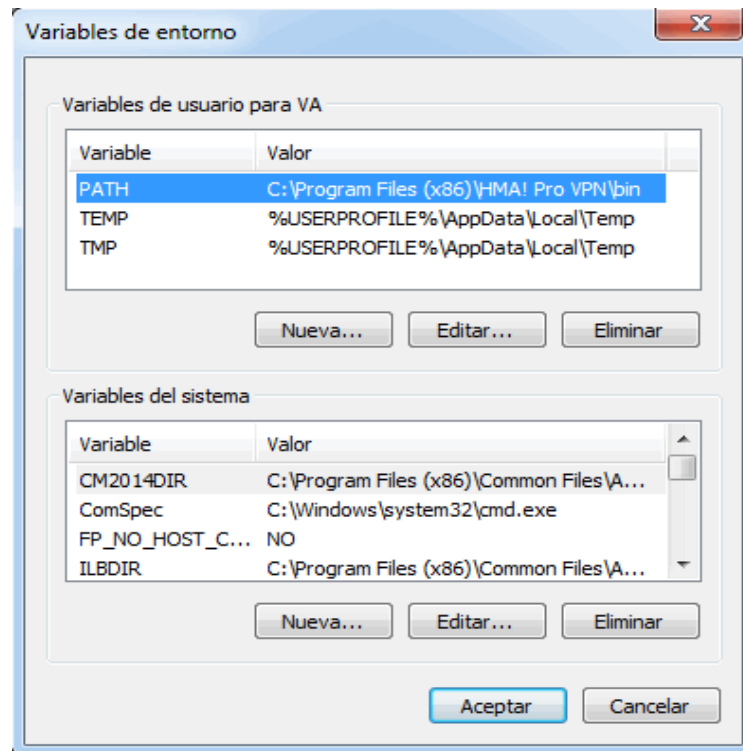
El programa se inicia llamando al archivo ejecutable correspondiente con los parámetros necesarios.

### 5.16.2 Guía como compilar el programacion.

Para compilar podemos usar el Microsoft Visual Studio 2008 o Sharp Develop 5.1. He usando compilador que incluye Windows.

- **Cambiar la configuración del entorno**

C:\Windows\Microsoft.NET\Framework64\v4.0.30319



*Figura 64 Cambiar la configuración del entorno.*

- **Verifique la versión del compilador, debe 4.7 y mas.**

```
C:\Program Files (x86)\Crypto Pro\.NET SDK \Encrypt\cs>csc
Microsoft (R) Visual C# Compiler version 4.7.2558.0
```

- Buscamos el camino con **CryptoPro.Sharpei.Base**



C:\Windows\Microsoft.NET\assembly\GAC\_MSIL\CryptoPro.Sharpei.Base\v4.0\_1.4.0.10\_\_473b8c5086e795f5

**- Compilar el archive**

```
C:\Program Files (x86)\Crypto Pro\.NET SDK\Encrypt\cs>csc
/r:C:\Windows\Microsoft.NET\assembly\GAC_MSIL\CryptoPro.Sharpei.Base\v4.0_1.4.0.10__473b8c5086e795f5\CryptoPro.Sharpei.Base.dll
EncryptDecryptRandomFile.cs
```





**- Compruebe si el archivo compilado**

```
C:\Program Files (x86)\Crypto Pro\.NET SDK\Encrypt\cs>dir
```

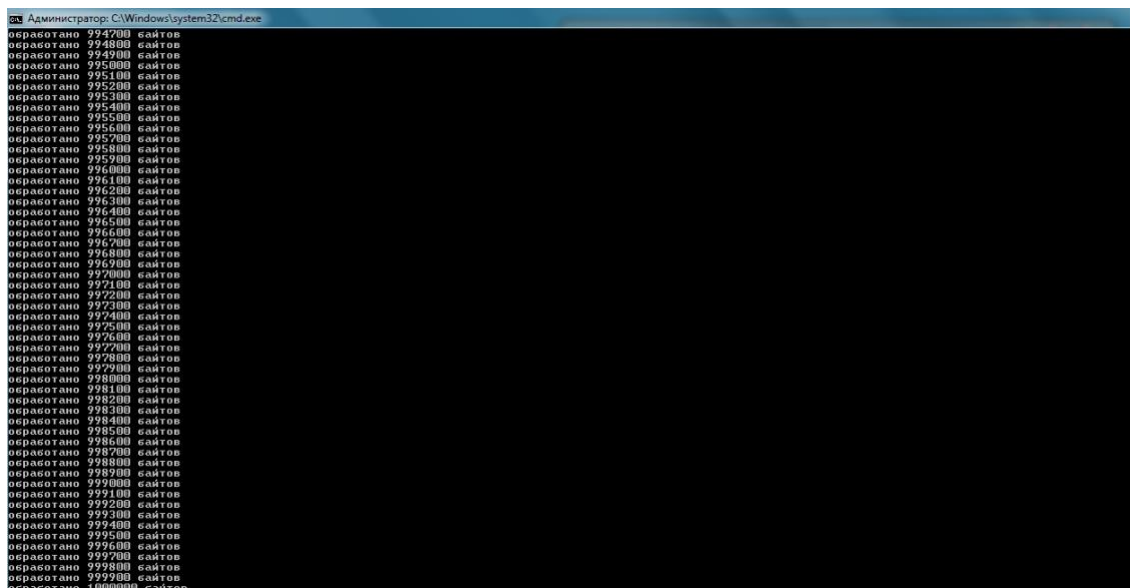
### 5.17 Los datos obtenidos durante el experimento:

Cifrado/descifrado los datos con tamaños :1Mb,10Mb,100Mb,200Mb  
Cifrado/descifrado con GOST 28147 - 89

**- Experimento №1- 1 Mb**

 EncryptDecryptRandomFile.exe	04.12.2018 10:31	5 КБ
 test.dec	05.12.2018 9:54	977 КБ
 test.enc	05.12.2018 9:54	977 КБ
 test.txt	05.12.2018 9:38	977 КБ

*Figura 65 Los ficheros de experimento №1.*



*Figura 66 Procesamiento de datosde experimento №1.*

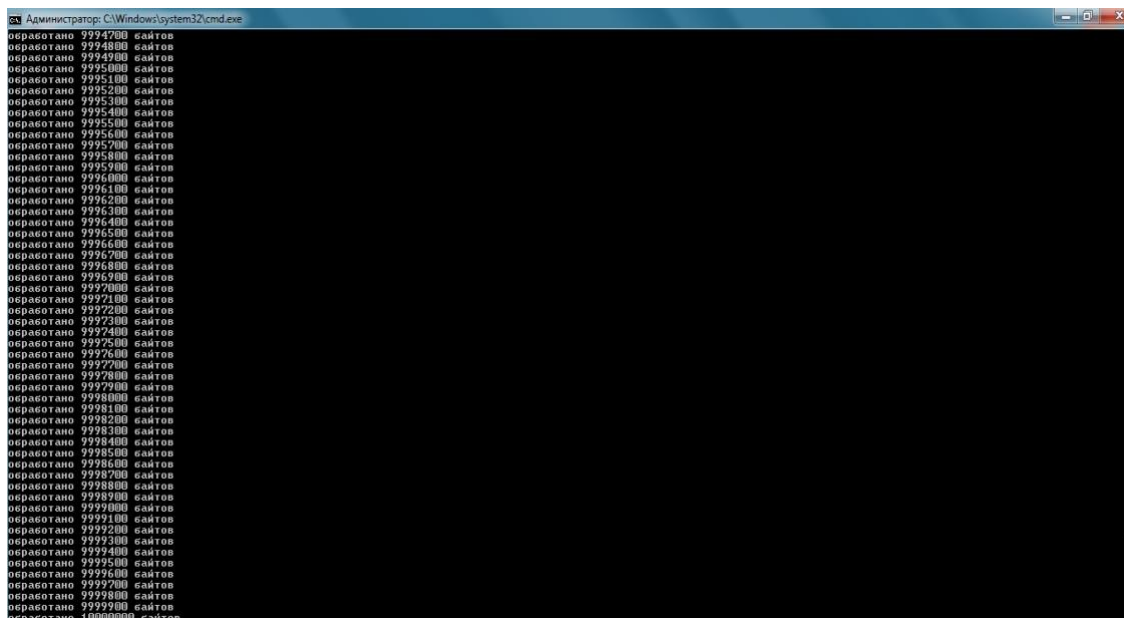
Cifrado	Descifrado
63,441 Ms	75,677 Ms

*Tabla 3 Cifrado/descifrado experimento №1.*

**- Experimento №2- 10 Mb**

 EncryptDecryptRandomFile.exe	04.12.2018 10:31	5 KB
 test.dec	05.12.2018 10:00	9 766 KB
 test.enc	05.12.2018 10:00	9 766 KB
 test.txt	05.12.2018 9:37	9 766 KB

*Figura 67 Los ficheros de experimento №2.*



*Figura 68 Procesamiento de datos de experimento №2.*

Cifrado	Descifrado
268,192 Ms	293,439 Ms

*Tabla 4 Cifrado/descifrado experimento №2.*

- Experimento №3- 100 Mb

EncryptDecryptRandomFile.exe	04.12.2018 10:31	5 КБ
test.dec	05.12.2018 10:07	97 657 КБ
test.enc	05.12.2018 10:06	97 657 КБ
test.txt	05.12.2018 9:38	97 657 КБ

Figura 69 Los ficheros de experimento №3.

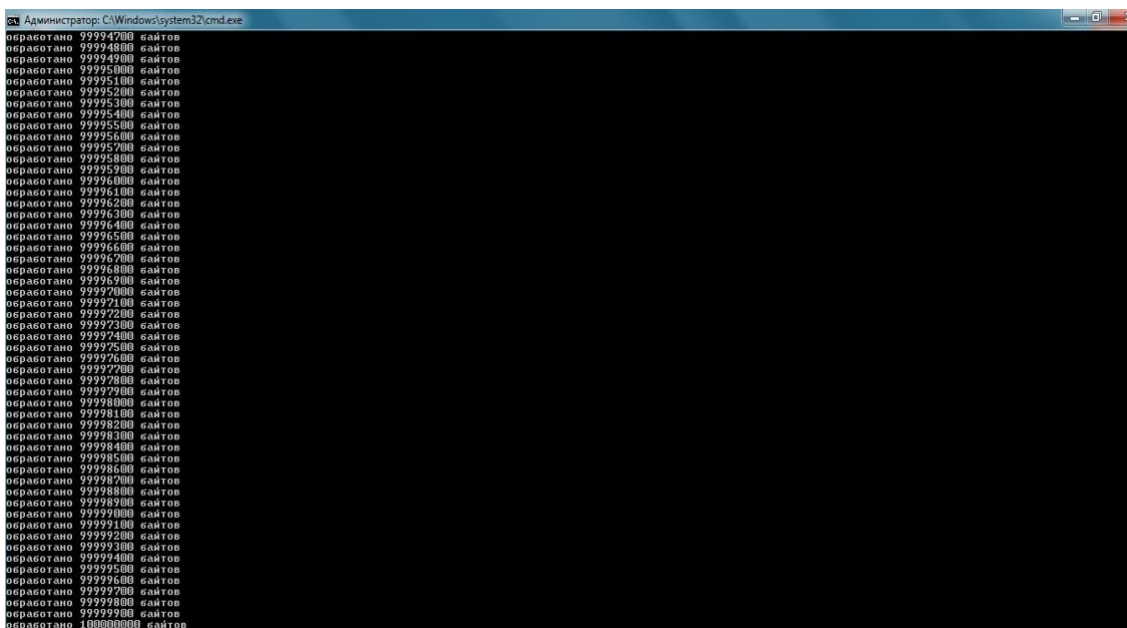


Figura 70 Procesamiento de datos de experimento №3.

Cifrado	Descifrado
2911,88 Ms	3455,959 Ms

Tabla 5 Cifrado/decifrado experimento №3.

- Experimento №4- 200 Mb

 EncryptDecryptRandomFile.exe	04.12.2018 10:31	5 KB
 test.dec	05.12.2018 10:14	195 313 KB
 test.enc	05.12.2018 10:12	195 313 KB
 test.txt	05.12.2018 9:38	195 313 KB

Figura 71 Los ficheros de experimento №4.

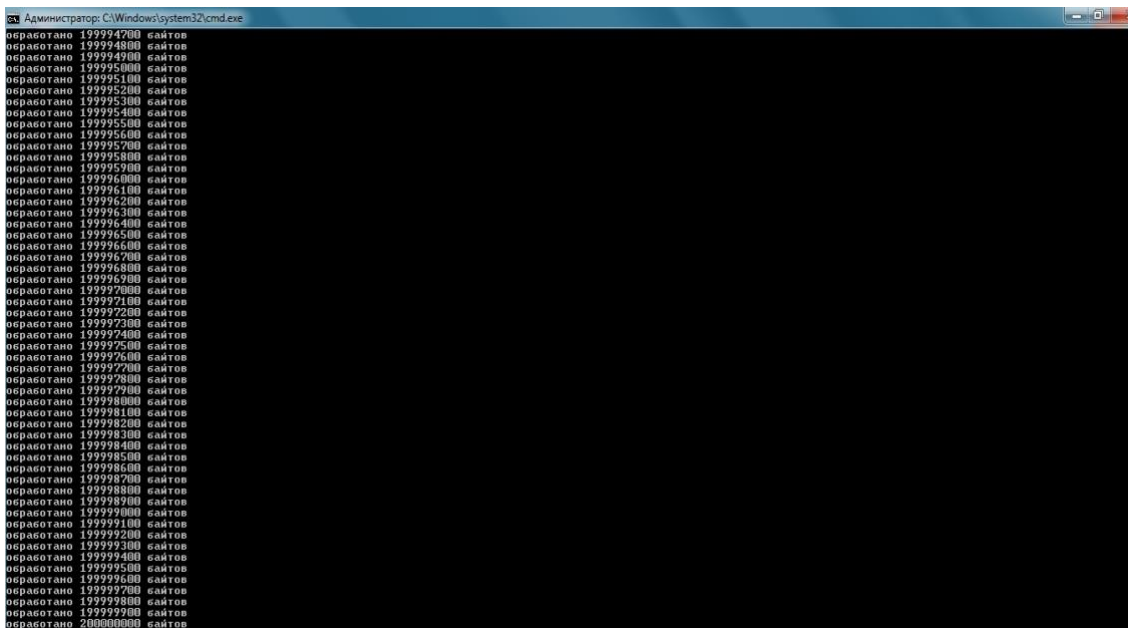


Figura 72 Procesamiento de datos de experimento №4.

Cifrado	Descifrado
7882,77 Ms	8185,619 Ms

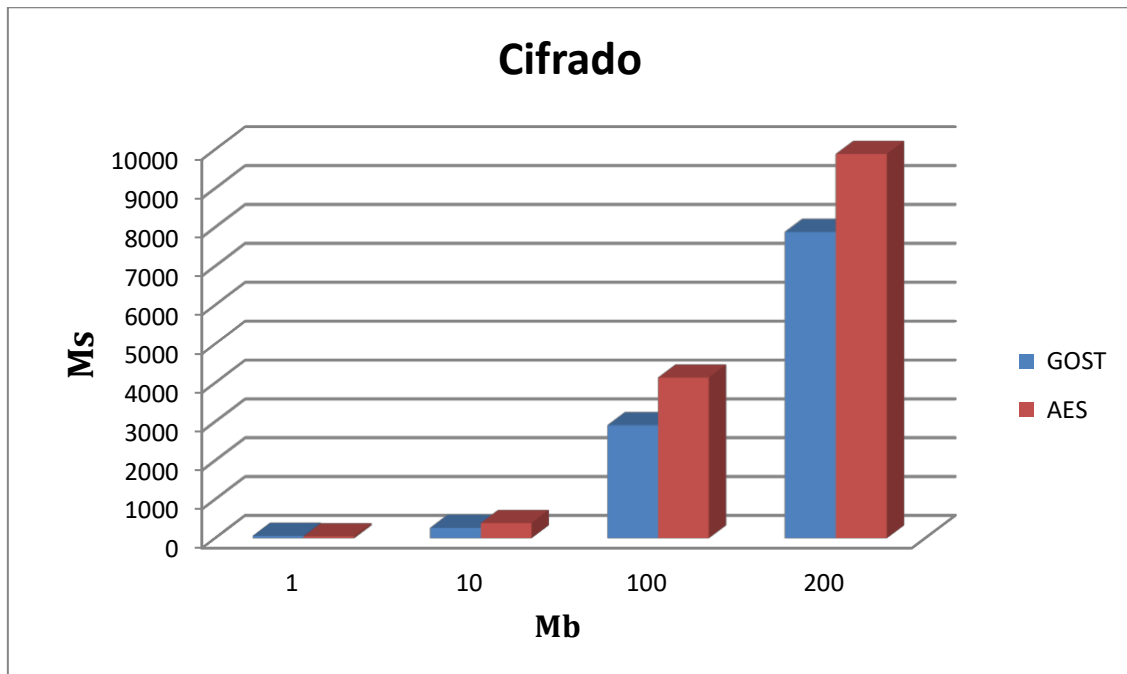
Tabla 6 Cifrado/descifrado experimento №4.

**Tabla de datos obtenido durante en el experimento**

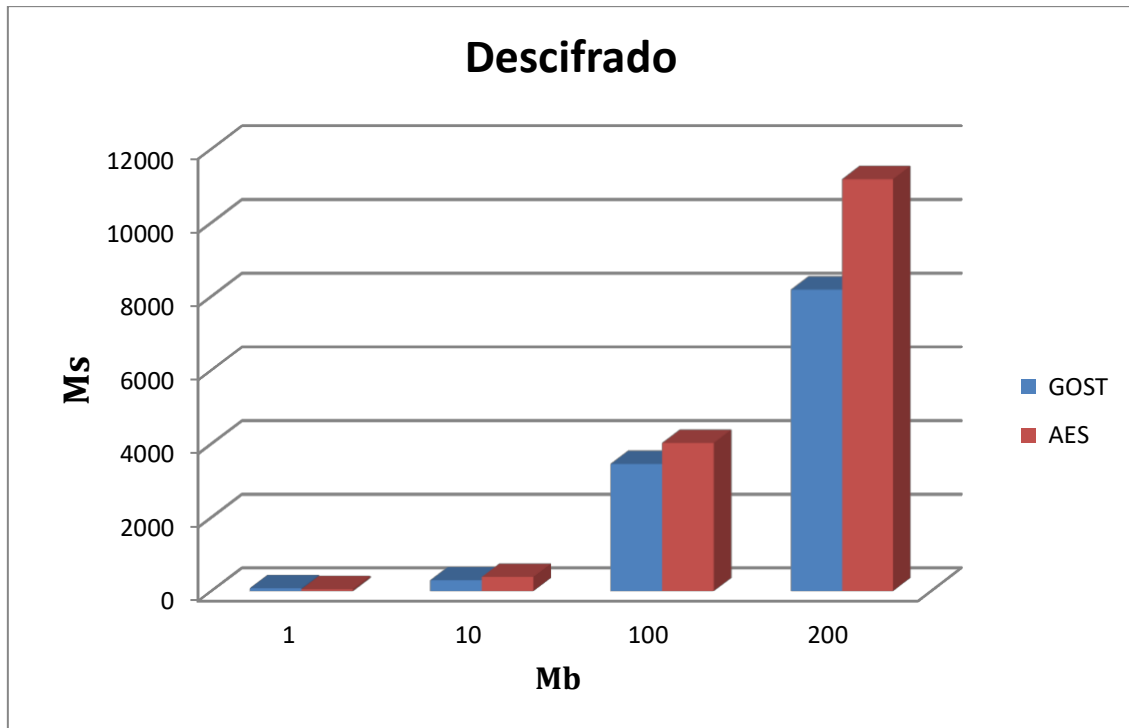
Tiempo (Ms)		
GOST 28147-89		
	<b>Cifrado</b>	<b>Descifrado</b>
1 Mb	63,441	75,677
10 Mb	268,192	293,439
100 Mb	2911,88	3455,959
200 Mb	7882,77	8185,619
XTS -AES		
	<b>Cifrado</b>	<b>Descifrado</b>
1 Mb	50,1408	51,2849
10 Mb	391,7939	384,8279
100 Mb	4139,421	4027,037
200 Mb	9886,859	11182,816

*Tabla 7 Cifrado/descifrado de experimentos.*

**Gráficos para la comparación de dos algoritmos de cifrado/descifrado.**



**Gráfico 3.** *Comparación descifrado entre GOST y XTS-AES*



**Gráfico 4.** Comparación descifrado entre GOST y XTS-AES

### 5.18 Conclusiones:

He desarrollado el programa para cifrar descifrar los datos con algoritmo GOST 28147-89.

Se han comprobado los datos obtenidos durante el experimento. El algoritmo GOST 28147-89 de cifrado/descifrado de datos es más rápido que XTS-AES.

## 5.19 Los experimentos con Azure.

El propósito del experimento es comparar la velocidad de búsqueda de palabras clave utilizando Azure y un ordenador normal.

Usamos en este experimento el fichero Excel con tamaño 576 Mb con 5 folios y 744870 columnas con información.

### Batería de prueba:

#### Ordenador normal:

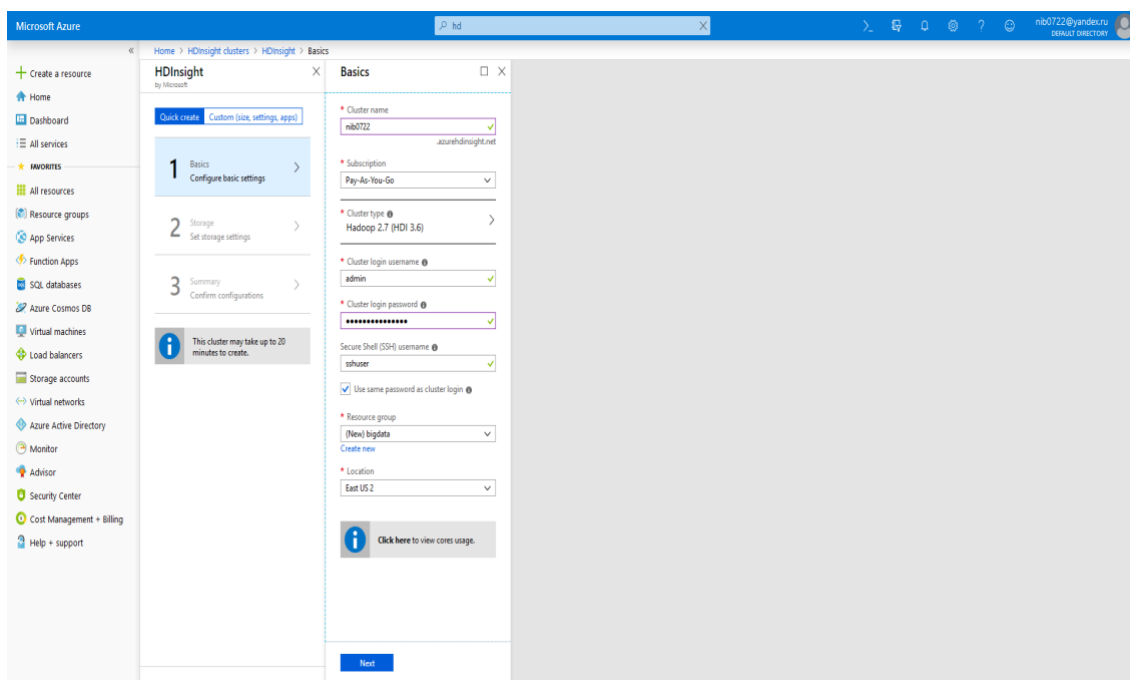
- 1.CPU i5 4460 -3,2Ghz
- 2.Memoria operativa - 8 Gb DDR3
- 3.SSD Kingston V300

#### Azure:

- Apache Hadoop y Azure HDInsight

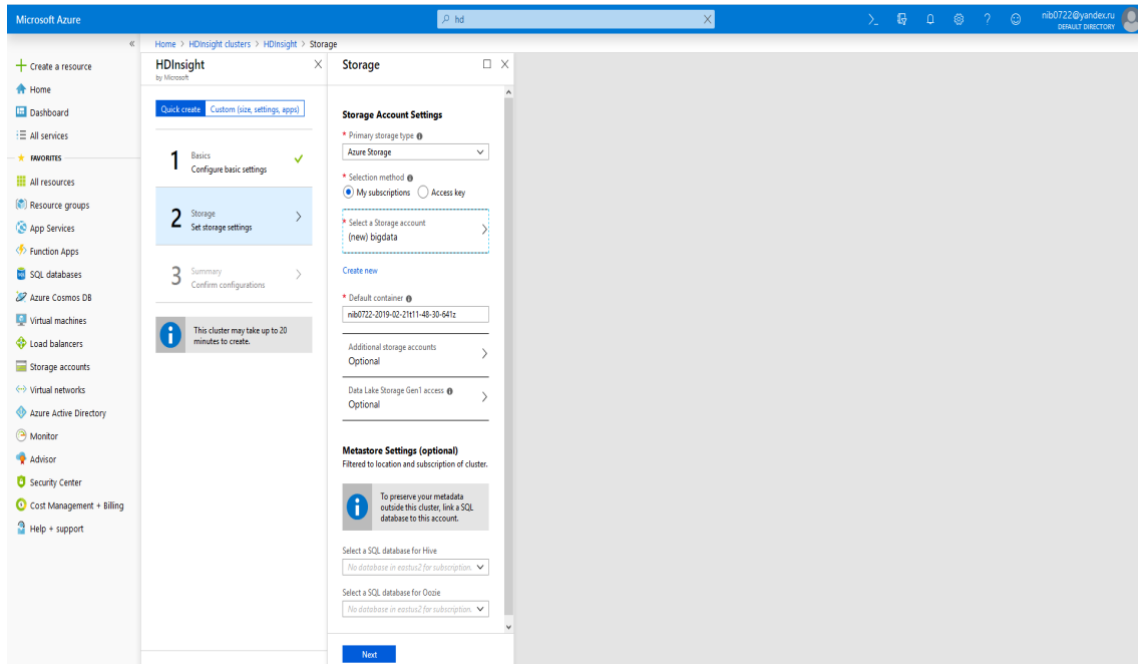
#### Instalamos HDInsight:

- **Primero paso:**



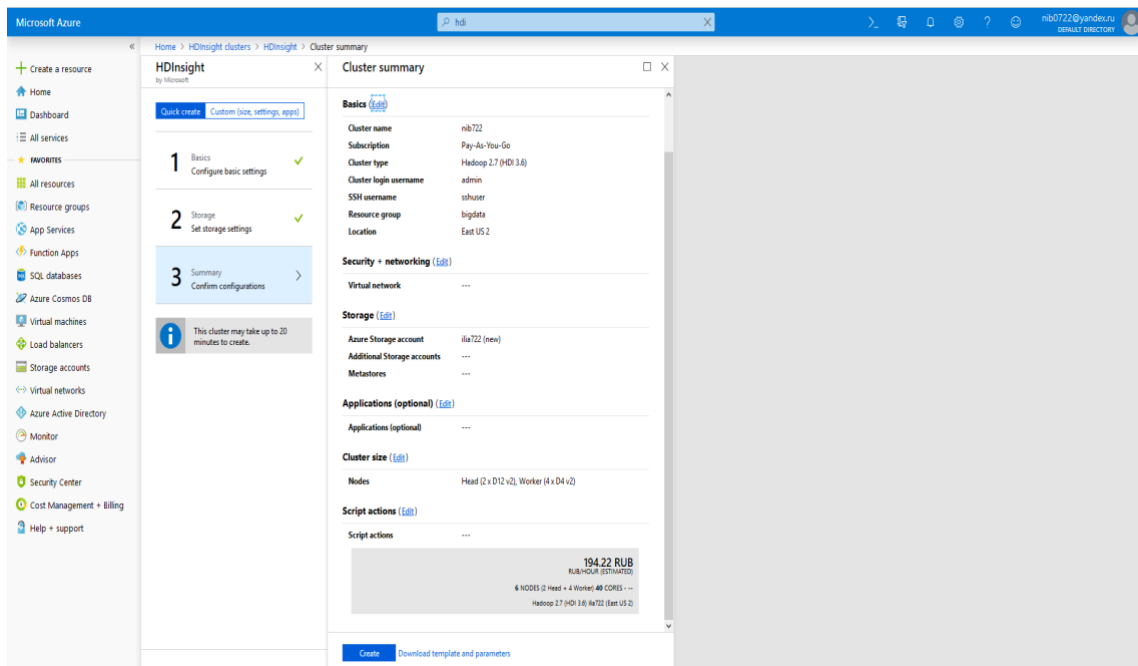
*Figura 73 Configurado HDInsight en Azure.*

- Segundo paso:



*Figura 74 Configurado HDInsight en Azure.*

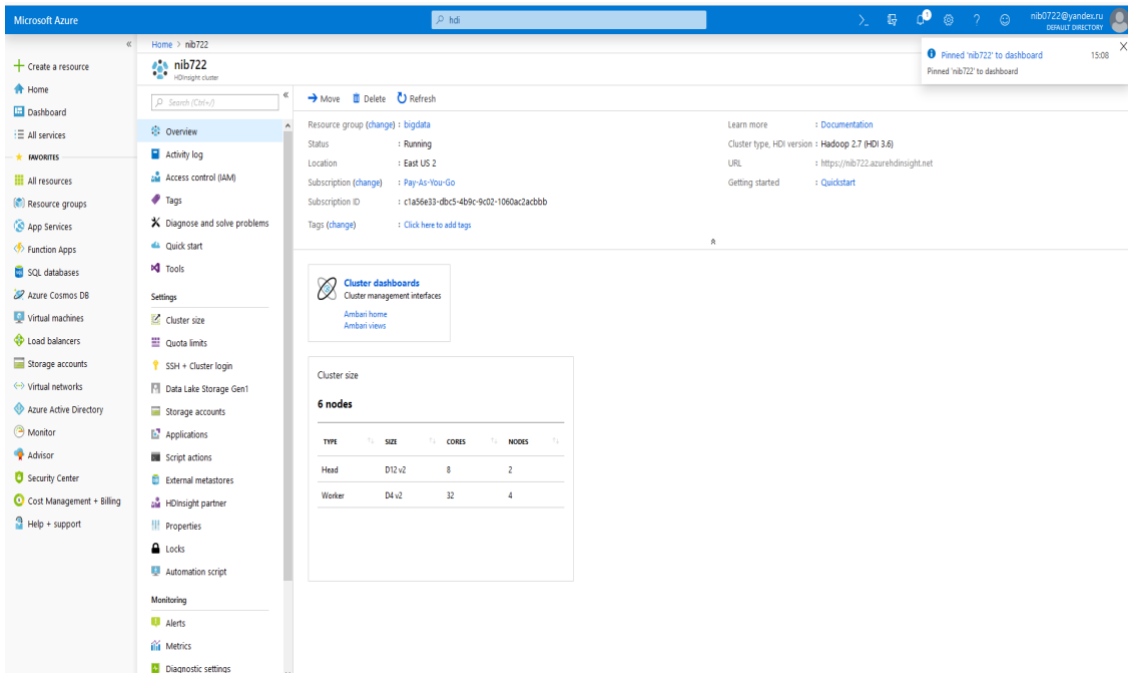
- Tercero paso:



*Figura 75 Configurado HDInsight en Azure.*

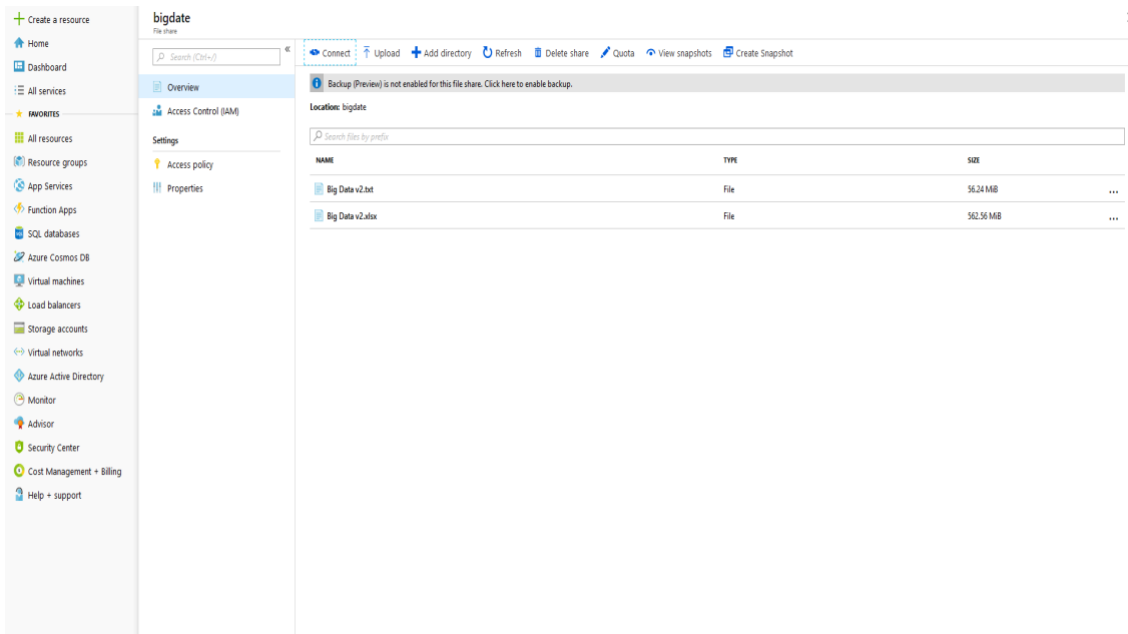


- **Paso Final:**



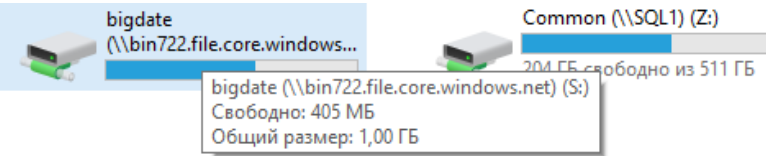
*Figura 76 Configurado HDInsight en Azure.*

- **Hacemos el contenedor y cargamos los ficheros con datos estructurados y sin estructura**



*Figura 77 Configurado HDInsight en Azure.*

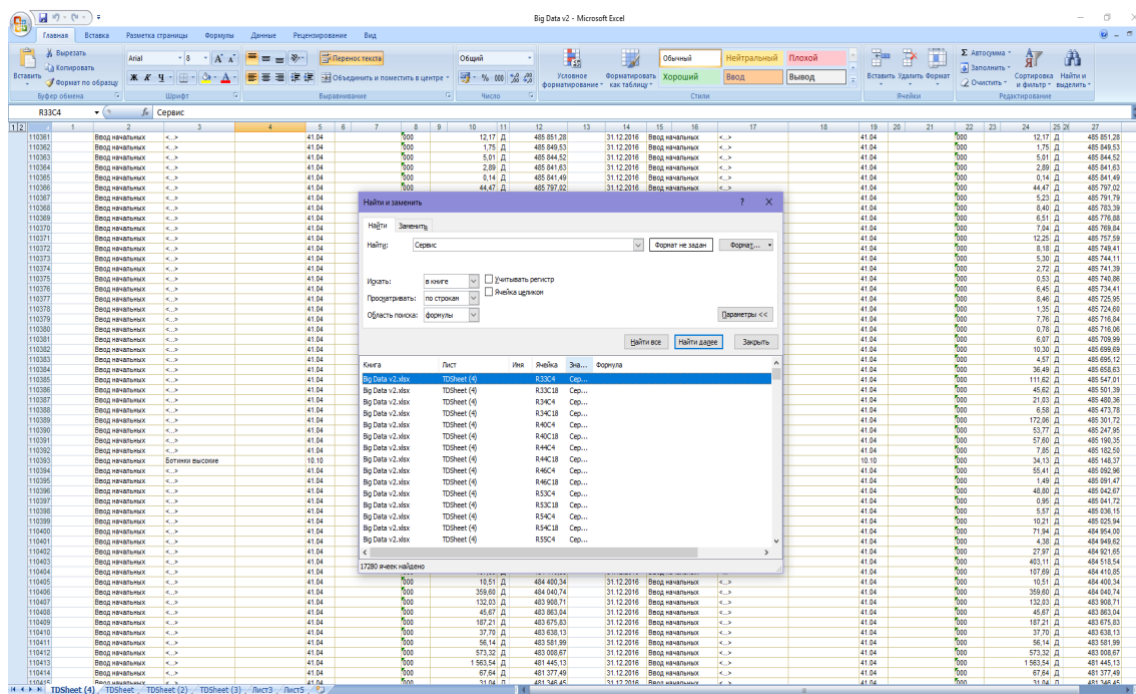
**Podemos iniciar el disco remoto**



*Figura 78 Iniciando el disco remoto.*

**5.20 Los experimentos con Excel:**

- Experimento №1 Buscar en el fichero en Azure una palabra:



*Figura 79 Buscar en el fichero en Azure una palabra.*

**El tiempo de búsqueda 725 sec. Las palabras iguales 27120 y células 17280**

- Experimento №2 Buscar y cambiar una palabra en el fichero en Azure:

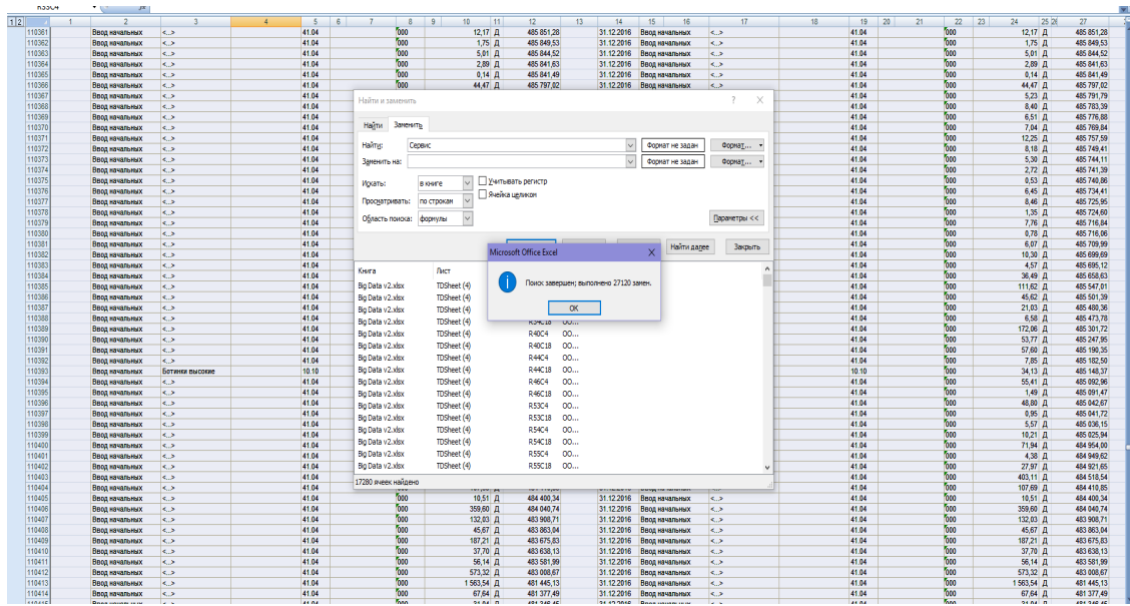


Figura 80 Buscar y cambiar una palabra el fichero en Azure.

**El tiempo de cambio 740 Ms. Las palabras iguales 27120 y células 17280**

- Experimento №3 Buscar en el fichero un folio sin estructura en Azure una palabra:



Figura 81 Buscar en el fichero en Azure una palabra (sin estructura de datos).

**El tiempo de busqueda 43 sec.**

- Experimento №4 Buscar y cambiar una palabra en el fichero un folio sin estructura en Azure:

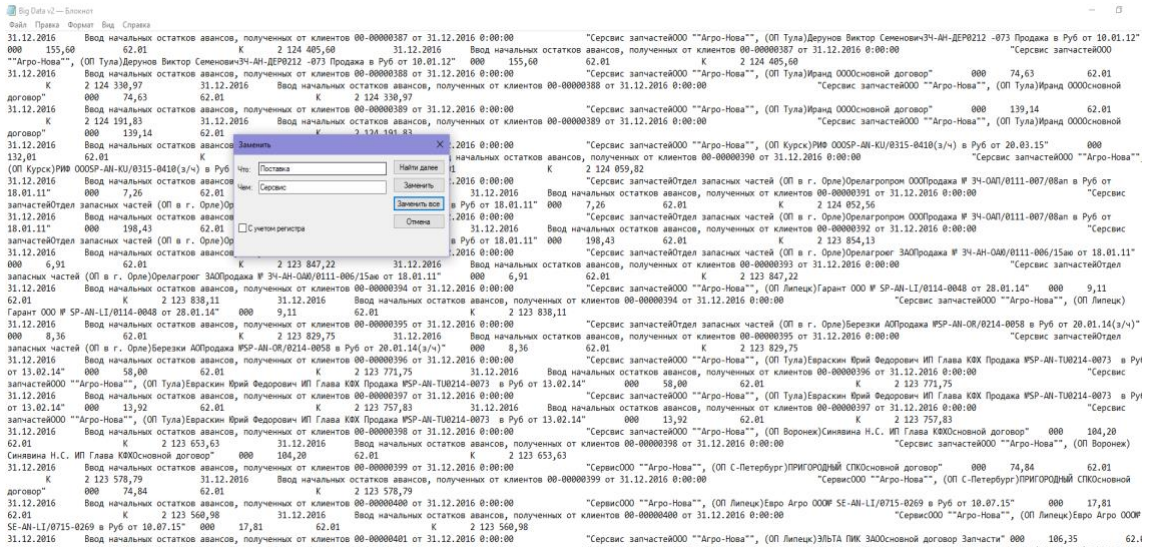


Figura 82 Buscar y cambiar en el fichero un folio sin estructura de datos.

El tiempo de cambiado 52 sec.

- Experimento №5 Buscar en el fichero en ordenador normal una palabra:

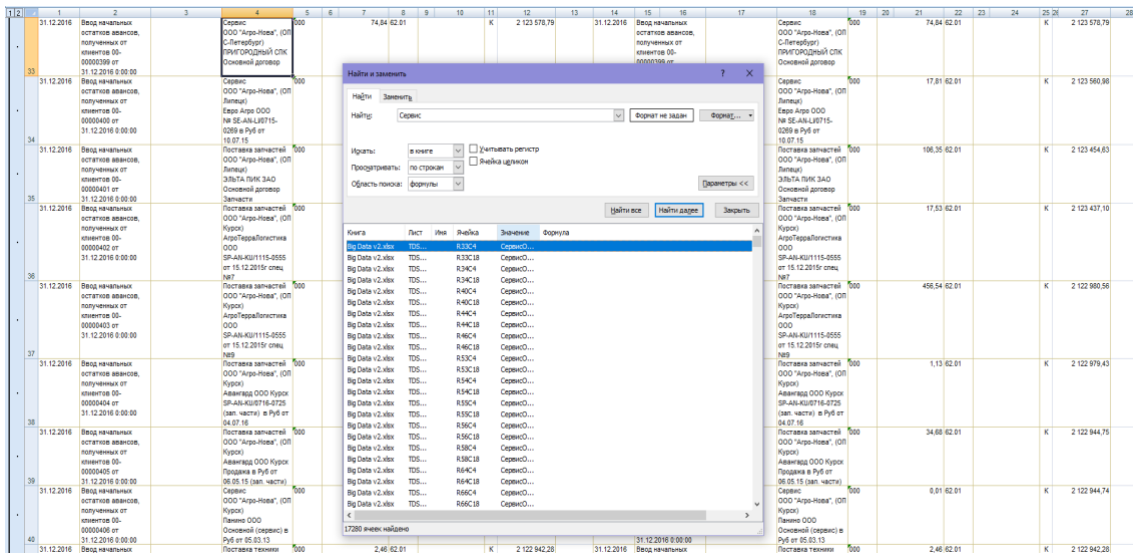


Figura 83 Buscar en el fichero en ordenador normal una palabra.

El tiempo de busqueda 5344 sec.



- Experimento №6 Buscar y cambiar una palabra en el fichero en ordenador normal:

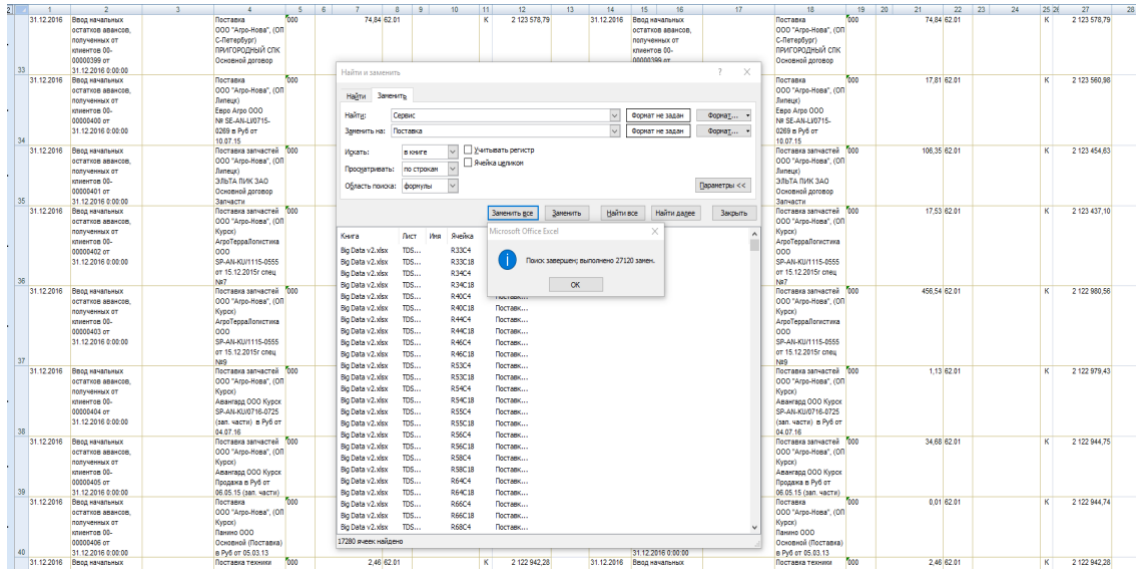


Figura 84 Buscar en el fichero en ordenador un palabra.

El tiempo de búsqueda 5652 sec. Las palabras iguales 27120 y células 271280

- Experimento №7 Buscar en el fichero una palabra en un folio sin estructura en ordenador:



Figura 85 Buscar en el fichero un palabra en ordenador.

El tiempo de búsqueda 253 sec.

- Experimento №8 Buscar y cambiar una palabra en el fichero un folio sin estructura en ordenador:

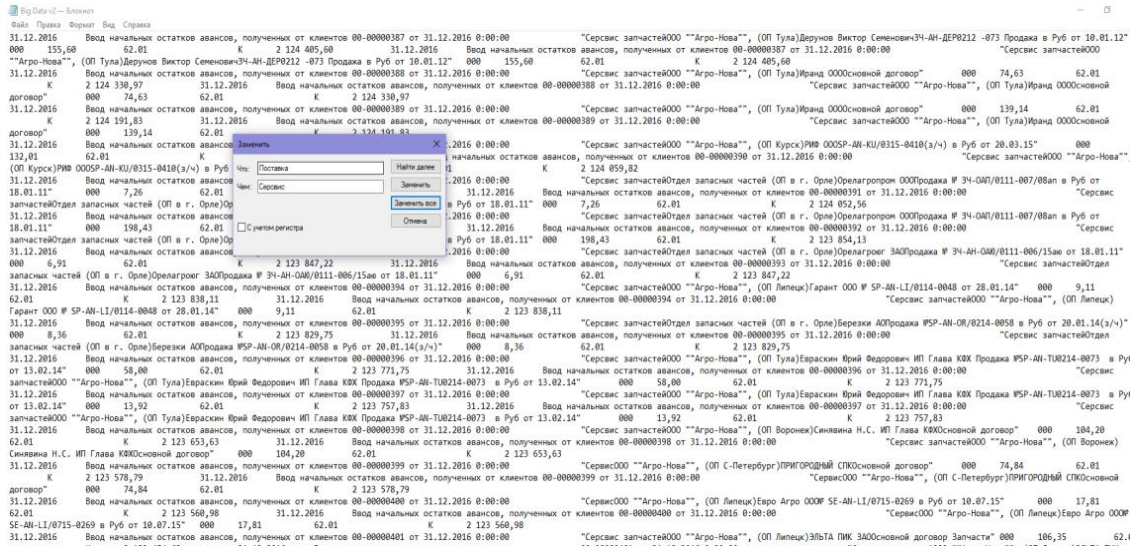


Figura 86 Buscar y cambiar en el fichero un folio sin estructura de datos.

El tiempo de cambiado 331 sec.

### 5.21 Comparación entre Azure y ordenador.

	Buscar una palabra	Cambiar un palabra	Buscar una palabra en los datos sin estructura	Buscar una palabra en los datos sin estructura
<b>Azure</b>				
<b>El tiempo (sec)</b>	725	740	43	52
<b>Ordenador</b>				
<b>El tiempo (sec)</b>	5344	5652	253	331

Tabla 8 Comparación el tiempo de búsqueda entre Azure y Ordenador.

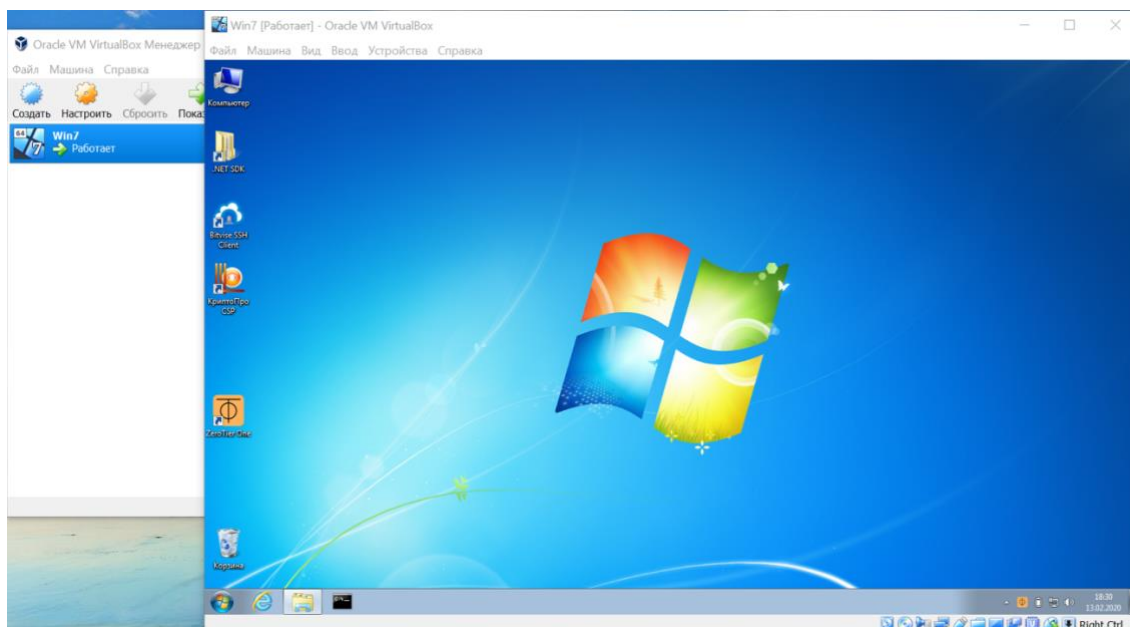
## 5.22 Conclusiones:

- Datos obtenidos durante el experimento muestran que la velocidad de búsqueda en Azure HDInsight funciona más rápido que en el ordenador. En este experimento he usado el fichero con tamaño 562 Mb de datos estructurados y el fichero con tamaño 56,24Mb sin estructura. Si consideramos los datos de un gran tamaño, el aumento en la velocidad de procesamiento de datos será enorme.
- Azure HDInsight mejor usar en grandes empresas que generan cada día los datos de un gran tamaño porque necesito el especialista para Azure también necesito pagar por cada aplicación en Azure.

## 5.23 El experimento con máquina virtual de AWS Amazon.

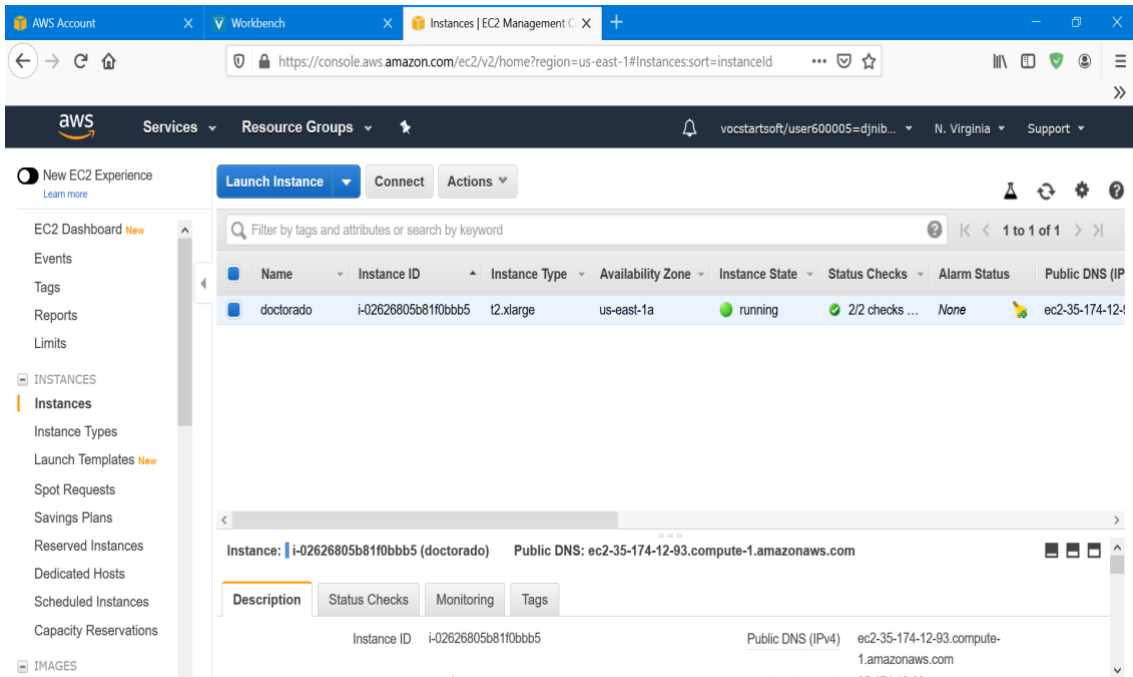
Una idea del experimento es transferir un archivo cifrado a un servidor virtual y descifrarlo. En este experimento uso las aplicaciones: Virtual Box, Bitwise, Zero one y la aplicación que yo he desarrollado.

Creamos una máquina virtual Windows 7 con ayuda de Virtual Box. También instalamos las aplicaciones Cripto pro, Cripto SDK.

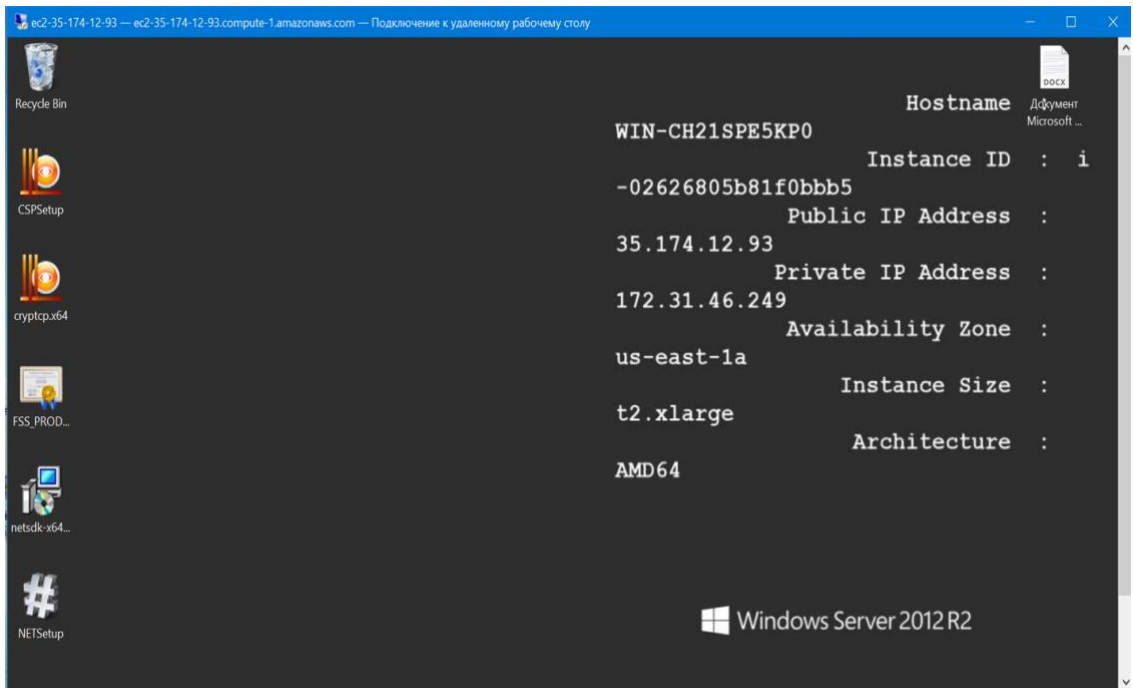


*Figura 87 Máquina virtual Virtual Box.*

- Creamos una maquina virtual Windows Server 2012 al Amazon AWS. También instalamos las aplicaciones Cripto pro, Cripto SDK.



*Figura 88 Pagina web de maquina virtual Amazon AWS.*



*Figura 89 Maquina virtual Amazon AWS.*



- Instalamos en todos los servidores ZeroTier para hacer VPN y ejecutamos por la pagina web

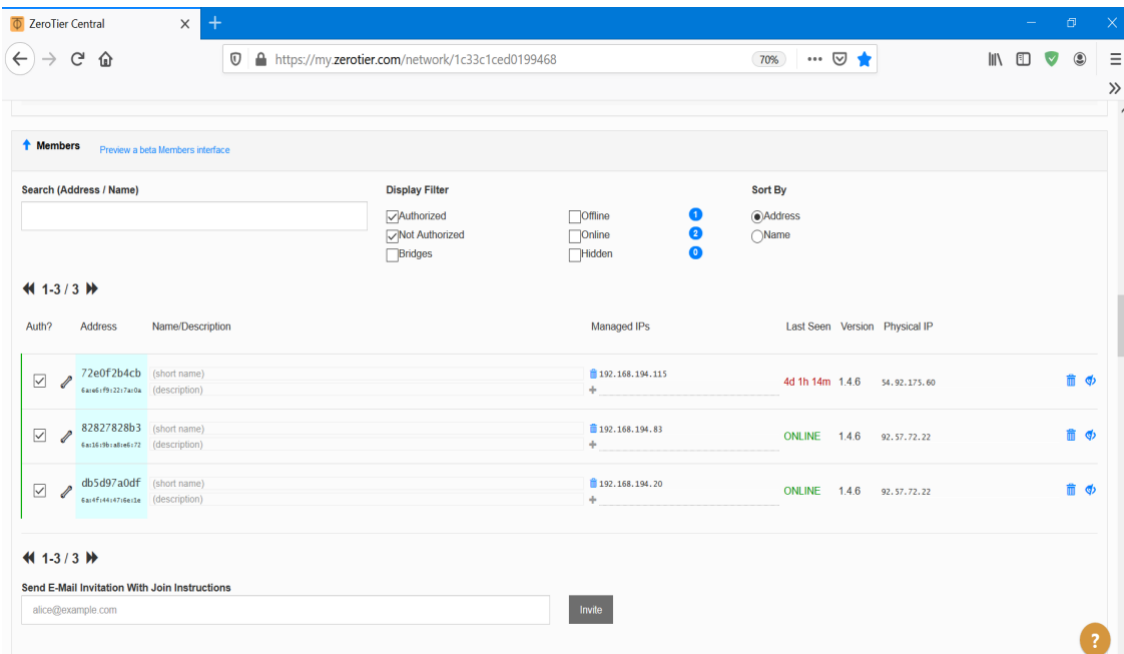


Figura 90 Interface de ZeroTier.

- Instalamos el cliente de Bitwise al Amazon AWS y Virtual Box:

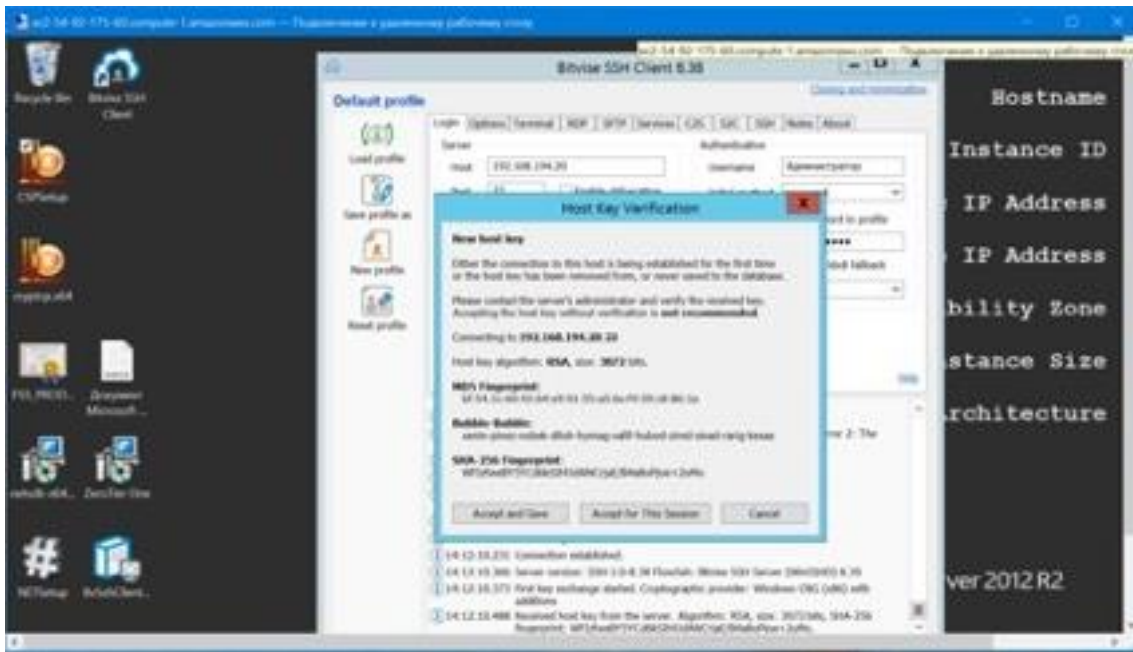
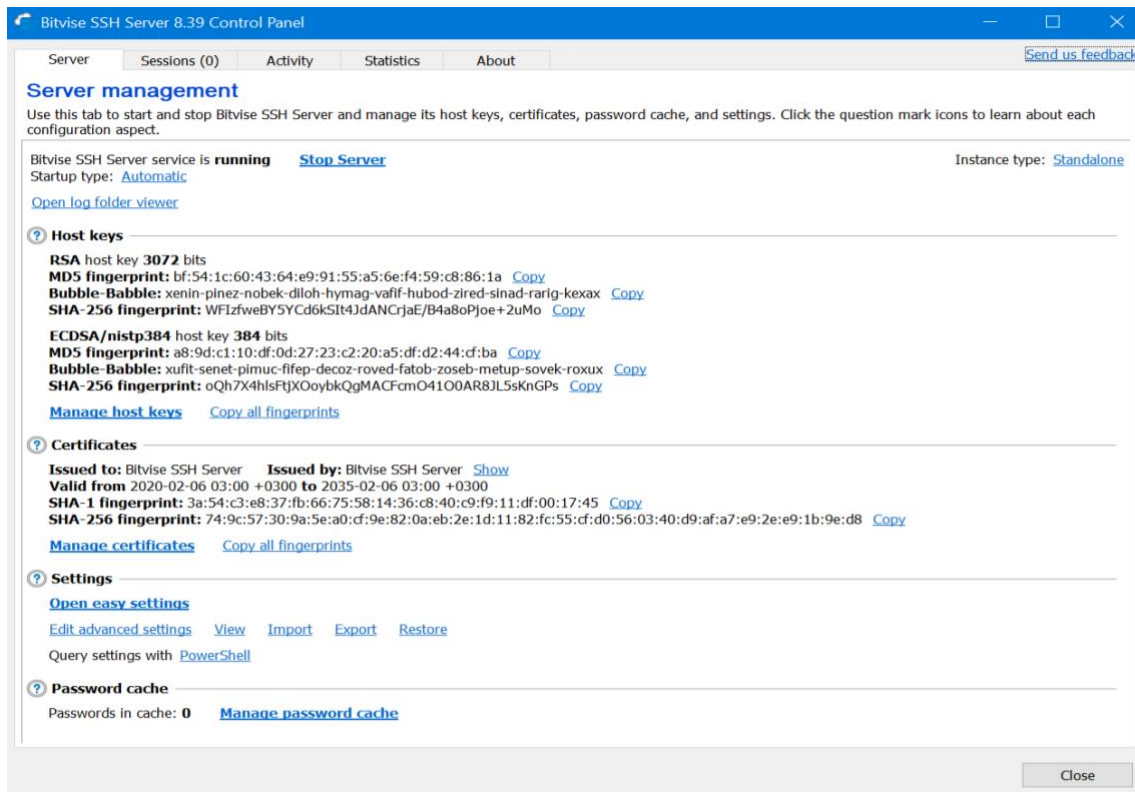


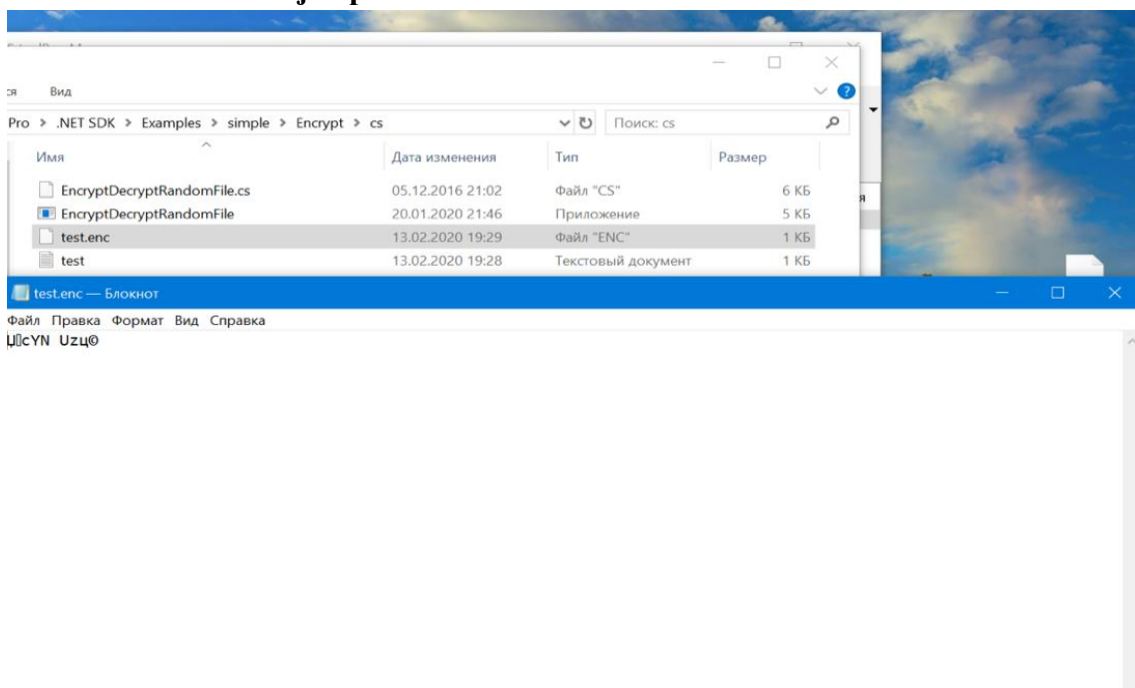
Figura 91 Interface del cliente de Bitwise.

- **Instalamos el Bitvise servidor al maquina normal con sistema operativo Windows 10:**



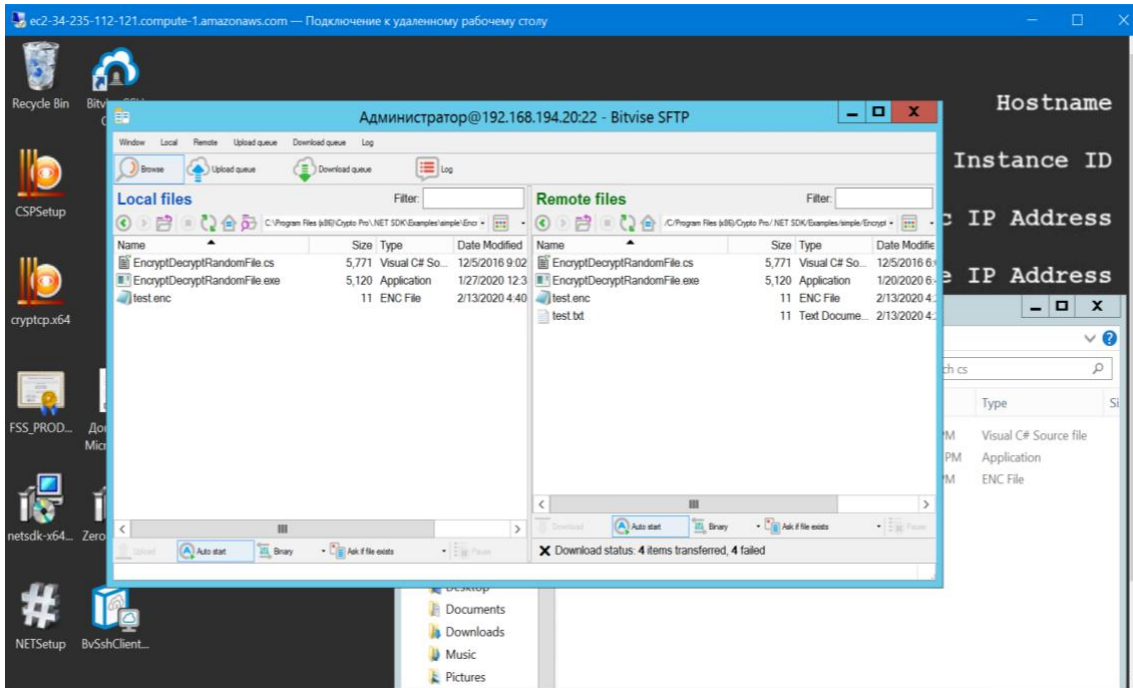
*Figura 92 Interface del servidor de Bitvise.*

- **Ciframos un ejemplo con se llama “test”**

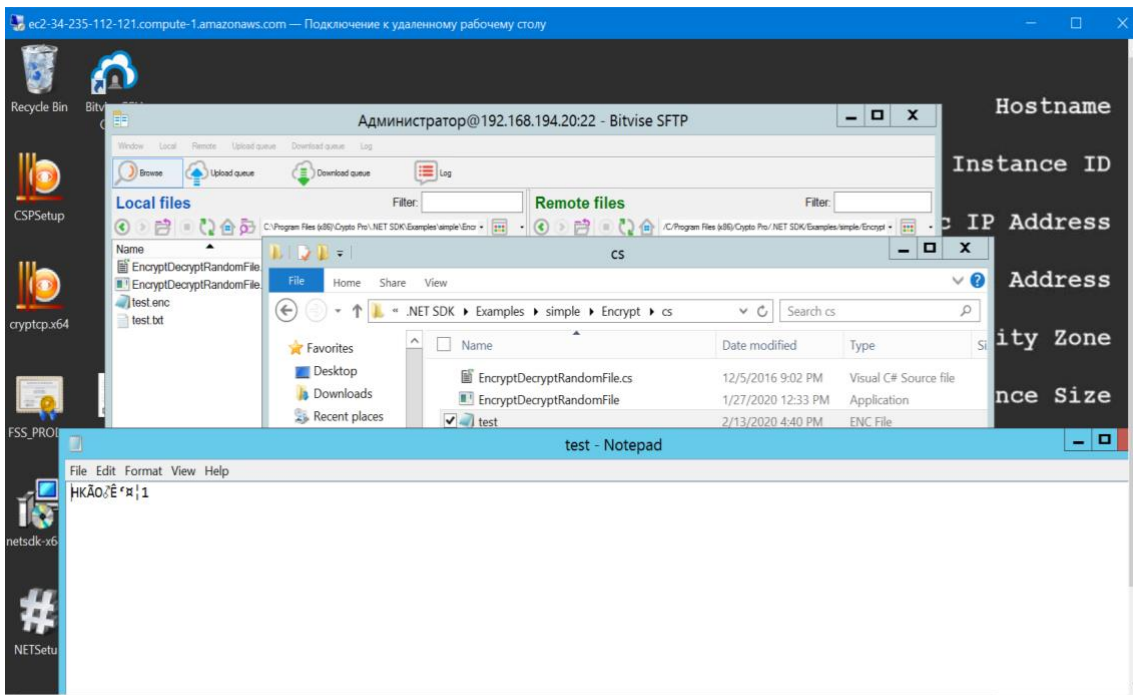


*Figura 93 Cifrado de datos.*

- **Transmitimos el fichero cifrado al maquina virtual de AWS Amazon**

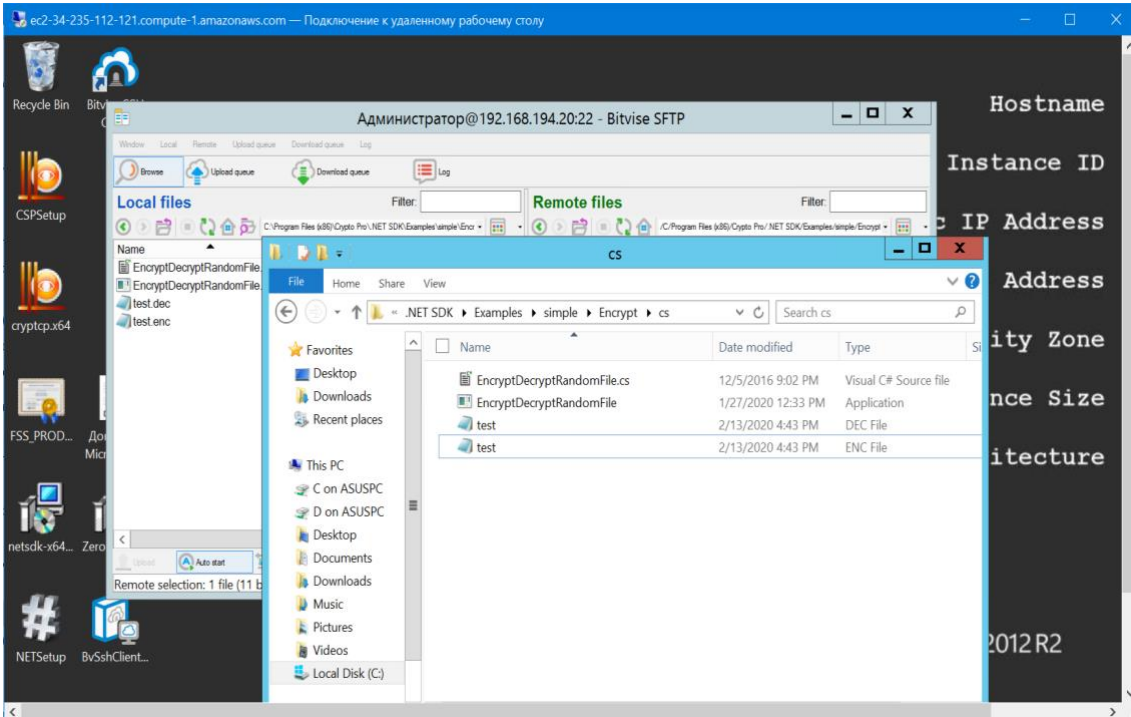


*Figura 94 Migración de datos.*



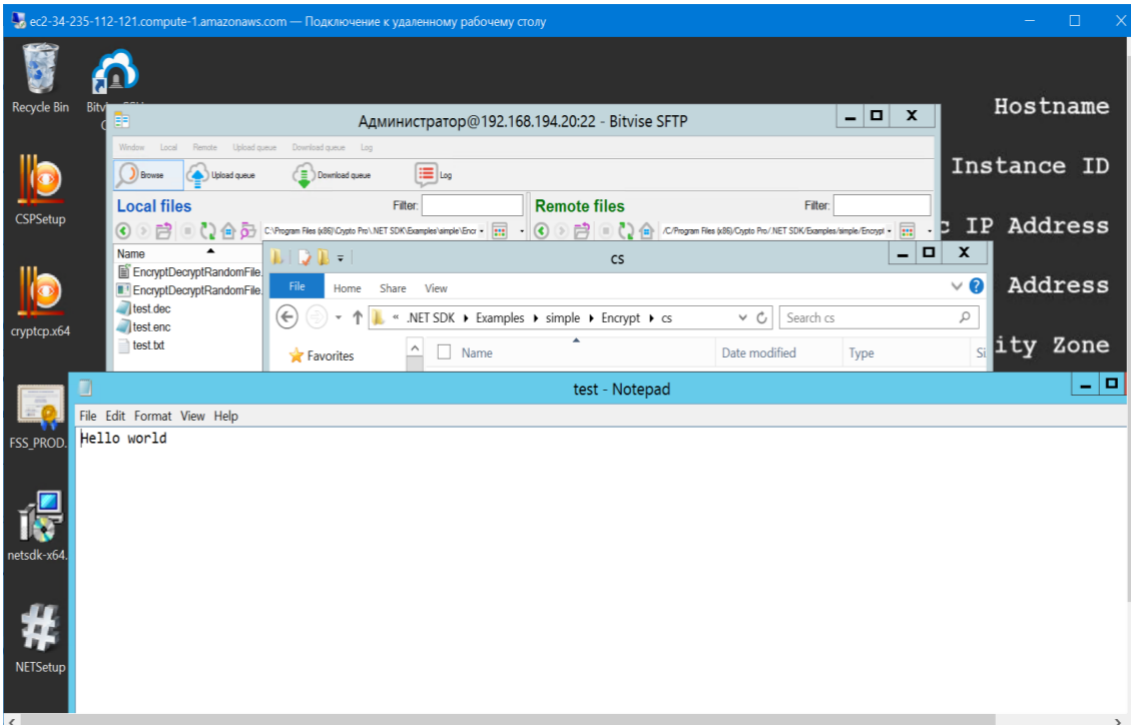
*Figura 95 Fichero cifrado.*

- **Desciframos el fichero**



*Figura 96 Fichero descifrado.*

- **El éxito.**



*Figura 97 Fichero descifrado.*

## **5.24 Los experimentos con Amazon AWS.**

Una idea del experimento es cifrar /descifrar los datos en la maquina virtual Amazon AWS y comparar los datos del experimento con ordenador portátil.

### **Ordenador portátil:**

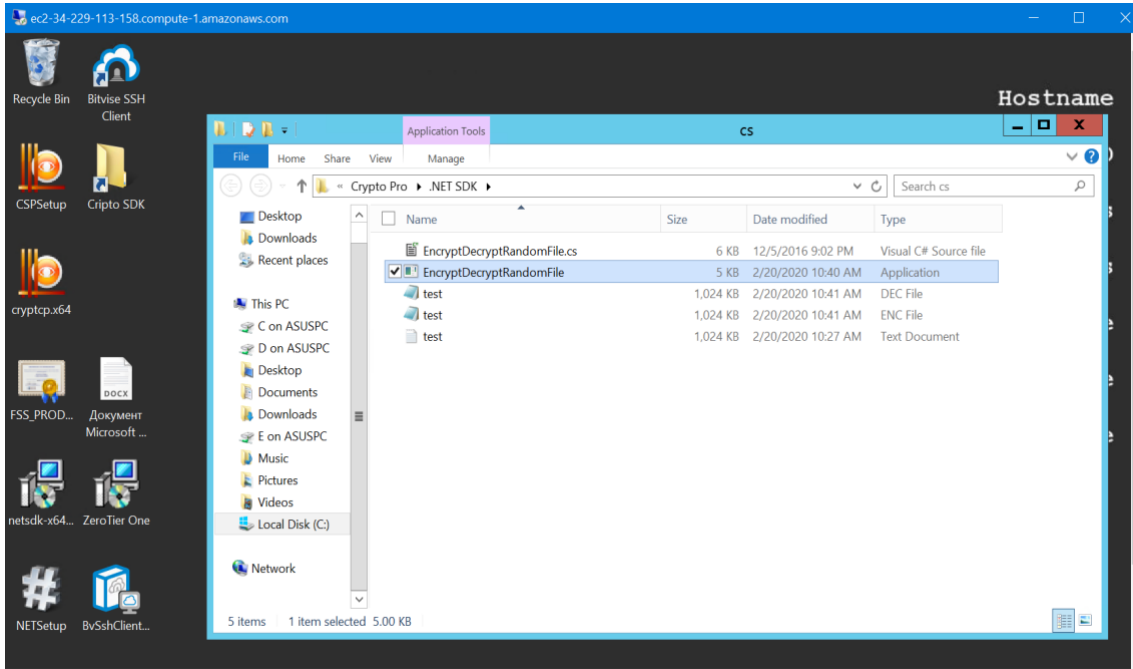
- 1.CPU i5 4460 -3,2Ghz
- 2.Memoria operative - 8 Gb DDR3
- 3.SSD Kingston V300
- 4.Windows 7x64

### **Amazon AWS:**

- 1.CPU Intel Xeon E5 2686 v.4 -2,3GHz
- 2.Memoria operative - 16 Gb DDR3
- 3.HDD SAS
- 4.Windows Server 2012 R2 x64

## Cifrado/descifrado los datos con tamaños :1Mb,10Mb,100Mb,200Mb Cifrado/descifrado con GOST

### - Experimento №1- 1 Mb:



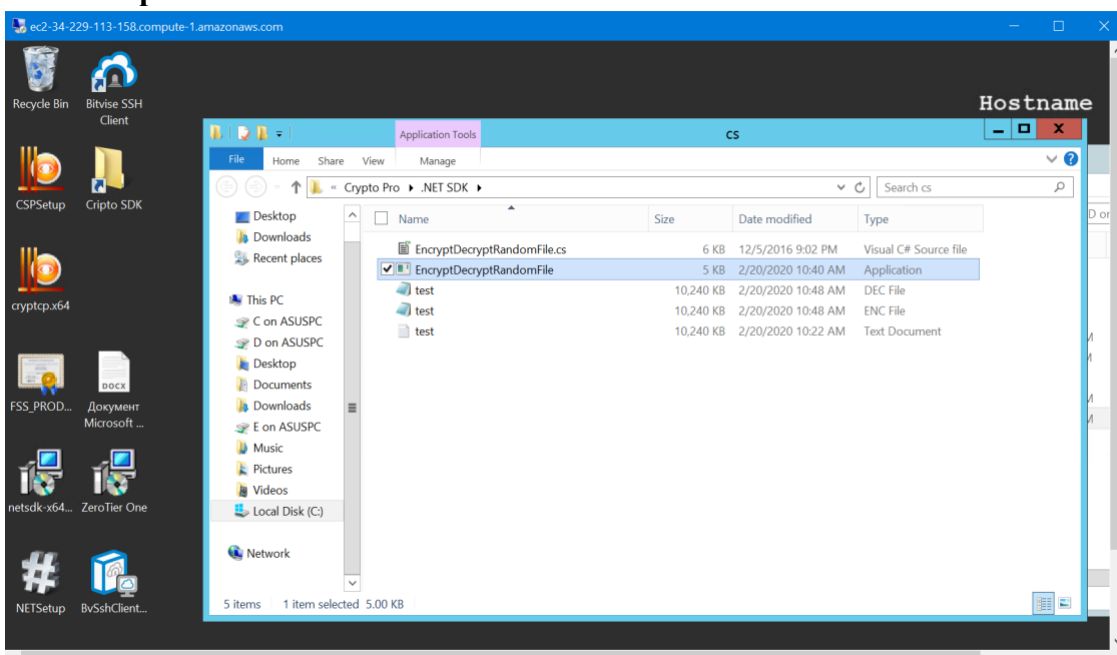
*Figura 98 Experimento №1.*

### Los datos obtenidos durante del experimento:

Tamaño	Cifrado	Descifrado
1 Mb	98,194 Ms	93,02 Ms

*Tabla 9 Los datos del experimento №1.*

### - Experimento №2- 10 Mb:



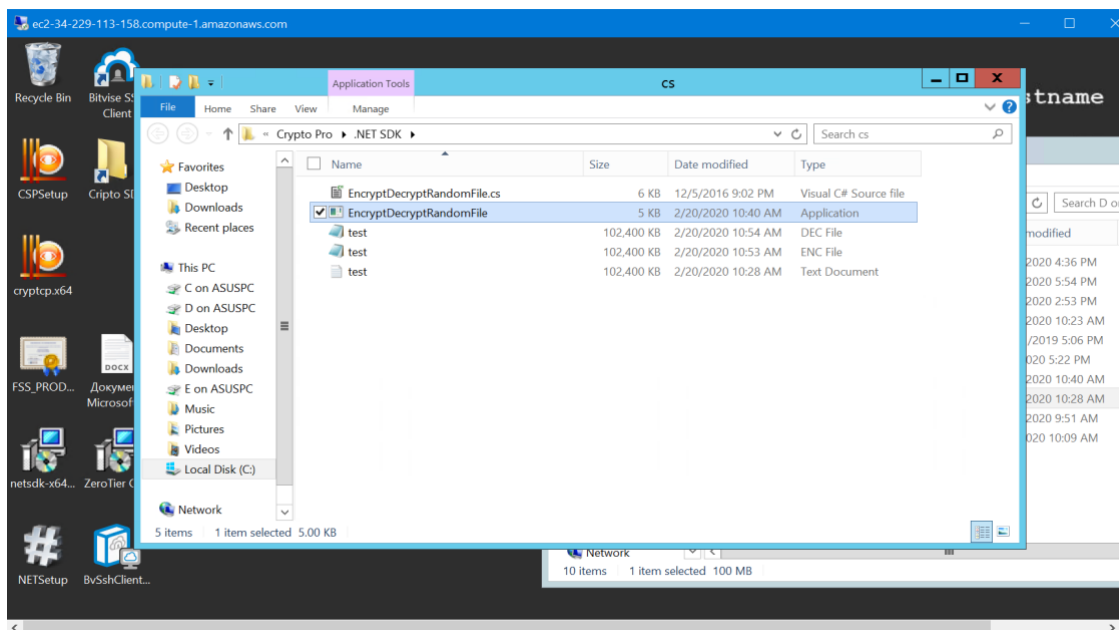
*Figura 99 Experimento №2.*

**Los datos obtenidos durante del experimento:**

Tamaño	Cifrado	Descifrado
10 Mb	784,31 Ms	776,11 Ms

*Tabla 10 Los datos del experimento №2.*

**- Experimento №3- 100 Mb:**



*Figura 100 Experimento №3.*

**Los datos obtenidos durante del experimento:**

Tamaño	Cifrado	Descifrado
100 Mb	5667,1 Ms	5388,22 Ms

*Tabla 11 Los datos del experimento №3.*

- Experimento №4- 200 Mb:

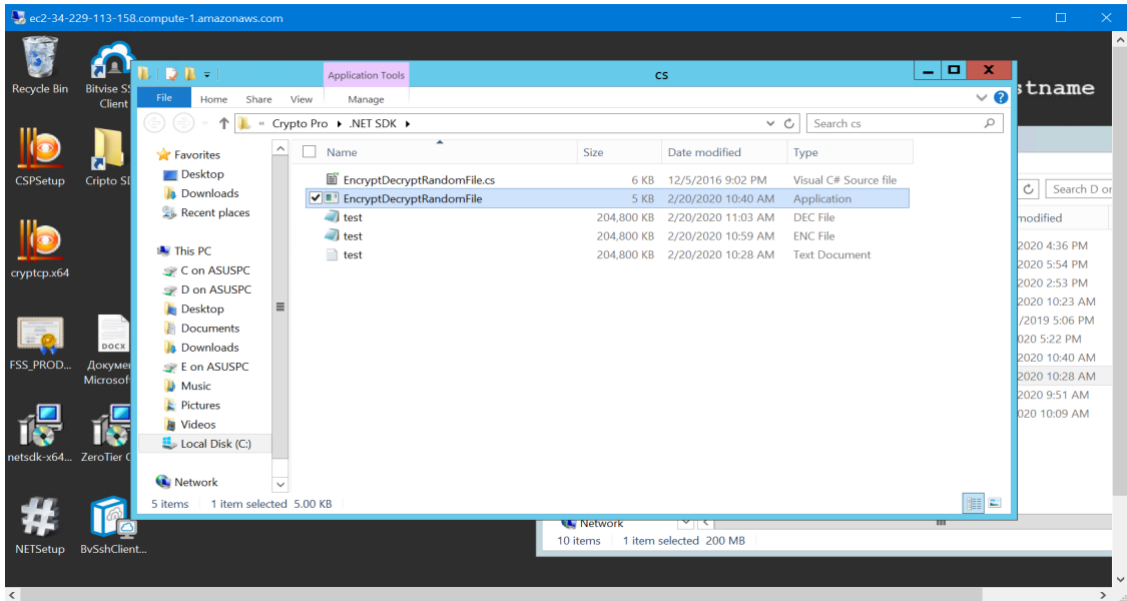


Figura 101 Experimento №4.

**Los datos obtenidos durante el experimento:**

Tamaño	Cifrado	Descifrado
200 Mb	17165,73 Ms	16341,7 Ms

Tabla 12 Los datos del experimento №4.

Comparemos los datos obtenidos durante el experimento con los datos obtenidos en el experimento anterior:

**Amazon AWS:**

Tamaño	Cifrado	Descifrado
1 Mb	98,194	93,02
10 Mb	784,31	776,11
100 Mb	5667,1	5388,22
200 Mb	17165,73	16341,7

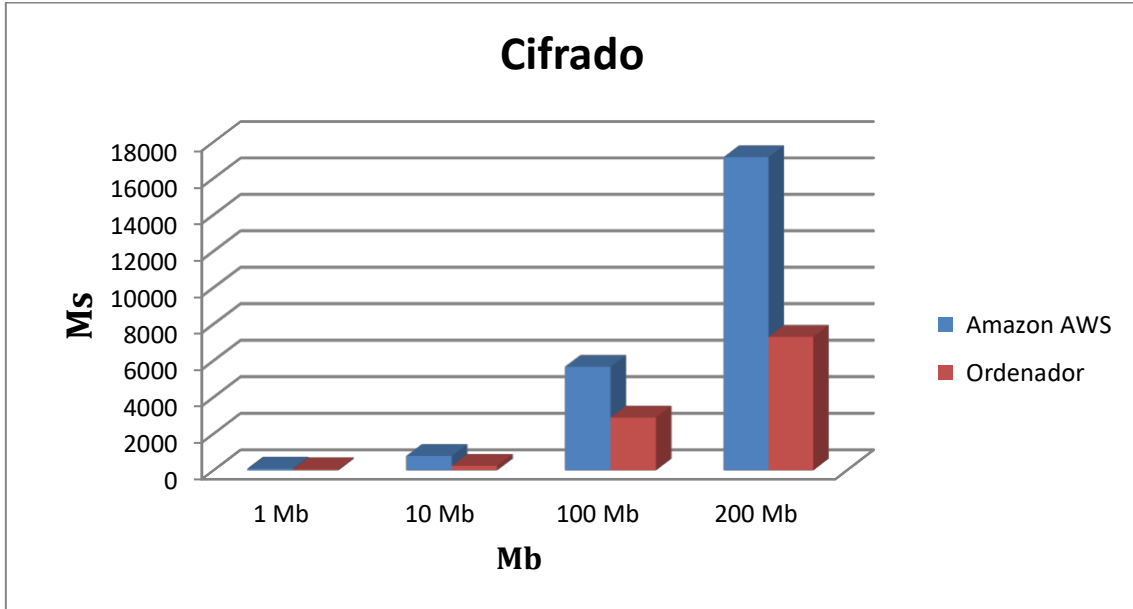
**Ordenador portátil:**

Tamaño	Cifrado	Descifrado
1 Mb	61,8093	61,8093
10 Mb	247,3746	247,3746
100 Mb	2887,4432	2887,4432
200 Mb	7302,7723	7302,7723

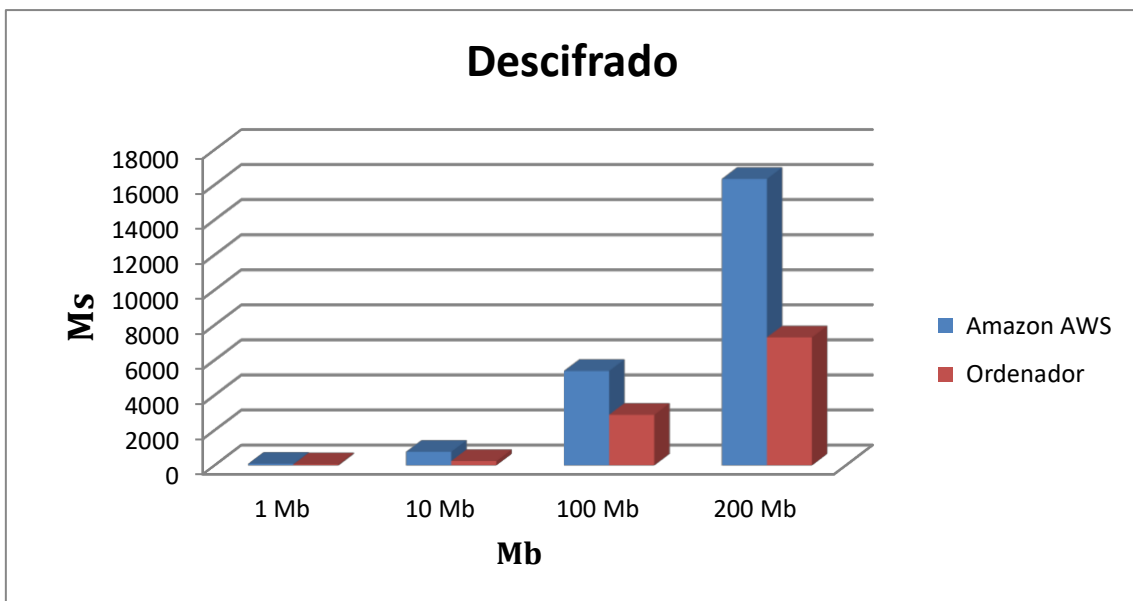
Tabla 13 Comparación el tiempo de búsqueda entre Amazon AWS y Ordenador portátil.



**Gráficos para la comparación cifrado/descifrado entre AWS y ordenador:**



**Gráfico 5. Comparación cifrado entre Amazon y Ordenador**



**Gráfico 6. Comparación descifrado entre Amazon y Ordenador**

## **5.25 Conclusiones.**

Durante el experimento, se cifró el archivo y se transfirió a través del protocolo SFTP a un servidor remoto, seguido de un descifrado correcto. El software desarrollado se puede utilizar para el cifrado remoto seguido de la transmisión para el descifrado.

Los datos obtenidos durante el experimento se mostró que ordenador portátil era mejor para cifrar/descifrar que la maquina virtual de Amazon AWS debido a la configuración utilizada de la máquina virtual que usa un disco duro.



## 6. Sistemas de autenticación para Hadoop.

En este capítulo se describen los sistemas de autenticación desarrollado para Hadoop basado en un circuito electrónico (eToken) y con cifrado XTS-AES, así como un modelo basado en Cryptopro con cifrado GOST 28147-89.

### 6.1 eToken

Para implementar un modelo robusto de seguridad se utiliza un componente que no sea duplicable y criptográficamente robusto. Para ello se ha utilizado un eToken que se conecta mediante USB al nodo que actúa como servidor de licencia de uso de datos. El eToken esta basado en el circuito de criptoautenticacion de Atmel ATECC508A que incorpora diversos elementos de seguridad:

- Ejecución rápida algoritmo de algoritmos de clave publica:
- ECDSA (Elliptic Curve Digital Signature Algorithm) FIPS186-3.
- ECDH (Elliptic Curve Diffie-Hellman Algorithm) NIST SP800-56A.
- Estándar NIST P256.
- Hash SHA-256 con opcion HMAC.
- Longitud de clave 256/283 bits.
- Numero de serie de 72 bits.
- Generador de números aleatorios FIPS (Federal Información Procesan Estándares) de gran calidad.
- Soporte cliente/servidor en el chip.
- Almacenamiento hasta 16 claves.
- Circuito integrado robusto diseñado contra manipulación y anti sabotaje.

La ventaja de este circuito es que permite una comunicación segura entre sus elementos internos y el software de verificación en el computador, evitando posibles ataques Man-In-the-Middle por captura del intercambio de datos entre ambos. El circuito puede ejecutar el ECDSA y el ECDSH (que se describen brevemente a continuación).

La siguiente Figura muestra los elementos principales de la comunicación entre el ordenador y el chip. Incluye el algoritmo de cifrado asimétrico basado en criptografía de curva elíptica (ECC) que es bastante mas eficiente que el RSA. Según la recomendación de ECRYPT II y publicado por ENISA un cifrado de 256 bits en ECC requería unos 3248 bits en RSA y este sería mas lento debido al tamaño de las claves, por lo que la industria se esta decantando por sistemas basados en ECC. En un Intel Xeon, la firma ECDSA de 256 bits es aproximadamente unas 9 veces más rápida que una firma RSA de 2048 bits. [CHA12]

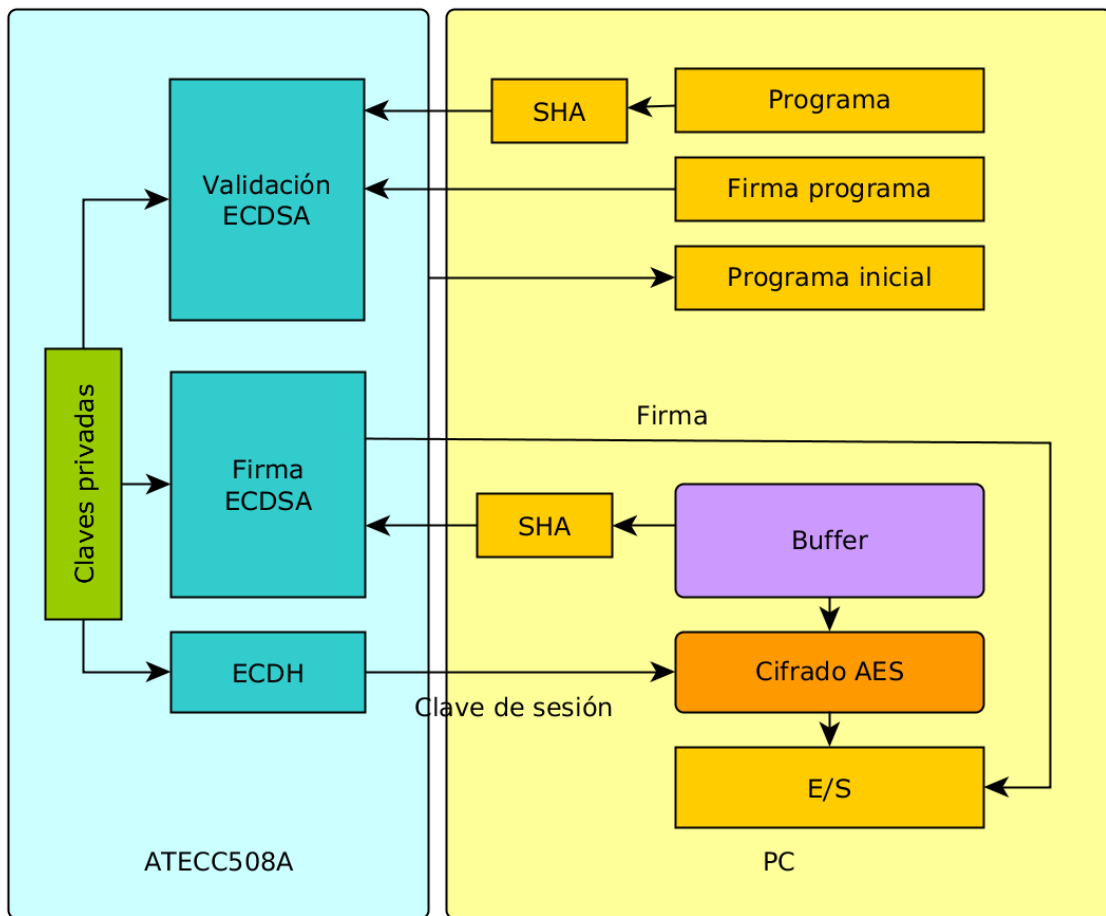


Figura 102 Intercambio de datos entre el PC y el ATECC508A.

## 6.2 ECDSA.

El algoritmo de firma digital de curva elíptica (ECDSA) es variante del algoritmo DSA basado en curva elíptica. Fue propuesto por Scott Vanstone en 1992 y fue aceptado como un estándar ISO en 1998 (ISO 14888-3), estándar ANSI (ANSI X9.62) en 1999, estándar IEEE 163-2000 en 2000 y estándar FIPS (FIPS 186-2) en 2000. Como ejemplos de uso, el ECDSA se utiliza en Bitcoin y en uno de los mecanismos de autenticación que implementa TLS (Transport Layer Security). En particular se utiliza el FIPS186-3 en este circuito, aunque existe una versión mas actualizada del estándar FIPS186-4 de 2013. Veamos como se utiliza el algoritmo para firmar: Supongamos que Alicia quiere firmar un mensaje con su clave privada ( $d_A$ ) y Bob confirmara la firma con la clave publica de Alicia ( $H_A$ ). De esta forma solo Alicia podrá firmar y todos podrán verificarlo. Se realizan los siguientes pasos:

- Se calcula un hash del mensaje y se trunca al tamaño de  $n$  bits (que es el orden del subgrupo), valor que denotaremos como  $z$ .
- Se genera un valor aleatorio  $k$ .
- Al valor le aplicamos el algoritmo con la clave privada  $d_A$  obteniendo el par  $(r; s)$  que representa la firma.

Para verificar la firma se toma dicho par  $(r; s)$  y el hash truncado del mensaje,  $z$ , que tras aplicarse el algoritmo de verificación con la clave pública  $HA$  podrá validarse la autenticidad de la firma. Aunque el algoritmo es criptográficamente muy robusto hay que utilizarlo de forma correcta ya que si no se siguen las recomendaciones de su uso puede ser fácilmente vulnerable. El número  $k$  es muy importante que sea un buen número aleatorio y que cambie cada vez que se utilice. De hecho, están documentadas diversas vulnerabilidades, como ejemplo: La videoconsola PlayStation 3 podía ejecutar solo juegos firmados por Sony con ECDSA evitando así que pudiera haber juegos en el mercado sin su firma, pero utilizaban la misma  $k$ . Aunque inicialmente se desconoce, partiendo de dos juegos distintos podía calcularse  $k$  y a partir de aquí se podía obtener la clave privada. Un problema parecido se detectó en diversas aplicaciones de monederos de Bitcoin para Android y OpenSSL. El problema se corrigió en la versión 1.0.0e .

### 6.3 ECDH.

El algoritmo Diffie - Hellman de curva elíptica (ECDH) es un protocolo de intercambio de claves de forma segura a través de un canal inseguro. El circuito implementa el estándar NIST SP800-56A, aunque existe una versión más actualizada (Rev.2) del documento.

### 6.4 Como evitar el uso del eToken de forma remota.

Un posible uso fraudulento es contar con un eToken en una localización y tratar de utilizarlo de forma remota (USB over Network). Para ello el código que se ejecuta en el servidor de uso de datos incluye diversos elementos de seguridad como son:

- Mecanismos para dificultar la depuración del código.
- Parte de la API está implementada en el Kernel de Linux para mayor dificultad de depuración del código.
- Detectar la virtualización.
- Comprobar que físicamente está conectada.
- Control de tiempos de acceso al chip.

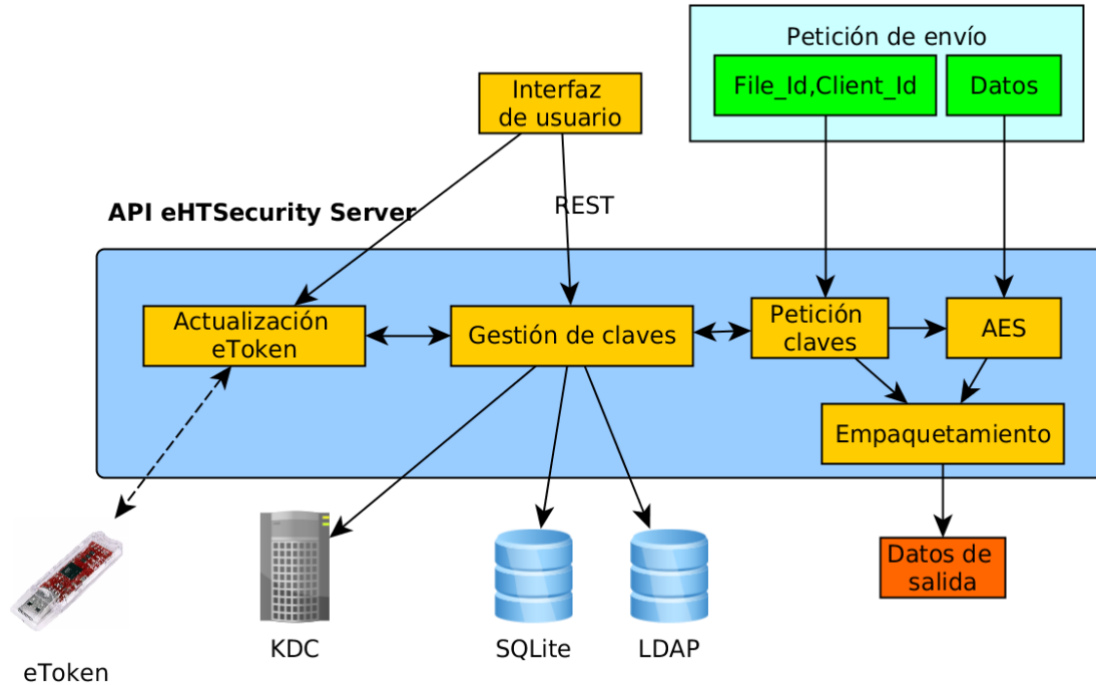
### 6.5 API eHTSecurity.

Para facilitar el diseño modular se ha definido una API eHadoop Token Security (eHTSecurity) que estable un mecanismo robusto de cifrado y autenticación basado en el chip y está disponible para Linux. La API eHTSecurity consta de dos módulos independientes:

- Servidor de datos: eHTSecurity Server.
- Receptor de datos: eHTSecurity Client.

La API servidor de datos dispone de tres módulos que se muestran en la figura 102:

- Gestión de claves.
- Actualización de eTokens.
- Cifrado.



**Figura 103 API Server.**

La gestión de claves se encarga de crear y almacenar las claves de cifrado simétrico para cifrar los datos y las claves asimétricas para la autenticación de los clientes. La actualización del eToken permite introducirle las claves de forma segura, tanto localmente (si el eToken está en el servidor) como si se encuentra remotamente. Cuando se valida inicialmente un eToken se introduce las claves privadas que le permitirá:

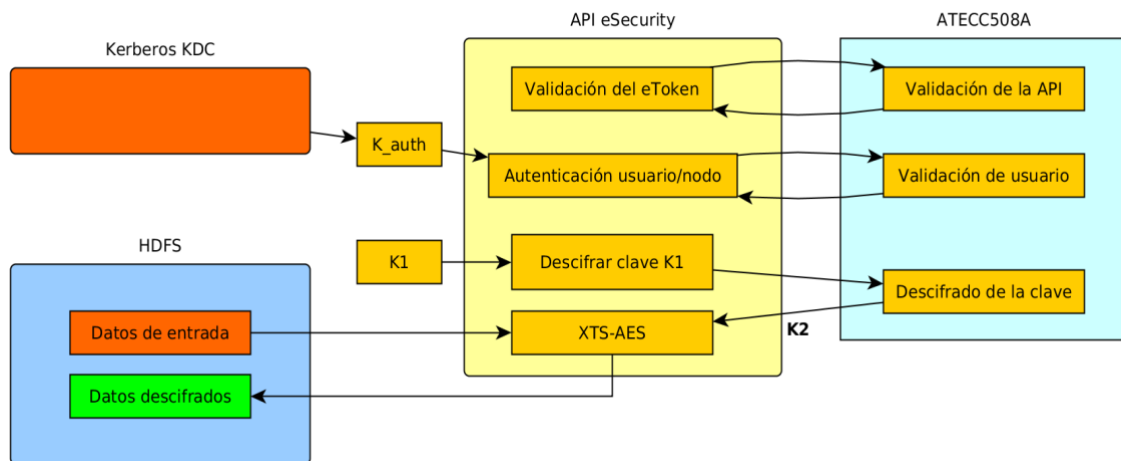
- Autenticarse.
- Descifrar la clave simétrica.

La API receptor de datos verifica que el usuario o nodo está autenticado y obtiene la clave simétrica para descifrar a partir de la clave privada contenida en el chip.

Se utiliza el ECDSA para que el módulo verifique al chip y que este a su vez compruebe que el código del módulo no ha sido manipulado. Se utiliza en ESSH para descifrar la clave enviada a través del canal inseguro I2C+USB de forma que se utiliza la clave para extraer los datos utilizando la clave privada que está dentro del chip.

En la Figura 103 se muestra el proceso de comunicación entre el chip y el computador en el cliente. La API se integra en el modelo de seguridad de Hadoop 2 niveles:

- Autenticacion.
- Cifrado.



**Figura 104** Comunicación entre API eSecurity y el ATECC508A.

La ventaja del modelo de Hadoop de utilizar una autenticación on basada en tokens es que integra los mecanismos básicos de comunicación, por lo que facilita su implementación. Aun que el HDFS permite el cifrado, esta orientado a que en el almacenamiento local los datos no queden accesibles. En nuestro caso, los datos se cifran en el origen y por lo tanto el cifrado propio de HDFS no es necesario, aunque no es incompatible.

Para el cifrado se utiliza el estándar IEEE 1619-2007 con el modo XTS (XEX-TCBCTS). El XTS ofrece mayor protección para el cifrado de bloques frente a otros sistemas como CBC y ECB. El XTS también se emplea en BitLocker, FileVault 2, TrueCrypt, FreeOTFE, dm-crypt entre otros, así como en el cifrado de dispositivos de almacenamiento.

Este sistema, mostrado en la Fig. 11, utiliza dos claves, una para realizar el cifrado AES de bloques y la otra para cifrar el "Tweak Value". Este valor ya cifrado se modifica además con una función polinómica de Galois GF (2128) y se realiza una XOR con el texto sin formato y el texto cifrado de cada bloque. La función GF proporciona una mayor aleatoriedad y asegura que los bloques de datos idénticos no producirán idénticos textos cifrado. De esta forma se evita utilizar vectores de inicialización y de encadenamiento. El descifrado de los datos se lleva a cabo mediante la inversión de este proceso. Dado que cada bloque es independiente y no hay un encadenamiento, si los datos almacenados cifrados están dañados, solo los datos de ese bloque en particular serán irre recuperables.

Con los modos de encadenamiento, estos errores se pueden propagar a otros bloques cuando descifra. Otra de las ventajas del XTS es que puede ejecutarse en paralelo, lo que permite acelerar el procesamiento de Big Data. [DIA16]



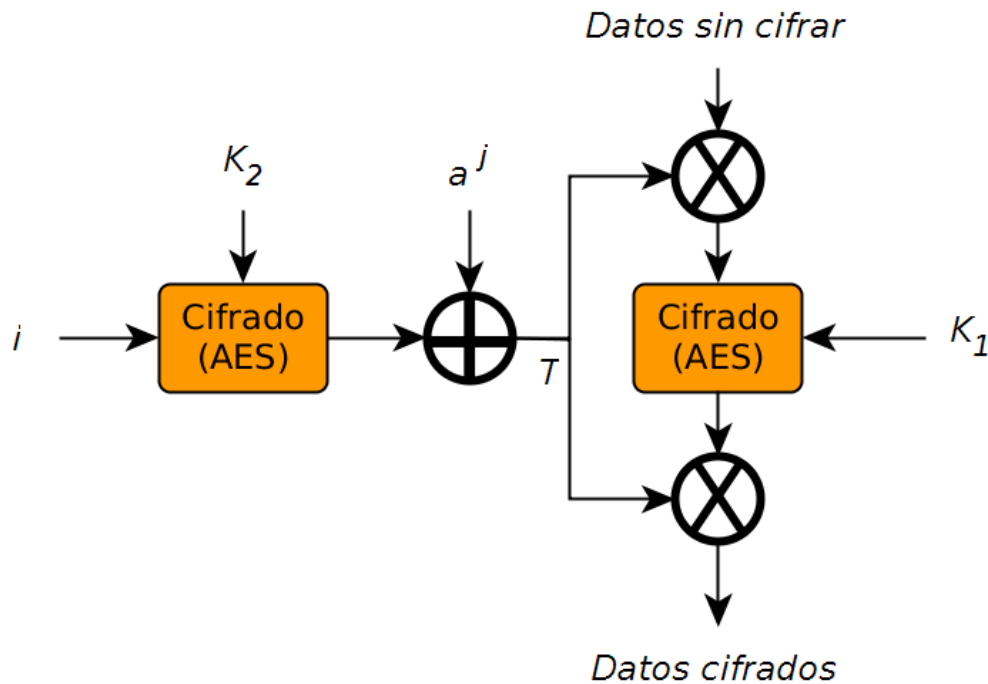


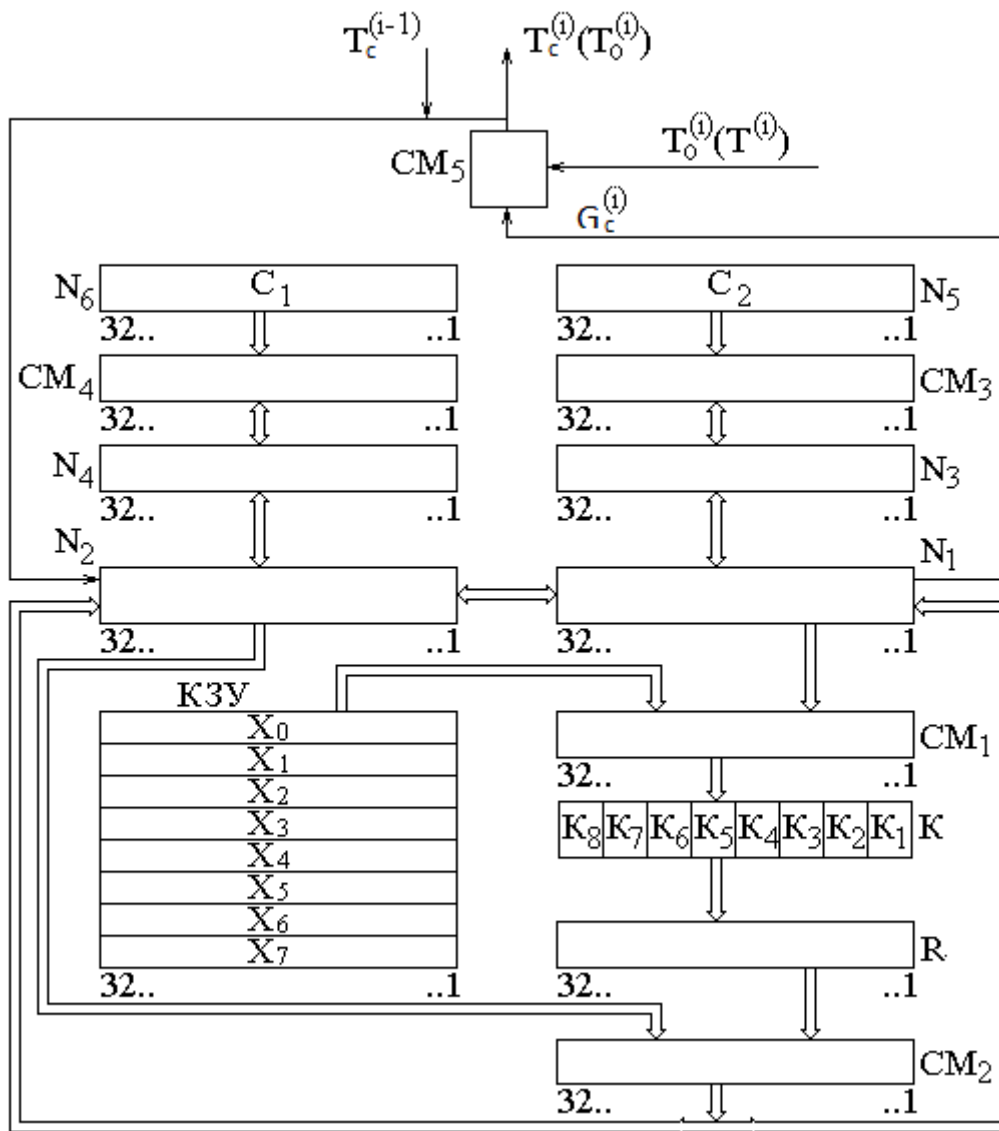
Figura 105 Diagrama del sistema XTS-AES.

## 6.6 GOST 28147-89.

GOST 28147-89 es un conocido cifrado de bloques de 256 bits que es una alternativa plausible para AES-256 y triple DES, que sin embargo tiene un costo de implementación mucho menor. GOST se implementa en bibliotecas criptográficas estándar como OpenSSL y Crypto. Hasta 2011, los investigadores acordaron unánimemente que GOST podría o debería ser muy seguro, lo que se resumió en 2010 con estas palabras: "a pesar de los considerables esfuerzos criptoanalíticos realizados en los últimos años 20, GOST aún no está roto".

### 1. Diagrama de bloques del algoritmo de transformación criptográfica

- Memoria de 256 bits, que tiene de ocho acumuladores de 32 bits ( $X_0, X_1, X_2, X_3, X_4, X_5, X_6, X_7$ );
- Cuatro acumuladores 32 bits ( $N_1, N_2, N_3, N_4$ );
- Dos acumuladores 32 bits ( $N_5, N_6$ ) con grabado ellos llenado permanente  $C_2, C_1$ ;
- Dos de 32 bits sumador de módulo  $2^{32}$  ( $CM_1, CM_3$ );
- 32-bit de módulo 2 sumador bit a bit ( $CM_2$ );
- 32-bit de módulo 2 sumador bit a bit ( $2^{32}-1$ ) ( $CM_4$ );
- Módulo 2 sumador ( $CM_5$ ), relativa a la limitación del sumador poco  $CM_5$  no se impone;
- Bloque de sustitución ( $K$ );
- Registrar los pasos cíclicos desplazamiento hacia once MSB ( $R$ ).



**Figura 106** Diagrama de bloques del algoritmo de transformación criptográfica.

Bloque de sustitución K se compone de ocho nodos de reemplazos  $K_1, K_2, K_3, K_4, K_5, K_6, K_7, K_8$  con memoria 64 bit cada nodo. Los candidatos al bloque de sustitución vector de 32 bits se cortan en ocho vectores consecutivos de 4 bits cada uno de los cuales se convierte en un vector de 4 bits reemplazando el nodo correspondiente es una tabla de tiene 16 líneas que tiene cuatro bits en la línea de llenado. El vector de entrada determina la dirección de la tabla de cadenas. Llenado de esta línea está emergiendo del vector. Después los vectores de salida de 4 bits sucesivamente combinan en un vector de 32 bits.

Cuando se llena la clave  $(W_1, W_2, \dots, W_{256}), W_q \in \{0,1\}, q=i \div 256$ , el valor  $W_1$  introducido en la 1ª categoría de la unidad  $X_0$ , el valor  $W_2$  introducido en la 2ª categoría de la unidad  $X_0, \dots$ , el valor  $W_{32}$  introducido en la 32ª categoría de la unidad  $X_0$ , el valor  $W_{33}$  introducido en la 1ª categoría de la unidad  $X_1$ , el valor  $W_{34}$  introducido en la 2ª categoría de la unidad  $X_1, \dots$ , el valor  $W_{64}$  introducido en la 32ª categoría de la

unidad  $X_1$ , el valor  $W_{65}$  introducido en la 1ª categoría de la unidad  $X_2$  y etc, el valor  $W_{256}$  introducido en la 32ª categoría de la unidad  $X_7$ .

El relleno constante  $C_1$  (constante) acumulador  $N_6$  se muestra en la tabla 1:

El valor $N_6$	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17
Valor de descarga	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
El valor $N_6$	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Valor de descarga	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0

*Tabla 14 El relleno constante  $C_1$  (constante) acumulador  $N_6$ .*

El relleno constante  $C_2$  (constante) acumulador  $N_5$  se muestra en la tabla 2:

El valor $N_6$	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17
Valor de descarga	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
El valor $N_6$	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Valor de descarga	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1

*Tabla 15 El relleno constante  $C_2$  (constante) acumulador  $N_5$ .*

Llenar unidad de tabla de consulta  $K$  es un elemento clave de largo plazo.

Organización de los diferentes tipos de comunicación alcanzados mediante la construcción del sistema de tecla correspondiente. Esto se puede utilizar la posibilidad de generar las claves en un modo de sustitución simple y las codifica en un modo de sustitución simple con la protección para la transmisión a través de canales de comunicación o el almacenamiento en la memoria del ordenador.

### **Cifrado en un modo de simple sustitución.**

El esquema, que implementa el algoritmo de cifrado en el modo de simple sustitución debe ser de la forma que se muestra en la figura siguiente:

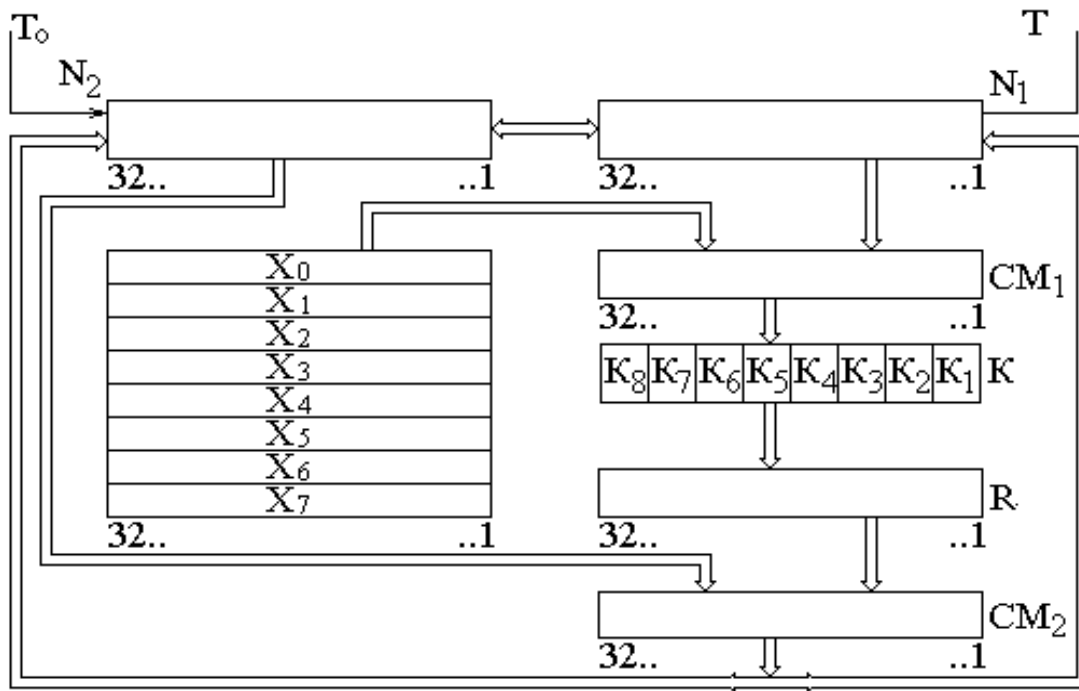


Figura 107 El algoritmo de cifrado en el modo de simple sustitución.

Los datos que se van cifrados, se dividen en bloques de 64 bits cada uno. El uso de cualquier bloque  $T_0=(a_1(0), a_2(0), \dots, a_{32}(0), b_1(0), b_2(0), \dots, b_{32}(0))$  un almacenamiento de información binaria.

$N_1$  y  $N_2$  hecho tan para que el valor  $a_1(0)$  grabado en primera categoría  $N_1$ , el valor  $a_2(0)$  grabado en segunda categoría  $N_1$  y etc, el valor  $a_{32}(0)$  grabado en 32- categoría  $N_1$ ; el valor  $b_1(0)$  grabado en primera categoría  $N_2$ , el valor  $b_2(0)$  grabado en segunda categoría  $N_2$  y etc, el valor  $b_{32}(0)$  grabado en 32- categoría  $N_2$ . El resultado es un estado  $(a_{32}(0), a_{31}(0), \dots, a_2(0), a_1(0))$  acumulador  $N_1$  y el estado  $b_{32}(0), b_{31}(0), \dots, b_2(0), b_1(0)$  acumulador  $N_2$ .

En memoria grabado 256 bit de llave. Contenido de ocho acumuladores de 32 bits  $X_0, X_1, \dots, X_7$ :

$$X_0=(W_{32}, W_{31}, \dots, W_2, W_1)$$

$$X_1=(W_{64}, W_{63}, \dots, W_{34}, W_{33})$$

.....

$$X_7=(W_{256}, W_{255}, \dots, W_{226}, W_{225})$$

El algoritmo de cifrado de bloques de datos de 64 bits en el modo de sustitución simple consta de 32 ciclos.

En el primer ciclo de llenado inicial  $N_1$  resumió por modulo  $2^{32}$  en sumador  $CM_1$  con llenado de acumulador  $X_0$ , el que llenado de acumulador  $N_1$  guardado. Se convierte en

un bloque de suma K y sustituyendo el vector resultante con la entrada de registro R que se hace girar en los once pasos en la dirección de las categorías de alto nivel. Resultado de cambios se añade bit a bit en módulo 2 sumador CM2 con acumulador de 32 bits de relleno N2. CM2 el resultado grabado en N1, el que el valor antiguo N1 grabado en N2. El primer ciclo termina. Los ciclos posteriores se llevan a cabo de manera similar el que en segundo ciclo de acumulador leer llenado X1. en tercer ciclo de acumulador leer llenado X2 y etc:

**X0, X1, X2, X3, X4, X5, X6, X7.**

En los últimos ocho ciclos del 25 al 32 th rellenos orden de lectura de acumulador retorno:

**X7, X6, X5, X4, X3, X2, X1, X0.**

Por lo tanto, al cifrar unas 32 ciclos llevado a cabo el siguiente procedimiento para seleccionar tanque de acumuladores.

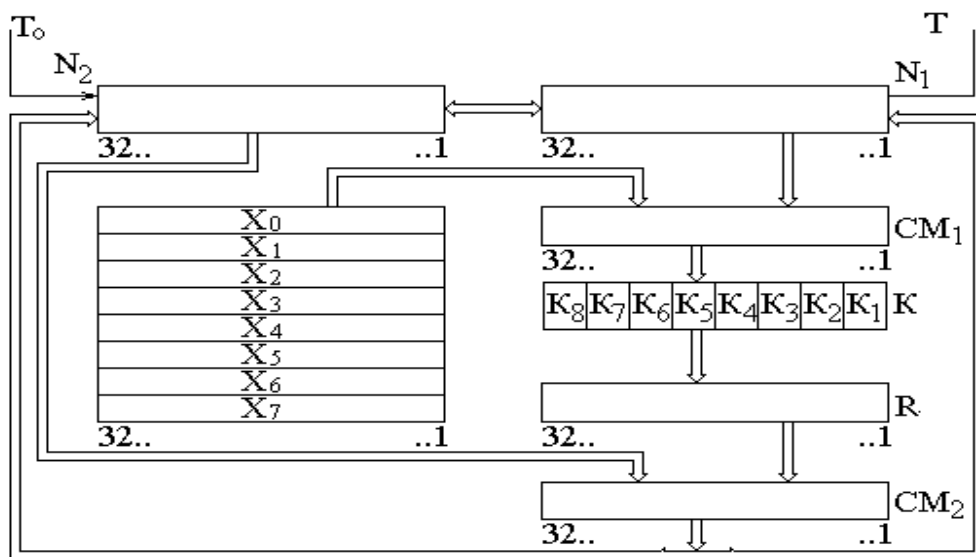
**X0, X1, X2, X3, X4, X5, X6, X7, X0, X1, X2, X3, X4, X5, X6, X7,**

**X0, X1, X2, X3, X4, X5, X6, X7, X7, X6, X5, X4, X3, X2, X1, X0.**

En el 32 ° ciclo del resultado de la sumador CM2 grabado en acumulador N2 pero en acumulador N1 conservado el viejo llenado.

**El descifrado de datos cifrados en un modo de simple sustitución**

Cripto sistema, que implementa el modo de cifrado de sustitución simple, tiene la misma forma que en la codificación:



*Figura 108 El modo de cifrado de sustitución simple, tiene la misma forma que en la codificación.*

En acumulador se graban 256 bit fragmentos de la misma tecla, que se llevó a cabo la codificación.

Descifrado de datos se divide en bloques de 64 bits cada uno. La entrada de cualquier unidad  $T_c=(a_1(32), a_2(32), \dots, a_{32}(32), b_1(32), b_2(32), \dots, b_{32}(32))$  en acumulador  $N_1$  y  $N_2$

está hecho para que el valor  $a_1(32)$  grabado en primero rango  $N_1$ , el valor  $a_2(32)$  grabado en segundo rango  $N_1$ , y etc el valor  $a_{32}(32)$  grabado en 32-th rango  $N_1$ ; el valor  $b_1(32)$  grabado en primero rango  $N_2$ , el valor  $b_2(32)$  grabado en segundo rango  $N_2$  y etc. El cifrado se lleva a cabo utilizando el mismo algoritmo que el de datos pública de cifrado, con la modificación de que el relleno acumulador  $X_0, X_1, \dots, X_7$  se leen en el siguiente orden:

$X_0, X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_7, X_6, X_5, X_4, X_3, X_2, X_1, X_0,$

$X_7, X_6, X_5, X_4, X_3, X_2, X_1, X_0, X_7, X_6, X_5, X_4, X_3, X_2, X_1, X_0.$

$T_0=(a_1(0), a_2(0), \dots, a_{32}(0), b_1(0), b_2(0), \dots, b_{32}(0))$ , bloque de datos cifrado correspondiente

el que valor  $a_1(0)$  el bloke  $T_0$  corresponde al contenido en primero range  $N_1$ , el valor  $a_2(0)$  corresponde al contenido en segundo range  $N_1$  y etc, el valor  $a_{32}(0)$  corresponde al contenido en 32 -th range  $N_1$ , el valor  $b_1(0)$  corresponde al contenido en primero range  $N_2$ , el valor  $b_2(0)$  corresponde al contenido en segundo range  $N_2$  y etc, el valor  $b_{32}(0)$  corresponde al contenido en 32-th range  $N_2$ .

Datos de la misma manera, los bloques restantes se descifran codificadas.

## 6.7 El método de autenticación del usuario en Cloudera Hadoop con GOST.

### Cripto Login

Proporciona la capacidad de organizar la autenticación de la autorización y el usuario el uso de certificados digitales. Cripto Login se basa en el algoritmo GOST 34.10 – 2001. Este es método nuevo que tiene nevil B-testado.

### El módulo que permite:

- Autenticación de usuario con ayuda certificados digitales
- Un canal de comunicación segura entre un cliente y un servidor a través de protocolos SSL y TLS
- Soporte de funcionar de Rutoken, JaCarta, eToken

## 6.8 El esquema de autenticación usuarios en Hadoop con certificado GOST.

1. Usuario con el certificado GOST dentro de Cripto Fox que solicita el acceso en Hadoop
2. Servidor, en el que el pre-instalado y configurado Cripto Pro y Cripto Login verifica el certificado de usuario GOST y certificado GOST que instalado en servidor
3. Cripto Login permite el acceso a la dirección previamente grabado.



**Figura 109** Esquema de autenticación usuarios en Hadoop con certificado GOST.

En el capítulo anterior he desarrollado un mecanismo de funcionar CriptoPro y Cripto ARM que permitió el uso de certificados electrónicos y eToken. En este apartado se desarrollará mediante el método de autenticación de dos factores el usuario con ayuda certificado electrónico GOST + nombre de usuario y contraseña en Cloudera Hadoop.

- Instala y configura CriptoPro + CriptoARM (metodología detallada de ejecutar fue desarrollado y descrito en el capítulo anterior)
- Instala una versión especial del navegador Web con se llama CriptoFox que soporta cifrado y certificados GOST
- Instala el modulo CriptoLogin para CriptoFox
- Solicitamos un certificado GOST de fabricación de criptoproveedor indicada todos los parámetros necesarios.
- Solicitamos una licencia para CroptoPro CSP R2 de criptoproveedor.
- Instala los certificados de CroptoPro y el certificado GOST a las carpetas apropiada. (metodología detallada de instalar fue desarrollado y descrito en el capítulo anterior)
- Ejecutar el modulo CriptoLogin

Создание web приложения

Описание  
Cloudera Manager

URL сайта\*  
cdhmaster.admin.com:7180/cmfd/login

URL контроллера авторизации\*  
cdhmaster.admin.com:7180/cmfd/login

Время действия ключа восстановления  
100

Время действия ключа доступа  
900

Пароль\*  
•••••

Подтверждение пароля  
•••••

ОТМЕНА СОЗДАТЬ

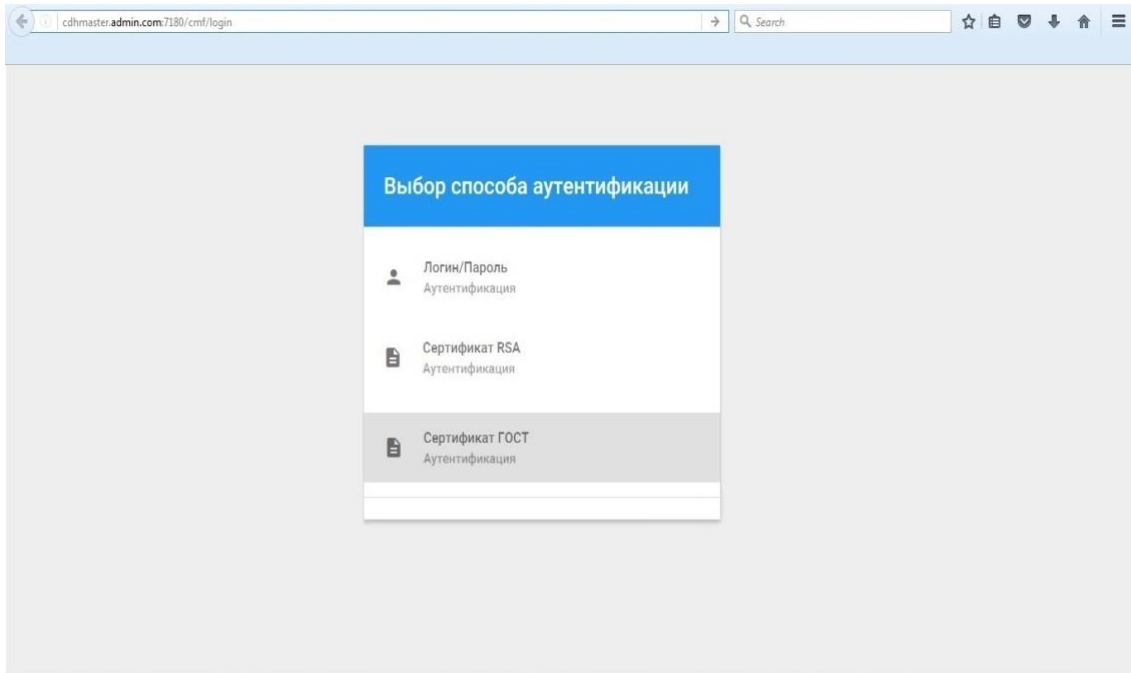
*Figura 110 Cripto Login.*

Rellene todos los puntos:

- Descripción
- URL
- URL de controlador de autorizaciones
- Tiempo de acción clave de recuperación
- La duración de la clave de acceso
- Contraseña
- Confirmación de la contraseña
- Pasamos en Cloudera Manager (La instalación detallada de Cloudera Manager se ha descrito en el capítulo anterior)
- Seleccione el método de autenticación:
- nombre de usuario / contraseña
- el certificado RSA
- el certificado GOST

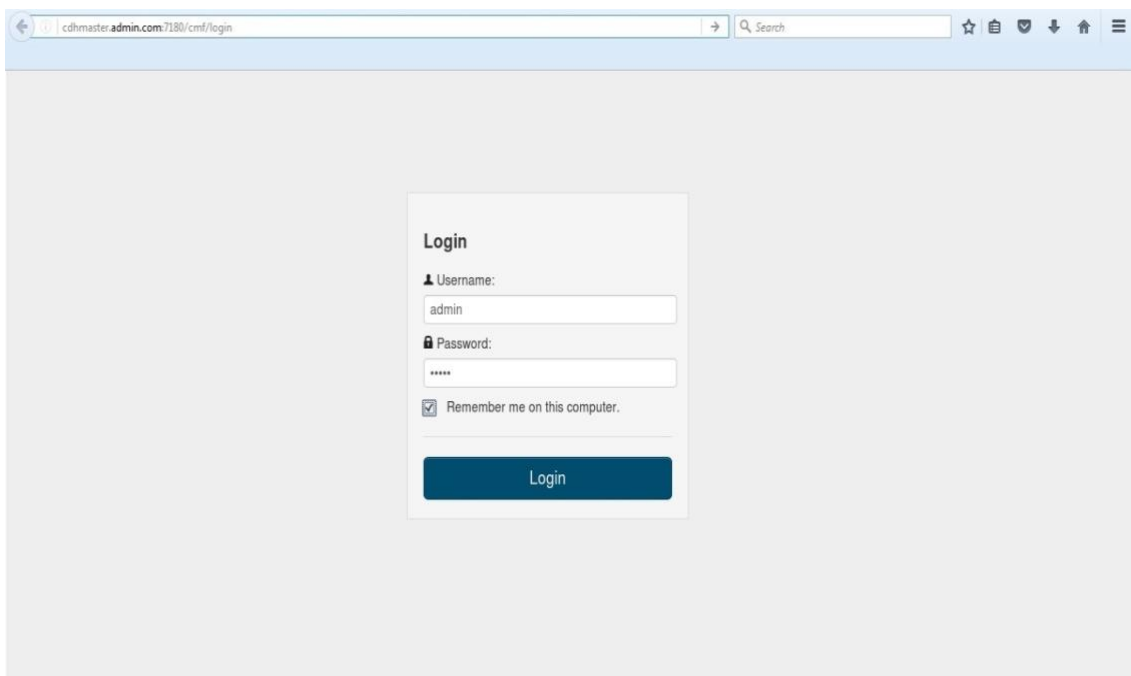
¡Importante! Sin un software específico y personalizado CriptoPro + CriptoARM los métodos de autorizaciones no funciona.





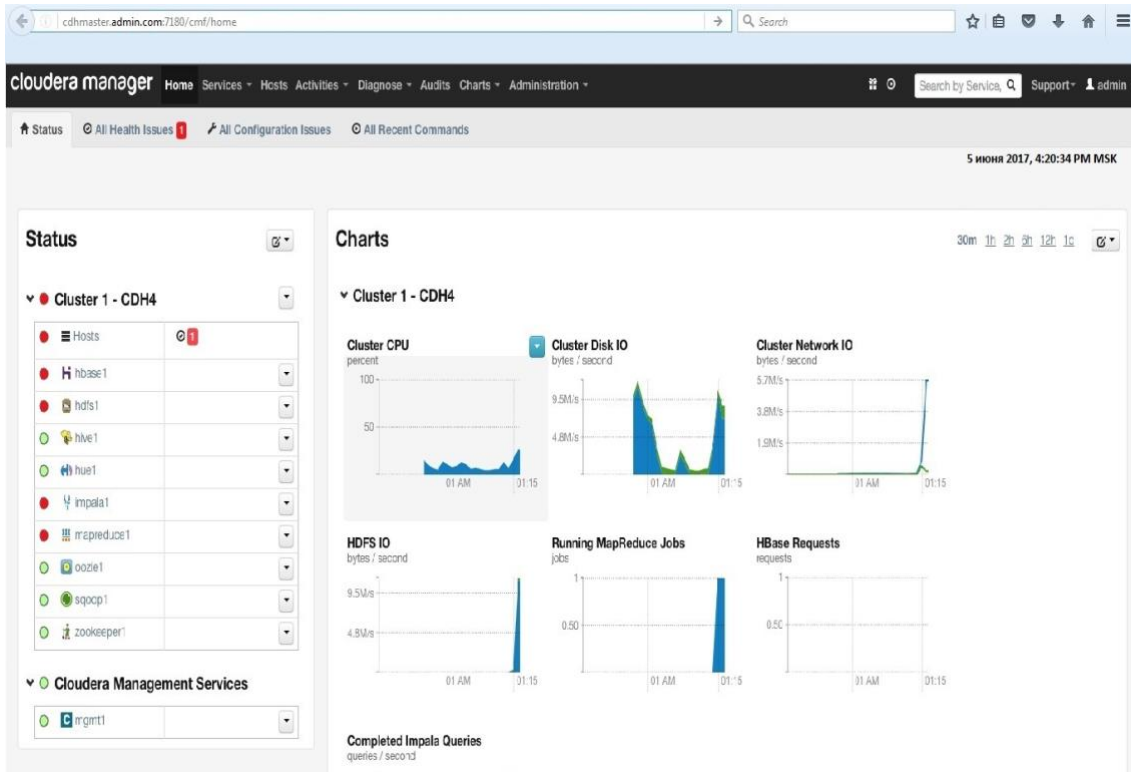
**Figura 111 El método de autenticación.**

- Seleccione el método de autenticación " El certificado GOST"
- Si todo el software está configurado correctamente, el certificado instalado, vaya al siguiente paso, el usuario autenticación.
- Introduzca el nombre de usuario y contraseña admin/admin



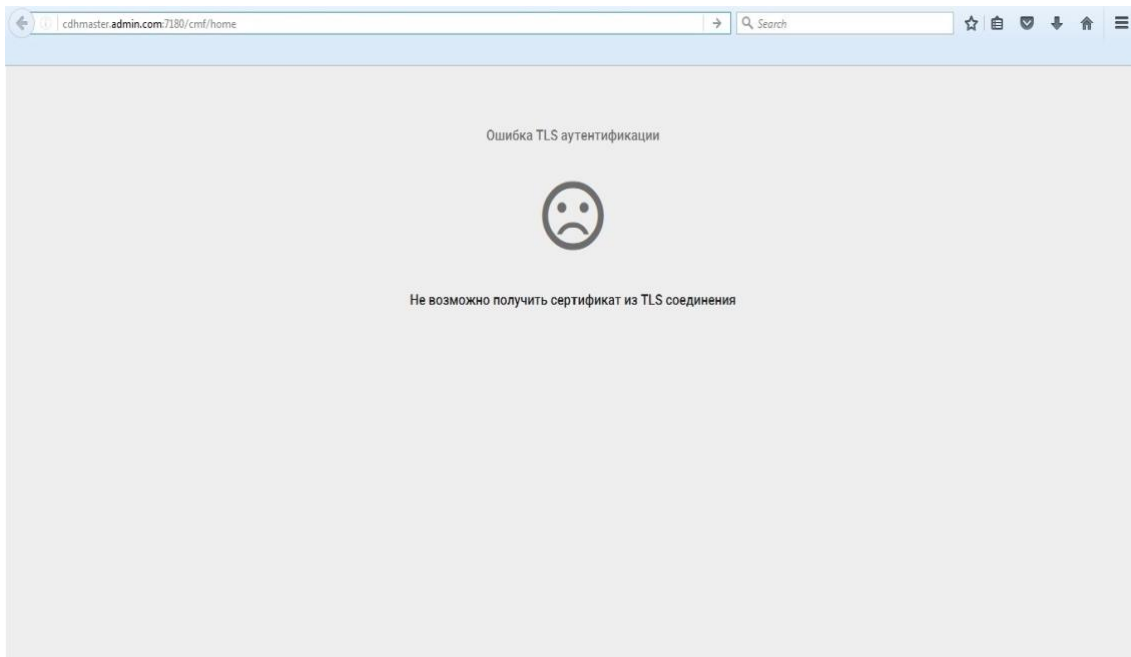
**Figura 112 Autorizacion en Cloudera Manager.**

- La autenticación en Cloudera Manager:



*Figura 113 Cloudera Manager.*

- Si algo se ha configurado de forma incorrecta, o la falta del certificado GOST no pasa autorización:



*Figura 114 No hay autorización.*

## **Conclusiones**

He estudiado y probado un nuevo método de autenticación doble de usuario en Cloudera Manager mediante Certificado GOST + CriptoARM +CriptoPro para el sistema operativo que permite al usuario mejorar la seguridad de acceso en Hadoop.

Ventajas de CriptoLogin:

- Soporte de certificado GOST
- Soporte de autenticación de acceder en distintos web interfaces con certificado GOST

Desventajas de CriptoLogin:

- Sin soporte políticas para acceder a archivos, carpetas, bases de datos, tablas o columnas
- Licencia GOST de pago

## 7. Sistema de autenticación multiprotocolo.

Para incrementar la seguridad en el acceso a sistemas se suele utilizar autenticación multifactor. Actualmente podemos encontrar SmartCards y dispositivos para ofrecer esta autenticación, pero tienen algunas limitaciones. Estos elementos generalmente tienen que estar conectados al equipo que tienen que autorizar y están orientados a plataformas Web o necesitan modificaciones en las aplicaciones para poder ser utilizados.

Gracias a los avances en los dispositivos basados en System-on-Chip, podemos integrar soluciones criptográficamente robustas y de bajo coste.

En este capítulo se presenta un dispositivo autónomo que permite la autenticación multifactor en los sistemas cliente-servidor de manera transparente, lo que facilita su integración en los sistemas HPC y en la nube, a través de una puerta de enlace genérica. El token electrónico propuesto (eToken), basado en ESP32, proporciona una capa adicional de seguridad basada en la criptografía de curva elíptica. Las comunicaciones seguras entre elementos utilizan el protocolo Message Queueing Telemetry Transport (MQTT) para facilitar su interconexión. Se han evaluado diferentes tipos de posibles ataques y el impacto en las comunicaciones. El sistema propuesto ofrece una solución eficiente para aumentar la seguridad en el acceso a servicios y sistemas.

### 7.1 Introduccion

Los servidores y servicios definen esquemas de autenticación para evitar el acceso no autorizado. En consecuencia, son elementos esenciales para implementar un sistema seguro. Una gran cantidad de ataques informáticos aprovechan las vulnerabilidades en los sistemas de autenticación. Common Vulnerabilities and Exposures (CVE) es una lista actualizada y creciente de vulnerabilidades de seguridad conocidas. En particular, dentro de ellos están aquellos relacionados con la autenticación [CVE20]. Es importante implementar mecanismos para fortalecer y garantizar la autenticación de usuarios y sistemas.

Un primer enfoque es usar contraseñas para controlar el acceso, pero es insuficiente, ya que puede ser susceptible a ataques basados en diccionarios, ingeniería social o acceso a bases de datos donde se almacenan las credenciales [KHA14].

Se han propuesto algunos sistemas para evitar el uso de claves de usuario, como Pico [STA11], donde se establece un protocolo de autenticación seguro.

Agregar elementos adicionales a una contraseña permite un componente de seguridad extra. Los sistemas con autenticación multifactor se utilizan a menudo para mejorar la autenticación, por lo que, si la seguridad de un elemento se ve comprometida, el otro elemento puede garantizar que el acceso permanezca seguro.

Los recursos criptográficos, como el cifrado simétrico y asimétrico, el hash y la firma de clave pública son elementos sólidos para resolver la autenticación de forma matemática. Además, es vital el uso de verdaderos generadores de números aleatorios seguros (true random number generators o TRNG).

Las vulnerabilidades a menudo están relacionadas principalmente con la forma en que se implementan las soluciones. Es crucial tener cuidado al elegir la combinación correcta de recursos para prevenir ataques. Podemos encontrar muchos sistemas de autenticación basados en los elementos indicados anteriormente. En particular, la criptografía de clave pública basada en curvas elípticas [JOH01] tiene un impacto significativo en el desarrollo de sistemas robustos actuales.

Podemos combinar contraseñas con algunos otros elementos para aprovechar la seguridad. Existen soluciones basadas en aplicaciones móviles [YIL15], o sistemas basados en el envío de claves por SMS como medida que Azure se integra en su portal [MIC20]. Pero este tipo de autenticación tiene defectos bien documentados [KRE20]. Si se obtiene una tarjeta SIM duplicada mediante ingeniería social o suplantando al propietario, la seguridad basada en el teléfono móvil puede verse comprometida.

Los usuarios están progresivamente sensibilizados con respecto a la privacidad de su teléfono móvil y no desean instalar aplicaciones para acceder a recursos relacionados con el trabajo. Además, si cambian sus teléfonos, implica reinstalar el software. Por esta razón, muchos usuarios prefieren tener un elemento de autenticación externo.

Los dispositivos electrónicos con recursos criptográficos permiten agregar seguridad adicional y son más difíciles de duplicar y manipular. Desde hace tiempo, podemos encontrar soluciones basadas en tarjetas inteligentes o tokens de acceso.

Los nuevos dispositivos System-on-Chip ofrecen capacidad de 32 bits, recursos criptográficos integrados y varios sistemas de comunicación, lo que permite la integración de nuevas alternativas de autenticación.

El proceso de autenticación se puede abordar de diferentes maneras, dependiendo de cómo se interconectan los recursos y qué nivel de seguridad queremos implementar. El objetivo del sistema propuesto es abordar algunos de estos problemas, utilizando un token electrónico (eToken) en un esquema de autenticación multifactor que evita la duplicación y facilita la comunicación entre elementos. El sistema incluye una puerta de enlace transparente que permite la integración con los sistemas actualmente en uso, lo que simplifica su integración global. También utiliza comunicaciones seguras basadas en el protocolo Message Queuing Telemetry Transport (MQTT). Este modelo es más flexible que los protocolos de autenticación habituales, ya que permite autenticar sistemas remotos o combinarse con la autorización multi-eToken.

## 7.2 Los sistemas de seguridad

En los entornos HPC y entornos Cloud suele confiarse en una pareja de clave pública/privada para acceder remotamente. Un problema es el uso descontrolado de claves públicas/privadas ya que, si vamos dejando claves públicas en diversos servidores, puede llegar un momento en que no sabemos exactamente en todas las máquinas donde están instaladas. Si alguien accede al ordenador original, queda comprometido el acceso a todos los nodos. Para resolver esto, Cloudflare utiliza certificados short-lived y autenticación basada en credenciales SSO.

Durante un tiempo se han utilizado sistemas basados en smartcard para identificar de forma segura. Estas tarjetas requerían un lector que podía estar integrado en un teclado, ordenador portátil o lectores USB. El problema es que, con el tiempo, algunos controladores han dejado de funcionar, porque han quedado obsoletos por las versiones de sistema operativo. Nuevas smartcards también soportan NFC lo que facilita el acceso, pero implica utilizar un equipo como elemento intermedio para autenticar. Si no disponemos de dicho elemento, no es posible realizar la autenticación. Además, el usuario no puede validar o cancelar una petición si la tarjeta está insertada.

El uso de sistemas basados en token ha crecido ya que son elementos sencillos de utilizar. Es un recurso que el usuario posee, por lo que no es suficiente disponer una clave de acceso o una pareja de clave pública/privada. Además, el usuario valida mediante alguna pulsación el acceso solicitado.

## 7.3 Sobre el sistema propuesto.

El sistema propuesto ofrece un sistema multifactor de autenticación que puede utilizarse en diversas configuraciones. Desde un modelo básico cliente/servidor, un entorno HPC donde se puede acceder a múltiples servicios y servidores, o bien un entorno cloud aprovechando diversos recursos como el envío de mensajes o servidores virtualizados. Aunque podría utilizarse en sí como un único elemento de autenticación, debido al incremento en el número de ciberataques, es recomendable disponer de múltiples elementos que refuercen la autenticación. Por eso se propone este sistema como un elemento adicional de autenticación que puede trabajar de forma transparente.

El sistema propuesto permite hacer que cualquier aplicación pueda utilizar este sistema de autenticación, no sólo SSH, sino cualquier protocolo basado en TCP/IP (bases de datos en red, sistemas de big data, navegadores). El sistema permite la autenticación remota, es decir, no es necesario que el cliente acceda a un sistema para que autentique, el acceso a un servidor. Simplemente puede ser que cuando un usuario quiere acceder a un sistema remoto, se solicite la autenticación.

El sistema supervisa en tiempo real las comunicaciones, manteniendo un registro de las comunicaciones activas, por lo que si es necesario se puede cortar una comunicación específica de forma instantánea.

Es escalable, ya que se basa en sistemas MQTT ampliamente extendidos, con gran cantidad de implementaciones que ofrecen soluciones de alta disponibilidad y redundantes. Los servidores MQTT son ampliamente utilizados en infraestructuras IoT, e incluso proveedores cloud también ofrecen servicios basados en MQTT por lo que puede aprovecharse los servicios que ofrecen.

Es independiente del dispositivo, ya que accede mediante una petición MQTT. Así, distintos diseños pueden ofrecer una solución de autenticación sin tener que instalar ningún controlador adicional.

Permite la autenticación entre sistemas remotos, no sólo de la sesión que tiene abierta el cliente con un servidor.

El servidor evita ataques más fácilmente ataques de denegación de servicio ya que los puertos están cerrados y sólo se activan cuando se requiere la petición.

El protocolo es abierto, lo que permite el desarrollo e implementación de diversas soluciones, con distintos modelos de clave pública/privada.

El eToken puede verificar el contexto, es decir, puede detectar que recursos Bluetooth o Wifi están próximos. Para que eToken autorice debe detectar un entorno válido. Esta operación se puede activar remotamente por el administrador que puede controlar que recursos son válidos.

El sistema separa el servicio de autenticación del sistema al que se accede, como ocurre con Kerberos, aunque en nuestro caso no necesitamos marcas de tiempo y la autenticación está basada en los algoritmos de clave pública.

Características	Modelo Propuesto	Kerberos + biometric auth [HOA15]	Kerberos + TPM → Hadoop	FIDO U2F Yubikey	Oauth + OpenID Connect+ SecSign	Smart Cards
Criptografía clave pública/privada	Público	Privado	Público (TPM)/ Privado(Kerberos)	Público	Público	Público/ Privado
Autorización de nodos remotos	Sí	Sí	No	Sí	Sí	No
Escalable	Sí	No	No	Sí	Sí	No
Multiprotocolo / Fácil de adaptar a nuevos protocolos	Sí	Kerberización/ Sólo Windows	Sólo para Hadoop	Sí	No	No
Reprogramable	Sí	Yes	No (TPM)	No	Sí	No
Activación con Password	Sí	No	No	No	No	Pin
Autorización Combinada	Sí	No	No	No	No	No
Indirect Client to Auth-server	Sí	No	No	No	No	No

connection						
Control de revocación en tiempo real	Sí	No	No	No	No	No
Time Synchronisation independent	Yes	No	No	Yes	Yes	Yes

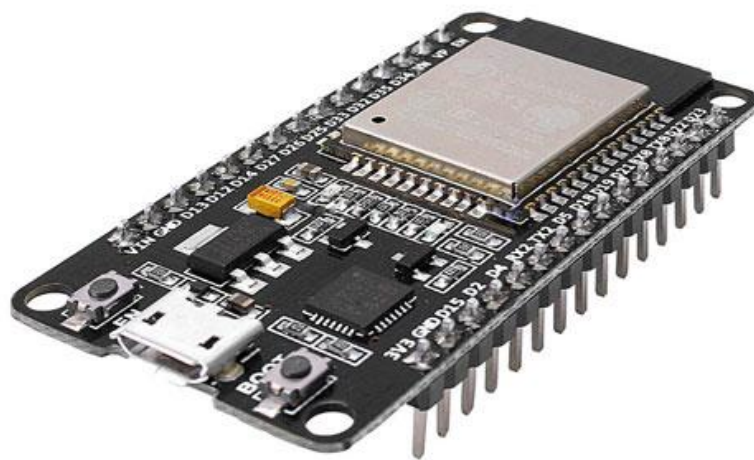
*Tabla 16 Comparación del sistema propuesto con otros esquemas de autenticación.*

## 7.4 eToken y ESP 32.

El eToken desarrollado está basado en el ESP32, un dispositivo de bajo coste con un procesador con 2 núcleos, soporte para funciones criptográficas e incluye comunicaciones WiFi y Bluetooth. Aunque tiene conexión USB, puede trabajar standalone, no necesita estar conectado a USB, ideal para llevarlo de forma portable y trabajar con equipos a los que no se tiene acceso USB. Una de las ventajas de estos sistemas frente a chips como el ECC608 es que, si se detecta una vulnerabilidad, el sistema puede ser reprogramado. El sistema es abierto, la idea es tener un sistema abierto de bajo coste que pueda utilizarse para una autenticación segura. El sistema también permite la conexión de sistemas de forma que puede utilizarse para autenticar cualquier aplicación.

El hardware de la plataforma IoT está hecho en el módulo ESP-WROOM-32 ESPresif ESP32-D0WDQ6. Chip ESP32-D0WDQ6: hecho con tecnología SoC (System-on-a-Chip).

ESP32-WROOM es un módulo con un chip ESP32-D0WDQ6, 4 MB de memoria Flash y todos los periféricos que están ocultos debajo de una carcasa de metal. Al lado de la carcasa hay una antena en miniatura de la pista en la capa superior de la placa de circuito. La carcasa metálica protege los componentes del módulo y, por lo tanto, mejora las propiedades electromagnéticas.



*Figura 115 ESP32.*

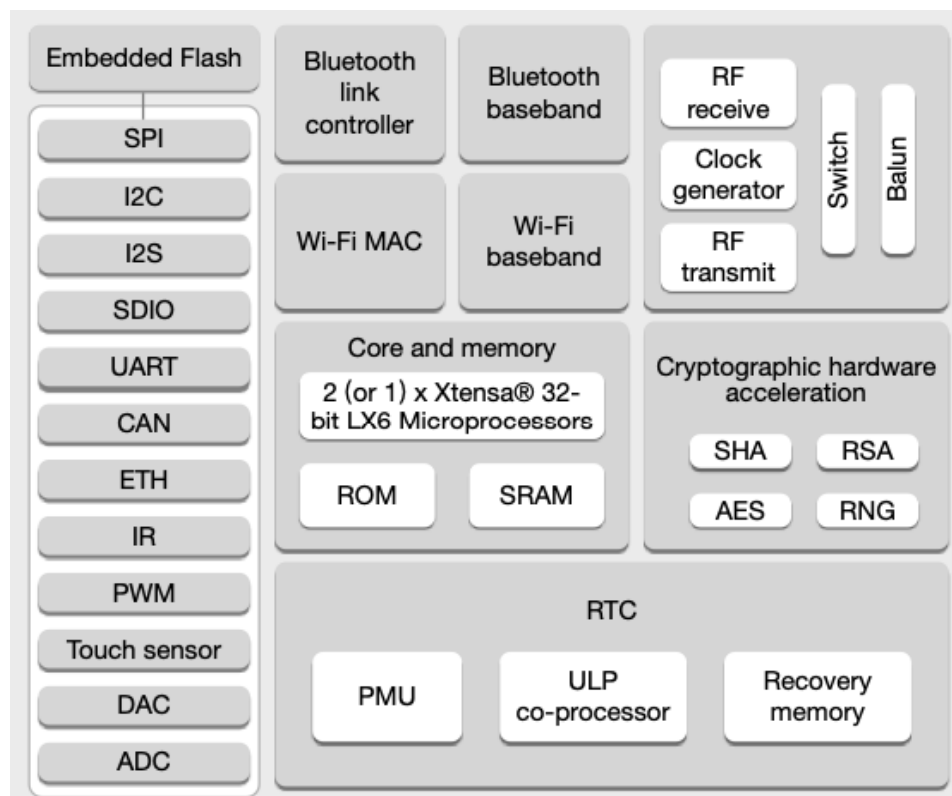


El cristal incluye un procesador Tensilica Xtensa LX6 de 2 núcleos de 32 bits, 520 KB de SRAM y 448 KB de memoria flash, 4 MB de memoria flash externa. La frecuencia del reloj se establece en 240 MHz, dependiendo del modo de consumo de energía.

Hay un sensor de temperatura incorporado, un sensor Hall, un controlador de infrarrojos para recibir y transmitir, un controlador de botón táctil, Bluetooth (BLE v4.2 BR / EDR), Wi-Fi (Wi-Fi 802.11 b / g / n (2.4 GHz)).

El convertidor USB-UART en el chip CP2102 permite que ESP32-WROOM se comunique con el puerto USB de la computadora. Cuando se conecta a una PC, la plataforma ESP32 DevKit se define como un puerto COM virtual.

El conector micro-USB está diseñado para programar y alimentar la plataforma ESP32 DevKit usando una computadora.



*Figura 116 Diagrama de bloques del microcontrolador ESP32.*

## Contactos (PinOut)

A ambos lados del circuito hay pines de 15 pines.

25 pines de uso general disponibles. Todos los contactos admiten interrupciones. Corriente máxima en pines: 12 mA



## **Alimentación.**

El regulador de voltaje reductor lineal AMS1117-3.3 proporciona energía al microcontrolador. El voltaje de salida es de 3.3 voltios con una corriente máxima de 1 A.

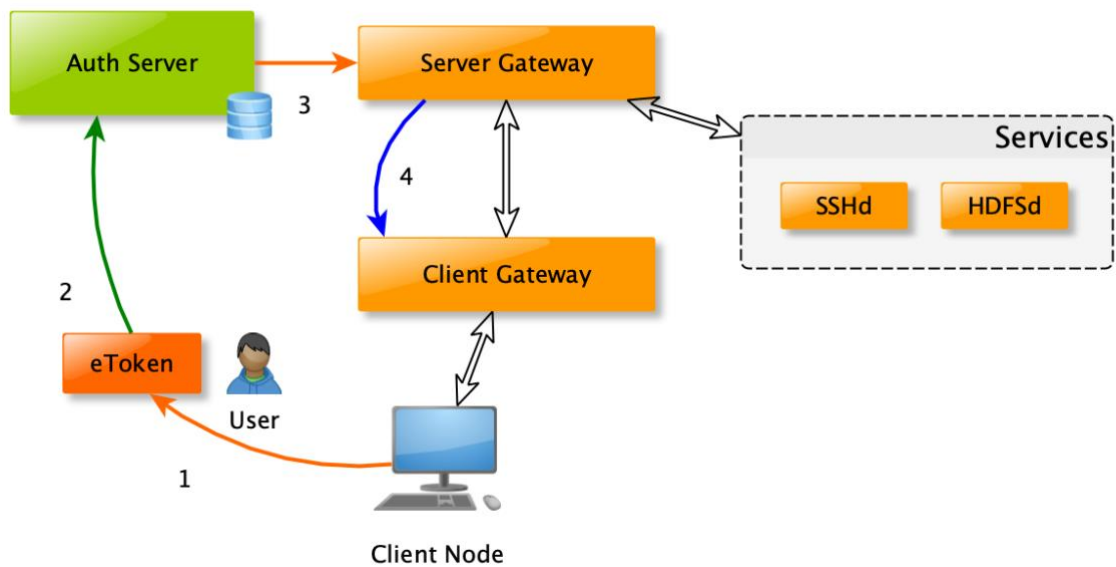
La alimentación se suministra a través de un conector micro-USB o un pin Vin. La fuente se determina automáticamente. Cuando se alimenta a través de USB, use un cargador de 5V con un cable Micro USB. En caso de suministro de energía a través de Vin, se recomienda el voltaje de entrada de 5 a 14 V. El convertidor de potencia en la placa igualará el voltaje de entrada a los 3.3 V.

## **7.5 Arquitectura del sistema.**

El sistema propuesto está orientado a ser distribuido y escalable, permitiendo el control de elementos que pueden colocarse en infraestructuras privadas, entornos de nube o modelos híbridos. Además, permite que cualquier aplicación use este esquema de autenticación. Por lo tanto, actúa como un puente entre el cliente y el servidor, evitando tener que modificar las aplicaciones originales y ofreciendo un impacto reducido en el rendimiento global, como se describe en la Sección 5.

El sistema de comunicación se implementa con los siguientes elementos:

- Servicio de solicitud de cliente (CRS): realiza la solicitud de servicio al eToken.
- eToken (ET): valida la solicitud del usuario y la firma.
- Servidor de autenticación (AS): este servidor verifica la autenticidad del mensaje y solicita acceso al servidor de puerta de enlace. Otro proceso de autorización verifica las reglas de ACL para acceder a los servicios.
- Gateway Server (GS): acepta la comunicación entrante al servidor y monitorea la comunicación en caso de revocación.
- Cliente de puerta de enlace (GC): permite la comunicación con el servidor de puerta de enlace y se integra con el servicio de solicitud del cliente.



**Figura 118 Elementos principales del sistema de autenticación y comunicación.**

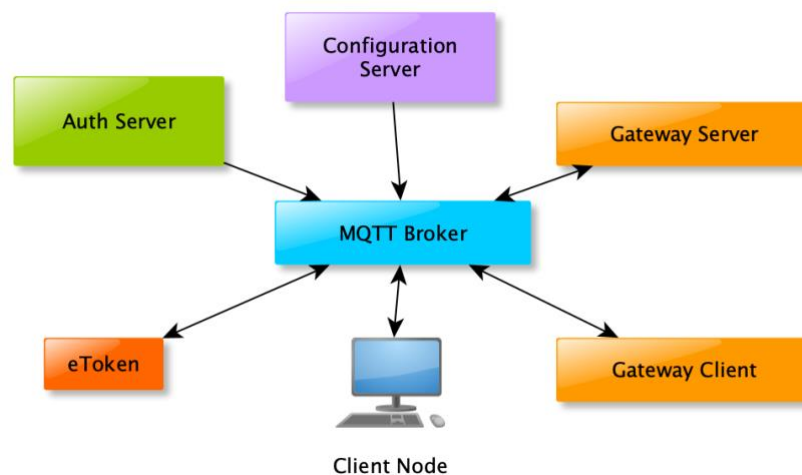
La ventaja del servidor de autenticación centralizado es que simplifica la administración de las reglas de acceso y evita la administración de la revocación de certificados. Además, AS puede forzar la cancelación de la comunicación al instante, informando a la SG que debe cerrar una conexión específica.

CRS, AS, GS y GC están escritos en Golang. Es un lenguaje que permite que las aplicaciones sean portátiles, rápidas, robustas y optimizadas para la concurrencia en cada arquitectura. Además, el sistema puede funcionar con contenedores como Docker. El código ESP32 está escrito en C ++, utilizando el conjunto de herramientas PlatformIO. La Figura 1 muestra cómo los diferentes elementos están interconectados.

Además, hay dos elementos adicionales:

- MQTT Broker (MB): establece la comunicación entre todos los componentes de forma transparente.
- Servidor de configuración (CS): el único servidor autorizado para administrar cualquier configuración en el sistema

El protocolo MQTT basado en un sistema de publicación y suscripción. Gracias al modelo centralizado con un agente MQTT, todos los elementos están interconectados de forma segura.



*Figura 119 El broker MQTT comunica todos los elementos.*

## 7.6 Algoritmos Usados

Los algoritmos que utilizamos se basan principalmente en: funciones hash unidireccionales, criptografía de curva elíptica, esquema de cifrado integrado de curva elíptica (ECIES) y generadores de números aleatorios verdaderos (TRNG). Para no extender la duración de este trabajo, no se incluye la base matemática de estos algoritmos, ya que se describe en la bibliografía aquí incluida.

### 7.6.1 Funciones hash

Las funciones hash unidireccionales se usan ampliamente en sistemas criptográficos [STI06][SAR10]. Ofrecen un valor determinista para una secuencia de entrada. Dado un valor de salida, es computacionalmente imposible encontrar el valor de entrada original. Por defecto, el sistema usa la función hash SHA-256 que se considera criptográficamente robusta, pero también puede funcionar con SHA-384 y SHA-512. De acuerdo con Lu et al. [LU11], algunas implementaciones de SHA-256 son vulnerables a los ataques, por lo que se recomienda utilizar funciones hash más fuertes. En cualquier caso, otras funciones seguras de un solo sentido podrían implementarse y utilizarse gracias a la capacidad de reprogramación del eToken.

### 7.6.2 Criptografía de curva elíptica

Los algoritmos de clave pública basados en la curva elíptica también se usan con frecuencia en sistemas de comunicación criptográficos robustos como Transport Layer Security (TLS) [BLA06]. En particular, el algoritmo de firma digital de curva elíptica (ECDSA) [JOH01] es una variante del algoritmo de firma digital (DSA) basado en la criptografía de curva elíptica. Estos algoritmos son una alternativa atractiva que está reemplazando los sistemas basados en RSA, en parte porque se utilizan tamaños de clave más pequeños en la curva elíptica para proporcionar un nivel de seguridad equivalente. En los algoritmos de clave simétrica, existe una correspondencia directa

entre el nivel de seguridad y el tamaño de la clave utilizada. En contraste, en los algoritmos de clave asimétrica puede cambiar dependiendo de los algoritmos utilizados. La siguiente Tabla muestra una comparación de diferentes tamaños de clave para los algoritmos RSA y de curva elíptica y sus referencias de acuerdo con el Marco de Desarrollo IoT de Espressif.

Una diferencia entre RSA y ECDSA es que RSA permite la firma y el cifrado, mientras que ECDSA solo permite la firma. Alternativamente, la curva elíptica Diffie-Hellman (ECDH) se basa en el algoritmo Diffie-Hellman [DIF76] usando la curva elíptica y permite el intercambio de un valor seguro entre dos elementos que se pueden usar más adelante en el cifrado simétrico entre ambos.

RSA key Size	ECDSA Key Size	ESP_IDF Curve
1024 bits	160-223 bits	secp192r1, sec192k1
2048 bits	224-255 bits	secp224r1, sec224k1
3072 bits	256-383 bits	secp256r1, secp256k1, bp256r1
7680 bits	384 bits-511 bits	secp384r1, bp384r1
15360 bits	512 >= bits	sepc512r1, bp512r1

**Tabla 17 Comparación del tamaño de clave en RSA Y ECDSA.**

El Grupo de estándares para la criptografía eficiente recomienda los parámetros de dominio para cada curva [SEC20]. La función de firma utilizada se basa en cualquiera de las curvas elípticas utilizadas actualmente implementadas en MBED TLS, tales como:

- FIPS 186-4: secp192r1, secp224r1, secp256r1, secp384r1, secp512r1.
- Brainpool: bp256r1, bp384r1, bp512r1
- Koblitz: secp192k1, secp224k1, secp256k1.

Podemos encontrar algunas implementaciones de algoritmos de curva elíptica según lo descrito por Liu et al. [LIU15] utilizando una TI MSP430. Es un microcontrolador de 16 bits con una velocidad máxima de 25 MHz. El eToken propuesto es de bajo costo, basado en un microcontrolador de 32 bits de doble núcleo a 240 MHz, supera esta implementación con tiempos de ejecución reducidos incluso con tamaños de bits más grandes. En la subsección 4.1 hemos estudiado los tiempos de diferentes funciones para estas curvas elípticas utilizadas en el eToken.

### 7.6.3 Esquema de cifrado integrado de curva elíptica

El esquema de cifrado integrado de curva elíptica (ECIES) es un sistema híbrido que combina cifrado simétrico con criptografía de clave pública basada en curva elíptica para la entrega segura de mensajes. La clave secreta utilizada para el cifrado simétrico se oculta mediante una función de derivación de clave (KDF) y una función de intercambio de clave (ECDH) de KA (acuerdo de clave). En [GAY10], se muestra en detalle cómo funcionan estos esquemas. La siguiente Figura ilustra cómo se cifra la clave de sesión (SKk) usando este esquema.

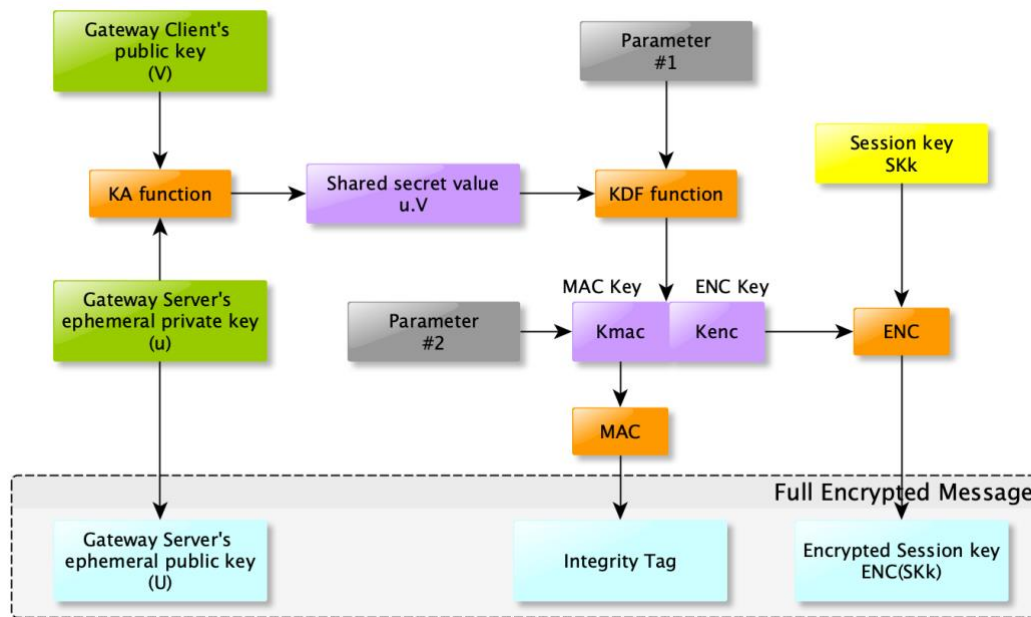


Figura 120 Esquema utilizado para cifrar la clave de sesión.

### 7.6.4 Generador de números aleatorios verdaderos

El ESP32 incluye un generador de números aleatorios de hardware. Si el Bluetooth o WiFi está habilitado (como en nuestro caso), utiliza el ruido de RF como fuente de entropía y los valores obtenidos pueden considerarse verdaderos números aleatorios. Además, las funciones generadoras de números aleatorios utilizadas por todos los elementos pasan la suite Dieharder Random Number Test [BRO20].

### 7.6.5 Combinando Algoritmos

La función hash se calcula para todos los mensajes enviados y el elemento emisor la firma con la clave privada para que el receptor pueda validarla.

Dado que el sistema se basa en una clave pública, no es necesario almacenar claves de cifrado simétricas que puedan comprometer la seguridad del sistema. Además, la autenticación no requiere múltiples mensajes entre el cliente y el servidor, ya que la autenticación se verifica en cada paso.

Utilizamos la notación que se muestra en la siguiente Tabla para identificar elementos y funciones primarias utilizadas durante las diferentes etapas. Cada una de las fases involucradas se describe a continuación.

Símbolo	Significado
$SKk$	Clave de Sesión
$Clc$	Cliente
$Sn$	Sesion
$Ui$	Usuario
$Tt$	eToken t
$Tt2$	Posible 2º eToken
$Sj$	Servidor

$GWCgc$	Cliente pasarela gc
$GWSgs$	Servidor pasarela gs
$Pub(Ei)$	Clave Pública Elemento i
$Priv(Ei)$	Clave Privada Elemento i
$ESk$	Clave Sesión Efímera
$AS$	Servidor Autenticación
$SVCs$	Servicio
$SEQq$	Número de secuencia
$CS$	Servidor Configuración
$h()$	Función Hash
$Sg(hash,privK)$	Función firma hash con clave privada privK
$Vf(msg,pubK)$	Función verificación msg con clave pública pubK

*Tabla 18 Notación de los elementos utilizados.*

## 7.7. Registro de recursos

El sistema funciona bajo un modelo de confianza basado en un esquema de clave pública, por lo que es esencial almacenar las claves secretas de forma segura y correcta. Para crear una nueva configuración, definimos 3 pasos:

### 7.7.1 Registro del servidor de configuración

Inicialmente, inicializamos el servidor de configuración. Este es el único que puede modificar la configuración del sistema de cualquier recurso del sistema. Es responsable de:

- Se crean un par de claves pública y privada para firmar los mensajes de configuración. La clave pública se instala en todos los elementos y, por lo tanto, es la única autorizada para realizar cambios.
- Se crea la autoridad de certificación raíz para firmar los certificados utilizados para acceder al agente MQTT. La clave pública de esta Autoridad de Certificación (CA) debe instalarse en todas las computadoras y eTokens para establecer una comunicación TLS segura entre los elementos.

### 7.7.2 Registro del broker MQTT

Las comunicaciones entre los diferentes elementos se realizan a través de mensajes MQTT utilizando comunicaciones seguras basadas en TLS y evitando posibles ataques Man-in-the-Middle (MitM).

Este agente MQTT necesita un certificado que debe estar firmado por la CA, es decir, por el servidor de configuración. Las comunicaciones MQTT pueden basarse en dos modelos:

- Usar solo el certificado del servidor MQTT, que es verificado por los clientes que deben tener la clave pública.
- Uso de certificados de cliente y servidor firmados por la CA.



### 7.7.3 Registro de elementos funcionales

Todos los elementos funcionales (servidores, clientes, usuarios y eTokens) deben seguir el siguiente paso para ser incluidos en el sistema de autenticación:

- Genere un UUID único de 128 bits.
- Un par de claves pública / privada para firmar los mensajes. La clave privada debe almacenarse de forma segura en cada uno de ellos.
- Cada elemento tiene un perfil de configuración creado por el servidor de configuración para acceder al agente MQTT con un usuario definido para cada uno. Este mecanismo de control permite cancelar el acceso si es necesario eliminando el usuario asociado con ese elemento.
- Cada recurso envía su UUID y clave pública al Servidor de configuración. Por lo tanto, cada elemento tiene un conjunto de claves públicas autorizadas. Por lo tanto, cualquier solicitud que no provenga de ninguna de estas claves se descarta directamente. En el caso del eToken, la conexión USB puede actuar como una puerta de enlace inicial para configurarlo, y la autorización inicial solo es posible cuando el dispositivo elimina todas las claves o se reinicia por completo, dejándolo vacío.

El servidor de configuración también debe almacenar la información de forma segura, principalmente para evitar que se modifiquen o agreguen elementos falsos. Por lo tanto, cuando se inicia el sistema, el servidor de configuración genera la configuración de los nodos para conectarse mediante el agente MQTT.

Cuando se cambia la configuración, el servidor de configuración envía los cambios a los elementos que actualizan su información. Todos los elementos solo necesitan almacenar la configuración básica y mantenerla actualizada correctamente.

En el caso de una nube pública, podemos tener el Servidor de autenticación y el Servidor de configuración en un lugar seguro, fuera de la nube si es necesario, para que podamos implementar físicamente algunas medidas de seguridad adicionales.

## 7.8 Petición de servicio

Una vez establecida la configuración, el sistema ahora está disponible para establecer comunicaciones de forma segura. Por lo tanto, el proceso de solicitud de servicio se puede dividir en 5 pasos:

### 7.8.1 El usuario solicita eToken para acceder al servicio

El cliente crea una nueva sesión ( $S_n$ ) identificada con un UUID, con un número de secuencia ( $SEQ_q$ ) para el servicio ( $SVC$ ) y con las firmas de validación hash de solicitud correspondientes. Esta solicitud está firmada por el usuario ( $U_i$ ) (1) y por el nodo del cliente ( $CL_c$ ) (2) que solicita acceso:

$$Sgn\_user(S_n) = Sg(h(S_n, SEQ_q, U_i, CL_c, SVCs), Priv(U_i)) \quad (1)$$

$$Sgn\_client(S_n) = Sg(h(S_n, SEQ_q, U_i, CL_c, SVCs), Priv(CL_c)) \quad (2)$$

Se utiliza una función hash robusta para obtener un valor único para cada solicitud. Además, cada solicitud tiene un número de serie, que tiene 2 ventajas:

- Son mensajes diferentes con firmas diferentes.
- El número es incremental, por lo que no se puede volver a usar el mismo número.

Cuando se inicia un elemento, el Servidor de configuración envía un número aleatorio desde el cual continúa el recuento. Si el número es menor que el valor anterior o mayor que una ventana pequeña, envía un mensaje de revocación de acceso que informa al Servidor de configuración, por lo que el elemento debe identificarse nuevamente.

La nueva solicitud de acceso a un servicio se envía al eToken (Tt). Esta es una diferencia importante con respecto al modelo U2F, donde la clave de hardware no identifica al usuario. Por lo tanto, un eToken que no sea reconocido por un usuario autorizado rechazará la solicitud.

### 7.8.2 Etoken solicita acceso al servidor de autenticación

Después de una condición de encendido o después de un período de tiempo, el usuario debe iniciar sesión en eToken para permitirle enviar una contraseña como un nivel de seguridad adicional.

El eToken tiene una lista de claves públicas autorizadas para que pueda confirmar si un usuario (Ui) de un nodo (CLc) tiene acceso al sistema. Esto simplifica la configuración del eToken ya que solo requiere la lista de claves públicas autorizadas.

El eToken verifica la firma del nodo (CLc) usando Vf(msg, pub (CLc)), y si es válido, muestra en la pantalla la solicitud para que el usuario lo valide presionando un botón o usando la huella digital sensor si está habilitado. Los tiempos necesarios para la validación de la firma se muestran en la siguiente sección, en función de la curva elíptica seleccionada.

Además, el eToken puede verificar el contexto; es decir, puede detectar qué recursos Bluetooth o Wifi están cerca. Debe detectar un entorno válido para autorizar. Esta operación puede ser activada de forma remota por el servidor de configuración que puede controlar qué recursos son válidos.

Debido a que el identificador de sesión (Sn) es único usando un UUID de 128 bits, la información no se repite en cada firma. Solo es necesario firmar Sn.

Una vez autorizado por el usuario y el contexto, se agrega la firma eToken (3):

$$Sgn\_eToken(Sn) = sg(h(Sn), Priv(Tt)) \quad (3)$$

Si definimos una configuración con autenticación multi-usuario y multi-eToken, la solicitud autorizada por el primer eToken se envía al segundo eToken para ser validada por el segundo usuario. En este caso, también se incluye una firma adicional del segundo eToken (4).

$$Sgn\_eToken2(Sn) = sg(h(Sn), Priv(Tt2)) \quad (4)$$

Esta solicitud se envía al servidor de autenticación para la autorización global y para verificar todas las ACL.

### 7.8.3 El servidor de autenticación solicita acceso a la puerta de enlace

El eToken envía la solicitud al servidor de autenticación. Esto tiene todas las reglas de ACL que permiten decidir si el usuario  $U_i$  del nodo  $CL_c$  con el eToken  $T_t$  tiene acceso al servicio SVC. Las posibles reglas se describen más adelante en la subsección 3.7.

El servidor de autenticación verifica todas las firmas para que pueda confirmar que se trata de una solicitud válida. Luego, el proceso de autorización verifica las reglas de ACL y, cuando valida la solicitud, firma con su clave privada (5):

$$\text{Sgn\_AS}(S_n) = \text{sg}(h(S_n), \text{Priv}(AS)) \quad (5)$$

Esta solicitud se envía al servidor de puerta de enlace para habilitar el acceso.

### 7.8.4 El servidor de puerta de enlace habilita el servicio

Se crea una clave de sesión aleatoria de 256 bits (SKk) que será válida durante un tiempo configurable que, de forma predeterminada, se establece en 60 segundos. Dado que el servidor es el que genera la clave de sesión, no hay necesidad de marcar la hora, por lo que evita el problema de la sincronización horaria entre nodos.

El servidor abre el puerto de comunicación que permite el acceso al servicio y envía la clave de sesión SKk utilizando el esquema de cifrado híbrido ECIES (6) al Gateway Client que está integrado con el sistema de solicitud de servicio inicial.

$$\text{GwMsg} = \text{ECIES\_Crypt}(SKk, \text{Pub}(GWCgc)) \quad (6)$$

El mensaje incluye una clave ESk efímera para un intercambio seguro utilizando ECIES. Gateway Client recibe el mensaje (6) para obtener la clave de sesión SKk (7).

$$SKk = \text{ECIES\_Decrypt}(\text{GwMsg}, \text{Priv}(GWCgc)) \quad (7)$$

### 7.8.5 Comunicación a través del portal

Cuando el cliente recibe la clave de sesión (SKk) (7), solicita comunicación con el servidor a través de la puerta de enlace.

El servidor de puerta de enlace comprueba la clave de sesión en la lista de claves válidas y decide si se acepta la conexión. Las conexiones no válidas se rechazan después de unos segundos, evitando un ataque de fuerza bruta. Si la clave de sesión es válida, establece comunicación con el servicio y el ID de sesión ( $S_n$ ) se incluye en la lista de comunicaciones activas.

Si el servidor de autenticación recibe una solicitud de revocación de comunicación, puede enviar el mensaje al servidor de puerta de enlace para cortar inmediatamente la comunicación. Esta revocación puede provenir del supervisor global.

La siguiente figura muestra el gráfico de secuencia de mensajes (MSC) entre los diferentes elementos:

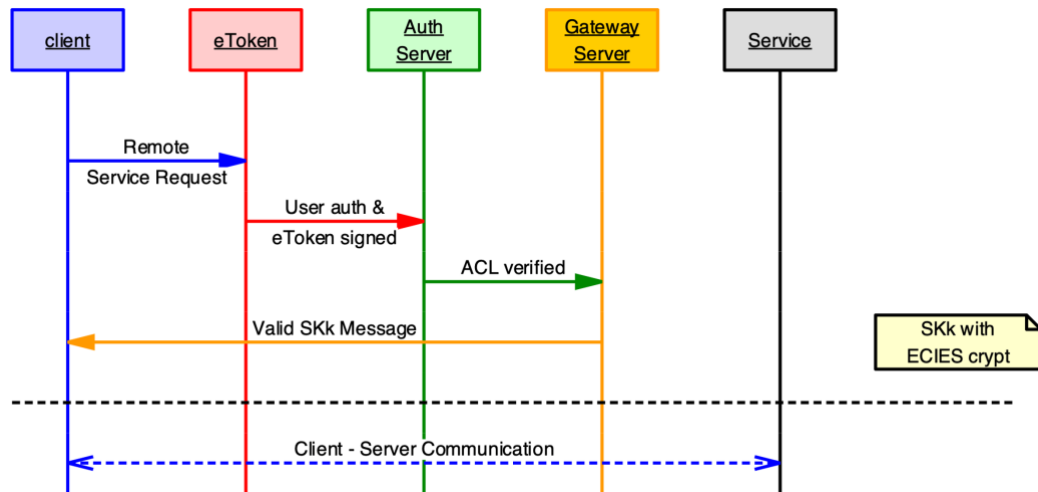


Figura 121 Diagrama MSC entre los elementos del sistema.

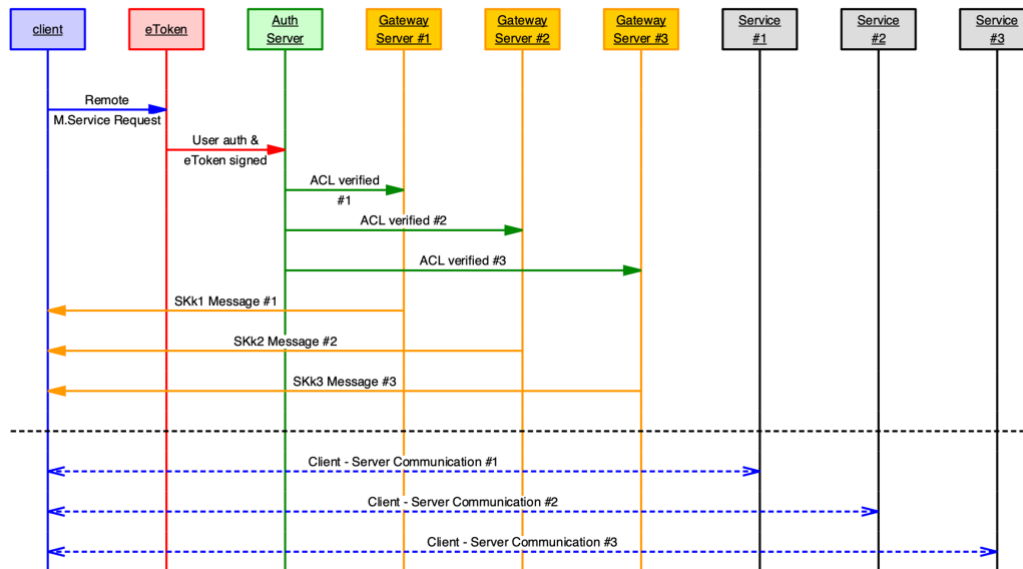
## 7.9 Autorización Multiservicio

Una de las ventajas de este modelo es que permite abrir múltiples canales de comunicación simultáneamente con la autenticación.

En este caso, el cliente solicita el servicio como de costumbre. Gracias al modelo MQTT, el servidor de autenticación envía múltiples solicitudes a diferentes servidores de puerta de enlace. Por lo tanto, cada servidor reenvía un mensaje de apertura desde su canal de comunicación, confirmando el acceso del cliente para los sockets TCP.

Suponga una aplicación Big Data basada en Hadoop donde un cliente accede a múltiples servidores Hadoop. El cliente crea múltiples solicitudes para acceder a los servidores. En este caso, se crean varias claves de sesión SKk1, SKk2, ..., una para cada cliente de Gateway que accede al sistema.

Las sesiones se pueden mantener abiertas durante un tiempo, lo que evita tener que autenticarse repetidamente. La siguiente Figura muestra cómo el sistema realiza una solicitud multiservicio con múltiples servidores. Los mensajes amarillos tienen las claves de sesión que el cliente recibe para acceder a diferentes servicios.



**Figura 122 Autorización multiservicio.**

### 7.10 Formato de mensaje

Los mensajes se envían en formato JSON para facilitar la comunicación a través de MQTT y los elementos los procesan rápidamente.

Cada solicitud tiene un número de secuencia que aumenta con cada solicitud. La verificación confirma que el número debe ser mayor que el último recibido y dentro de una ventana de 16 valores. Los valores fuera de esta ventana se consideran anómalos.

Uno de los parámetros de configuración es el nivel de control de secuencia. En el modelo de seguridad relajado, se genera un mensaje de advertencia para el administrador. En modo estricto, el administrador debe aceptar el reinicio de la numeración.

Por lo tanto, después de cada verificación y autorización, los elementos agregan nuevos campos JSON con la firma correspondiente.

Como cada elemento tiene su UUID, se utiliza para distribuir la información en la jerarquía de temas utilizada en el agente MQTT. Cada solicitud de servicio también tiene un UUID único, por lo que no hay superposición en las solicitudes determinadas de forma exclusiva, y no hay colisión en los mensajes enviados a través de MQTT. El agente MQTT se encarga de enviar el mensaje correspondiente a cada nodo.

Las comunicaciones principales se realizan utilizando el siguiente esquema de temas:

*anb/domain/dst\_element\_uuid/request/src\_element\_uuid/*

Por lo tanto, el elemento que recibe los mensajes tiene acceso a todos los subtemas:

*anb/domain/dst\_element\_uuid/*

Y en cambio, el elemento de envío solo tiene acceso a los subtemas:

*anb/domain/dst\_element\_uuid/request/src\_element\_uuid/*

## 7.11 Definición de reglas de acceso

Se pueden establecer reglas flexibles para acceder a los servicios en función de varias combinaciones de clientes, usuarios y nodos eToken. Se pueden definir grupos de elementos para facilitar la configuración. El archivo de configuración YML consta de 3 partes principales:

- Definición del elemento. (Proceso de autenticación)
- Definición de elementos grupales y combinados. (Proceso de autenticación)
- Definición de regla de acceso (ACL). (Proceso de autorización)

Cada elemento se identifica mediante un alias: su UUID y el nombre del archivo que contiene la clave pública en formato PEM. Se pueden crear nuevos alias y pueden contener cualquier combinación de otros alias con los operadores "AND" y "OR".

En las reglas de acceso, por un lado, están los elementos que acceden: usuarios, nodos de clientes y eTokens y, por otro lado, los elementos a los que se accede: Servidores y servicios.

Es posible establecer cualquier combinación de reglas que se evalúen para autorizar el acceso a un servicio en un servidor. Como se mencionó anteriormente, se pueden combinar varios eTokens para definir una autorización en cascada en la que dos usuarios deben autorizar el acceso a un servicio, lo que aumenta la seguridad en el sistema.

Los procesos de autenticación y autorización funcionan de forma independiente y, si es necesario, se pueden separar en diferentes servidores y contextos. YML se puede dividir en cada parte, pero por razones prácticas, el mismo servidor puede resolver las solicitudes de los clientes.

## 7.12 Análisis de seguridad

Esta sección estudia la resistencia de diferentes elementos en el sistema propuesto a ataques frecuentes en entornos de autenticación.

### 7.12.1 Resistencia en el broker MQTT

En el modelo de entrega de mensajes basado en MQTT, solo el agente MQTT tiene un puerto TCP de escucha y todos los elementos se conectan a él mediante sockets TCP bidireccionales. Este diseño permite que los elementos no tengan un puerto de escucha y, por lo tanto, no hay una forma directa de acceder a ellos.

En cualquier caso, es el broker MQTT el que debe ser más robusto contra los ataques. El protocolo MQTT es bastante seguro y confiable; Aunque MQTT se puede considerar robusto porque varias implementaciones de agente MQTT tienen su seguridad bien evaluada, la forma en que se implementa y configura puede causar problemas.

En el modelo propuesto, utilizamos MQTT con certificados, comunicaciones TLS y usuarios con reglas de acceso, lo que ofrece varias ventajas:

- Las comunicaciones están cifradas.
- Solo se aceptan certificados firmados por la CA. En nuestro caso, los certificados son generados por el servidor de configuración.
- Los usuarios de MQTT para cada elemento permiten el control de la comunicación.

- Las ACL se utilizan para acceder a los temas del agente MQTT, de modo que cada elemento tenga acceso limitado solo a su información.

Todos estos recursos agregan un nivel de seguridad adicional a la comunicación entre todos los elementos. Incluso si el agente MQTT se ve comprometido y un atacante tiene acceso para leer todos los mensajes o puede enviar mensajes nuevos, el sistema sigue siendo confiable porque:

Si el atacante puede leer todos los mensajes, conoce las solicitudes de servicio de los usuarios, pero estos mensajes tienen información irrelevante. En relación con la clave de sesión SKk, se cifra utilizando el esquema ECIES, por lo que el atacante no puede acceder al servidor de puerta de enlace en una fase posterior. SKk solo se puede descifrar con Priv (GWCgc).

Si el atacante puede enviar mensajes, también necesita una clave privada válida Priv (Ei) para firmar sus solicitudes falsas, pero todos los mensajes con firmas no válidas son rechazados.

### 7.12.2 Resistencia a la manipulación del EToken

El eToken almacena la clave privada y las claves públicas autorizadas en su memoria flash ESP32 a través de la biblioteca de almacenamiento no volátil (NVS) y utilizando cifrado NVS basada en AES-XTS de 256 bits.

ESP32 usa tablas de particiones internas para la memoria Flash. Dado que la partición está marcada como cifrada y la opción de cifrado Flash está habilitada, el gestor de arranque cifrará esta partición utilizando la clave de cifrado flash en el primer arranque. Cuando ESP32 ejecuta el programa eToken, guarda la información en forma cifrada, por lo que solo el programa puede interpretar correctamente el contenido. Por lo tanto, no es posible ningún intento externo de modificar la configuración en el eToken NVS.

Si ESP32 se reprograma parcialmente para intentar acceder más tarde a la partición Flash que almacena las claves, no es legible ya que está cifrado.

Cambiar la configuración ESP32 solo es posible si recibe los mensajes firmados por el servidor de configuración.

### 7.12.3 Resistencia al robo de EToken

En el esquema propuesto, si un adversario A roba un eToken, este no puede extraer ninguna información de la memoria de los dispositivos robados, como se describe en la subsección 4.2.

Además, un eToken robado se puede bloquear en varios niveles:

- Después de una condición de encendido o un tiempo establecido, se necesita una contraseña para iniciar sesión en eToken y habilitarla. Después de varios intentos fallidos, el dispositivo está bloqueado.
- Validación de contexto: (cuando está activo) El eToken solo funciona dentro del rango de detección de dispositivos WiFi o Bluetooth validados. Entonces, fuera de su ubicación, el eToken no está activo.
- En el servidor MQTT: Eliminar la cuenta de acceso.
- En el servidor de autenticación: cancelación de permisos de autenticación.

- En el acceso WiFi eToken: el eToken solo puede comunicarse con las redes WiFi preconfiguradas.

El eToken puede reprogramarse, pero como se mencionó en la sección anterior, la configuración operativa está cifrada y no se puede usar para acceder al entorno seguro.

#### **7.12.4 Resistencia de los ataques MitM**

MQTT establece comunicaciones entre nodos a través de comunicaciones TLS. El agente MQTT y los clientes MQTT usan certificados firmados por la CA, evitando así que un nodo se haga pasar por el agente MQTT.

TLS incluye comprobaciones de integridad, por lo que la comunicación a nivel TCP / IP no se puede modificar. Incluso, si el MQTT se ve comprometido, los mensajes no pueden modificarse porque están firmados con las claves privadas respectivas en cada etapa.

En relación con la clave de sesión SKk, el atacante no puede obtenerla porque incluso si puede obtener el mensaje, se cifra con el esquema ECIES, y se necesita la clave privada Priv (GWCgc) para descifrar.

El número de secuencia SEQq evita el uso de solicitudes anteriores, por lo que debe crear nuevas solicitudes firmadas adecuadamente.

En el caso de la comunicación entre puertas de enlace, si el servicio ofrece comunicaciones seguras como SSH, es el servicio mismo el que puede detectar un ataque MitM, evitando así agregar una capa adicional de verificación de integridad que podría ralentizar las comunicaciones.

#### **7.12.5 Resistencia de los ataques al nodo del cliente**

Supongamos un sistema con un cliente SSH donde usamos clave pública / privada para acceder a un servidor. Si un atacante accede al nodo del cliente, podría acceder al servidor SSH en una configuración típica, pero con el sistema propuesto, el acceso al servidor SSH requiere la autorización eToken.

Además, el modelo propuesto puede usar una clave privada para el nodo del cliente y una clave privada para el usuario. El administrador puede seleccionar qué teclas se pueden usar, una de ellas o ambas, si es necesario.

Para ajustar la clave privada al nodo del cliente, la aplicación del cliente almacena la clave privada del nodo cifrada con una clave que se deriva en parte de las propiedades de la CPU, como: modelo, familia, líneas de caché y extensiones. Por lo tanto, esta clave privada solo puede funcionar en una plataforma con las mismas características físicas, lo que dificulta el uso de la clave privada en otro nodo de cliente.

Por lo tanto, si un adversario accede al nodo del cliente podría usar la clave privada, pero necesita conocer la contraseña para habilitar el eToken y tener que acceder a ella para validar la solicitud de servicio. Si el usuario tiene el eToken en su poder, no autorizará la solicitud y, por lo tanto, no es posible acceder al servicio.

Las credenciales utilizadas para acceder a MQTT están cifradas, por lo que el atacante no tiene acceso directo al agente MQTT. Además, debido a la segmentación de los



temas utilizados, este nodo tiene acceso a un número limitado de temas, por lo que el MQTT no se ve comprometido.

Del mismo modo, si hay un ataque al nodo del cliente que copia la clave privada, esto por sí solo no es suficiente para acceder al servidor, ya que se requiere autorización de eToken.

### **7.12.6 Resistencia de los ataques en el servidor de autenticación**

El servidor de autenticación es un elemento vital porque decide si una solicitud es válida y puede solicitar a GS una nueva conexión. En el modelo propuesto, el AS es más seguro que otros sistemas porque no necesita tener puertos abiertos afuera. Es una diferencia significativa porque está protegido contra ataques directos desde el exterior. AS debe estar debidamente aislado, ya que es responsable de admitir autorizaciones. El sistema propuesto tiene dos ventajas:

- Las solicitudes llegan a través de mensajes MQTT, por lo que el servidor de autenticación no tiene un puerto abierto explícitamente para recibir solicitudes.
- Este modelo permite tener el servidor en una ubicación diferente y mucho más segura. Por lo tanto, en una configuración de HPC, el servidor de autenticación puede estar en otra ubicación con un mayor nivel de seguridad. Del mismo modo, en una configuración en la nube, las instancias se pueden virtualizar y el servidor de autenticación puede estar fuera de la nube, en un entorno físico seguro.

### **7.12.7 Resistencia de ataques en el servidor de configuración**

El servidor de configuración, como el servidor de autenticación, es un elemento esencial, ya que es el único que establece la cadena de confianza en todos los elementos. Por lo tanto, se recomienda que se coloque en un lugar seguro.

Del mismo modo, el servidor de configuración se comunica mediante mensajes MQTT, por lo que tampoco tiene un puerto abierto para recibir ataques. Almacena su clave privada cifrada con una clave que se deriva en parte de las propiedades de la CPU, como lo hacen los clientes con sus claves privadas, y en parte de la contraseña del administrador. Además, el administrador puede usar su propio eToken para realizar modificaciones válidas en la configuración global, por lo que no es posible un ataque directo sin ese eToken.

### **7.12.8 Resistencia de los ataques que acceden a las puertas de enlace**

Los servidores de puerta de enlace solo aceptan solicitudes a través de mensajes MQTT firmados por AS y tienen puertos cerrados. Solo cuando se solicita acceso entre puertas de enlace, el puerto de acceso en el servidor se abre temporalmente.

La clave de sesión SKk garantiza que solo un cliente con ese valor puede establecer una comunicación con el servidor. El esquema ECIES permite enviarlo de Gateway Server a Gateway Client de manera segura. Ni el administrador ni el servidor de configuración pueden acceder a ese valor.

Este puerto puede permanecer fijo para facilitar las reglas del firewall, o puede ser dinámico, es decir, en diferentes solicitudes, puede abrir diferentes puertos, por lo que un atacante no sabría qué puerto usar para acceder al sistema. Esta información se transmite en el mensaje que las puertas de enlace han intercambiado.

Cuando el servidor de puerta de enlace recibe una solicitud, el puerto escucha durante un tiempo limitado y, después de recibir el inicio de una comunicación, se verifica la clave de sesión SKk. Si es incorrecto o no se recibe por un tiempo, la comunicación se cierra.

### **7.12.9 Resistencia de los ataques de denegación de servicio**

Como se mencionó anteriormente, solo hay dos puntos de entrada: el agente MQTT y el servidor de puerta de enlace que solo está abierto después de una solicitud.

Si el servidor de puerta de enlace permanece abierto durante un tiempo porque se solicitó un servicio, el servidor espera la clave de sesión SKk y, si no es válido, cierra la conexión después de unos segundos, evitando el acceso repetido al sistema. Entonces, el elemento que puede soportar la mayor presión es el agente MQTT, pero estos servidores también están diseñados para evitar ataques de denegación de servicio (DoS). Además, se pueden tomar otras medidas a nivel de firewall para limitar el número de intentos de acceso por segundo o bloquear las direcciones IP después de algunos intentos de acceso fallidos.

### **7.12.10 Resistencia de cortes de red**

Aunque las redes locales actuales son muy robustas, pueden experimentar atropellos inesperados durante períodos cortos. También es esencial considerar estos escenarios para evitar posibles ataques en condiciones de inestabilidad de la red.

Todas las comunicaciones con el agente MQTT usan TLS, por lo que un atacante no puede aprovechar el envío de mensajes falsos debido a la integridad incluida en las comunicaciones TLS. MQTT es un protocolo bastante robusto, incluso bajo redes poco confiables. En nuestro caso, todos los elementos, incluidos los servidores de autenticación y configuración, son clientes MQTT e incluyen la reconexión automática, por lo que tan pronto como la red se restablezca, estarán listos para transmitir y recibir. El sistema usa la Calidad de servicio (QoS) 1 en los mensajes MQTT, lo que garantiza que un mensaje se envíe al menos una vez al receptor, de modo que los paquetes no se pierdan, y el remitente sabe si tiene que retransmitir el mensaje debido a la pérdida de conexión.

### **7.12.11 Comunicación de bloque selectivo en tiempo real**

El sistema propuesto facilita el control de todas las comunicaciones que están en uso. Si se descubre un acceso no autorizado, es posible cortar cualquiera de las comunicaciones que están en progreso, gracias al control en tiempo real de todas las comunicaciones. Si es necesario, el Configuration Server puede enviar mensajes de configuración

actualizados bloqueando las reglas de acceso de ACL y evitando el acceso posterior al sistema al elemento involucrado.

En otros sistemas, una vez que se autoriza el acceso, se establece la comunicación. Para cancelar la comunicación, se debe identificar el socket que tiene la comunicación o detectar la dirección IP y el número de puerto para bloquear la conexión con las reglas del firewall.

## 7.13 Análisis de rendimiento

En nuestras pruebas, hemos estudiado el tiempo de ejecución de las principales funciones criptográficas utilizadas en diferentes entornos, y el impacto de la puerta de enlace en algunas transferencias de datos.

### 7.13.1 Tiempos de ejecución de ECC en EToken

El eToken se basa en ESP32, un sistema de bajo costo que incorpora un procesador de doble núcleo de 240 MHz, con recursos de aceleración criptográfica y comunicaciones integradas.

Una cuestión importante es saber si este dispositivo podría procesar las funciones criptográficas de curva elíptica requeridas en un tiempo razonable.

Hemos evaluado diferentes curvas elípticas con la biblioteca MBed TLS. La siguiente Tabla muestra los tiempos de ejecución obtenidos para la generación de claves, firma y verificación.

	Genkey (ms)	Sign (ms)	Verify (ms)
MBEDTLS ECP DP SECP192R1	112	52	175
MBEDTLS ECP DP SECP224R1	140	67	231
MBEDTLS ECP DP SECP256R1	224	95	347
MBEDTLS ECP DP SECP384R1	331	129	497
MBEDTLS ECP DP SECP521R1	574	229	842
MBEDTLS ECP DP BP256R1	2057	726	2784
MBEDTLS ECP DP BP384R1	3792	1132	4990
MBEDTLS ECP DP BP512R1	6897	2067	9361
MBEDTLS ECP DP SECP192K1	163	69	239
MBEDTLS ECP DP SECP224K1	193	85	311
MBEDTLS ECP DP SECP256K1	230	99	349

*Tabla 19 Tiempos de generación, firma y verificación para diversas curvas elípticas en el ESP32.*

La curva SECP256R1 ofrece tiempos de respuesta razonables por debajo de 0,35 segundos para verificar y 0,1 segundos para firmar.

### 7.13.2 Tiempos de funciones ECDSA

Se han evaluado varias arquitecturas para obtener los tiempos de ejecución de la generación de claves, la verificación y la clave de firma de un valor hash de 256 bits utilizado en diferentes funciones de cliente y servidor. Preferimos esta forma de análisis

porque este modelo puede implementarse con cualquier combinación de elementos ubicados en diferentes arquitecturas y sistemas operativos. Las siguientes 4 tablas muestran los tiempos promedio obtenidos en microsegundos utilizando la biblioteca Golang ECIES con diferentes procesadores, tales como: Intel i7-4980HQ @ 2.80GHz, Intel Xeon Silver 4116 CPU @ 2.10GHz, AMD EPYC 7571 @ 2.4 GHz y ARMv7 en Raspberry PI 4 modelo B.

	Genkey(us)	Sign (us)	Verify (us)
Elliptic.P224	763	804	1751
Elliptic.P256	17	59	114
Elliptic.P384	4417	4724	8430
Elliptic.P521	7594	7955	15903

**Tabla 20 Tiempos de ejecución ECC en Intel i7-4980HQ @ 2.80GHz MacOS 10.15.4.**

	Genkey(us)	Sign (us)	Verify (us)
Elliptic.P224	2478	2327	4406
Elliptic.P256	59	182	309
Elliptic.P384	14401	12176	27282
Elliptic.P521	27484	29694	50466

**Tabla 21 Tiempos de ejecución ECC en Intel Xeon Silver 4116 CPU @ 2.10GHz Centos 7.7 kernel 3.10.0.**

	Genkey(us)	Sign (us)	Verify (us)
Elliptic.P224	1178	1228	2353
Elliptic.P256	26	65	147
Elliptic.P384	6315	6613	12953
Elliptic.P521	10854	11271	20077

**Tabla 22 Tiempos de ejecución ECC en AMD EPYC 7571 @ 2.4 GHz RHEL 8.2.**

	Genkey(us)	Sign (us)	Verify (us)
Elliptic.P224	5798	5937	7935
Elliptic.P256	2329	3525	9015
Elliptic.P384	112167	117424	229602
Elliptic.P521	211553	209649	416941

**Tabla 23 Tiempos de ejecución ECC en Raspberry PI4 ARMv7 Raspbian GNU/Linux 10.**

Puede comprobarse que, en general, la curva elíptica P256 es la que requiere menor tiempo de procesamiento, y que, además, en las arquitecturas de Intel y AMD también tiene tiempos de ejecución considerablemente menores debido a las optimizaciones hardware.

### 7.13.3 Tiempos del algoritmo ECIES en diversos clientes y servidores.

Para garantizar el acceso entre los gateways de acceso cliente y servidor intercambian una clave secreta de 256 bits. Esta clave se envía de forma segura basada en el esquema ECIES. Es importante que

A continuación, se muestran los tiempos de cifrado y descifrado del token de 256 bits basado en Golang 1.14 y utilizando la biblioteca ECIES. En este caso se utiliza la curva secp256k1 para generar la clave efímera. En la Tabla siguiente se muestran los tiempos de generación, cifrado y descifrado para distintos procesadores.

	Genkey (ms)	Cifrado (ms)	Descifrado (ms)
Intel i7-4980HQ @ 2.80GHz MacOS 10.15.4	2,55	4,54	2,39
Intel Xeon E5-2676 v3 @ 2.40GHz Amazon Linux 2 kernel 4.14.173	3,03	5,79	2,92
AMD EPYC 7571 @ 2.4 GHz RHEL 8.2	3,14	6,34	2,83
ARM 64 bits AWS Graviton @ 2.3 GHz	4,48	8,62	4,42
Intel Xeon Silver 4116 CPU @ 2.10GHz Centos 7.7 kernel 3.10.0	8,26	13,97	6,62
Raspberry PI4	61,93	86,06	43,21

*Tabla 24 Tiempos de ejecución ECIES en diferentes plataformas clientes y servidores.*

### 7.13.4 Tiempos de comunicación y de acceso a los servicios

Se ha estudiado el tiempo medio empleado en cada etapa.

En la siguiente tabla se muestran los tiempos para cada una de las etapas descritas en el apartado 3.5 y considerando el siguiente escenario.

Podríamos tener los siguientes tiempos medios:

Estado	Tiempo medio
Primera petición → eToken	57 ms
Verificación firma eToken	347 ms
Validación usuario	Espera validación usuario (toque)
eToken → Servidor Autenticación	4ms
Validación firma & comprob.ACL en Serv.Aut.	1ms .. 5 ms (depende de reglas ACL)
Envío al servidor pasarela	0.67 ms
Validación & Apertura serv. Pasarela	0.75 ms
Cifrado ESCIES, envío & descifrado clave sesión	9 ms
Conexión pasarela Cliente-Servidor	0.1 ms

*Tabla 25 Tiempos medios de ejecución de cada etapa.*

Aunque el proceso conlleva varias etapas, y aunque el tiempo de verificación de firma es mayor en el ESP32, en realidad, el mayor tiempo se debe a la espera a que el usuario valide el acceso. En general, los tiempos de autenticación son razonables.

Para estas medidas se ha utilizado la curva elíptica SECP256R1 que ofrece un elevado nivel de seguridad y tiempos óptimos de ejecución.

Una de las ventajas de este modelo es que puede funcionar en paralelo, de forma que simultáneamente pueden establecerse múltiples procesos de autenticación.

### 7.13.5 Sobrecarga en el ancho de banda y latencia.

En este caso utilizamos 2 programas, un eco de TCP, que retransmite todo lo que recibe y otro programa que transmite y recibe secuencias de paquetes de diversos tamaños. De esta forma se puede evaluar las latencias y los anchos de banda considerando la comunicación de forma simétrica.

Tiempos de envío y vuelta.

	Conexión directa	Conexión a través de los gw	% ancho de banda efectivo
1	77 us	148 us	52%
10	136 us	256 us	53%
100	202 us	369 us	55%
1000	295 us	514 us	57%
10000	495 us	762 us	65%
100000	1148 us	1490 us	77%
1000000	7491 us	8424 us	89%
10000000	70140 us	75519 us	93%

*Tabla 26 Tiempos de transferencia en Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, CentOS 7.7.*

### 7.13.6 Impacto en las comunicaciones SSH

También hemos estudiado el impacto en las comunicaciones SSH. En nuestro caso, hemos definido un túnel SSH con el programa de eco TCP al final. Hemos medido el tiempo de ida y vuelta dividido por 2, como la prueba anterior con conexión directa y utilizando nuestra puerta de enlace.

Packet Size (Bytes)	Direct connection (time in $\mu$ s)	Connection through Gateways (time in $\mu$ s)	% Effective BW
1	169	205	82%
10	283	350	81%
100	406	493	82%
1,000	632	718	88%

10,000	1,043	1,207	86%
100,000	2,160	2,450	88%
1,000,000	13,797	15,785	87%
10,000,000	126,017	127,589	99%

*Tabla 27 Tiempos de transferencia en SSH para Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, CentOS 7.7.*

### 7.13.7 Impacto en el acceso a datos en Hadoop HDFS

Algunas HPC usan Hadoop para el procesamiento de Big Data. Hemos estudiado el impacto de acceder a los datos almacenados en un clúster de Hadoop utilizando HDFS. La versión de Hadoop es 3.2.1, y los nodos de la computadora tienen Intel E5-2620 v4 @ 2.10GHz, CentOS 7.7, 500 GB HDD y están conectados con una red de 1 Gbps. El nodo del cliente utiliza un cliente Golang para HDFS.

#### Acceso a metadatos

Hemos probado 4 funciones básicas: crear archivo, crear directorio, eliminar archivo y eliminar directorio.

Todas las funciones tienen casi el mismo tiempo de ejecución y varían de 3.37 ms a 4.01 ms dependiendo de la carga del servidor y los procesos internos de Hadoop. La penalización debida a la puerta de enlace es de solo 0.05 ms, que está prácticamente oculta en la fluctuación normal de los tiempos obtenidos. Solo representa 1.25%. Por lo tanto, la puerta de enlace tiene un impacto reducido en estas operaciones.

#### Acceso a los datos

Como hemos mencionado antes, Hadoop tiene cierta fluctuación en el tiempo de ejecución debido a cómo maneja los buffers de descarga internos y su política de acceso al disco. En este caso, hemos ejecutado cada prueba 100 veces y hemos obtenido el valor medio para evitar valores extremos. El tiempo de ejecución incluye: crear / abrir, escribir / leer y cerrar operaciones.

La siguiente Tabla muestra el tiempo de ejecución obtenido con estas pruebas para la conexión directa y el uso de la puerta de enlace. Podemos observar que el tiempo obtenido para un archivo con tamaños desde 1 byte hasta 10,000 bytes es casi el mismo, debido a la sobrecarga de las funciones remotas, por lo que las memorias intermedias no pueden funcionar correctamente y optimizar sus transferencias de datos.

Packet Size (Bytes)	Read Direct connection (time in $\mu$ s)	Read through Gateways (time in $\mu$ s)	Write Direct connection (time in $\mu$ s)	Write through Gateways (time in $\mu$ s)
1				
10	1,617	1,837	19,916	21,014
100				

1,000				
10,000				
100,000	2925	2,977	19,929	21,104
1,000,000	10,578	10,949	27,913	29,849
10,000,000	87,488	88,157	109,151	110,587
100,000,000	861,559	861,737	906,371	907,872
1,000,000,000	8,581,931	8,681,764	9,625,042	9,938,063

**Tabla 28 Tiempos de transferencia en Hadoop HDFS para lectura y escritura en Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, CentOS 7.7.**

El impacto de las puertas de enlace es reducido como puede apreciarse con diferentes operaciones de lectura/escritura y distintos tamaños de operación.

## 7.14 Conclusiones

Es necesario analizar el impacto de las principales funciones criptográficas implicadas para determinar si el modelo propuesto puede funcionar en un entorno HPC real y la sobrecarga de las puertas de enlace. Como ya se discutió, este modelo se puede implementar con cualquier combinación de elementos ubicados en diferentes arquitecturas y sistemas operativos.

Los resultados obtenidos confirman que las funciones utilizadas en clientes y servidores basados en criptografía de curva elíptica tienen un tiempo de ejecución reducido en diferentes arquitecturas. Además, las funciones del eToken también pueden ejecutarse en décimas de segundo.

En cuanto a las puertas de enlace, simplifican el proceso de integración con diferentes programas y servicios. En general, tienen una sobrecarga reducida como podemos observar en las pruebas de comunicaciones realizadas.

El uso de circuitos criptográficos está proliferando como sistemas de autenticación multifactor para el acceso a sistemas de forma segura. La ventaja estos sistemas es que es el usuario tiene “físicamente” un elemento no duplicable para autorizar el acceso. Aunque existen estándares como los propuestos por FIDO, estos son objeto de posibles mejoras. En este capítulo se propone un modelo alternativo, basado en el ESP32, un System-on-Chip de bajo coste que permite la ejecución de funciones criptográficas basadas en curva elíptica y diversas interfaces de comunicación.

En combinación con el eToken, se propone un sistema de comunicación que permite la autenticación en programas cliente-servidor de forma transparente, lo que facilita la integración en sistemas. Este sistema ofrece un nivel extra de seguridad, flexibilidad en la configuración, tiempos reducidos en el proceso de autenticación y bajo impacto en las comunicaciones, como se ha podido comprobar tanto en túneles SSH como en el sistema de ficheros HDFS que utiliza Hadoop.



## 8. Conclusiones.

### 8.1 Principales aportaciones y conclusiones.

Las principales aportaciones y conclusiones obtenidas pueden resumirse en los siguientes puntos:

- Se ha estudiado Big Data. Por Big Data nos referimos a conjuntos de datos cuyo tamaño es superior a las capacidades de bases de datos típicos enumerados, almacenamiento, gestión y análisis de la información. El reto de los "grandes datos" no es sólo el análisis de grandes cantidades de información sino también procesar aquella que no está suficientemente estructurada. En dicho procesamiento pueden existir gran cantidad de personas que pueden intervenir por lo que la seguridad no es sólo externa (accesos no autorizados desde Internet) sino también interna.
- Se ha estudiado Hadoop. Hadoop es una plataforma ampliamente utilizada para el procesamiento de grandes volúmenes información y por lo tanto resulta ideal para el ámbito de Big Data. Los modelos de seguridad en Hadoop han ido evolucionando debido a los requerimientos de seguridad necesarios en ámbitos donde existen gran cantidad de recursos y usuarios, por lo que se definen diversos niveles: Administración, autenticación, autorización, auditoría y protección de datos.
- Se ha estudiado Rhino. Es un proyecto de código abierto para mejorar la plataforma Hadoop con mecanismos de protección adicionales. El objetivo de este proyecto es eliminar los agujeros de seguridad en la pila de Hadoop. Para este propósito, Intel está aprovechando los repertorios de instrucciones criptográficas de sus procesadores para mejorar la velocidad en el acceso a datos cifrados. Entre la variedad de trabajos realizados en el marco del Proyecto de Rhino están las nuevas características más interesantes para el cifrado/descifrado de los archivos a través de múltiples modelos, así como añadir una capa de abstracción común para definir una API criptográfica y definir entorno adecuado para la distribución y gestión de claves.
- Se ha hecho un experimento con Hadoop. Los datos obtenidos del trabajo hecho experimental incluyen dos ejecuciones de MapReduce. Durante la primera cuentan registros únicos y se llevó a cabo en la segunda la clasificación. Estos procedimientos no pueden ser paralelizados y llevan a cabo en un solo paso. Se procesó más de 180 millones de registros, por un total de más de 32 GB. La ejecución de todo llevó unos 15 minutos.

- Se ha desarrollado un código para hacer alfa testado Cripto ARM con el algoritmo GOST 28147-89 y después comprobar los datos con algoritmo XTS-AES. Los datos obtenidos durante el experimento mostraron que algoritmo GOST 28147 cifra/descifrar más rápido que XTS-AES.
- Se ha probado un nuevo método de autenticación del usuario con algoritmo GOST 34.10-2001 llamado Cripto Login con Cloudera Hadoop. Este experimento ha mostrado que el algoritmo GOST 34.10-2001 funciona conjuntos con Cloudera Hadoop.
- Se ha desarrollado una aplicación para cifrar/descifrar los datos con algoritmo GOST 28147.
- Se ha hecho un experimento con Azure. Los datos para experimento esta en formato Excel. Los datos obtenidos durante el experimento mostraron que la maquina virtual Azure procesa datos varias veces más rápido que un ordenador normal.
- Se ha hecho un experimento con Amazon AWS donde los datos se cifran en un ordenador normal y después se transmiten a Amazon AWS y descifran. Este esquema funciona muy bien.
- Se ha hecho un experimento para cifrar/descifrar los datos en Amazon AWS.
- Se ha diseñado un sistema de autenticación para Hadoop basado en un circuito electrónico (eToken) con el circuito de criptoautenticacion de Atmel ATECC508A.
- Se ha desarrollado un sistema de autenticación multiprotocolo que puede ser aplicado en entornos Big Data basado en el ESP32, un System-on-Chip de bajo coste que permite la ejecución de funciones criptográficas basadas en curva elíptica con diversas interfaces de comunicación.
- En combinación con el eToken, se propone un sistema de comunicación que permite la autenticación en programas cliente-servidor de forma transparente, lo que facilita la integración en sistemas. Este sistema ofrece un nivel extra de seguridad, flexibilidad en la configuración, tiempos reducidos en el proceso de autenticación y bajo impacto en las comunicaciones, como se ha podido comprobar tanto en túneles SSH como en el sistema de ficheros HDFS que utiliza Hadoop.

## 8.2. Trabajo futuro.

El trabajo descrito en esta memoria abre nuevas perspectivas de investigación y desarrollo. Esta tesis aborda diferentes líneas de aplicación, y cada una de ellas puede ampliarse de forma más profunda.

En las posibles actividades para continuar con la línea de mejora de prestaciones del sistema cifrado/descifrado con algoritmos GOST propuesto en esta tesis, y conocer los beneficios que puede ofrecer a un sistema de seguridad real se destacan:

- Implementar un sistema que permita el uso de un certificado electrónico personal para acceder a los datos cifrados.
- Implementar un sistema que permita automáticamente cifrar/descifrar los datos entre servidor a cliente con algoritmo GOST.
- Integrar el sistema de cifrado de datos con algoritmo GOST en Hadoop.
- Implementar un sistema que permite usar eToken con GOST para acceso en Hadoop.
- Continuar mejorando los sistemas propuestos agregando compatibilidad con otros métodos de autenticación como el Módulo de autenticación conectable de Linux (PAM) u otros protocolos de autenticación.

Finalmente, hay que indicar que apenas existen en la bibliografía implementaciones los algoritmos GOST ya que estos sólo se usan en Rusia pero no en la Unión Europea ni en Estados Unidos. Esto hace posible desarrollar un sistema de cifrado alternativo y la posterior integración en varias aplicaciones.



## Bibliografía.

- [APA13] Apache Hadoop and eToken.Forum.2013  
<https://issues.apache.org/jira/browse/HADOOP-9392>
- [BIS15] Biswas N., Moorthy J., Nanath K. Big Data: Prospects and Challenges // Vikalpa. pp. 70 – 80. 2015.
- [BIF12] A. Bifet “Mining Big Data in Real Time” Informatica Journal pp. 15–20 Dec. 2012.
- [BIR17] M. Birjali, A. Beni-Hssane.Procedia Computer Science. pp. 280- 282. 2017.
- [BEL15] V. Belov. Altiscale Data Cloud. 2015.
- [BLA06] Blake-Wilson, S.; Bolyard, N.; Gupta, V.; Hawk, C.; Moeller, B. Elliptic Curve Cryptography (ECC) Cipher Suites for Transport Layer Security (TLS); RFC Editor, 2006; p. RFC4492;.
- [BRO20] Brown, R.G. Dieharder: A Random Number Test Suite Available online: <https://webhome.phy.duke.edu/~rgb/General/dieharder.php> (accessed on Jun 15, 2020).
- [BUY11] R. Buyya, J. Broberg, GoscinskiDaille A. Cloud Computing: Principles and Paradigms: Wiley, No 2. pp. 3-41.2011.
- [CIO19] Ciolino, S.; Parkin, S.; Dunphy, P. Of Two Minds about Two-Factor: Understanding Everyday {FIDO} U2F Usability through Device Comparison and Experience Sampling.; 2019.
- [CHE11] L. Chernyak. Big Data. 2011.  
<https://www.osp.ru/os/2011/10/13010990/>
- [CAR16] T. Carskaya. Big Data en medecine.2016  
[https://medaboutme.ru/zdorove/publikacii/stati/sovety\\_vracha/big\\_data\\_v\\_medicine\\_tekushchaya\\_situatsiya\\_i\\_perspektivy/](https://medaboutme.ru/zdorove/publikacii/stati/sovety_vracha/big_data_v_medicine_tekushchaya_situatsiya_i_perspektivy/)
- [CHI15] Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Big data: The next frontier for innovation, competition, and productivity.2015  
<http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation>

- [CHA12] Chak L., Hadoop. M.: DMK Press, - 424 c.2012.
- [CHA19] Chadwick, D.W.; Laborde, R.; Oglaza, A.; Venant, R.; Wazan, S.; Nijjar, M. Improved Identity Management with Verifiable Credentials and FIDO. IEEE Comm. Stand. Mag. 2019, 3, 14–20, doi:10.1109/MCOMSTD.001.1900020.
- [CHA99] Chae, C.-J.; Kim, K.-B.; Cho, H.-J. A study on secure user authentication and authorization in OAuth protocol. Cluster Comput 2019, 22, 1991–1999, doi:10.1007/s10586-017-1119-6.
- [CHU14] V. Chuprin, A. Chernishov, N. Gubenko.NoSQL and Hadoop. 2014.  
[http://masters.donntu.org/2014/fknt/chuprin/library/\\_hadoop-security.htm](http://masters.donntu.org/2014/fknt/chuprin/library/_hadoop-security.htm)
- [CNE14] CNews.Big Data.2014.
- [COO10] W. Coombs, Timothy and Sherry J Holladay. The Handbook of Crisis Communication. 1st ed. Chichester, U.K.: Wiley-Blackwell, 2010.
- [COO19] W. Coombs, W. Timothy. Ongoing Crisis Communication. 1st ed. Thousand Oaks: Sage Publications,1999.
- [COU11] N. Courtois.Security Evaluation of GOST 28147-89 In View Of International Standardisation.IACR Cryptology ePrint Archive.pp. 1 - 2.2011
- [CVE20] Common Vulnerabilities and Exposures (CVE). MITRE Corporation <https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=authentication> (accessed on Jun 15, 2020).
- [DIA16] A. Díaz, I. Blokhin, J. Ortega, Hernández-Palacios, Raúl; Rodríguez-Quintana, Cristina; Díaz-García, Juan. Secure Data Access in Hadoop Using Elliptic Curve Cryptography. Lecture Notes in Computer Science. 10049, pp. 136 - 145. 2016.
- [DJO11] M. Djons. Hadoop. 2011.  
<https://www.ibm.com/developerworks/ru/library/l-hadoop-1/>
- [DUB14] N. Dubnova. Big Data. Variety of sources.2014  
<https://www.osp.ru/news/articles/2014/14/13040541/>
- [DJO14] M. Djons. IBM Hadoop and Sentry. 2014.

- <https://www.ibm.com/developerworks/ru/library/se-hadoop/>
- [DON10] B. Dong, Jie Qiu, Qinghua Zheng, Xiao Zhong, Jingwei Li, Ying Li. A Novel Approach to Improving the Efficiency of Storing and Accessing Small Files on Hadoop. 65-72, 2010. [www.comsis.org/hadoop.pdf](http://www.comsis.org/hadoop.pdf)
- [DEE16] Deepak Vohra. Practical Hadoop Ecosystem. Chapter Apache Flume, Book 10.1007/978-1-4842-2199-0 pp 287-300, September 2016.
- [DEB12] C. Debians, P.A.T. Togores, and F. Karakusoglu, “Hdfs replication simulator,2012. ” <https://github.com/peteratt/HDFS-Replication-Simulator>
- [DEA14] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Google, Inc. 2014.
- [DIF76] Diffie, W.; Hellman, M. New directions in cryptography. IEEE Transactions on Information Theory 1976, 22, 644–654.
- [DOU18] Dou, Z.; Khalil, I.; Khreishah, A. A Novel and Robust Authentication Factor Based on Network Communications Latency. IEEE Systems Journal 2018, 12, 3279–3290, doi:10.1109/JSYST.2017.2691550.
- [ELE11] El-Emam, E.; Koutb, M.; Kelash, H.M.; Faragallah, O.S. An Authentication Protocol Based on Kerberos 5. I. J. Network Security 2011, 12, 159–170.
- [EVA05] A. Evangeli. ITWeek.2005.  
<https://www.itweek.ru/infrastructure/article/detail.php?ID=70657>
- [ESA15a] A. Esaulenko. Big Data.2015.  
<https://www.osp.ru/news/articles/2015/13/13045468/>
- [ESA15b] A. Esaulenko Big Data, EMS Isilion. 2015.  
<https://www.osp.ru/news/articles/2015/12/13045419/>
- [FOR14] M. Foru. Big Data, SafeNet.2014.  
<http://www.mforum.ru/news/article/105957.htm>
- [FED13] A. Fedotov, I. Sulkis.Big Data.2013  
<http://wiki.dataved.ru/knol/technology/enterprise-integration/bigdata/big-data-explained>

- [GAP15] D. Gapotchenko. Big Data.2015  
<https://www.osp.ru/news/articles/2015/14/13045492/>
- [GAY10] Gayoso Martínez, V.; Hernandez Encinas, L.; Sánchez Ávila, C. A Survey of the Elliptic Curve Integrated Encryption Scheme. *Journal of Computer Science and Engineering* 2010, 2, 7–13.
- [GAT13] A. Gates. *IEEE Data Eng. Bull.* pp. 34-35. 2013.
- [GAT09] A. Gates, et al.: Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experience. *Proceedings of the VLDB Endowment*, v.2 n.2, August 2009.
- [GOO20] Google Authenticator Available online:  
[https://play.google.com/store/apps/details?id=com.google.android.apps.authenticator2&hl=en\\_US](https://play.google.com/store/apps/details?id=com.google.android.apps.authenticator2&hl=en_US) (accessed on Jun 15, 2020).
- [HAR14] S. Harshawardhan, Prof. P. Devendra. “A Review Paper on Big Data and Hadoop” *International Journal of Scientific and Research Publications*, Volume 4, Issue 10, October 2014.
- [HAR20] Hardt, D. The OAuth2.0 Authorization Framework. Internet Engineering Task Force (IETF) RFC 6749 (accessed on Jun 15, 2020).
- [HOA15] Hoa Quoc Le; Hung Phuoc Truong; Hoang Thien Van; Thai Hoang Le A new pre-authentication protocol in Kerberos 5: biometric authentication. In *Proceedings of the 2015 IEEE RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*; IEEE: Can Tho, Vietnam, 2015; pp. 157–162.
- [HOL12] A. Holmes, *Hadoop in Practice*. M.: Manning Publications, pp 536 .2012
- [HOA15] Hoa Quoc Le; Hung Phuoc Truong; Hoang Thien Van; Thai Hoang Le A new pre-authentication protocol in Kerberos 5: biometric authentication. In *Proceedings of the The 2015 IEEE RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*; IEEE: Can Tho, Vietnam, 2015; pp. 157–162.
- [HUW13] Hu W.-C. *Big Data Management, Technologies, and Applications (Advances in Data Mining and Database Management)*. - 1. - Hershey: IGI Global – 70 c.2013.
- [ISL12] N. Islam, X. Lu, M. Wasi-ur-Rahman, J. Jose, and D. K. Panda, “A micro-benchmark suite for evaluating HDFS operations on modern



- clusters,” in Specifying Big Data Benchmarks. Berlin, Germany, pp. 129–147, 2012.
- [JOH01] Johnson, D.; Menezes, A.; Vanstone, S. The Elliptic Curve Digital Signature Algorithm (ECDSA). *IJIS* 2001, 1, 36–63, doi:10.1007/s102070100002.
- [JON12] M. Jones. Understand Representational State Transfer (REST) in Ruby. 2012.  
<https://www.ibm.com/developerworks/linux/library/os-understand-rest-ruby/>
- [KAN19] Kang, W. U2Fi: A Provisioning Scheme of IoT Devices with Universal Cryptographic Tokens. arXiv:1906.06009 [cs] 2019.
- [KHA14] Khalil, I.; Khreishah, A.; Azeem, M. Cloud Computing Security: A Survey. *Computers* 2014, 3, 1–35, doi:10.3390/computers3010001.
- [KHA15] Khalil, I.; Dou, Z.; Khreishah, A. TPM-Based Authentication Mechanism for Apache Hadoop. In *International Conference on Security and Privacy in Communication Networks*; Tian, J., Jing, J., Srivatsa, M., Eds.; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer International Publishing: Cham, 2015; Vol. 152, pp. 105–122 ISBN 978-3-319-23828-9.
- [KRE20] Krebs, B. Reddit Breach Highlights Limits of SMS-Based Authentication Available online: <https://krebsonsecurity.com/2018/08/reddit-breach-highlights-limits-of-sms-based-authentication/> (accessed on Jun 15, 2020).
- [KRU14] I. Kruglenko. Cifrado los datos en Linux. 2014. <https://cryptoworld.su/data-linux/>
- [KRU12] K. Kruglenko. *Jetinfo* №1-2/2012.  
[https://www.jetinfo.ru/jetinfo\\_arhiv/big-data/bolshie-dannye-bolshaya-problema/2012](https://www.jetinfo.ru/jetinfo_arhiv/big-data/bolshie-dannye-bolshaya-problema/2012)
- [KOR15] N. Korotovsky. Single Sign on and Symphony2. 2015.  
<https://habr.com/ru/post/260183/>
- [LAW14] Lawal Muhammad Aminu, “Implementing Big Data Management on Grid Computing Environment”, *International Journal of Engineering and*

- Computer Science ISSN: 2319-7242, Volume 3, Issue 9, pp. 8455-8459, September 2014.
- [LI18] Li, W.; Mitchell, C.J.; Chen, T. Mitigating CSRF attacks on OAuth 2.0 and OpenID Connect. CoRR 2018, abs/1801.07983.
- [LIN13] J. Lin “MapReduce Is Good Enough?” The control projects. IEEE Computer 32 ,2013.
- [LU115] Liu, Z.; Seo, H.; Hu, Z.; Hunag, X.; Großschädl, J. Efficient Implementation of ECDH Key Exchange for MSP430-Based Wireless Sensor Networks. In Proceedings of the Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security - ASIA CCS '15; ACM Press: Singapore, Republic of Singapore, 2015; pp. 145–153.
- [LU11] X. Lu; W. Wang; Z. Lu; Jianfeng Ma From security to vulnerability: Data authentication undermines message delivery in smart grid. In Proceedings of the 2011 - MILCOM 2011 Military Communications Conference; 2011; pp. 1183–1188.
- [LYN08] C. Lynch. Big data: How do your data grow? Nature. №7209. – pp. 81 – 84. 2008.
- [MIC12] I. Michailenko. Cifrado los datos en Linux. 2012.  
<https://cryptoworld.su/data-linux/>
- [MAR15] C. Marshall, Julian S., Anthony P., Mike M., David G., “Overview of Microsoft Azure Services”, Microsoft Azure, Part 1, 2015.
- [MAY13] V. Mayer -Shenberger, K. Kuker. Big Data. pp. 134 .2013.
- [MIC20] Microsoft SMS-based authentication using Azure Active Directory  
<https://docs.microsoft.com/en-us/azure/active-directory/authentication/howto-authentication-sms-signin> (accessed on Jun 15, 2020).
- [MIL88] Miller, S.P.; Neuman, B.C.; Schiller, J.I.; Saltzer, J.H. Kerberos Authentication and Authorization System. In Proceedings of the In Project Athena Technical Plan; 1988.
- [NEU10] L. Neumeyer, B. Robbins, A. Nair, A. Kesari. S4: Distributed Stream Computing Platform. Data Mining Workshops (ICDMW), IEEE International Conference, 2010.

- <https://www.crn.ru/news/detail.php?ID=105995>
- [OTI12] M. Oti. What is special about the Hadoop system?2012.  
<https://www.osp.ru/winitpro/2012/06/13033252/>
- [OFF15] Official web site. Cripto Pro. <https://www.cryptopro.ru/>
- [OPE20] OpenID Foundation OpenID Connect Available online:  
<https://openid.net/connect/> (accessed on Jul 6, 2020).
- [ORL15] S. Orlov. NEC.2015  
<https://www.osp.ru/lan/2015/12/13047907/>
- [ORA14] Oracle Advanced Security Administrator's Guide Release 2 (9.2) Part Number A96573-01 Configuring Kerberos Authentication. 2014.
- [OLS08] C. Olston, B. Reed, U. Srivastava, R. Kumar, A. Tomkins: Pig Latin: A Not-So-Foreign Language for Data Processing. Proceedings of the 2008 ACM SIGMOD international conference on Management of data, June 09-12, Vancouver, Canada, 2008.
- [PIL14] M. Pilin. What is Hadoop? 2014.  
<https://www.xelent.ru/blog/chto-takoe-hadoop/>
- [POP06] V. Popov, I. Kurepkin, S. Leontie: RFC 4357: Additional Cryptographic Algorithms for Use with GOST 28147-89, GOST R 34.10-94, GOST R 34.10-2001, and GOST R 34.11-94 Algorithms, IETF January 2006.  
<http://tools.ietf.org/html/rfc4357>
- [POK04] P. Pokrovsky, Ilia Chetvertakov. Authentication methods. «Windows IT Pro», № 06, 2004.  
<http://www.osp.ru/win2000/2004/06/177152/>.
- [PAL12] A. Palladin. Big Data: big potential, high priority.2012.  
[https://www.cisco.com/c/ru\\_ua/about/press/2013/04112013a.html](https://www.cisco.com/c/ru_ua/about/press/2013/04112013a.html)
- [PET16] A. Petrov. ITU News of Big Data №1.2016

- <https://itunews.itu.int/Ru/Note.aspx?Note=4879>
- [PAT15] D. Patgiri, Performance evaluation of HDFS in big data management. ICACNI Vol.2. pp 128.2015.
- [PRA13] K. Praisberger. Security of data. 2013.
- <https://www.itweek.ru/security/article/detail.php?ID=150345>
- [PAT16] D. Patgiri, Ripon.Dr. Hadoop: an infinite scalable metadata management for Hadoop.pp 15-31. 2016.
- [PER13] Pereñíguez-García, F.; Marín-López, R.; Kambourakis, G.; Ruiz-Martínez, A.; Gritzalis, S.; Skarmeta-Gómez, A.F. KAMU: providing advanced user privacy in Kerberos multi-domain scenarios. Int. J. Inf. Secur. 2013, 12, 505–525, doi:10.1007/s10207-013-0201-1.
- [QIU09] F., Qiu, J., Yang, J., et al., Hadoop high availability through metadata replication. Proc. 1st Int. Workshop on Cloud Data Management, p.37-44. 2009.
- [REY15] Jorge L. Reyes-Ortiz, Luca Oneto, Davide Anguita; Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf, Procedia Computer Science, Volume 53, 2015, Pages 121-130, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2015.07.286>.
- [RFC20] RFC4120: The Kerberos Network Authentication Service (V5) Available online: <https://tools.ietf.org/html/rfc4120> (accessed on Jul 6, 2020).
- [RHE20] Rhea, Sam; Johnson, E. Cloudflare: Public keys are not enough for SSH security Available online: <https://blog.cloudflare.com/public-keys-are-not-enough-for-ssh-security/> (accessed on Jun 15, 2020).
- [ROM13] N. Romodanov. Encrypt en Unix Systems.2013
- <http://rus-linux.net/MyLDP/sec/encrypt.html>
- [ROK11] J. Rokoty. NoSQL databases: a step to database scalability in web environment. Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, p. 278-283, ACM New York, NY, USA, 2011.
- [SAA20] saaspass Multi-factor authentication (mfa) with OpenID Connect protocol Available online: <https://blog.saaspass.com/multi-factor->

- authentication-mfa-with-openid-connect-protocol-d6b64c49c99c  
(accessed on Jul 6, 2020).
- [SAR10] Sarkar, P. A Simple and Generic Construction of Authenticated Encryption with Associated Data. *ACM Trans. Inf. Syst. Secur.* 2010, 13, 1–16, doi:10.1145/1880022.1880027.
- [SEC20] SEC 2: Recommended Elliptic Curve Domain Parameters. Available online: <http://www.secg.org/SEC2-Ver-1.0.pdf> (accessed on Jun 15, 2020).
- [SEC20b] Secsign: OAuth 2.0 Integration Available online: <https://www.secsign.com/developers/oauth-2-two-factor-authentication/> (accessed on Jul 6, 2020).
- [SHW18] T. Shwe, M. Aritsugi. *IEICE Transactions on Information and Systems*. pp. 2-3 -2018.
- [SIM13] G. Sims. *Encrypt en Unix Systems*.2013.  
  
<http://rus-linux.net/MyLDP/sec/encrypt.html>
- [SHL12] E. Shlik. *Big Data news*. 2012  
  
<http://www.iksmedia.ru/news/4659923-Big-Data-dlya-bolshix-dannyx.html> [SVI12] S. Svinarov. *PC Week/Russian Edition*.2012
- [SER14] N. Sergeev. *IT News.Big Data №1*. 2014.  
  
<https://itunews.itu.int/Ru/Note.aspx?Note=4879>
- [SMI14] E. Smirnov. *Insurance companies are embarking on the Big Data path*.2014.  
  
[https://www.cnews.ru/news/top/strahovye\\_kompanii\\_vstayut\\_na\\_put\\_big\\_data](https://www.cnews.ru/news/top/strahovye_kompanii_vstayut_na_put_big_data)
- [SEV15] N. Sevtinuk. *What is Big Data Analysis?* 2015.  
  
<https://lab-music.ru/chto-takoe-analiz-bolshih-dannyh-chto-takoe-big-data-harakteristiki-klassifikaciya-primery-analiz/>
- [SHA10] K. Shvachko, Haining Kuang, Sanyjy Radia, et al. *The Hadoop Distributed File System*. pp.1-10, 2010.  
<http://home.apache.org/~shv/Publications.html>

- [SHA10] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system.” in Proc. MSST, vol. 10, pp. 1–10, 2010.
- [SAS18] G. Sasubilli, U. Sekhar. Proceedings of the First International Conference on Information Technology and Knowledge Management pp. 132-133. 2018.
- [SHR04] V. Shramko. Combined Identification and Authentication Systems. 2004.  
<https://www.itweek.ru/infrastructure/article/detail.php?ID=69114>
- [SHV10] K. Shvachko, H. Kuang, S. Radia, R. Chansler. The Hadoop Distributed File System. MSST '10 Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1-10,2010.
- [SRI15] Srinivas, S.; Balfanz, D.; Tiffany, E.; Czeskis, A.; Alliance, F. Universal 2nd factor (U2F) overview. FIDO Alliance Proposed Standard 2015, 1–5.
- [STA11] Stajano, F. Pico: No More Passwords! In Security Protocols XIX; Christianson, B., Crispo, B., Malcolm, J., Stajano, F., Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; Vol. 7114, pp. 49–81 ISBN 978-3-642-25866-4.
- [STI06] Stinson, D.R. Some Observations on the Theory of Cryptographic Hash Functions. Des Codes Crypt 2006, 38, 259–277, doi:10.1007/s10623-005-6344-y.
- [SUL15] I. Sulkis. News en Big Data. 2015  
[http://www.tadviser.ru/index.php/\\_\(Big\\_Data\)](http://www.tadviser.ru/index.php/_(Big_Data))
- [SEK00] H. Seki, T. Kaneko: Differential Cryptanalysis of Reduced Rounds of GOST. In SAC 2000, Selected Areas in Cryptography, Douglas R. Stinson and Stafford E. Tavares, editors, LNCS 2012, pp. 315323, Springer, 2000.
- [SCH96] B. Schneier: Section 14.1 GOST, in Applied Cryptography, Second Edition, John Wiley and Sons. ISBN 0-471-11709-9, 1996.
- [SHO01] V. Shorin, V. Jelezniakov and Ernst M. Gabidulin: Linear and Differential Cryptanalysis of Russian GOST, Preprint submitted to Elsevier Preprint, 4 April 2001.
- [SHO00] V. Shorin, V. Jelezniakov, E.M. Gabidulin Security of algorithm GOST 28147- 89, (in Russian), 2000.

- [TBA17] Tbatou, Z.; Asimi, A.; Asimi, Y.; Sadqi, Y.; Guezzaz, A. A New Mutual Kerberos Authentication Protocol for Distributed Systems. *I. J. Network Security* 2017, 19, 889–898.
- [TIA17] M. Tian, T. Feng. Ordinal Optimization-Based Performance Model Estimation Method for HDFS. *IEEE Access* pp 1-2.2017
- [TSA08] Tsay, J.K. Formal Analysis of the Kerberos Authentication Protocol, University of Pennsylvania, 2008.
- [VIK13] S. Vikram Phaneendra & E. Madhusudhan Reddy “Big Data- solutions for RDBMS problems a survey” In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) Osaka, Japan, 2013.
- [VIR14] D. Virostkov. Solutions Architect of DataArt. 2014.  
<https://habr.com/ru/company/dataart/blog/262817/>
- [W3C20] W3C Rec., “Web Authentication: An API for Accessing Public Key Credentials Level 1” Available online:  
<https://www.w3.org/TR/webauthn/> (accessed on Jun 15, 2020).
- [WIK12] Wikipedia. GOST 34.11-2012.  
[https://ru.wikipedia.org/wiki/GOST\\_34.11-2012](https://ru.wikipedia.org/wiki/GOST_34.11-2012)
- [WIK89] Wikipedia. GOST 28147-89. [https://ru.wikipedia.org/wiki/GOST\\_28147-89](https://ru.wikipedia.org/wiki/GOST_28147-89)
- [WIK01] Wikipedia. GOST 34.10-2001.  
<http://dic.academic.ru/dic.nsf/ruwiki/261586>
- [WEI18] J. Weipeng, T. Danyu. *Computer Science and Information Systems* pp.1-2 -2018.
- [WHI12] T. White: *The Definitive Guide*. M.: O'Reilly Media / Yahoo Press, pp 628 c. 2012.
- [WHI11] T. White. *Hadoop: the definitive guide*. Beijing: Tsinghua University Press, 2011.
- [WHI09] T. White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc. 2009.
- [WSA14] Web site Apache. Token based authentication and Single Sign On. 2014.

<https://issues.apache.org/jira/browse/HADOOP-9392>

- [YAN16] S. Yanni. Application and Realization of Improved Apriori Algorithm in Hadoop Simulation Platform for Mass Data Process. *Int. J. Online Eng.* pp. 16 - 19. 2016
- [YAR15] N. Yaruson. GOST 28147-89. 2015.  
<https://habr.com/ru/post/256843/>
- [YIL15] Yildirim, N.; Varol, A. Android based mobile application development for web login authentication using fingerprint recognition feature. In *Proceedings of the 2015 23rd Signal Processing and Communications Applications Conference (SIU)*; IEEE: Malatya, Turkey, 2015; pp. 2662–2665.
- [YLO19] Ylonen, T. SSH Key Management Challenges and Requirements. In *Proceedings of the 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*; IEEE: CANARY ISLANDS, Spain, 2019; pp. 1–5.
- [YUB20] Yubico: Protect your digital world with YubiKey Available online: <https://www.yubico.com/> (accessed on Jun 15, 2020).
- [YUB20b] Yubico: Security advisory 2019-06-13 Available online: <https://www.yubico.com/support/security-advisories/ysa-2019-02/> (accessed on Jun 15, 2020).
- [ZAN15] Zan Mo, Yanfei Li Research of Big Data Based on the Views of Technology and Application *American Journal of Industrial and Business Management*, pp. 192-197 Published Online April 2015.
- [ZAB89] I. Zabotin, G. Glazkov, V. B. Isaeva: Cryptographic Protection for Information Processing Systems, Government Standard of the USSR, GOST 28147-89, Government Committee of the USSR for Standards, 1989.



## Acrónimos.

<b>ACL</b>	Access Control List
<b>AD</b>	Active Directory
<b>AENOR</b>	La Asociación Española de Normalización y Certificación
<b>AES</b>	Advanced Encryption Standard
<b>API</b>	Application Programming Interface
<b>AS</b>	Authentication Server
<b>CA</b>	Certification Authority
<b>CFS</b>	Cluster File Server
<b>CRS</b>	Client Request Service
<b>CS</b>	Configuration Server
<b>CVE</b>	Common Vulnerabilities and Exposures
<b>DFS</b>	Distributed File System
<b>DLL</b>	Dynamic Link Library
<b>DMZ</b>	Demilitarized Zone
<b>DoS</b>	Denial of Service
<b>DSA</b>	Digital Signature Algorithm
<b>ECC</b>	Elliptic curve cryptography
<b>ECDH</b>	Elliptic Curve Diffie-Hellman
<b>ECDSA</b>	Elliptic Curve Digital Signature Algorithm
<b>ECIES</b>	Elliptic Curve Integrated Encryption Scheme
<b>EMC</b>	European Metallurgical Conference
<b>ESP_IDF</b>	Espressif IoT Development Framework
<b>FIDO</b>	Fast Identity Online Alliance
<b>FIPS</b>	Federal Information Processing Standard
<b>GC</b>	Gateway Client
<b>GFS</b>	Google File System
<b>GOST</b>	Estándar del estado
<b>GS</b>	Gateway Server
<b>HD</b>	High Definition
<b>HDFS</b>	Hadoop File System
<b>HPC</b>	High performance Computing
<b>HTTP</b>	Hyper Text Transfer Protocol
<b>IBM</b>	International Business Machines
<b>IP</b>	Internet Protocol Address
<b>IT</b>	Information Technology
<b>JSON</b>	JavaScript Object Notation
<b>JWT</b>	JSON Web Token
<b>LDAP</b>	Lightweight Directory Access Protocol
<b>MAC</b>	Message Authentication Code
<b>MB</b>	MQTT Broker
<b>MD</b>	Merkla Damgarda
<b>MitM</b>	Man-in-the-Middle
<b>MPI</b>	Message Passing Interface
<b>MQTT</b>	Message Queuing Telemetry Transport
<b>MSC</b>	Message Sequence Chart

<b>NASSCOM</b>	India's National Association of Software and Services Companies
<b>NDFS</b>	Nutch Distributed File System
<b>NFC</b>	Near Field Communication
<b>NLP</b>	Natural language processing
<b>NVS</b>	Non-Volatile Storage
<b>OEIN</b>	Organización Europea para la Investigación Nuclear
<b>OTP</b>	One Time Password
<b>PAM</b>	Linux Pluggable Authentication Module
<b>PEM</b>	Privacy Enhanced Mail
<b>PTR</b>	Pointer
<b>QoS</b>	Quality of service
<b>QR</b>	Quick Response Code
<b>RAID</b>	Redundant Array of Independent Disks
<b>RBAC</b>	Role-based Access Control
<b>RPC</b>	Remote Procedure Call
<b>RSA</b>	Rivest, Shamir & Adleman
<b>RST</b>	Representational State Transfer
<b>SAML</b>	Security Assertion Markup Language
<b>SAP</b>	Systeme Anwendungen und Produkte
<b>SASL</b>	Simple Authentication and Security Layer
<b>SDK</b>	Software Development Kit
<b>SGI</b>	Silicon Graphics Inc
<b>SIM</b>	Subscriber Identity Module
<b>SMS</b>	Short Message Service
<b>SNE</b>	Servicio Nacional de Estadística
<b>SOAP</b>	Simple Object Access Protocol
<b>SQL</b>	Structured Query Language
<b>SSH</b>	Secure Shell
<b>SSO</b>	Single Sign-On
<b>SWT</b>	Simple Web Token
<b>TBC</b>	Tweakable Block Cyphers
<b>TCP</b>	Transmission Control Protocol
<b>TI</b>	Texas Instruments
<b>TLS</b>	Transport Layer Security
<b>TMS</b>	Token Management System
<b>TPM</b>	Trusted Platform Module
<b>TRNG</b>	True Random Number Generators
<b>U2F</b>	Universal Second Factor protocol
<b>UDP</b>	User Datagram Protocol
<b>UFC</b>	Universal Authentication Framework
<b>URL</b>	Uniform Resource Locator
<b>UUID</b>	Universally Unique Identifier
<b>VLAN</b>	Virtual Local Area Network
<b>W3C</b>	World Wide Web Consortium
<b>XML</b>	eXtensible Markup Language
<b>XTS</b>	XEX Tweakable Block Cipher with Ciphertext Stealing
<b>YML</b>	YAML Ain't Markup Language

