# Un estudio de las amenazas de seguridad para Big Data y Hadoop

**Sandra D. ORANTES JIMÉNEZ**
**Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC-IPN), Departamento de Investigación en Ciencias de la Computación, CDMX, C.P. 07738, México**

**Graciela VÁZQUEZ ÁLVAREZ**
**Instituto Politécnico Nacional, SEPI-Escuela Superior de Ingeniería Mecánica y Eléctrica (SEPI-ESIME-ZACATENCO-IPN), CDMX, C.P. 07738, México**

**María E. AGUILAR JÁUREGUI**
**Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC-IPN), Departamento de Investigación en Ingeniería de Cómputo, CDMX, C.P. 07738, México**

**RESUMEN**

Una noticia que resulta ser frecuente en estos días, es acerca de violaciones de seguridad a servicios de Big Data, que como consecuencia deriva en la exposición de enormes cantidades de datos privados. Por ello, con esta investigación se estudian las amenazas de seguridad actuales para Big Data y Hadoop. Big Data es almacenado y analizado por servicios de la nube, en consecuencia, este trabajo, también puede ubicarse en los problemas de seguridad en la computación en la nube que están asociados con Big Data. Las aplicaciones de Big Data proporcionan un valioso beneficio para las organizaciones, empresas y muchas industrias de grande y pequeña escala. En este trabajo, debido a que Hadoop se ha empleado de manera prominente para manejar una gran cantidad de información sobre su estructura de programación dispersa, también se examinan varias soluciones posibles para los problemas de seguridad en Hadoop. Por otra parte, la seguridad informática en la nube se está incrementando incluyendo seguridad de red, seguridad de la información y privacidad de datos y en consecuencia desempeña un papel vital en la protección de datos, aplicaciones e infraestructura relacionada, apoyándose en políticas, tecnologías, controles y herramientas de Big Data. Es por ello, que el estudio presentado en este documento resulta de interés, ya que proporciona una revisión de la literatura sobre amenazas de seguridad y violaciones de privacidad de Big Data y procesamiento simultáneo de Hadoop y MapReduce, debido a que HDFS y MapReduce, son la base de un Hadoop robusto y permiten trabajar de forma paralela con grandes datos, ofreciendo una solución integrada y fácilmente escalable.

**Palabras Claves:** Big Data, MapReduce, Hadoop, seguridad, computación en la nube, HDFS.

## 1. INTRODUCCIÓN

Pero, ¿qué es exactamente "Big Data"?. En palabras sencillas, se puede describir como enormes volúmenes de datos que pueden ser tanto estructurados como no estructurados. En general, es tan gigantesco que ofrece un desafío para procesar el uso de técnicas de base de datos y software convencionales.

En escenarios empresariales, se puede inferir que:
- Los datos son sorprendentes en términos de volúmenes.
- Cambian rápidamente.
- Superan la capacidad de almacenaje de su información.

Existen estudios que han demostrado que para 2020 el mundo habrá aumentado 50 veces la cantidad de datos que tenía en 2011, que era de 1.8 zettabytes o 1.8 billones de gigabytes de datos [8]. Para satisfacer las demandas de la industria, se están gestionando nuevos manifiestos de manipulación de Big Data.

Por otra parte, Big Data no tiene que ver solamente con el Volumen de los datos, sino que también incluye la Variedad y la Velocidad. De tal manera, que juntos estos tres atributos forman las tres “V” de Big Data. No obstante, también pueden reconocerse otras dos dimensiones el Valor y la Veracidad, formando entonces las 5 “V” de Big Data [7] (véase Figura 1).

Cada una de las “V” representadas en la Figura 1 se describen a continuación:
El Volumen hace referencia al tamaño de los datos, que pueden provenir de múltiples fuentes y que actualmente, es una cantidad mayor que terabytes y petabytes. Los datos provienen de máquinas, redes e interacción humana en sistemas como los medios sociales, donde el volumen de datos a analizar es muy grande. [10] La Velocidad define la rapidez del procesamiento de datos, se requiere no solo para Big Data, sino también para todos los procesos e implica el procesamiento en tiempo real y el procesamiento por lotes. La Variedad se refiere a que la información proviene de diferentes tipos de datos de diversas o muchas fuentes, tanto estructuradas como no estructuradas. En el pasado, los datos se almacenaban de fuentes como hojas de cálculo y bases de datos, ahora, vienen en forma de correos electrónicos, imágenes, audio, videos, dispositivos de monitoreo, archivos PDF, etc. Esta diversidad de datos no estructurados crea problemas para el almacenamiento, la extracción y el análisis de los datos. [10] La Veracidad corresponde al sesgo, el ruido y la alteración de datos. Los responsables del proyecto Big Data han de preguntarse honestamente si los datos que se almacenan y extraen son

directamente relacionados y significativos al problema que se trata de analizar. Esta característica puede ser el mayor reto cuando se comparan con otras como el volumen o la velocidad. Cuando se valore el alcance en una estrategia de Big Data es necesario contar en el equipo con participantes imparciales que ayuden a mantener los datos limpios y asegurarse que los procesos no acumulen "datos sucios" en sus sistemas.


**Figura 1**. Las 5 "V" de Big Data [Elaboración propia].

Además, el Valor en Big Data es de gran importancia ya que es posible extraer información significativa que puede ayudar a tomar mejores decisiones y a realizar movimientos estratégicos de negocios ya que una vez recuperada puede visualizarse a través de gráficas estadísticas, se pueden hacer correlaciones, identificar eventos, establecer hipótesis como ¿son los datos correctos y precisos para el uso previsto? Este aspecto es clave para poder tomar las decisiones correctas.

Por otra parte, para procesar el gran volumen de datos de diferentes fuentes, para un procesamiento rápido se usa Hadoop; el cual, es un marco de programación gratuito basado en Java que admite el procesamiento de grandes conjuntos de datos en un entorno informático distribuido. Hadoop permite ejecutar aplicaciones en sistemas con miles de nodos con miles de terabytes de datos [14]. Su sistema de archivos distribuidos admite velocidades de transferencia de datos rápidas entre nodos y permite que el sistema continúe operando ininterrumpidamente en los momentos donde falla el nodo.

Hadoop consta de sistemas de archivos distribuidos, almacenamiento de datos y plataformas de análisis y una capa que maneja el cómputo paralelo, la tasa de flujo (flujo de trabajo) y la administración de la configuración [10]. HDFS (Hadoop Distributed File System, Sistema de Archivos Distribuido de Hadoop) se ejecuta en los nodos de un clúster de Hadoop y juntos conectan los sistemas de archivos en muchos nodos de datos de entrada y salida, para convertirlos en un gran sistema de archivos [14]. La arquitectura actual de Hadoop (como se muestra en la Figura 2) consta del núcleo de Hadoop, Map- Reduce, HDFS y varios componentes relacionados, como Apache Hive, HBase, Oozie, Pig y Zookeeper, estos componentes se explican a continuación [8, 9]:

- HDFS: un sistema de archivos distribuido altamente tolerante a fallos que es responsable de almacenar datos en los clústeres.
- MapReduce: Es una técnica de programación paralela para el procesamiento distribuido de una gran cantidad de clústeres de datos.
- HBase: una base de datos NoSQL distribuida orientada a columnas para el acceso aleatorio de lectura/escritura.
- Pig: un lenguaje de programación de datos de alto nivel para analizar datos de cómputo de Hadoop.
- Hive: una aplicación de almacenamiento de datos que proporciona un modelo relacional y de acceso de tipo SQL.
- Sqoop: un proyecto para transferir/importar datos entre bases de datos relacionales y Hadoop.
- Oozie: una gestión de orquestación y flujo de trabajo para trabajos dependientes de Hadoop.

En este documento, se describe la seguridad tradicional de Big Data Hadoop y también se analiza la debilidad de la misma, es decir, las amenazas a la seguridad, para poder exponer posteriormente una posible solución de seguridad para cada uno de los componentes de Hadoop y también, proporcionar un estudio de las tecnologías de seguridad utilizadas para proteger Hadoop. Finalizando, este trabajo con algunas conclusiones.

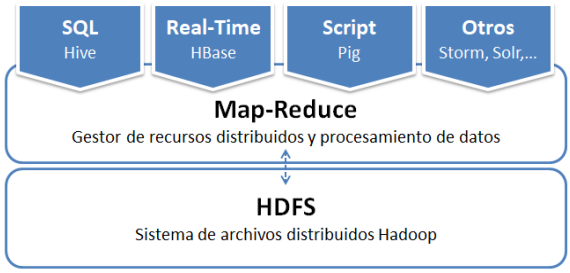


**Figura 2**. Arquitectura de Hadoop [14].

## 2. SEGURIDAD TRADICIONAL DE BIG DATA HADOOP

**Visión general de la seguridad de Hadoop**

Originalmente, Hadoop se desarrolló sin tener en cuenta la seguridad, ni el modelo de seguridad, ni la autenticación de usuarios y servicios, ni la privacidad de los datos, por lo que cualquier persona podía enviar un código arbitrario para ser ejecutado. Aunque los controles de auditoría y autorización (permisos de archivos HDFS y ACL) se usaron en distribuciones anteriores, dicho control de acceso se evadió fácilmente porque cualquier usuario podría hacerse pasar por otro. Debido a que la suplantación era frecuente la mayoría de los usuarios lo hacían y las medidas de control de seguridad que subsistían, no eran muy efectivas.

Posteriormente se agregó la autorización y la autenticación, pero con alguna debilidad; debido a que había muy pocas medidas de control de seguridad dentro del ecosistema de Hadoop, ocurrieron muchos incidentes fortuitos y de seguridad en tales entornos. Los usuarios, incluso aquellos sin malas intenciones, pueden cometer errores (por ejemplo, borrar cantidades masivas de datos en segundos con una eliminación distribuida). Otro problema de seguridad, descubierto en ese momento es que todos los usuarios y programadores tenían el mismo nivel de privilegios de acceso a todos los datos del clúster, cualquier trabajo podía acceder a cualquiera de los

datos del clúster y cualquier usuario podía leer cualquier conjunto de datos [4]. Debido a que MapReduce no tenía un concepto de autenticación o autorización, un usuario malintencionado podría reducir las prioridades de otros trabajos de Hadoop para que su trabajo se completara más rápido o para ser ejecutado primero, o peor, podía eliminar a los otros trabajos. Hadoop es una arquitectura completa de aplicaciones que involucra a Hive, HBase, Zookeeper, Oozie y Job Tracker. Cada una de estas aplicaciones requiere endurecimiento. Para agregar potencialidades o capacidades de seguridad en un entorno de Big Data, las funciones deben escalarse con los datos. La seguridad suplementaria no se escala bien y simplemente no puede mantenerse al día. [15] La comunidad de Hadoop admite algunas características de seguridad a través de la implementación actual de Kerberos, el uso de cortafuegos y los permisos y ACL básicos de HDFS [13]. Kerberos no es un requisito obligatorio para un clúster de Hadoop, lo que hace posible ejecutar clústeres completos sin implementar ninguna seguridad. Kerberos tampoco es muy fácil de instalar y configurar en el clúster e integrarse con los servicios AD (Active Directory, Directorio Activo) y LDAP (Lightweight Directory Access Protocol, Protocolo ligero de acceso a directorios). [15] Esto dificulta la implementación y limita la adopción de las funciones de seguridad más básicas para los usuarios de Hadoop. La seguridad de Hadoop no es tratada adecuadamente por los *firewalls*, una vez que éste está dañado; entonces, el clúster, está abierto para el ataque. Los *firewalls* no ofrecen protección para los datos en reposo o en movimiento dentro del clúster y tampoco ofrecen protección contra fallas de seguridad que se originan dentro de su perímetro [15].

Un atacante que puede ingresar al centro de datos, ya sea física o electrónicamente, puede robar los datos que desee, ya que los datos no están cifrados y no se tiene una autenticación para el acceso [6, 11].

**Amenazas de seguridad**

Las categorías comunes que se han identificado de violación de seguridad son:

- Divulgación no autorizada de información
- Modificación no autorizada de información y
- Denegación de recursos.

Las siguientes son las áreas relacionadas de amenaza que identificar en Hadoop [3]:

- Un usuario no autorizado puede acceder a un archivo HDFS a través de RPC (Remote Procedure Call, Llamada a procedimiento remoto) o mediante protocolos HTTP y podría Ejecutar código arbitrario o realizar nuevos ataques.
- Un cliente no autorizado puede leer y/o escribir en el bloque de datos de un archivo de un protocolo de transferencia de datos.
- Un cliente no autorizado puede obtener privilegios de acceso y puede enviar un trabajo a una cola o eliminar o cambiar la prioridad del trabajo.
- Un usuario no autorizado puede acceder a datos intermedios del trabajo a través de sus rastreadores de protocolo aleatorio de tareas HTTP.
- Una tarea en ejecución puede utilizar las interfaces del sistema operativo huésped, para acceder a otras tareas o a datos locales que incluyen el mapa intermedio de salida para el almacenamiento local del nodo de datos que se ejecuta en el mismo nodo físico.
- Un usuario no autorizado puede escuchar a escondidas y/o rastrear datos o enviar paquetes por los nodos de datos al cliente.
- Una tarea o nodo puede hacerse pasar por un componente de servicio Hadoop, tal como un nodo de datos, nombre de un nodo, rastreador de trabajos, rastreador de tareas, etc.
- Un usuario puede enviar un flujo de trabajo a Oozie como otro usuario.
- Los nodos de datos no impusieron el control de acceso, un usuario no autorizado podría leer bloques de datos arbitrarios desde nodos de datos, sin pasar por el control de acceso, es decir el mecanismo y/o restricciones o escribir datos basura a nodos de datos. [11].

## 3. PROBLEMAS DE SEGURIDAD

Hadoop presenta un conjunto único de problemas de seguridad para Gestores de centros de datos y profesionales de la seguridad. Los problemas de seguridad se describen a continuación [5, 6]:

- Datos fragmentados: los clústeres de Big Data contienen datos que retratan la calidad de la fluidez de la información, permitiendo múltiples copias moviendo de un lado a otro de varios nodos, asegurando redundancia y resiliencia. Los datos están disponibles para la fragmentación y se puede compartir en múltiples servidores; como resultado, se agrega más complejidad como consecuencia de la fragmentación, que plantea un problema de seguridad debido a la ausencia de un modelo de seguridad.
- Computación distribuida: la disponibilidad de recursos lleva al procesamiento virtual de datos en cualquier instante o instancia donde esté disponible y esto progresa. A grandes niveles de computación paralela, como resultado, se crean ambientes complicados que están a la altura de riesgos, como ataques que sufren a su vez sus contrapartes de repositorios y que son gestionadas centralmente y monolíticamente, permitiendo implicaciones de seguridad fáciles.
- Control de acceso a datos: los entornos proporcionan acceso a nivel de esquema, sin mayor granularidad; ya que muchos de los esquemas de seguridad de bases de datos disponibles, proporcionan acceso basado en roles. Es suficiente accesar, crear escenarios de acceso relacionados.
- Comunicación de nodo a nodo: una preocupación con Hadoop es que no implementan una comunicación segura; ya que, por ejemplo, ponen en uso el RPC (Remote Procedure Call, Llamada a procedimiento remoto) sobre TCP/IP.
- Interacción con el cliente: la comunicación del cliente tiene lugar con gestores de recursos y nodos de datos. Sin embargo, a pesar de que la comunicación eficiente es facilitada por este modelo, resulta engorroso proteger los nodos de los clientes y viceversa y también el nombre de los servidores desde los nodos. Así, clientes que han sido comprometidos tienden a propagar datos maliciosos o enlaces a cualquiera de los dos servicios.
- Prácticamente no hay seguridad: las pilas de datos grandes se diseñaron con poca o ninguna seguridad en mente cuando predominan grandes volúmenes de datos. Las instalaciones están construidas en el modelo de servicios web, con poca o ninguna facilidad para prevenir amenazas comunes en la web, por lo que es altamente susceptible.

## 4. SOLUCIONES DE SEGURIDAD HADOOP

Hadoop es un sistema distribuido que permite almacenar grandes cantidades de datos y procesar datos en paralelo y se utiliza, como un servicio multiusuario y de almacenamiento de datos confidenciales, como información de identificación personal. Otras organizaciones, incluidas las financieras, utilizan Hadoop y están empezando a almacenar datos sensibles sobre clústeres de Hadoop. Como resultado, es sumamente necesaria la autenticación y autorización. [3]

La arquitectura Hadoop consta de varios componentes que necesitamos asegurarse. En esta parte del documento, se intenta examinar cada uno de los componentes y la solución de seguridad propuesta.

Cada uno de estos componentes Hadoop, tienen sus propios desafíos de seguridad, problemas y necesidades que deben configurarse adecuadamente basándose en su arquitectura para asegurarlos y tienen usuarios finales, que usan directamente a un servicio *backend* accediendo a componentes principales (HDFS y MapReduce).

Se ha realizado un análisis de seguridad de componentes Hadoop y un breve estudio de la seguridad incorporada en su arquitectura, observándose que la seguridad de Hadoop no es muy fuerte, por lo que en este documento se intenta proponer una solución de seguridad alrededor de la Autenticación, Autorización, Encriptación y Auditorías (AAEA), para cada uno de los componentes del ecosistema, para ayudar a asegurar el clúster de Hadoop.

**Autenticación**
La autenticación es verificar la identidad del usuario o del sistema al acceder a él. Hadoop proporciona Kerberos como autenticación primaria, inicialmente se utilizó SASL (Simple Authentication and Security Layer, Autenticación simple y capa de seguridad) que soporta los mecanismos GSS-API (Generic Security Service Application Programming Interface, Interfaz de programación de aplicaciones de servicio de seguridad genérica) para implementar Kerberos y autenticar mutuamente a los usuarios, sus aplicaciones y los servicios de Hadoop a través de conexiones RPC [3]. Hadoop también soporta autenticación "Pluggable" para las consolas web HTTP, lo que significa que los implementadores de aplicaciones y consolas web podrían hacer su propio mecanismo de autenticación para HTTP. Esto incluye pero no está limitado a la autenticación HTTP SPNEGO (Simple and Protected GSS-API Negotiation, Negociación GSS-API simple y protegida). Los componentes de Hadoop son compatibles con SASL Framework, es decir, la capa RPC se puede cambiar a soportar la autenticación mutua basada en SASL. Autenticación SASL Digest-MD5 o SASL GSSAPI/ autenticación Kerberos.

MapReduce admite la autenticación Kerberos, la SASL Digest MD-5 y también *token* de delegación de autenticación en conexiones RPC. En HDFS Las comunicaciones entre el nombre del nodo y los nodos de datos son sobre conexión RPC y autenticación Kerberos, realizado mutuamente entre ellos [9]. HBase soporta autenticación SASL segura del cliente Kerberos a través de RPC, HTTP. Hive soporta autenticación Kerberos y LDAP (Lightweight Directory Access ProtocoL, Directorio ligero de acceso a directorios) para el usuario a través de Apache Knox.

Pig usa las credenciales de usuario para enviar el trabajo a Hadoop, así que no hay necesidad de ninguna seguridad adicional de Kerberos, pero antes de iniciar Pig el usuario debe autenticarse con KDC (Key Distribution Center, Centro de distribución de llaves) y obtener un acceso a Kerberos válido [9]. Oozie proporciona autenticación de usuario a sus servicios web. También proporciona Kerberos HTTP Simple y un mecanismo de negociación GSSAPI protegido (SPNEGO) para autenticación de clientes web. Se utiliza el protocolo SPNEGO cuando una aplicación cliente quiere autenticarse en un servidor remoto, pero no está seguro de los protocolos de autenticación a utilizar.
Zookeeper admite la autenticación Kerberos SASL en conexiones RPC. Hue ofrece autenticación SPNEGO. La autenticación LDAP, ahora también es compatible con SSO SAML autenticación [9].
Hay una serie de flujos de datos involucrados en autenticación Hadoop - mecanismo de autenticación Kerberos RPC, que se utiliza para la legitimación de usuarios, aplicaciones y servicios.

La autenticación HTTP SPNEGO se utiliza para consolas web y usa tokens de delegación [11]. Un token de delegación es un protocolo de autenticación de dos partes utilizado entre el usuario y el nombre del nodo para autenticar usuarios, es simple y más eficaz que el protocolo de tres partes utilizado por Kerberos [8, 9]. Oozie y HDFS, soporta el token de delegación MapReduce.

**Autorización**
La autorización es un proceso de especificación de privilegios de control de acceso para el usuario o el sistema. En Hadoop, los controles de acceso se implementan mediante el uso de permisos basados en archivos que siguen el modelo de permisos UNIX. El control de acceso a los archivos en HDFS podría ser aplicado por el nombre del nodo (NameNode) con base en los permisos de archivos y las ACL (Access Control List, Lista de control de acceso) de usuarios y grupos. MapReduce proporciona ACL para colas de trabajos, que definen qué usuarios o grupos pueden enviar trabajos a una cola y cambiar las propiedades de la cola. Hadoop ofrece una autorización detallada utilizando permisos de archivo en HDFS y control de acceso a nivel de recursos utilizando ACL para MapReduce y un control de acceso de grano más grueso a nivel de servicio [5, 12]. HBase ofrece autorización de usuario en tablas y familias de columnas. La autorización del usuario se implementa utilizando coprocesadores.

Los coprocesadores son como activadores de base de datos en HBase [9], interceptan cualquier solicitud a la tabla antes y después, para después emplear el proyecto Apache Rhino para ampliar el soporte de HBase para las ACL de nivel de celda. En Hive, la autorización se implementa utilizando Apache Sentry. Pig proporciona la autorización mediante ACL para las colas de trabajos; Zookeeper también ofrece autorización utilizando el nodo ACL. Hue proporciona control de acceso a través del permiso del sistema de archivos y también ofrece ACL para la cola de trabajos.

Aunque Hadoop puede configurarse para realizar el control de acceso a través de los permisos de usuario y grupo y las listas de control de acceso (ACL), esto puede no ser suficiente para todas las organizaciones.

Actualmente, muchas organizaciones utilizan políticas de control de acceso dinámicas y flexibles basadas en XACML (eXtensible Access Control Markup Language, Lenguaje de marcado de control de acceso extensible) y ABAC (Attribute Based Access Control, Control de acceso basado en atributos) [5, 11, 12]. Hadoop ahora puede configurarse para admitir RBAC (Role-Based Access Control, Control de acceso basado en roles) y control de acceso ABAC usando algún marco o herramienta de terceros.

Algunos de los componentes de Hadoop, como HDFS, pueden ofrecer ABAC con Apache Knox y también Hive puede soportar el control de acceso basado en roles con Apache Sentry.

Zettaset Orchestration, un producto de Zettaset, proporciona soporte de control de acceso basado en roles y permite que Kerberos se integre a la perfección en la arquitectura de Hadoop. [6, 9].

**Encriptación**
El cifrado o encriptación, garantiza la confidencialidad y privacidad de la información del usuario y protege los datos confidenciales en Hadoop; el cual, es un sistema distribuido que se ejecuta en máquinas distintas, lo que significa que los datos deben transmitirse a través de la red de manera regular y existe una creciente necesidad de demanda, para mover información sensible a la arquitectura Hadoop que necesita salvaguardarse. Los datos confidenciales dentro del clúster requieren un tipo especial de protección, tanto en reposo como en movimiento [11]. Estos datos deben protegerse durante la transferencia hacia y desde el sistema Hadoop. El marco de autenticación simple y la capa SASL (Simple Authentication and Security Layer, Autenticación simple y capa de seguridad) se utiliza para cifrar los datos en movimiento en el ecosistema Hadoop. La seguridad de SASL garantiza que los datos se intercambian entre el cliente y los servidores y se asegura de que no puedan leerse por un atacante MITM (Man-In-The-Middle). SASL admite varios mecanismos de autenticación, por ejemplo, DIGEST-MD5, CRAM-MD5, etc.

Por otra parte, los datos en reposo se pueden proteger de dos maneras: en primer lugar, cuando el archivo se almacena en Hadoop, el archivo completo se puede cifrar primero y después almacenarse. En este enfoque, los bloques de datos en cada nodo de datos (DataNode) no se pueden descifrar hasta que vuelvan a colocarse todos los bloques y se cree el archivo cifrado completo. En segundo lugar, debe aplicarse el cifrado a los bloques de datos una vez que se cargan en el sistema Hadoop [9].

Hadoop admite la capacidad de cifrado para varios canales, como RPC, HTTP y el Protocolo de transferencia de datos para datos en movimiento. Se han introducido el marco de trabajo de *Hadoop Crypto Codec* y *Crypto Codec Implementation* para respaldar el cifrado de datos en reposo. HDFS es compatible con AES (Advanced Encryption Standard, Estándar de cifrado avanzado), encriptación a nivel de Sistema Operativo para datos en reposo. *Zookeeper, Oozie, Hive, HBase, Pig, Hue* no ofrecen una solución de encriptación de datos en reposo, pero para este componente, el encriptado se puede implementar a través de técnicas de cifrado personalizadas o herramientas de terceros como el *zNcryptor* de *Gazzang* utilizando el marco de códec criptográfico. La seguridad a nivel de sistema de archivos y el cifrado y descifrado de archivos se pueden realizar sobre la marcha utilizando *eCryptfs* y las herramientas *zNcrypt* de *Gazzang*, que son soluciones de seguridad comercial disponibles para los clústeres de Hadoop [9, 11, 14].

Para proteger los datos en tránsito y en reposo, se pueden implementar técnicas de encriptado y enmascaramiento. Herramientas como *Optim* y *Dataguise* de IBM proporcionan enmascaramiento de datos para datos empresariales [9]. Así también, Intel ofrece cifrado y compresión de archivos [9]. *Project Apache Rhino* permite el cifrado a nivel de bloque similar a *Dataguiseand Gazzang*. [13]

**Auditorías**
El clúster de Hadoop aloja información confidencial y la seguridad de estos datos, es sumamente importante para que las organizaciones tengan un trayecto exitoso y seguro de Big Data, ya que siempre existe la posibilidad de que ocurran violaciones de seguridad por acceso no autorizado o acceso inadecuado por parte de usuarios privilegiados. [12] Por lo tanto, para cumplir con los requisitos de cumplimiento de seguridad, debe auditarse todo la arquitectura de Hadoop periódicamente e implementar un sistema que realice el monitoreo de registros.

HDFS y MapReduce proporcionan soporte de auditoría. Metastore de Apache Hive mantiene información de auditoría (quién/cuándo) para las interacciones de Hive [9, 12]. Apache Oozie, el motor de flujo de trabajo, proporciona un seguimiento de auditoría para los servicios, donde el envío del flujo de trabajo se mantiene en los archivos de registro de Oozie. Hue también admite registros de auditoría. Para los componentes de Hadoop que no proporcionan el registro de auditoría, es posible emplear las herramientas de monitoreo de registros. *Scribe* y *LogStash* son herramientas de código abierto que se integran en la mayoría de los entornos de Big Data, como lo hacen muchos productos comerciales. Por lo tanto, solo hay que encontrar una herramienta compatible, instalarla, integrarla con otros sistemas como la administración de registros y luego revisar los resultados y lo que podría haber fallado. *Cloudera Navigator* de Cloudera es una popular herramienta comercial que proporciona registros de auditoría para entornos Big Data. Zettaset Orchestration proporciona administración de configuración centralizada, registro y soporte de auditoría. [9, 15].

## 5. TECNOLOGÍAS DE SEGURIDAD: SOLUCIÓN PARA LA SEGURIDAD HADOOP

A continuación, se proporciona una descripción general de las diversas tecnologías comerciales y de código abierto que están disponibles para abordar los diferentes aspectos de seguridad de Big Data Hadoop [9].

**Apache Sentry**
Es un proyecto de código abierto de Cloudera, es un módulo de autorización para Hive, Search, Impala etc. Desbloquea los requisitos clave de RBAC (role-based access control, control de acceso basado en roles): autorización, granularidad, autorización basada en roles, administración multiusuario. Es de código abierto y es un proyecto de incubadora de Apache. Proporciona soporte para ecosistemas: Apache SOLR, HiveServer2 e Impala 1.1+. Ofrece beneficios clave como: Almacenar datos sensibles en Hadoop, extiende Hadoop a más usuarios y cumple con las regulaciones [2, 16].

**Apache Rhino**
Proporciona una solución integrada de seguridad de datos de extremo a extremo para el ecosistema de Hadoop. Proporciona una autenticación basada en token y una solución de SSO (Single Sign-On, Autenticación Única). Ofrece el marco de códec criptográfico Hadoop y la implementación del códec criptográfico para proporcionar cifrado a nivel de bloque para los datos almacenados en Hadoop. Admite la distribución y administración de claves para que MapReduce pueda descifrar el bloque de datos y ejecutar el programa según los requisitos. También mejora la seguridad de HBase al ofrecer autenticación a nivel de celda y cifrado transparente para la tabla almacenada en Hadoop. Es compatible con el marco de registro de auditoría para facilitar las marcas de auditoría. [16]

Es un proyecto de código abierto para mejorar la plataforma Hadoop con mecanismos de protección adicionales y tratando de eliminar los posibles agujeros de seguridad en la pila de Hadoop. Entre la variedad de trabajos realizados en el marco de este proyecto resulta interesante destacar una capa de abstracción común para fijar con claridad, exactitud y precisión una API criptográfica, así como la definición de un entorno adecuado para la distribución y gestión de claves. [1]

**Apache Knox Gateway**
Es un sistema que proporciona un único punto de autenticación y acceso para varios servicios de Hadoop en un clúster. Proporciona una solución de seguridad perimetral para Hadoop. La segunda ventaja es que admite varios escenarios de autenticación y verificación de token. Administra la seguridad en varios clústeres y versiones de Hadoop. También proporciona soluciones de SSO y permite la integración de otras soluciones de gestión de identidad, como LDAP (Lightweight Directory Access Protocol, Protocolo Ligero/Simplificado de Acceso a Directorios), AD (Active Directory, Directorio Activo) y SSO basado en SAML (Security Assertion Markup Language, Lenguaje de marcado para confirmaciones de seguridad) y otros sistemas de SSO [1, 6].

## 6. RETOS DE LA SEGURIDAD EN HADOOP PARA LAS EMPRESAS ACTUALES

Como ya se ha mencionado, la seguridad no era parte del desarrollo inicial de Hadoop y no tenía autenticación de usuarios o servicios y en 2009, Yahoo se enfocó en agregársela; pero Hadoop, todavía tenía capacidades de autorización limitadas.

En 2013, Apache Software Foundation lanzó Project Rhino para agregar funciones de seguridad a Hadoop.

Para muchas organizaciones, Hadoop se ha convertido en una plataforma de datos empresariales que almacena información confidencial y ha habilitado, la administración de datos que es masivamente escalable, ágil, rica en características y rentable. Pero a medida que los datos que alguna vez estuvieron en los repositorios se reúnen en un vasto depósito de datos y se ponen a disposición de una variedad de usuarios en toda la organización, surgen nuevos desafíos de seguridad [4].

Abordar estos desafíos debe garantizar que:

- Los usuarios que acceden a Hadoop están debidamente autenticados.
- Los usuarios autorizados de Hadoop solo pueden acceder a los datos a los que tienen derecho.
- Los historiales de acceso a datos para todos los usuarios se registran de acuerdo con las normas de cumplimiento y para fines definidos.
- Los datos están protegidos, tanto en reposo como en tránsito, mediante cifrado de nivel empresarial.

Los grandes datos que residen en un entorno de Hadoop pueden contener datos financieros confidenciales, información corporativa patentada e información de identificación personal, como los nombres, direcciones y números de Seguridad Social de clientes y empleados.

Debido a la naturaleza sensible de estos datos y al daño potencial que puede ocurrir si cae en las manos equivocadas, debe protegerse adecuadamente.

Es posible por ejemplo, que una empresa que vaya a implementar un Big Data, trabaje con requisitos de seguridad propios y específicos, sobre todo con respecto al manejo y almacenaje de información restringida.

Con ese fin, cada empresa según sus necesidades, puede desarrollar una estrategia de seguridad de Hadoop junto con las mejores prácticas para ayudar a mantener seguros los datos confidenciales.

Por tanto, las empresas actuales deben construir la base tecnológica para el análisis en toda la infraestructura de su(s) centro(s) de datos; deben buscar las herramientas de seguridad de hardware y software adecuadas, basadas en alguna arquitectura, que ayuden a bloquear amenazas conocidas, identificar compromisos y acelerar la reparación para mejorar significativamente la seguridad de la información. Entre las características de seguridad que las organizaciones deben buscar están aquellas que les permitan aumentar el rendimiento de la criptografía y que les permita, almacenar, transmitir y analizar de forma segura más datos.

La integración de las características de seguridad de Hadoop ha planteado algunos desafíos sustanciales. Las pautas de seguridad empresarial se busca que sean cada vez más estrictas en términos de autorización, autenticación, cifrado, administración y gobierno.

En implementaciones iniciales de Hadoop y Hive (el almacén de datos de Hadoop), no incluía controles de seguridad: los usuarios podían acceder a los objetos de datos sin autenticación o autorización y en las empresas, se tenía que crear una solución temporal donde se aislaba explícitamente cada aplicación para almacenar datos comerciales en carpetas separadas en el sistema de archivos. Luego, podían aplicarse controles de la lista de control de acceso (ACL) de la interfaz del sistema operativo portátil (POSIX) a las carpetas; aunque definitivamente, esta solución temporal era difícil de administrar desde el principio.

Esto trajo como consecuencia, que los grupos empresariales comenzaron a solicitar el control de acceso basado en roles de soporte, por lo cual, los desafíos de la administración se multiplicaron. Los permisos POSIX dieron un control de todo o nada sobre los usuarios y grupos. Si un usuario pertenecía a

múltiples grupos y asumía múltiples "roles" (como administradores de datos, propietarios de datos, ingenieros de datos, analistas de negocios, etc.), el modelo de todo o nada simplemente no era suficiente.

Además, los desafíos se complicaron cuando la autenticación fue manejada por los protocolos Lightweight Directory Access Protocol (LDAP) mientras la autorización se gestionaba en la capa de almacenamiento.

Con ese enfoque, no era posible cumplir con los requisitos del negocio, razón por la cual, fue necesario pensar en otras estrategias.

En la siguiente sección, se plantea a manera de ejemplo una propuesta de una posible implementación de seguridad de Hadoop, que pretende solventar los retos que hasta no hace mucho enfrentaban las organizaciones con el manejo y control de su información.

## 7. PROPUESTA EJEMPLO DE UNA IMPLEMENTACIÓN DE SEGURIDAD DE HADOOP

Parte de los desafíos en la implementación de las características de seguridad de Hadoop [16] se deriva de su apertura y flexibilidad.

Hadoop puede alojar datos de cualquier tipo y esto, puede llevar a que se agreguen datos confidenciales a un conjunto de datos ya existente, sin activar el gobierno de seguridad apropiado; además, el Sistema de Archivos Distribuidos de Hadoop (HDFS) puede usar almacenamiento de menor costo en servidores en clúster, en lugar de usar sistemas de almacenamiento heredados.

La desventaja es que los controles de cumplimiento integrados que proporcionan los sistemas de almacenamiento heredados ya no están disponibles.

Una estrategia de seguridad de la plataforma Hadoop puede enfocarse en reforzar la seguridad en cuatro áreas: perímetro, acceso, visibilidad y datos.

Esta estrategia utiliza una solución lista para ser empleada para minimizar la carga de soporte del equipo, en lugar de crear una solución personalizada; puede usarse para ello el software Cloudera Enterprise [2], Apache Sentry y Cloudera Navigator en sus versiones más actuales (mencionados en la sección 5 de este documento) para proporcionar los principales controles de seguridad.

Luego, debe realizarse una evaluación técnica del producto Apache Sentry y un análisis de impacto para determinar los costos de refactorización de los proyectos existentes en la organización.

Este esfuerzo puede darse como un enfoque gradual, como se ilustra la hoja de ruta de la Figura 3, dividiendo el perímetro y los controles de acceso que se sugiere por ejemplo, se aborden en el primer semestre de 2020 y los controles de visibilidad y datos en el segundo semestre de 2020; definiendo claramente el alcance de cada capa de control y detallando una entrega trimestral contra cada capa.

En el primer semestre de 2020, debería lanzarse un proyecto para habilitar LDAP para la autenticación y controles específicos para la autorización con Apache Sentry.

Durante este proyecto, debe mencionarse que es posible enfrentar varios desafíos críticos, que incluso pueden amenazar un posible reinicio y un impacto en el mapa de ruta de implementación. Por ejemplo, para admitir las características de Apache Sentry, es posible tener que incorporar la autenticación multi-dominio con Kerberos y Apache Sentry, para establecer la confianza del certificado de capa de sockets seguros (SSL) entre el sistema front-end y el clúster back-end, y reemplazar Apache Beeline [17] con la interfaz de línea de comandos de Hive.

Cada una de estas tareas implica un importante re-trabajo de los desarrolladores; por lo que es conveniente contar con un equipo de trabajo bien consolidado, que tendrá que trabajar en estrecha colaboración con Cloudera [2] para identificar causas raíz de varios problemas y desarrollar parches apropiados para la implementación de seguridad propuesta en la empresa.

También es posible enfrentarse a desafíos al probar nuevas características de Apache Sentry en un entorno con recursos limitados (recordando que cada organización es diferente).

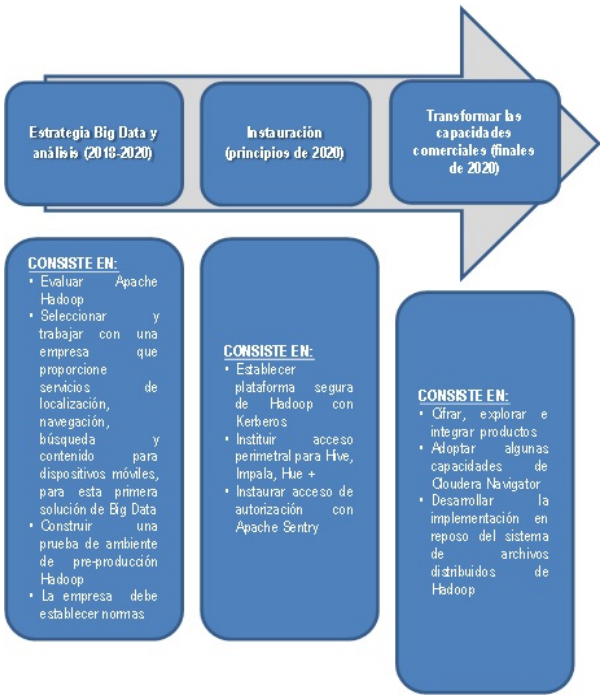


**Figura 3**. Hoja de ruta para la implementación de seguridad de Hadoop de TI empresarial. [Elaboración propia]

## 8. CONCLUSIONES

Debido a que los datos se reúnen de varias fuentes, es decir, no hay una fuente de datos fija; en el campo de Big Data, la seguridad es requisito crítico y una preocupación importante para las entidades que manejan información delicada y confidencial.

Con el aumento de la aceptación de Hadoop dentro de la industria, también se ha extendido la intranquilidad natural por la seguridad; por lo tanto, ha surgido una creciente necesidad de aceptar y asimilar soluciones y características de seguridad comerciales.

Debido a que uso inicial de Hadoop fue experimental, no requería seguridad y no había suficientes recursos para diseñarlo e implementarlo e incluir mecanismos de protección de la información. Sin embargo, a medida que el uso de Hadoop creció y dejó de ser una tentativa, la seguridad se volvió el foco de atención. Mudar datos confidenciales de clientes y empresas en clústeres privados no fue práctico ni rentable para algunos. Como resultado, la seguridad tuvo que agregarse a Hadoop.

Por otra parte, HDFS y MapReduce, son la base de un Hadoop robusto. El uso de Hadoop con programas como HDFS o MapReduce permite trabajar de forma paralela con grandes datos, ofreciendo una solución integrada y fácilmente escalable. Tom White en su libro "Hadoop The definitive guide" [14], define a Hadoop como linealmente escalable (se pueden añadir nodos en función de las necesidades), con una alta disponibilidad (los archivos son replicados tantas veces como sea necesario, dotándolo de fiabilidad) y tolerancia a fallos y esta tolerancia a fallos, se traduce en una robustez para Hadoop.

En este documento, se ha intentado cubrir alguna información sobre la solución de seguridad que se ofrece en el medio, para proteger el ecosistema de Hadoop.

Se presenta además, a manera de ejemplo, una propuesta de proyecto que utiliza Apache Sentry para aumentar la seguridad del entorno Hadoop de Big Data dentro de una organización de TI, tratando de establecer las mejores prácticas, orientación y proceso para fortalecer la seguridad de la plataforma Hadoop, tratando de compartir esta propuesta que puede ser útil para la industria en general.

Debe aclararse, que la organización deberá realizar un exhaustivo análisis de impacto y entender el alcance del proyecto; donde es necesaria una comunicación y una articulación clara de la recompensa del negocio contra el riesgo; de hecho, esa es la única forma en que es posible obtener el soporte de administración de la empresa que decida implementarla. Deberán evaluarse los proyectos, documentarse claramente los hitos, trabajar en estrecha colaboración con Cloudera y continuar siendo ágiles para cambiar rápidamente el rumbo cuando sea necesario. Además, debe tomarse en consideración que también tendrán que desarrollarse scripts de automatización, migración y pruebas reutilizables que pueden usarse para proyectos futuros.

Las empresas deben trabajar cada vez más para lograr la seguridad de sus Big Data y si apuestan por Hadoop, deben trabajar para que esta seguridad se haga realidad.

## 9. REFERENCIAS


[1] Blokhin I., Díaz A.F., Ortega J., Palacios, R.H., Navas, S., (2016). "Acceso seguro a datos en Hadoop mediante criptografía de curva elíptica". JNIC2016 Sesión I4: Privacidad / Control de Acceso

[2] CLOUDERA (2019). **Cloudera Security. Cloudera Enterprise 6.1.x**. Cloudera, Inc. All rights reserved. JNIC2016 Sesión I4: Privacidad/Control de Acceso.

[3] Das D., O'Malley O., Radia S., and Zhang K., (2011). "Adding Security to Apache Hadoop". Hortonworks Technical Report 1.

[4] Ferguson, M., (2013). "Enterprise Information Protection - The Impact of Big Data". INTELLIGENT BUSINESS STRATEGIES LIMITED. White Paper IBM.

[5] Gawali, P.R., Talmale, R., Babu, R. (2015). "Hadoop Security- A Review". International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET). Vol. 3, Issue 10, October 2015. ISSN (Online): 2320-9801. ISSN (Print): 2320-9798. DOI:10.15680/IJIRCCE.2015.0310017

[6] HORTONWORKS (2018). **Hortonworks Data Platform: Security**. Copyright © 2012-2017 Hortonworks, Inc. Some rights reserved. May 17, 2018. docs.hortonworks.com.

[7] Marr B., (2015). "Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance". WILEY.

[8] Mearian, L., (2018). "World's data will grow by 50x in next decade, IDC study predicts". COMPUTERWORLD. https://www.computerworld.com/article/2509588/data-center/world-s-data-will-grow-by-50x-in-next-decade--idc-study-predicts.html, Web Accessed on December 2018.

[9] Narayana, S. (2013). **Securing Hadoop-Implement robust end-to-end security for your Hadoop ecosystem**. ISBN 978-1-78328-525-9. Published by Packt Publishing Ltd. Birmingham-Mumbai.

[10] Sagiroglu, S. and Sinanc, D. (2013) "Big Data: A Review, Collaboration Technologies and Systems (CTS)". 2013 International Conference on Digital Object Identifier, 42-47. 978-1-4673-6404-1/13/ ©2013 IEEE.

[11] Sharma1 V., Joshi N.K. (2015). "The Evolution of Big Data Security through Hadoop Incremental Security Model". International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET). Vol. 4, Issue 5, May 2015. ISSN (Online): 2319-8753. ISSN (Print): 2347-6710. DOI: 10.15680/IJIRSET.2015.0405097.

[12] Sirisha. N., Kiran K.V.D., Karthik, R. (2018). "Hadoop Security Challenges and Its Solution Using KNOX". Indonesian Journal of Electrical Engineering and Computer Science. Vol. 12, No. 1, October 2018, pp. 107-116. ISSN: 2502-4752, DOI: 10.11591/ijeecs.v12.i1.pp107-116.

[13] VORMETRIC (2016). "Securing Big Data: Security Recommendations for Hadoop Environments", Marzo 21, 2016, Versión 2.0, Securosis, L.L.C.

[14] White, T., (2009). **Hadoop The definitive guide. Yahoo!** Press. OREILLY, ISBN 978-0- 596-521974.

[15] ZETTASET (2016). "The Big Data Security Gap: Protecting the Hadoop Cluster". Zettaset. White Paper.

[16] Zhang, X., (2014). "Secure Your Hadoop Cluster With Apache Sentry (Incubating)". April 07, 2014. Software Engineer, Cloudera.

[17] Kniveton, J., Llanas, A., Williams, L., (2000). **Beeline Plus 2 - Sb**. ISBN-10: 0333972554. ISBN-13: 978-0333972557. Published by MACMILLAN