Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

INTREGRACIÓN DE HADOOP A BASES DE DATOS RELACIONALES Y COMPARATIVA DE PROCESAMIENTO DE DATOS

Ever Estuardo Lux Pérez
Asesorado por el Ing. Luis Fernando Alonzo Jerónimo

Guatemala, octubre de 2019

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

INTEGRACIÓN DE HADOOP A BASES DE DATOS RELACIONALES Y COMPARATIVA DE PROCESAMIENTO DE DATOS

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

EVER ESTUARDO LUX PÉREZ
ASESORADO POR EL ING. LUIS FERNANDO ALONZO JERÓNIMO

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, OCTUBRE DE 2019

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Luis Diego Aguilar Ralón
VOCAL V	Br. Christian Daniel Estrada Santizo
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Pedro Antonio Aguilar Polanco
EXAMINADOR	Ing. César Augusto Fernández Cáceres
EXAMINADOR	Ing. Herman Igor Véliz Linares
EXAMINADOR	Ing. Marlon Francisco Orellana López
SECRETARIA	Inga. Lesbia Magalí Herrera López

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

INTEGRACIÓN DE HADOOP A BASES DE DATOS RELACIONALES Y COMPARATIVA DE PROCESAMIENTO DE DATOS

Tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha 05 de marzo de 2018.

Ever Estuardo Lux Pérez

Guatemala, 24 de octubre de 2018

Ingeniero
Carlos Azurdia
Escuela de Ciencias y Sistemas
Facultad de Ingeniería

Respetable Ingeniero Azurdia:

Por este medio le informo, que como asesor del trabajo de graduación del estudiante universitario de la carrera de Ingeniería en Ciencias y Sistemas, **EVER ESTUARDO LUX PÉREZ**, carné 200714566, hago constar que ha finalizado todos los capítulos del trabajo de investigación titulado: **INTEGRACIÓN DE HADOOP A BASES DE DATOS RELACIONALES Y COMPARATIVA DE PROCESAMIENTO DE DATOS**, el cuál he tenido la oportunidad de revisar y doy mi aprobación al mismo.

Agradeciendo su atención a la presente,

Atentamente,

Luis Fernando Alonzo Jerónimo
Ingeniero en Ciencias y Sistemas
Colegiado No. 8871

Ing. Luis Fernando Alonzo Jerónimo
Asesor de trabajo de graduación
Colegiado: 8871

Universidad San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 31 de octubre de 2018

Ingeniero
Marlon Antonio Pérez Türk
Director de la Escuela de Ingeniería
En Ciencias y Sistemas

Respetable Ingeniero Pérez:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación del estudiante **EVER ESTUARDO LUX PÉREZ** con carné **200714566 y CUI 1831 16682 0101** titulado **"INTEGRACIÓN DE HADOOP A BASES DE DATOS RELACIONALES Y COMPARATIVA DE PROCESAMIENTO DE DATOS"** y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo aprobado.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,

Ing. Carlos Alfredo Azurdia
Coordinador de Privados
y Revisión de Trabajos de Graduación

ESCUELA DE INGENIERÍA EN CIENCIAS Y SISTEMAS

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÌA
ESCUELA DE INGENIERÍA EN
CIENCIAS Y SISTEMAS
TEL: 24767644

El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, del trabajo de graduación ***"INTEGRACIÓN DE HADOOP A BASES DE DATOS RELACIONALES Y COMPARATIVA DE PROCESAMIENTO DE DATOS",*** *realizado por el estudiante, EVER ESTUARDO LUX PÉREZ aprueba el presente trabajo y solicita la autorización del mismo.*

"ID Y ENSEÑAD A TODOS"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DIRECCION DE INGENIERIA EN CIENCIAS Y SISTEMAS

Ing. Carlos Gustavo Alonzo
Director
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 14 de octubre de 2019

Universidad de San Carlos
De Guatemala

Facultad de Ingeniería
Decanato

Ref. DTG.453.2019

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al trabajo de graduación titulado: **INTEGRACIÓN DE HADOOP A BASES DE DATOS RELACIONALES Y COMPARATIVA DE PROCEDIMIENTO DE DATOS,** presentado por el estudiante universitario **Ever Estuardo Lux Pérez,** y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, se autoriza la impresión del mismo.

IMPRÍMASE.

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, Octubre de 2019

/cc

ACTO QUE DEDICO A:

Dios	Por la salud y fortaleza brindada y por permitirme la oportunidad de lograr esta meta.
Mi familia	Por todo el apoyo incondicional brindado en las diversas etapas académicas. Definitivamente este logro no hubiera sido posible sin ellos.
Ing. Luis Alonzo	Por los consejos y el ánimo brindado para completar este trabajo.
Mis amigos	Por el apoyo y la motivación durante toda la carrera. La solidaridad brindada en los distintos proyectos y cursos, así como la amistad durante y después de concluir la carrera.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala — Por la formación académica, los valores y las experiencias transmitidas en las aulas de clase. Id y enseñad a todos.

Facultad de Ingeniería — Por la dedicación de formar buenos profesionales, por fomentar el aprendizaje y la excelencia.

Mi padre — Efraín Lux, por el apoyo incondicional, buenos consejos y cuidados en todo momento. Por el buen ejemplo que ha sido en mi vida.

Mi madre — Marta Pérez de Lux, el cariño y cuidados de siempre, por esa dedicación ejemplar e invaluables consejos.

Mis hermanas

Por su compañía y cuidados en todo momento, por contagiar alegría en momentos difíciles.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
$	Dólar estadounidense
DB	Base de datos
GB	Gigabyte
MB	Megabyte
MHz	Megahercio
PB	Petabyte
SSD	Disco de estado sólido
TB	Terabyte

GLOSARIO

ACID Indica las cuatro propiedades necesarias para que una serie de instrucciones en una base de datos sea constituida como una transacción. Dichas propiedades son atomicidad, consistencia, aislamiento y durabilidad.

API Siglas de *application programming interface*. Término que se refiere a software que por medio de reglas y procesos permite la interconexión o comunicación entre distintas aplicaciones.

ASM *Automatic storage management*, es un sistema de archivos y administrador de almacenamiento propiedad de Oracle, optimizado para sus bases de datos.

Clúster Conjunto de dos o más computadoras que comparten recursos para conformar una unidad lógica de procesamiento y almacenamiento.

Código abierto Término con que se denomina al software que se desarrolla y distribuye de forma libre o gratuita.

CSV	Del inglés *comma separated values*, son archivos de texto plano que presentan datos en forma de tabla; las columnas son separadas por comas y las filas por saltos de línea.
EULA	Concepto utilizado en licenciamiento de software, en el cual los propietarios establecen términos de uso al usuario final, de ahí la denominación de *end-user license agreement* o en español acuerdo de licencia de usuario final.
Ext4	Sistema de archivos para sistemas operativos basados en Linux.
Framework	Esquema de trabajo que define y facilita métodos, prácticas y criterios de desarrollo e implementación de una aplicación, así como herramientas para acoplar diferentes módulos de un proyecto.
HDFS	*Hadoop distributed file system* por sus siglas en inglés; es un sistema de archivos distribuidos sobre el cual se basa Apache Hadoop para procesar, almacenar y administrar grandes volúmenes de datos de forma eficiente.
JFS	Sistema de archivos para sistemas operativos AIX de IBM.

NFS	Sistema de archivos en red que por medio de un protocolo permite compartir acceso a un espacio de almacenamiento.
NTFS	*New technology file system*, es un sistema de archivos para sistemas operativos Microsoft Windows.
OLTP	Tipo de procesamiento especializado para administrar aplicaciones transaccionales, típico de base de datos relacionales.
OLAP	Proceso analítico en línea, es la contraparte de OLTP. Describe una base de datos dimensional diseñada y ajustada para procesar análisis de tendencias y previsiones.
Paradigma	Sinónimo de modelo o patrón. Indica una forma de pensar, una tendencia o idea, empleadas para solventar problemas o situaciones determinadas que se planteen.
RAID	Acrónimo de matriz redundante de discos independientes, esta forma de almacenamiento permite incorporar varios discos duros para distribuir o replicar datos entre ellos como contingencia a fallas físicas de los mismos.

RAM	Memoria de acceso aleatorio, define la memoria que es utilizada por un procesador para recibir instrucciones y guardar resultados, dentro de sus características destaca el acceso rápido y su volatilidad.
Servidor	Equipo físico o computadora cuyo objetivo es suministrar información o servicios a una serie de clientes, que pueden ser personas u otros equipos conectados a la red.
SQL	Lenguaje de consulta estructurada, es un lenguaje estándar para acceso y manipulación de bases de datos.
Transacción	Serie de instrucciones u operaciones que se deben completar de forma atómica o indivisible.
XML	*Extensible markup language*, metalenguaje que permite definir estructuras o etiquetas personalizadas para descripción y organización de datos.

RESUMEN

En el primer capítulo se trata sobre el surgimiento, las características y el valor agregado que genera Big Data. También, se describe las características y principales componentes de Apache Hadoop. Adicional, se presenta una comparativa entre Hadoop y las bases de datos relacionales en cuanto a la funcionalidad, el almacenamiento y el esquema de búsqueda.

Como es de esperar existen elementos que se deben considerar para integrar o emplear Apache Hadoop. De tal forma, en el segundo capítulo se desarrolla, entre otros, los aspectos de adaptación a la tecnología, el tiempo de aceptación y la etapa de aprendizaje que deben tomarse en cuenta. Y se presenta algunos casos de éxito para referencia.

En el tercer capítulo se abarcan los costos por considerar para la integración de Hadoop, se da detalles de costos de implementación, licenciamiento, soporte o mantenimiento y capacitación de personal. A modo de comparación, también, se presentan los mismos datos referentes a una base de datos relacional. En cuanto al costo de infraestructura se expone dos escenarios a considerar según el caso, uno es en premisa y otro en la nube.

El desempeño de Apache Hadoop en un escenario controlado se desarrolla en el cuarto capítulo; en este se describe el ambiente y las métricas a evaluar, así como un análisis de tiempos de respuesta y consumo de recursos al integrar Hadoop con una base de datos relacional.

OBJETIVOS

General

Analizar las características de Apache Hadoop y sus componentes, además el beneficio de una implementación, a través de pruebas en un escenario simulado, en conjunto con una base de datos relacional Oracle, para incrementar el rendimiento en cuanto a almacenamiento y procesamiento se refiere.

Específicos

1. Efectuar un análisis de las características generales de Hadoop, que pueda ser utilizado como base para estudiar la viabilidad de implementar uno o varios diseños, que hacen uso de este modelo de almacenamiento y procesamiento de datos, según sea el caso particular.

2. Aportar un punto de referencia para el análisis de factibilidad de la implementación de Hadoop en conjunto con una base de datos relacional, a partir del rendimiento que esta combinación produce, así como los costos de implementación que esto conlleva.

3. Medir el impacto, mediante la comparación de costos y de rendimiento, de la implementación del modelo relacional o de ambas tecnologías en conjunto (Hadoop y relacional), para determinar cuál alternativa es la solución más viable y satisfactoria.

INTRODUCCIÓN

Las empresas u organizaciones se enfrentan a la creciente necesidad de analizar información para la toma de decisiones, la creación de estrategias, la evaluación de productividad, la reportería, entre otros. Es por ello que es indispensable contar con un sistema de alta calidad y de bajo coste que ofrezca, aun de forma mínima, disponibilidad, respuesta a fallos, velocidad y escalabilidad, para cumplir con los propósitos del negocio.

Aunque las bases de datos relacionales han evolucionado para responder a las demandas del mercado, también han ido surgiendo nuevas tecnologías bajo el concepto de Big Data, cuyo enfoque facilita trabajar con grandes volúmenes de información, variedad de tipo de datos y también maniobrar con distintas velocidades de flujo de registros. Una de estas plataformas es Apache Hadoop que contiene dos componentes principales: HDFS (*hadoop distributed file system*) y el *framework* de MapReduce.

En tal sentido, Apache Hadoop representa una solución factible ante la problemática que enfrentan muchos sistemas actualmente. HDFS es un sistema de archivos distribuido y MapReduce un *framework* para procesamiento de la información. Gracias a la combinación de dichos componentes es factible implementar soluciones de procesamiento en paralelo para grandes conjuntos de datos, además de la escalabilidad, pues Hadoop puede ser conformado por miles de nodos.

Como se ha mencionado, el modelo relacional de bases de datos propuesto por E. F. Codd, aunque fue paulatinamente adoptado, es muy utilizado dadas sus prestaciones y características que garantizan la correcta administración de los datos almacenados. Empero, frente a las demandas crecientes, una buena opción es trasladar tareas a alguna de las tecnologías ofrecidas por Big Data, específicamente Apache Hadoop.

Por tal razón, es importante identificar las características de aquellas tareas, actividades o aplicaciones que sean aptas para migrar o implementar en Apache Hadoop, y combinar así las fortalezas de cada tecnología para obtener el máximo beneficio posible para el negocio. De tal forma, en este trabajo de investigación se busca demostrar a través del análisis de costos monetarios que un sistema que combina las tecnologías de Apache Hadoop y bases de datos relacionales es factible económicamente y además representa la mejor solución para el almacenamiento y procesamiento de información, tras analizar los tiempos de respuesta y consumo de recursos obtenidos mediante pruebas en un escenario simulado.

1. ¿QUÉ ES HADOOP?

Hadoop consiste en una solución a la problemática abordada por Big Data. Por ello este capítulo se centra en describir no solo el surgimiento sino también las principales características de cada uno de dichos términos. A su vez, se plantea una comparativa entre las bases de datos relacionales y Hadoop, para comprender sus diferencias y áreas complementarias.

1.1. Surgimiento de Big Data

Big Data no representa una tecnología o software en particular, sino se refiere a metodologías de almacenamiento y procesamiento de datos que surgen a inicios del 2000 impulsadas por los motores de búsquedas, principalmente, Google y Yahoo!.

La inmensa cantidad de páginas web y su creciente número plantearon un reto de gran escala para los motores de búsqueda, pues cada vez sería más complicado indexar documentos y su contenido para brindar resultados significativos a los usuarios. En tal sentido, los fundadores de Google y Yahoo! de forma independiente identificaron que los enfoques tradicionales carecían de las herramientas necesarias para solventar tales demandas.

Lo anterior fue el detonante de las propuestas presentadas por Google en 2003 (The Google File System[1]) y 2004 (MapReduce: Simplified Data

[1] GHEMAWAT, Sanjay; GOBIOFF, Howard; LEUNG, Shun-Tak. *The Google file system*. https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf. Consulta: 06 de marzo de 2018.

Processing on Large Clusters[2]) así como la implementación de Yahoo! en 2006 liderada por Doug Cutting, colaborador en proyectos sobre motores de búsqueda como Apache Lucene y Apache Nutch, quien tomó como base técnica las publicaciones de Google para iniciar lo que ahora es Apache Hadoop, que posteriormente se abordará a más detalle sobre sus características.

1.1.1. El valor de Big Data

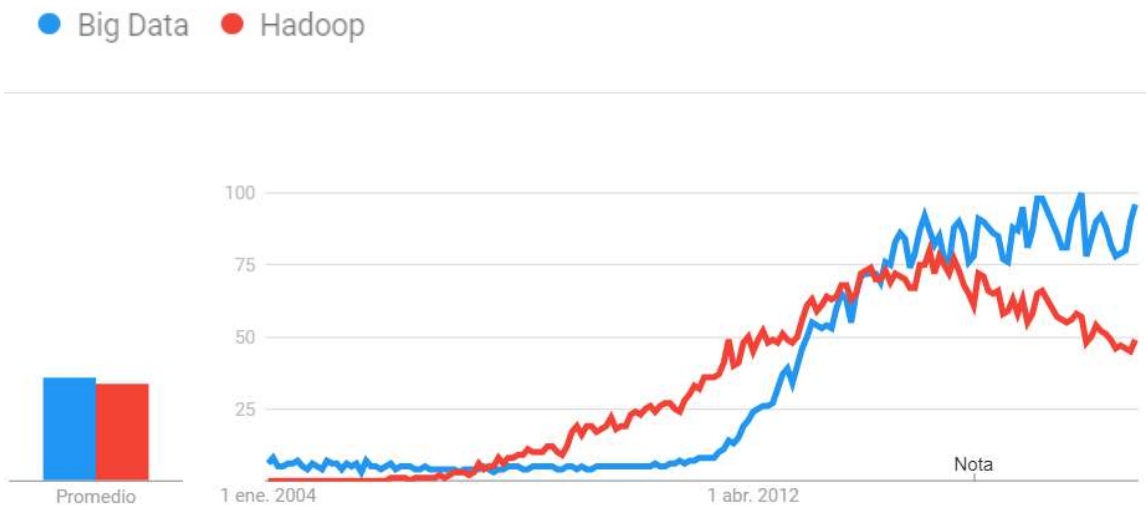
El desafío para almacenar y procesar una creciente cantidad de datos ya no es exclusivo de los motores de búsqueda, también las empresas de otras industrias se han percatado de lo valioso que es contar con una estrategia y una plataforma confiables para solventar estos requerimientos.

Además, sin importar el ámbito industrial, las empresas buscan una solución rentable y que genere valor agregado a los diversos modelos de negocios. Aprovechando de esta manera el cúmulo de información que disponen para incorporarla a la toma de decisiones, creación de estrategias, evaluación de productividad, reportería, etc.

En tal sentido, como se muestra en la figura 1, el interés tanto en Big Data como en Hadoop ha ido en aumento, dado que son términos estrechamente vinculados desde sus inicios, esto debido a que ofrece soluciones de alta calidad para contar con disponibilidad, respuesta a fallos, velocidad, escalabilidad y un bajo coste para cumplir con los propósitos del negocio.

[2] DEAN Jeffrey; GHEMAWAT Sanjay. *MapReduce: Simplified Data Processing on Large Clusters*. static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf. Consulta: 06 de marzo de 2018.

Figura 1. **Tendencia de Big Data y Hadoop**



Fuente: Google Trends. https://trends.google.com/trends/explore?date=all&q=Big%20Data,Hadoop. Consulta: 06 de marzo de 2018.

1.2. Características de Big Data

Previo a describir las características de Big Data, es importante aclarar a que se refiere el término dato:

"Información dispuesta de manera adecuada para su tratamiento por una computadora"[3].

De tal modo Big Data es la propuesta para gestionar, almacenar y procesar datos que cumplen con una o varias de las siguientes particularidades.

[3] Real Academia Española. http://dle.rae.es/srv/fetch?id=Bskzsq5%7CBsnXzV1. Consulta: 06 de marzo de 2018.

1.2.1. Volumen

Es quizá la principal característica, pues como su nombre indica, Big Data sugiere procesamiento de volúmenes masivos que suelen medirse en petabytes (10^{15} bytes) de información.

Empero, cabe mencionar que lo anterior no siempre es el caso para todas las organizaciones. Tales situaciones son discutidas en el documento titulado *Nobody ever got fired for using using Hadoop on a cluster*[4], en el cual se plantea que en por lo menos dos ambientes analíticos de producción de Microsoft y Yahoo! procesan en promedio cantidades menores a 14 GB; y en el caso de Facebook el 90 % de sus sistemas analíticos manejan cantidades inferiores a los 100 GB de datos.

De tal modo, aun cuando no se cuente con grandes volúmenes de datos, las empresas podrían tener requerimientos que se beneficiarían de las metodologías aportadas por Big Data.

1.2.2. Variedad

Este punto consiste en dos aspectos, siendo el primero de ellos la fuente de los datos, que podría ser diversa y no necesariamente relacionadas entre sí. Algunos ejemplos podrían ser archivos de texto, hojas de cálculo, bases de datos, audio, entre otros.

[4] ROWSTRON, Antony; DUSHYANTH, Narayanan; DONNELLY, Austin; O'SHEA, Greg; DOUGLAS, Andrew. *Nobody ever got fired for using Hadoop on a cluster*. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/hotcbp1220final.pdf. Consulta: 07 de marzo de 2018.

La naturaleza de los datos, el segundo aspecto, implica determinar si los mismos son estructurados (con formato fijo, como es el caso de una tabla en las bases de datos relacionales), no estructurados (con formato desconocido, combinación de varias fuentes como imágenes, vídeos, entre otros.) o semiestructurados (con formato conocido, pero no necesariamente fijo, como es el caso de archivos XML).

Estos dos aspectos son clave para determinar si alguna solución de Big Data se ajusta a la problemática de los negocios para una correcta administración y análisis.

1.2.3. Velocidad

Se refiere a cuán rápido se generan los datos. Acá radica la importancia de contar con una plataforma que logre responder a esta alta demanda de procesamiento para obtener el máximo potencial de la información.

De tal manera, Big Data ofrece métodos para lidiar con la alta producción de datos para casos como logs de aplicaciones, procesos de negocio, redes sociales, sensores, dispositivos móviles, entre otros.

1.2.4. Otros

Además de los aspectos anteriores, conocidos como las V[5] de Big Data, se suele incluir un atributo más que se refiere a la variabilidad, esto debido a lo variable, inconsistente e incompleta que puede ser la totalidad o una porción de los datos que se procesan.

[5] LANEY, Doug. *Deja VVVu: Others Claiming Gartner's Construct for Big Data*. http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/. Consulta: 07 de marzo de 2018.

Figura 2. **Crecimiento de Internet**

Fuente: elaboración propia.

Como se observa en la figura 2, el alcance de internet se ha expandido de forma considerable con el paso de los años, alcanzando para diciembre de 2017 la cantidad de 4 156 millones de usuarios[6]. Junto a esta evolución es evidente que la generación de información también va en aumento, lo cual recalca el valor de Big Data para todas las áreas de negocio.

[6] Internet Growth Statistics. http://www.internetworldstats.com/emarketing.htm. Consulta: 07 marzo de 2018.

1.3. Hadoop

Es un *framework* que permite procesamiento distribuido de grandes conjuntos de datos a través de clúster de computadoras. Está diseñado para escalar desde un solo servidor hasta miles de máquinas, cada una de ellas ofreciendo almacenamiento y procesamiento local para análisis de Big Data.

El administrar datos de forma local es un concepto clave en Hadoop, pues consiste en efectuar cálculos, únicamente, sobre la información existente en cada servidor. Lo cual varía respecto al paradigma de primero solicitar el conjunto de datos a cada servidor, para luego ser trasladado y procesado por un solo sistema o servidor en particular. Un ejemplo de esto serían los administradores de bases de datos relacionales que primero centralizan los datos para luego filtrarlos y retornarlos al cliente.

De tal forma, con el procesamiento local se garantiza un método eficiente pues se evita trasladar enormes cantidades de información por medio de la red dado que, de forma independiente, cada nodo del clúster analiza porciones pequeñas de todo el conjunto de datos sin comunicarse con el resto de los servidores.

1.3.1. Componentes de Hadoop

Apache Hadoop cuenta con dos componentes principales, siendo el primero el sistema de archivos distribuido denominado HDFS (*hadoop distributed file system*). Lo cual facilita el almacenamiento de grandes volúmenes de información a través de clúster de servidores y provee alta disponibilidad por medio de replicación en vez de redundancia de los datos.

El segundo componente consiste en MapReduce, es decir, un *framework* de un modelo de procesamiento en paralelo, básicamente se compone de dos etapas: mapeo y reducción. El mapeo es usualmente algún proceso de filtrado u ordenamiento. Y la etapa de reducción es una compilación o unificación de los resultados procesados por cada proceso de mapeo.

La integración de MapReduce con HDFS garantiza que siempre que sea posible, cada tarea de MapReduce se ejecute en un solo nodo de HDFS para cumplir con el concepto de procesamiento local. A continuación, algunas cualidades que comparten estos dos componentes:

- Diseñados para ejecutarse en clústeres de equipos de baja o mediana capacidad denominados *commodity servers*.

- Escalabilidad o aumento de capacidad por medio de la incorporación de más servidores.

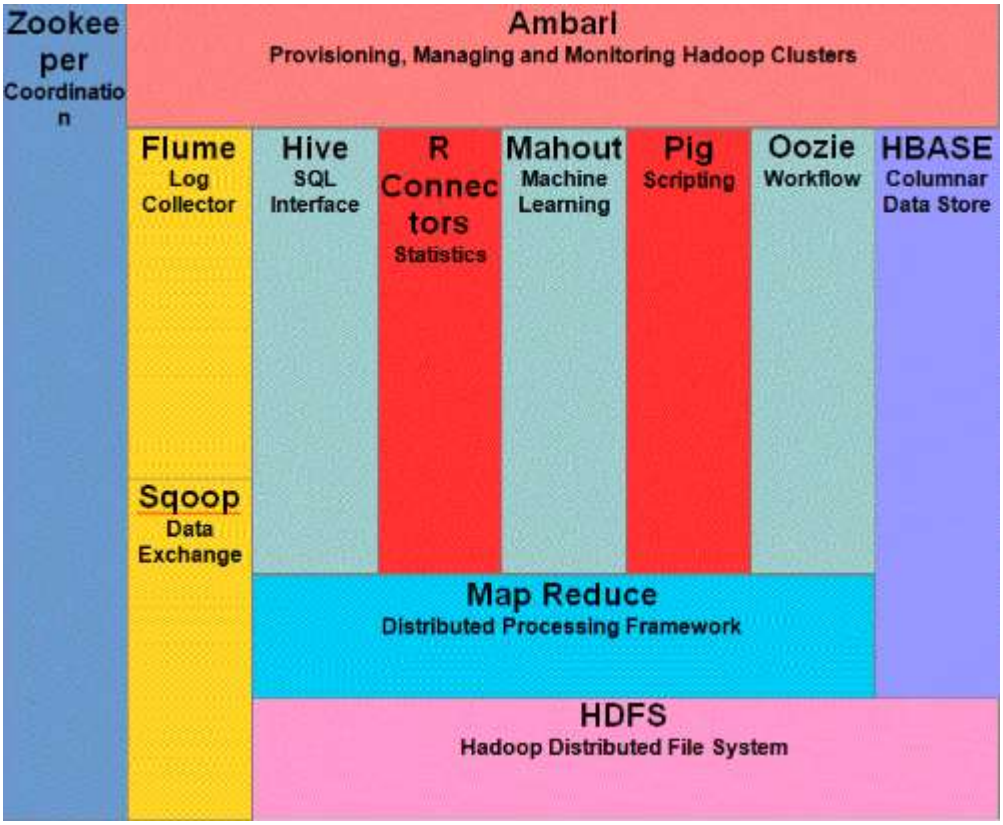
- Mecanismos para identificar y solventar fallas durante la ejecución de procesos.

- Herramientas necesarias para enfocarse en la resolución de los problemas o necesidades del negocio.

Es de mencionar que han ido surgiendo proyectos que interactúan con Hadoop para facilitar y mejorar el uso de los datos del negocio. A manera de indicar algunos casos, se tienen proyectos para carga de datos (Flume, Sqoop), herramientas para análisis (Hive, Pig), *machine learning* (Mahout), coordinación de flujos (Oozie), base de datos (HBASE) y para tareas administrativas (Zookeeper, Ambari).

Dichos proyectos forman parte del denominado ecosistema Hadoop, el cual se ilustra de forma parcial (pues es probable que sigan surgiendo nuevos proyectos) en la figura 3.

Figura 3. **Ecosistema Hadoop**



Fuente: RUNGTA, Krishna. *Learn hadoop in 1 day*. p. 13.

1.4. Comparativa de base de datos relacional versus Hadoop

Esta comparativa no pretende se exhaustiva, pero si se busca analizar los aspectos fundamentales de cada tecnología. En ese sentido se describirán la

funcionalidad, almacenamiento y esquema de búsquedas de cada una de dichas soluciones de administración de datos.

1.4.1. Funcionalidad

Cada una de estas herramientas está diseñada para cumplir con un objetivo en particular. Aunque técnicamente se pueda adaptar una base de datos relacional para una solución de Big Data o viceversa, es de esperar no solo inconvenientes en la implementación sino también una repercusión negativa en los recursos o rendimiento.

En este caso, se recalca que no están limitados a estos usos, pero las bases de datos relacionales son empleadas para procesamiento de transacciones en línea (OLTP por sus siglas en inglés). En tanto Hadoop y su ecosistema surgió para procesamiento analítico en línea (OLAP por sus siglas en inglés), sobre todo en aquellos casos donde existe volúmenes elevados de datos.

1.4.1.1. Función general de una base de datos relacional

Se define como base de datos relacional a la colección de elementos con relaciones predefinidas entre los mismos. Dichos elementos se disponen como un conjunto de tablas con filas y columnas. Las tablas agrupan características abstraídas de objetos a representar en la base de datos.

Cabe mencionar que cada columna almacena un solo tipo de dato (números enteros, números reales, cadena de caracteres de longitud variable, entre otros) cuyo valor o atributo es conocido como campo. Y cada fila o

registro representa el conjunto de características de cada uno de los objetos o entidades.

Es posible emplear uno o varios campos como identificadores únicos de cada fila, los cuales conforman las llaves primarias. Y las relaciones entre varias tablas se establecen por medio de llaves foráneas, es decir uno o varios campos compartidos entre las filas de las tablas involucradas. Esto último permite consultas complejas sobre los datos almacenados.

De tal manera, en términos generales las bases de datos relaciones siguen un modelo que garantiza la siguiente funcionalidad:

- Integridad de datos.
- Transaccionalidad.
- Conformidad con ACID (atomicidad, coherencia, aislamiento y durabilidad).
- Lenguaje de consulta (SQL).
- Evita duplicidad de datos por medio de reglas acordes al negocio.
- Recuperación de fallas.
- Optimización de consultas.
- Manejo de concurrencia.
- Seguridad y auditoría.

Los elementos anteriores interactúan entre sí para ofrecer un método confiable, seguro y ordenado para la administración, consulta, actualización, inserción y eliminación de los datos; además, de la integración sencilla a los aplicativos según las necesidades de los diversos negocios.

1.4.1.2. Función general de Hadoop

En cuanto a Hadoop se refiere, tal como se cubrió en secciones previas, está diseñado para:

- Manejo de grandes volúmenes de datos.
- Procesamiento distribuido de los datos.
- Estructura flexible de datos.
- Recuperación de fallas.
- Facilidad de escalamiento.
- Uso de servidores de mediana o baja capacidad (*commodity servers*).

A partir de tal diseño, con Hadoop es posible obtener tiempos bajos al procesar volúmenes masivos de datos con estructura definida o desconocida; además los *frameworks* preveen fallas y, por ende, implementan métodos para recuperarse ante las mismas (replicación de datos). Por la facilidad de escalamiento es factible crear estrategias para responder de forma rápida al aumento de usuarios o capacidad de procesamiento.

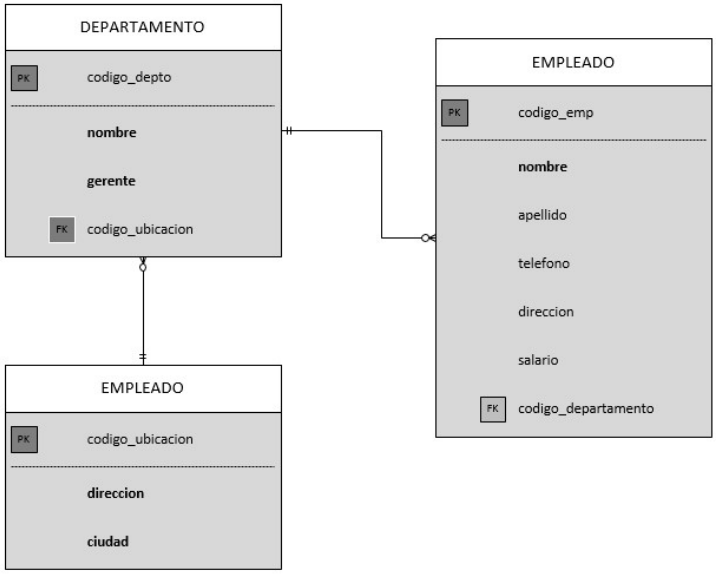
1.4.2. Almacenamiento

Como se verá a continuación, la forma de almacenamiento es una diferencia clave entre ambas tecnologías. Los clústeres basados en Apache Hadoop hacen uso del sistema de archivos distribuido llamado HDFS. Mientras que en el caso de las bases de datos relacionales utilizan sistemas de archivos dependientes del sistema operativo o implementaciones propias como es el caso de los clústeres Oracle.

1.4.2.1. Almacenamiento en una base de datos relacional

Existen múltiples propuestas o implementaciones de bases de datos relacionales cuyo sistema de archivos depende directamente del sistema operativo. Para los sistemas operativos basados en UNIX/Linux suele ser los que cumplen con las especificaciones POSIX (ext4, NTFS, ZFS, NFS, JFS). En el caso de Microsoft Windows típicamente es NTFS. Algunos en cambio han optado por crear un sistema de gestión propio como es el caso de Oracle que utiliza *automatic storage management* (ASM) para las infraestructuras en clúster.

Figura 4. **Modelo relacional**



Fuente: elaboración propia.

Independiente del sistema de archivos las bases de datos relacionales abstraen la forma de almacenamiento a través del modelo entidad relación del cual se observa un ejemplo en la figura 4. No se abordarán detalles en cuanto al esquema físico pues cada gestor de base de datos relacional (DB2, Oracle Database, Microsoft SQL Server, MySQL, etc.) ha ideado su propia solución.

Figura 5. **Tabla en modelo relacional**

Código	Nombre	Apellido	Teléfono	Dirección	Salario	Departamento
1	Carlos	López	34531234	Zona 5	5000	1
2	David	Ruiz	45678234	Zona 12	4500	2
3	Daniel	Ramirez	45782340	Zona 21	5200	3
4	Alejandr	Pérez	87964533	Zona 12	4500	2

Fuente: elaboración propia.

A modo de tener una idea más clara, en la figura 5 se da un ejemplo de representación de una tabla almacenada en una base de datos relacional.

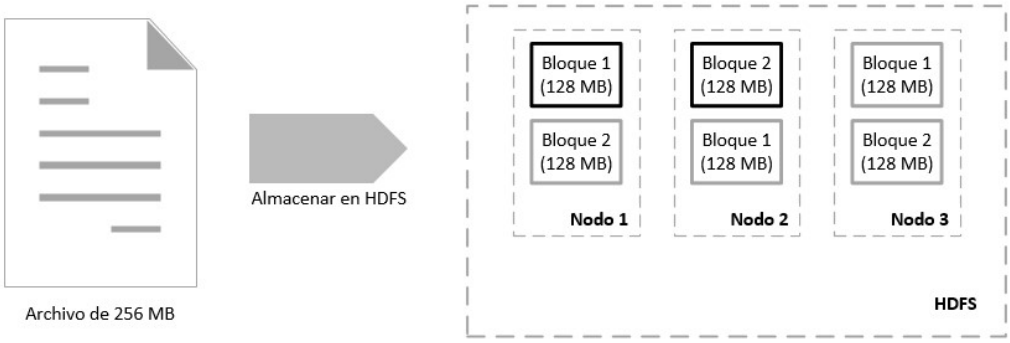
1.4.2.2. Almacenamiento en Hadoop

HDFS es un sistema de archivos para procesamiento de Big Data. Es muy eficiente para almacenar grandes volúmenes de datos debido a su diseño distribuido. Además de ser un sistema muy sencillo de escalar por medio de incorporación de más nodos o *commodity servers* y tolerante a fallas gracias al método de replicación de bloques.

El tamaño de los bloques de datos es típicamente de 64 MB o 128 MB, dicho tamaño es configurable. De tal modo, durante el proceso de almacenamiento los archivos son partidos y los bloques resultantes son

distribuidos en todos los nodos que conforman el clúster Hadoop. Al mismo tiempo los bloques son replicados, en clústeres de tres o más nodos de forma predeterminada se crean tres copias por bloque. En la figura 6 se ilustra la distribución y replicación de bloques en un clúster de tres nodos de HDFS.

Figura 6. **Distribución y replicación de bloques en HDFS**



Fuente: elaboración propia.

En los clústeres HDFS conviven dos tipos de nodos o roles para la administración de los datos:

- NameNode**:** único nodo maestro del sistema de archivos distribuido. En este nodo no se almacenan archivos o bloques de datos. Tiene como función principal la administración de la *metadata* o catálogo de propiedades, atributos y distribución/ubicación de los archivos; así como la creación, eliminación y replicación de bloques. Cabe mencionar que para lograr tiempos de respuesta eficiente dicha *metadata* reside en memoria RAM de este nodo, y se tiene una copia persistente que es actualizada o reconstruida a partir de los DataNodes cada vez que se reinicia el clúster.

- DataNode**:** uno o más nodos puede tomar este rol. Son denominados nodos esclavos y proveen el almacenamiento para los archivos o bloques de datos. Su propósito es responder ante las peticiones de escritura y lectura de los clientes o aplicaciones. Además, cada uno de estos nodos envía al NameNode de forma regular un inventario de bloques, lo cual es empleado para actualizar la *metadata* mencionada con anterioridad.

A continuación, se listan los aspectos más importantes de HDFS:

- Inmutabilidad, es decir, no es posible actualizar archivos una vez ya almacenados.

- El diseño de inmutabilidad (escribir una vez, leer varias veces) tiene como objetivo facilitar las peticiones de lectura de bloques.

- Es factible agregar o añadir archivos a los ya almacenados, pero no están soportadas las lecturas aleatorias. Y no existe caché de datos.

- El almacenamiento y procesamiento es de forma local. Es más eficiente la manipulación de datos de esta manera.

- Alta disponibilidad, debido a la replicación de bloques a través del clúster, es decir la falla de uno o varios nodos no interrumpe el sistema[7].

- Aunque los archivos están divididos en bloques, para el usuario o aplicativo se muestra como uno solo.

[7] EADLINE, Douglas. *Hadoop 2 quick-start guide*. p. 64.

- Los clústeres de HDFS consisten en un NameNode que administra la *metadata* del sistema de archivos distribuido y de uno o varios DataNodes que almacenan bloques de datos.

1.4.3. Esquema de búsquedas

Debido al propósito de cada tecnología y la diferencia en cuanto al método de almacenamiento, el proceso de recuperación de datos también varía significativamente. En las siguientes secciones se describe el método de búsqueda de las bases de datos relacionales y de Hadoop.

1.4.3.1. Esquema de búsquedas en una base de datos relacional

Las bases de datos relacionales típicamente siguen el proceso representado en la figura 7 para analizar y organizar los datos de una o varias tablas consultadas por medio del lenguaje SQL.

A continuación, se describe cada una de las etapas involucradas en las búsquedas:

- Análisis o *parsing* de la sentencia
    - Verificación de sintaxis: se valida que no existan errores de redacción en la escritura de la sentencia SQL.
    - Verificación semántica: aunque esté bien escrita la sentencia en este paso se valida la existencia de los objetos consultados.

- Verificación en caché: se evalúa si la sentencia se ha ejecutado con anterioridad. Si se encuentra en caché se ejecuta la sentencia.

- Optimización

  - Si la sentencia no se encuentra en caché, se calcula el plan de ejecución óptimo. Esto demora la ejecución de la sentencia pues podría producirse varios planes de ejecución.

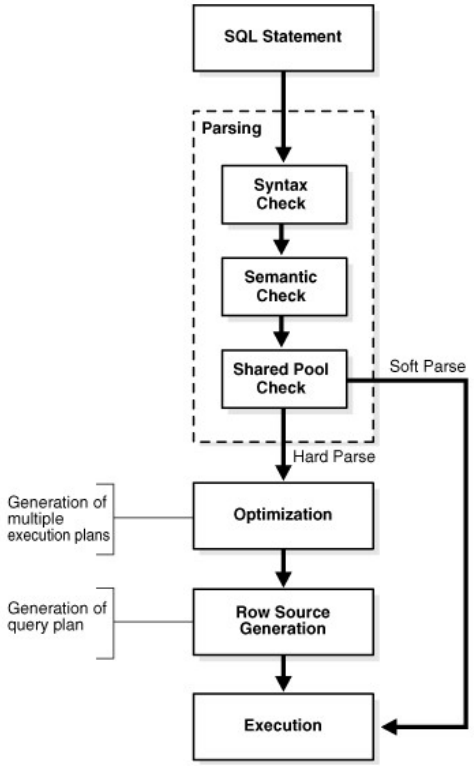
- Generación de plan de ejecución

  - Recibe plan óptimo de ejecución, el cual contiene de manera ordenada y detallada los pasos a seguir para responder a la sentencia. Así como los objetos a emplear: tablas, índices, método para unir tablas y operaciones como filtrado, ordenamiento o agregación.

- Ejecución de sentencia

  - Conforme al plan de ejecución, recupera de una manera óptima la información consultada.

Figura 7. **Etapas de procesamiento de consultas**



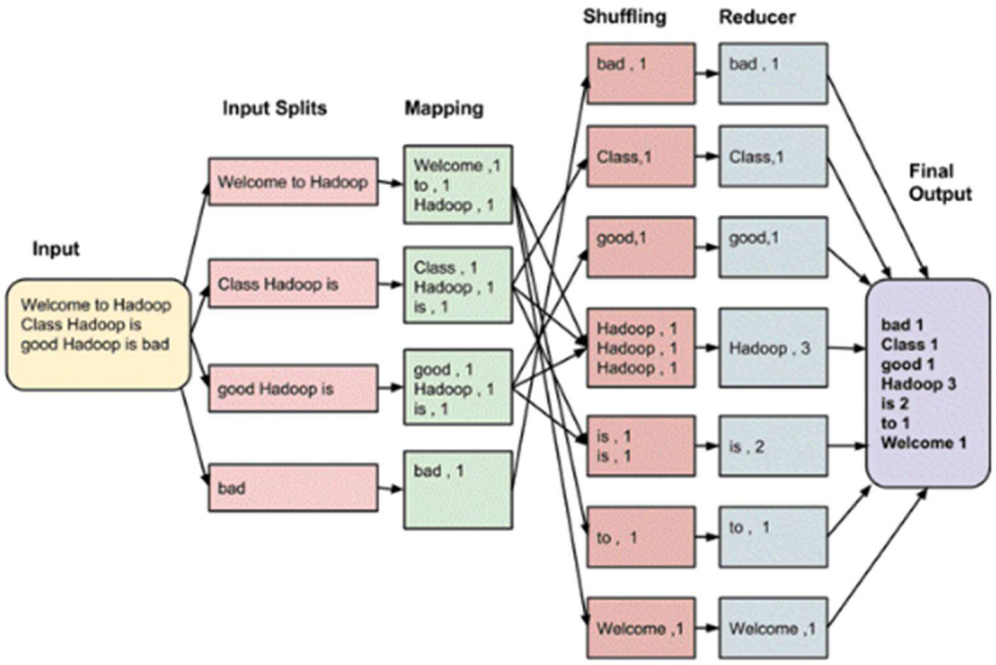
Fuente: Database SQL Tuning Guide: SQL Processing. https://docs.oracle.com/database/121/TGSQL/tgsql_sqlproc.htm. Consulta: 08 de marzo de 2018.

1.4.3.2. Esquema de búsquedas en Hadoop

Hadoop implementa el *framework* de MapReduce para el procesamiento de Big Data. Está basado en la idea de 'divide y vencerás', es decir partir un problema y resolverlo a través de múltiples subtareas. Dicho concepto toma mayor relevancia cuando las subtareas se ejecutan en paralelo; por ejemplo, una tarea que toma 1 200 minutos podría ser completada en 1 minuto por 1 200 subtareas ejecutadas en paralelo.

Por tal motivo, MapReduce resulta muy útil para procesamiento y análisis de datos a gran escala por medio de servidores en clúster. Los programas que siguen este modelo se ejecutan en dos fases: mapeo (*input splits* y *mapping*) y reducción (*shuffling* y *reducer*). Cabe mencionar que cada fase recibe entradas de la forma Llave-Valor.

Figura 8. **Procesamiento de MapReduce**



Fuente: Fuente: RUNGTA, Krishna. *Learn hadoop in 1 day*. p. 33.

La figura 8 representa un ejemplo del flujo en paralelo de un programa para contar palabras en MapReduce, los pasos básicos son:

- Input Splits: división de archivos o generación de entradas para las tareas de mapeo, son de tamaño menor o equivalentes a bloques de datos en HDFS.

- Mapping: cada tarea de mapeo procesa de forma paralela su respectiva entrada para producir salidas en formato llave-valor, que en el ejemplo de la figura sería palabra-frecuencia. MapReduce tratará de ejecutar las tareas en donde residan los bloques, debido a la replicación se seleccionará el nodo con menor carga.

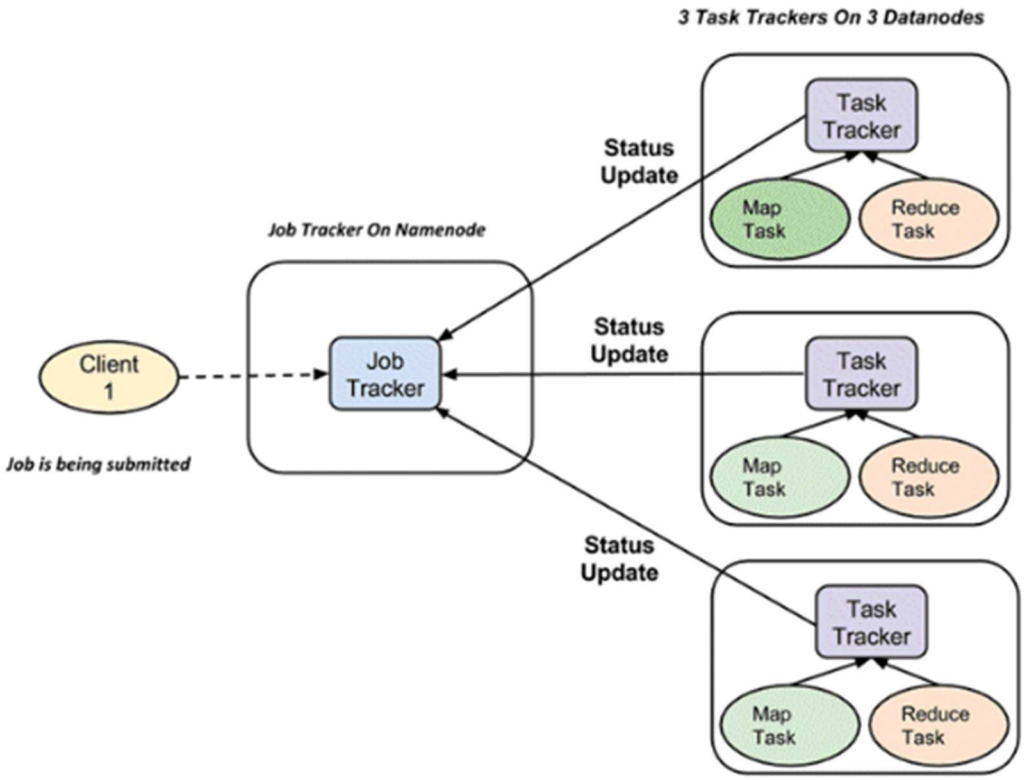
- Shuffling: tareas que consumen las salidas de la fase de mapeo. Consiste en consolidar la información. En el ejemplo de la figura se unifican las palabras repetidas y su frecuencia. Es de mencionar que si solo se emplea una tarea de reducción los procesos de *shuffling* no son necesarios.

- Reducer: en esta fase se unifica o combina todas las salidas de los procesos de *shuffling* en un solo resultado, el cual es escrito en HDFS. Acorde al ejemplo ilustrado en la figura, por cada palabra se calcula su número total de frecuencia.

Dichas tareas que intervienen en el flujo de búsquedas son coordinadas por los siguientes tipos de procesos:

- JobTracker: proceso maestro que se ubica en el NameNode de HDFS. Su objetivo es garantizar la completa ejecución de las tareas de MapReduce. Para ello crea múltiples tareas a realizarse en los nodos del clúster.
- TaskTracker: múltiples procesos esclavo, son ejecutados en los DataNodes que llevan a cabo las diversas subtareas. Envían mensajes de avance de forma periódica al JobTracker, en caso de falla el JobTracker programa una nueva ejecución.

Figura 9. **Coordinación de tareas de MapReduce**



Fuente: RUNGTA, Krishna. *Learn Hadoop in 1 Day*. p. 35.

2. ELEMENTOS POR CONSIDERAR PARA LA INTEGRACIÓN DE HADOOP CON BASES DE DATOS RELACIONAL

Adicional a conocer las características y el funcionamiento de Hadoop, es importante considerar su facilidad de implementación y de administración de datos, así como la dificultad que representa adaptarse a esta tecnología.

2.1. Adaptación de nueva tecnología

El adaptarse a una nueva tecnología involucra un proceso a mediano o largo plazo con el propósito de mejorar el funcionamiento o rendimiento de un sistema. Este periodo de adaptación usualmente comprende un tiempo de aceptación y una etapa de aprendizaje de las nuevas metodologías.

2.1.1. Tiempo de aceptación

La aceptación de Hadoop ha aumentado conforme ha ido madurando su arquitectura; pues como se observó en la figura 1, Hadoop cobró mayor interés a partir de 2015 en gran parte gracias a su bajo costo, pero también por la facilidad de configuración e instalación[8] de sus componentes.

Aun así, la variación del tiempo de aceptación depende de la empresa u organización que esté planificando implementar Hadoop. De tal modo, al personal le podría tomar semanas o meses acoplarse a esta tecnología.

[8] VAUGHAN Jack. *Aún se necesita "mucha madurez" para la arquitectura Hadoop*. http://searchdatacenter.techtarget.com/es/cronica/Aun-se-necesita-mucha-madurez-para-la-arquitectura-Hadoop. Consulta: 13 de marzo de 2018.

A continuación, se mencionan algunos de los principales factores que influyen en la adaptación de una nueva tecnología:

- Factores culturales: referente a la cultura organizacional de las empresas y de los individuos que las conforman. Esto establece ciertos patrones de conducta o metodologías en la resolución de problemas. Puede presentar inconvenientes al adoptar o facilitar la implementación de nuevas tecnologías.

- Factores ligados a la actitud: relativo a la disposición de los miembros de una empresa para enfrentar y superar los retos que trae consigo la adopción de una nueva tecnología. Es de alta importancia para las empresas enfocarse en crear un ambiente idóneo para que la actitud de sus trabajadores no suponga un freno en el avance de nuevos proyectos.

- Factores generacionales: como suele ser el caso, las empresas las integran personas de diversas edades, esto también es cierto para los equipos o departamentos de informática. No siempre se da el caso, pero las personas de mayor edad podrían ocasionar un incremento en el periodo de aceptación, lo cual se debe tomar en cuenta en la planificación inicial al desarrollar soluciones con nuevas tecnologías.

- Factores de costo-beneficio: uno de los puntos más críticos y de mayor importancia, por tal razón es necesario evaluar si los costos de implementación son justificados con base en los beneficios que se obtendrán de la nueva tecnología propuesta.

Para superar de forma exitosa los cuatro factores descritos es de suma importancia que los líderes del cambio o implementación de nuevas tecnologías

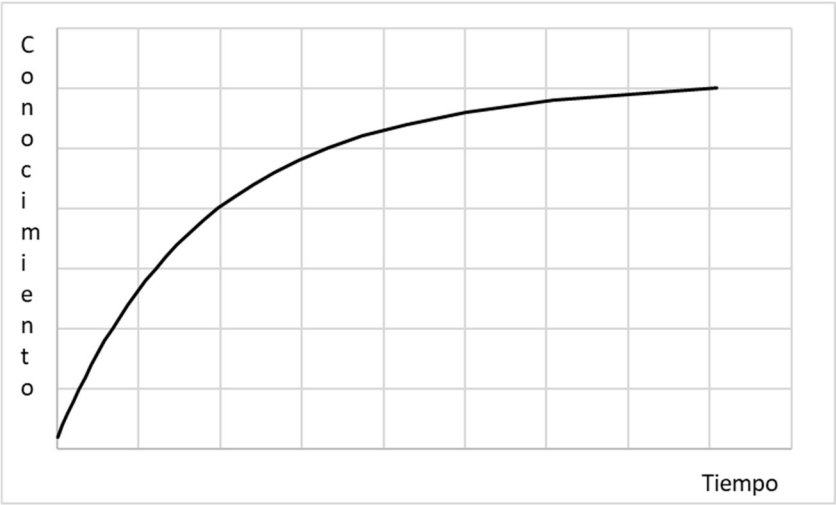
tengan la habilidad para comunicar de forma clara y concisa las ventajas y beneficios de adoptar o acoplar nuevas soluciones.

Los dirigentes deben ser capaces de gestionar y motivar al personal de manera adecuada, e identificar y solventar lo más pronto posible inconvenientes que puedan impedir el avance de la nueva solución propuesta. De esta forma, la etapa de aceptación se disminuye, lo cual garantiza el crecimiento y evolución de las empresas.

2.1.2. Etapa de aprendizaje

Esta fase se refiere a lo que se denomina curva de aprendizaje, término que describe el cúmulo de conocimiento adquirido durante un periodo determinado de aprendizaje. A través de una gráfica, el eje X representa el tiempo y el eje Y el conocimiento adquirido como se observa en la figura 10.

Figura 10. **Representación gráfica de curva de aprendizaje**



Fuente: elaboración propia.

Para el caso de Hadoop, al tratarse de una tecnología relativamente nueva y sobre todo en constante evolución, para muchos, el dominio de esta solución representa un reto muy complejo. La curva de aprendizaje de esta tecnología está relacionada con el conocimiento que se tenga del sistema operativo Linux para tareas de instalación y configuración. También, afecta la destreza que se posea del lenguaje de programación Java para manejo de MapReduce.

Cabe destacar que el conocimiento de Linux y Java es el punto de partida, dado el amplio ecosistema de Hadoop. Y la constante ampliación de características de los distintos componentes que interactúan entre sí para el almacenamiento y procesamiento de datos.

De tal manera, para reducir el tiempo de aprendizaje las empresas deben fomentar procesos graduales y continuos de formación. Así como crear métodos de apoyo entre el personal con mayor facilidad de aprendizaje y aquellos que enfrentan dificultades en la adopción de los cambios de tecnología.

2.2. Facilidad de implementación

La implementación de Hadoop se ha facilitado conforme ha ido madurando, es decir, debido al desarrollo y la constante evolución de herramientas para agilizar la instalación, configuración y administración de los recursos y procesos que puedan integrar un clúster de este tipo.

Además, Hadoop ofrece tres métodos de implementación, lo cual facilita el acercamiento a esta tecnología, pues existen dos escenarios que se pueden

desplegar de forma sencilla y con pocos recursos. A continuación, detalles de cada modo de despliegue[9]:

- Modo local o standalone: es la configuración predeterminada de Hadoop y bajo este diseño los procesos son no distribuidos, es decir, se ejecutan en una sola máquina virtual de Java. Esta forma de despliegue es útil para efectuar pruebas y tareas de *debugging* de código de aplicaciones.

- Modo pseudodistribuido: en este método de implementación cada proceso (NameNode, DataNode, JobTracker, TaskTracker) de Hadoop emplea su propia máquina virtual de Java en un mismo servidor. Este modo es utilizado para validar las interacciones que ocurren en un clúster completamente distribuido sin necesidad de emplear más de un servidor.

- Modo completamente distribuido: es el despliegue típico en un entorno de producción. Acá los diversos procesos de Hadoop se ejecutan por separado, es decir en distintos servidores o nodos. Existen varias estrategias para ello, un ejemplo usual es desplegar en un nodo los procesos de NameNode y JobTracker; mientras que en uno o varios nodos se ejecutan las tareas de DataNode y TaskTracker.

No existe un lineamiento estricto que determine las características en cuanto a servidores o hardware se trata, sino más bien la recomendación para las empresas es analizar de forma heurística la combinación de recursos apropiados para sus demandas.

[9] *Hadoop: setting up a single node cluster*. http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html. Consulta: 15 de marzo de 2018.

Un enfoque útil es hacer distinción de los requerimientos según el rol de cada nodo o servidor dentro del clúster Hadoop. A continuación, especificaciones de hardware a modo de guía para implementaciones medianas o grandes de entornos Hadoop[10].

- Nodos Maestro: Recursos para tareas administrativas, específicamente, NameNode y JobTracker.

    - 16 CPU como mínimo, 32 CPU de preferencia.
    - 128 GB de memoria RAM, idealmente contar con 256 GB.
    - Discos duros configurados en RAID, intercambiables en caliente.
    - Fuente de poder redundante.
    - Enlaces Ethernet Gigabit como mínimo.

- Hardware para nodos esclavo: Recursos para almacenamiento (DataNode) y procesamiento (TaskTracker).

    - De 16 a 32 procesadores.
    - De 64 a 512 GB de memoria RAM.
    - De 12 a 24 discos duros de entre 1 a 4 TB de capacidad. No es recomendado configuración en RAID.

Si se trata de un ambiente de pruebas en *standalone* o pseudodistribuido, las siguientes especificaciones cumplen con dicho propósito:

- 2 CPU
- 8 GB de RAM
- 30 GB de disco duro

[10] AVEN, Jeffrey. *Sams teach yourself hadoop in 24 hours*. p. 24-25.

Una consideración adicional es la adquisición de hardware que podría presentar dificultades logísticas, administrativas, entre otras. Sin embargo, como muchas otras tecnologías, se podría aprovechar las ventajas que ofrece la virtualización o bien soluciones en la nube ofrecidas por Google, IBM, Rackspace o Amazon Web Services por mercionar algunas opciones.

En cuanto a software se refiere, para instalar y configurar Hadoop se debe cumplir con los siguientes requerimientos en cada uno de los nodos que conformen el clúster.

- Sistema operativo Linux (Red Hat, Centos, Ubunto, entre otros.). Existe una versión de Hadoop compatible con Windows, pero la mayoría de las herramientas del ecosistema solo son soportadas por plataformas Linux.

- Java Runtime Environment (JRE) 1.7 o superior.

- Java Development Kit (JDK) 1.7 o superior, para compilación de aplicaciones de MapReduce.

- No utilizar Logical Volumen Manager (LVM) debido a la reducción de rendimiento que provoca en los recursos del almacenamiento.

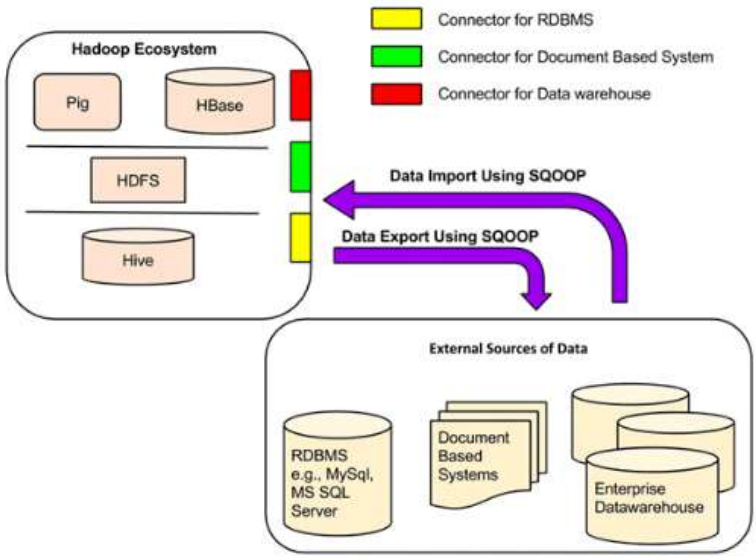
Lo anterior no contempla la instalación de alguno de los múltiples productos que se han agregado al ecosistema Hadoop, de tal modo se aconseja revisar la documentación oficial correspondiente para cerciorarse de los requisitos de instalación respectivos.

2.3. Facilidad de traslado y procesamiento de datos

Trasladar grandes volúmenes de datos a clústeres de Hadoop implica una serie de retos como: mantener y asegurar la consistencia de datos, uso eficiente de los recursos, trabajar con fuentes de datos heterogéneas. Por tal razón, previo a efectuar cargas de datos, se debe analizar el escenario adecuado para solventar esta problemática.

Actualmente se dispone de varias herramientas, cada una de ellas se ajusta a un tipo de fuente de datos en particular. Siendo Sqoop una de las herramientas más utilizadas para trasladar o importar datos a ecosistemas Hadoop, debido a que su diseño le permite efectuar cargas masivas a HDFS desde fuentes de datos estructurados, tales como bases de datos relacionales, *data warehouses* y entornos NoSQL.

Figura 11. **Arquitectura y funcionamiento de Sqoop**



Fuente: RUNGTA, Krishna. *Learn hadoop in 1 day*. p. 59.

La facilidad que ofrece Sqoop para la transferencia de datos es gracias a los diversos conectores que han sido desarrollados. En la figura 11 se ilustra la interacción de varias fuentes de datos con un ecosistema Hadoop por medio de su conector correspondiente.

2.4. Casos de éxito de sistemas que implementan Hadoop

Yahoo!, Facebook y Microsoft han sido de las primeras empresas en adoptar Hadoop para almacenamiento y procesamiento de datos. El éxito que cada uno ha tenido en sus diversos proyectos ha impulsado a otras organizaciones a recurrir a la incorporación e implementación de clúster Hadoop, también con muy buenos resultados.

Existen muchos ejemplos de compañías que han recibido con creces los beneficios de implementar Hadoop, como el aprovechamiento de los datos para generar valor agregado a sus negocios, además de una mejora significativa en el uso de recursos para almacenamiento y procesamiento de datos. A continuación, se describe algunos casos de éxito:

- Geinsinger[11]
    - Descripción de empresa: una de las organizaciones de servicios de salud más grandes en los Estados Unidos, cuenta con más de 3 millones de usuarios. Dispone de un sistema integrado por 30 000 empleados, de los cuales cerca de 1 600 son médicos; cuenta con 12 hospitales y 2 centros de investigación.

[11] Hortonworks. *Geisinger*. https://hortonworks.com/customers/geisinger/. Consulta: 23 de marzo de 2018.

- Caso: combinar los terabytes de datos recolectados de sus clínicas, pacientes y encuestas, para que su personal dispusiera de información útil para mejorar servicios. Enfrentó dos retos en particular: (i) gestionar de forma más rápida y eficiente los datos generados de forma masiva; (ii) reducir costos de almacenamiento de todos los datos para análisis profundos producto de la comparación de datos históricos.

- Solución: consolidar datos estructurados y no estructurados en Apache Hadoop con el objetivo de complementar la solución brindada por su *data warehouse* Teradata. Tras reconocer las ventajas de la tecnología de código abierto, incorporó la opción de Hortonworks Connected Data Platforms. De tal manera su sistema integró más de 30 terabytes de información sobre sus pacientes.

- Resultados: almacenamiento y procesamiento de 30 terabytes de datos. Ahorro de 2 millones de dólares al reemplazar el *data warehouse* Teradata y reducción de medio millón de dólares en costos de mantenimiento al eliminar dicho *data warehouse*.

- Cisco WebEx[12]

  - Descripción de empresa: se enfoca en servicios para conferencias a través de la web. Soporta más de 26 mil millones de minutos en conferencias cada mes. Su servicio de audio y vídeo ayuda en la conexión y colaboración entre personas alrededor del mundo.

[12] Cloudera. *Cisco WebEx improves the customer experience and customer ratings*. https://www.cloudera.com/more/customers/cisco.html. Consulta: 02 de abril de 2018.

- Caso: contar con datos sobre inconvenientes en el servicio, estadísticas de clientes y análisis de fraude en fuentes separadas. Esto complicaba la combinación y correlación de datos, lo cual limitaba la capacidad de la compañía para analizar la experiencia de los usuarios. Además de la complicación de gestionar el creciente volumen de datos recibidos por telemetría.

- Solución: implementación de un repositorio central de datos por medio de Apache Hadoop. De tal modo es posible combinar y correlacionar datos para mejorar la toma de decisiones. Y por medio de algoritmos de *machine learning* en Apache Spark se mejora el monitoreo y detección de fraudes.

- Resultados: con la nueva plataforma de datos centralizados, es posible indagar y profundizar hasta la raíz de algún inconveniente presentado. Esto permite aislar la causa y tomar acciones de forma más rápida para mejorar la experiencia de los usuarios. Por medio de los algoritmos en Apache Spark se automatizó la generación de reglas de detección de fraudes. Y se redujo a un décimo los costos de mantenimiento de la solución en Hadoop comparado con opciones tradicionales de *data warehouse* e inteligencia de negocios. Además, se tiene soporte para gestionar cerca de 1 petabyte de datos.

- Empresa dedicada a la fabricación de dispositivos móviles[13]

    - Descripción de la empresa: compañía encargada de fabricar teléfonos móviles y desarrollar una plataforma móvil popular. Su objetivo es producir dispositivos de alta calidad que garantice a los usuarios conexión y acceso al mejor contenido.

    - Caso: debido a la evolución de los teléfonos móviles, es necesario recolectar, almacenar, explorar y analizar cada vez más grandes volúmenes de datos no estructurados. Por tal razón, está compañía notó que su *data warehouse* en Oracle Real Application Clusters (RAC) no tenía la suficiente capacidad de procesamiento de datos. Pues era necesario comparar volúmenes masivos de datos sin estructura para determinar el rendimiento de los dispositivos.

    - Solución: integración de Hadoop y HBase para complementar Oracle. Por medio de Sqoop se traslada datos entre Hadoop y Oracle, y viceversa. En Oracle se mantiene información de algunos meses para tareas de reportería inmediata. Y en Hadoop se almacena todos los datos para análisis históricos por medio de Hive.

    - Resultados: identificación de interdependencias complejas los diversos componentes de los dispositivos móviles. Por ejemplo, a través de Hadoop se ha logrado analizar la influencia de un

[13] Cloudera. *Driving innovation in mobile devices with cloudera and oracle*. https://www.cloudera.com/content/dam/www/marketing/resources/case-studies/driving-innovation-in-mobile-devices-with-cloudera-and-oracle.pdf.landing.html. Consulta: 04 de abril de 2018.

componente de hardware sobre una anomalía en determinado software, lo cual ha permitido corregir la falla de forma rápida. El uso de Hadoop y Oracle permite contar con plataformas con propósitos y funcionalidades complementarias.

3. COSTOS A CONSIDERAR PARA LA INTEGRACIÓN DE HADOOP

En las siguientes secciones se detalla acerca de los costos de implementación y de infraestructura que se deben considerar para la integración de Hadoop a una plataforma existente o bien como parte de la planificación de pruebas de concepto de esta tecnología.

3.1. Costo de implementación

Al tratarse de un proyecto de código abierto la descarga, instalación, configuración y uso del software de Apache Hadoop o alguna herramienta relacionada con el ecosistema no representa costo alguno. Sin embargo, algunas compañías ofrecen plataformas basadas en Hadoop a través de licenciamiento o servicios de soporte.

3.1.1. Licenciamiento

El licenciamiento de software es definido como un acuerdo legal que estipula las cláusulas de uso del producto y define los derechos tanto del propietario como del usuario final del software[14]. Cualquier software se debe licenciar antes de implementar, sobre todo si su uso será en ambientes de producción, para no incurrir en irregularidades legales.

[14] UNCG. *Information technology services. software licensing*. https://its.uncg.edu/Software/Licensing/. Consulta: 05 de abril de 2018.

Hay diversos tipos de licenciamiento de software a los cuales las compañías podrían adherirse, tomar como base o bien combinar según consideren. Para las empresas es importante tener total claridad del tipo de licenciamiento al que se subscriben para aprovechar al máximo las características del software y respetar los límites estipulados en el acuerdo adquirido.

A modo de tener una noción básica, a continuación se listan y describen de forma breve los principales tipos de licenciamiento en la industria del software:

- Licencia propietaria

El desarrollador del software concede permiso para instalar, configurar y utilizar una o más copias del software. La mayoría de software se licencia bajo estos términos, en los cuales se estipula que la forma de uso, reproducción y distribución son facultades del propietario del software.

- Licencia pública

Términos bajo los cuales se publican los proyectos de código abierto. Permite ejecutar el software sin ninguna restricción. Aprueba la modificación de código bajo la premisa que el resultado debe ser también de acceso público, pero no impide que se cobre por tales mejoras o adaptaciones.

- Licencia indefinida

Licencias sin fecha de expiración, es decir, permite el uso indefinido de la versión del software adquirido sin pagar algún monto adicional al costo inicial. El

utilizar o actualizar el software a una versión más reciente representa un desembolso monetario adicional para adquirir nuevo licenciamiento.

- Licencia renovable

Estipulan el uso del software por un periodo de tiempo, típicamente, mensual o anual. Tras vencer el plazo se debe renovar la licencia por otro lapso o bien desinstalar el software. Es típico que al renovar se obtenga licenciamiento para una versión más actualizada del software.

El proyecto de Apache Hadoop se administra bajo licenciamiento público indefinido o de código abierto[15] y como se anotó en su momento, esto permite el uso o implementación en uno o más servidores sin incurrir en costos de software. Empero, compañías como Cloudera, MapR y Hortonworks ofrecen distribuciones comerciales de Apache Hadoop por medio de licencias de soporte o mantenimiento que se describirán más adelante.

En los manejadores de bases de datos relacionales, como Oracle, por ejemplo, dispone de varias ediciones del software de base de datos y dos modalidades de licenciamiento que se adaptan según las necesidades o limitaciones presupuestarias. En tal sentido es de vital importancia analizar la alternativa que mejor se adapta al propósito del sistema. A continuación, una descripción de cada edición de base de datos Oracle disponible:

[15] Apache Software Foundation. *Apache license*. http://www.apache.org/licenses/LICENSE-2.0. Consulta: 05 abril de 2018.

- Oracle Database Standard Edition 2

Disponible a partir de la versión de Oracle Database 12c Release 1 (12.1.0.2). Para versiones anteriores se dispone de Standard Edition One y Standard Edition. Dentro de sus principales características está la facilidad de uso, potencia, alto rendimiento para grupos de trabajo, para aplicaciones de departamento y aplicaciones Web.

- Oracle Database Enterprise Edition

Provee alto rendimiento, disponibilidad, escalabilidad y seguridad para aplicaciones críticas de alto volumen de transaccionalidad (OLTP), *data warehouse* con consultas intensivas y aplicaciones de internet de constante demanda. Esta edición dispone de todos los componentes del software de base de datos y cabe mencionar que se puede extender su capacidad por medio de opciones y paquetes descritos en la tabla III.

- Oracle Database Express Edition

Es una edición gratuita de entrada para explorar Oracle Database, fácil de instalar y administrar. Se puede instalar en cualquier equipo, con la limitante que utiliza como máximo 1 GB de RAM, un solo CPU y permite almacenar hasta 11 GB de datos. La versión actual de esta edición, también denominada XE, es 11g Release 2 y cabe mencionar que el soporte es a través de foros en línea.

- Oracle Database Personal Edition

Esta edición soporta entornos de desarrollo e implementación de un solo usuario. Compatibilidad completa con Standard y Enterprise Edition. Incluye todos los componentes y opciones de la edición Enterprise a excepción de Real Application Clusters (RAC).

La otra modalidad de licenciamiento de software de base de datos Oracle es por usuario nombrado. En este caso el precio es determinado por una cantidad fija o predeterminada de usuarios que harían uso de la base de datos. Se considera como usuario cualquier conexión a la base de datos, por ejemplo: sensores, máquinas o personas. Dependiendo de la edición se aplica ciertos mínimos de usuarios nombrados, a continuación, se detalla condiciones para las ediciones más utilizadas en la industria:

- Oracle Database Standard Edition 2: mínimo 10 licencias de usuarios nombrados por servidor. Aplicable en servidores con máximo de 2 sockets y 16 CPUs.

- Oracle Database Enterprise Edition: requiere un mínimo de 25 licencias por procesador o bien el número total de usuarios, cualquier que sea el máximo de estas dos cifras.

En la tabla I se observa el precio por usuario nombrado y su costo de actualización y soporte:

Tabla I. **Licenciamiento de base de datos Oracle por usuario nombrado**

Edición	Licencia por usuario nombrado ($)
Standard Edition 2	350,00
Enterprise Edition	950,00
Personal Edition	460,00

Fuente: Oracle Technology Global Price List. http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf. Consulta: 07 de abril de 2018.

La métrica de procesadores se usa con regularidad en entornos donde no es sencillo contabilizar a los usuarios. También, se emplea cuando la modalidad de usuarios nombrados no es rentable. No se entrará en detalle, pero existe un factor para determinar la cantidad de procesadores a licenciar[16]. De tal modo, en la tabla II se dan pormenores de este tipo de licenciamiento:

Tabla II. **Licenciamiento de base de datos Oracle por procesador**

Edición	Licencia por procesador ($)
Standard Edition 2	17 500,00
Enterprise Edition	47 500,00
Personal Edition	23 000,00

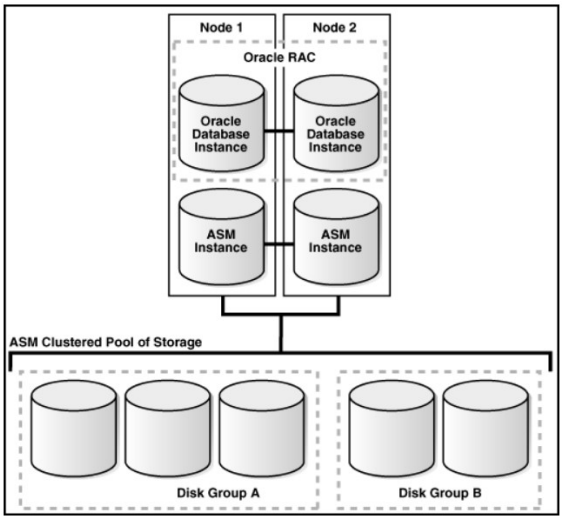
Fuente: Oracle Technology Global Price List. http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf. Consulta: 07 de abril de 2018.

[16] Database Licensing. http://www.oracle.com/us/corporate/pricing/databaselicensing-070584.pdf. Consulta: 07 de abril de 2018.

Algunas opciones y paquetes disponibles para la edición Enterprise son relevantes en cuanto al procesamiento y manejo de los datos, sobre todo cuando se trata de volúmenes considerables de información. A continuación, se describe de forma breve lo mencionado:

- Real Application Clusters: esta opción permite integrar varios servidores o nodos como parte de una misma base de datos, lo cual es posible por medio de almacenamiento compartido conformado por grupos de discos administrados por ASM (*automatic storage management*). Al implementarlo se obtiene alta disponibilidad y escalabilidad como parte de los beneficios principales de RAC. En la figura 12 se ilustra una arquitectura típica de Oracle RAC.

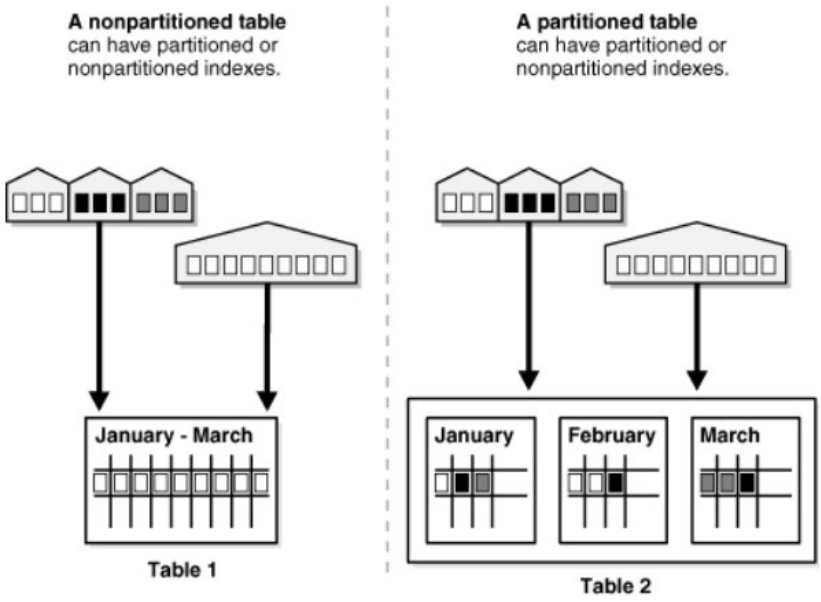
Figura 12. **Oracle RAC de dos nodos**



Fuente: Database 2 Day + Real Application Clusters Guide. https://docs.oracle.com/database/121/TDPRC/intro_tdprc.htm#TDPRC648. Consulta: 08 de abril de 2018.

- Particionamiento: permite dividir tablas o índices en varias piezas de menor tamaño, especialmente útil al manejar grandes volúmenes de datos. Al contar con subdivisiones se mejora tiempos de respuesta y permite plantear varias estrategias de administración.

Figura 13. **Tabla no particionada y particionada en Oracle Database**



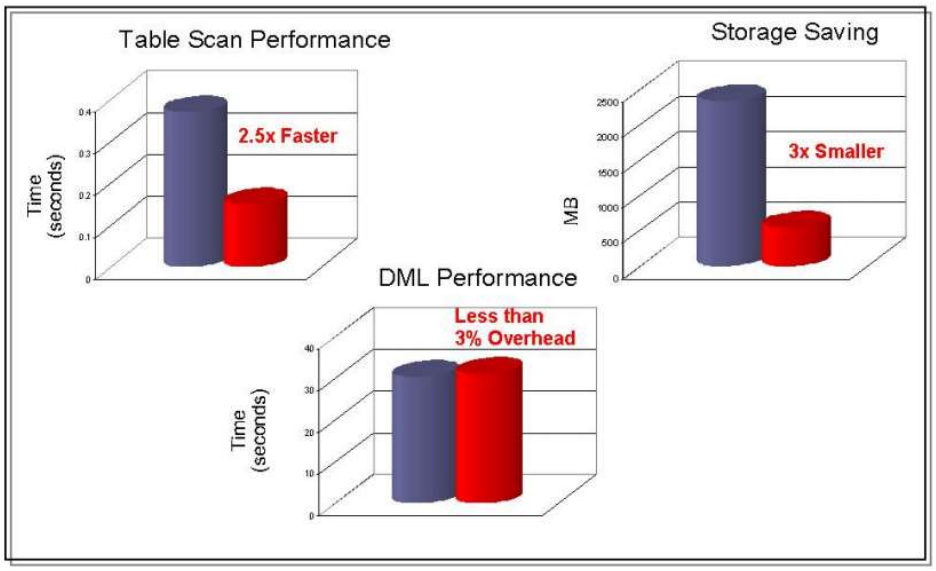
Fuente: *Database VLDB and Partitioning Guide.* https://docs.oracle.com/database/121/VLDBG/GUID-5AE7BBD6-02C1-4DB4-BB5B-B4E5B4C96FAD.htm#VLDBG14024. Consulta: 08 de abril de 2018.

En la figura 13 se ilustra que la tabla 1, no particionada, puede contar con índices particionados o no, esto también aplica cuando la tabla está particionada como es el caso de la tabla 2.

- Compresión avanzada: existen varios tipos de compresión que no se cubrirán a detalle, sin embargo, el propósito común es reducir cantidad

de almacenamiento empleado por tablas, índices, entre otros objetos de la base de datos. A su vez maximiza los accesos a disco a costa de incrementar uso de CPU en actividades de compresión.

Figura 14. **Impacto de compresión en el rendimiento de la base de datos**



Fuente: *Oracle Advanced Compression, an Oracle White Paper*. p 5. http://storage02.brainsonic.com/customers2/oracle/document/infra/pdf/Livre%20Blanc%20Oracle%20Advanced%20Compression%2011g%20UK.pdf. Consulta: 09 de abril de 2018.

Como se observa en la figura 14, el tiempo de barrido es dos y medio veces más veloces, se tiene un ahorro de espacio significativo y los tiempos de las transacciones no se ven afectados sobremanera.

- Paquete de diagnóstico: conformado por un conjunto de herramientas para monitorear y analizar en tiempo real el rendimiento de una base de datos. Permite generar una serie de reportes a partir de datos

almacenados en un repositorio centralizado. Con estos reportes se facilita la identificación de problemas de rendimiento a nivel de red, disco, CPU, RAM, identificar consultas costosas, entre otros elementos.

- Paquete de optimización: ofrece herramientas que automatizan el proceso de optimización, principalmente, de consultas. Facilita esta tarea por medio de asesores que, a través de estadísticas de rendimiento, recomienda acciones para disminuir tiempos de respuesta, por ejemplo, la creación de índices con los campos correspondientes. Este paquete al igual que el de diagnóstico se enfocan en presentar de forma clara y sencilla el comportamiento actual y el resultado de ejecutar las sugerencias de los diversos asesores.

Tabla III. **Licenciamiento de opciones y paquetes para Oracle Database Enterprise Edition**

Edición	Licencia por usuario nombrado ($)	Licencia por procesador ($)
Real Application Clusters	460,00	23 000,00
Particionamiento	230,00	11 500,00
Compresión Avanzada	230,00	11 500,00
Paquete de Diagnostico	150,00	7 500,00
Paquete de Optimización	100,00	5 000,00

Fuente: *Oracle Technology Global Price List.* http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf. Consulta: 09 de abril de 2018.

En la tabla III se presentan los costos de licenciamiento de las opciones y paquetes mencionados. Son de gran utilidad para facilitar medidas proactivas por medio de Oracle RAC, particionamiento y compresión; asimismo, permite

acciones reactivas por medio del análisis comparativo del rendimiento pasado y actual de la base de datos para identificar las acciones correctivas.

3.1.2. Soporte o mantenimiento

El soporte consiste en la asistencia en línea o presencial para solucionar errores, acceso a notas técnicas y parches para la resolución de brechas de seguridad en el software, en este caso de base de datos Oracle o Hadoop. El tiempo de respuesta depende de la criticidad del incidente o de la importancia del ambiente para el negocio.

Cabe destacar que no es obligatorio contratar servicios de soporte Oracle pero se corre riesgos de ser afectados por alguna vulnerabilidad o de no contar con asistencia para resolver alguna anomalía de la versión o edición del software licenciado. Existe una comunidad en línea donde se puede consultar o revisar, de forma gratuita, respuestas a casos planteados en caso no se tenga soporte directo con Oracle.

En la tabla IV se listan los costos de soporte Oracle por usuario nombrado según la edición de la base de datos:

Tabla IV. **Actualización y soporte de base de datos Oracle por usuario nombrado**

Edición	Actualización de licencia y soporte ($)
Standard Edition 2	77,00
Enterprise Edition	209,00
Personal Edition	101,20

Fuente: *Oracle Technology Global Price List*. http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf. Consulta: 10 de abril de 2018.

Para actualizar el software de base de datos Oracle es necesario adquirir licencias de la nueva versión. Si se da o no el caso anterior, en caso se tenga licenciamiento por procesador se debe considerar los costos por soporte presentados en la tabla V.

Tabla V. **Actualización y soporte de base de datos Oracle por procesador**

Edición	Actualización de licencia y soporte ($)
Standard Edition 2	3 850,00
Enterprise Edition	10 450,00
Personal Edition	5 060,00

Fuente: *Oracle Technology Global Price List*. http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf. Consulta: 10 de abril de 2018.

En dado caso se tenga licencias para hacer uso de opciones y paquetes se debe presupuestar conforme a lo expuesto en la tabla VI para su actualización y soporte Oracle.

Tabla VI. **Actualización y soporte de opciones y paquetes para Oracle Database Enterprise Edition**

Edición	Licencia por usuario nombrado ($)	Licencia por procesador ($)
Real Application Clusters	121,00	5 060,00
Particionamiento	50,60	2 530,00
Compresión avanzada	50,60	2 530,00
Paquete de diagnostico	33,00	1 650,00
Paquete de optimización	22,00	1 100,00

Fuente: *Oracle Technology Global Price List*. http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf. Consulta: 11 de abril de 2018.

En el caso de los proyectos Apache, como Hadoop, existe una comunidad muy sólida que facilita documentación y material para la instalación, configuración, ajustes y resolución de problemas debido a algún error en el código del software. Esto se encuentra disponible en línea en cualquier momento de forma gratuita.

Por ello, es usual que las comunidades de proyectos de código abierto empleen foros, blogs, correos y canales de internet para el intercambio de ideas, resolución de errores en el código y propuestas de nuevas funcionalidades o características en las aplicaciones. Esto conlleva registrarse en alguno de los canales de comunicación para plantear dudas o sugerencias.

El tiempo de respuesta o soporte de la comunidad es variable, podría tomar minutos o incluso días para solventar algún caso. A partir de esto surgen empresas que integran nuevas características o bien ofrecen una variante del software, en este caso Hadoop, acompañado de servicios de soporte o mantenimiento.

Tabla VII. **Costos de soporte técnico 24/7**

Distribuidor	**Costo anual por servidor ($)**
Cloudera	10 000,00
MapR	6 000,00
Hortonworks	3 500,00

Fuente: *Hortonworks, Cloudera or MapR?* https://thisdataguy.com/2015/10/01/hortonworks-cloudera-or-mapr/#price. Consulta: 12 de abril de 2018.

En la tabla VII se presenta las compañías cuyas distribuciones de Hadoop tienen mayor aceptación y utilización en la actualidad. Además, se observa el costo anual por cada servidor o nodo que conlleva el soporte técnico respectivo. Para tener una referencia, a continuación, se describe cada empresa y las opciones que ofrece en su propuesta de Hadoop.

- Cloudera

Fundada en 2008 por anteriores miembros de Google, Yahoo!, Oracle y Facebook. En 2009 se integró Doug Cutting, cocreador de Hadoop al equipo de trabajo. Esto convierte su propuesta o distribución en la más antigua de todas.

La visión de Cloudera es construir un ecosistema basado en tecnologías de código abierto. Para ello ha construido varias herramientas que automatizan

tareas y facilitan la gestión de volúmenes masivos de datos por medio de Hadoop, Spark, Impala, entre otros proyectos Apache.

Uno de los productos que más resalta es Cloudera Enterprise Data Hub. Esta herramienta permite desplegar y administrar clústeres Hadoop de forma simple. Facilita agregar o remover nodos de forma dinámica de los clústeres Hadoop. Brinda sistemas de monitoreo y la capacidad de generar reportes sobre el rendimiento del clúster Hadoop.

Otros productos que ha desarrollado Cloudera son Data Engineering y Data Science, que reúne las herramientas necesarias para procesamiento de datos por lotes o en tiempo real, análisis predictivo, e incluso implementación de algoritmos de *machine learning*.

Operational DB y Analytic DB también son alternativas que ha desarrollado Cloudera. La primera de estas opciones se basa en Apache Hbase y Apache Spark para procesamiento de datos estructurados y no estructurados. En tanto la segunda herramienta está enfocada en soluciones de inteligencia de negocios por medio de Apache Impala.

En la Tabla VIII se detallan los precios de licenciamiento por nodo de cada uno de los principales productos ofrecidos por Cloudera. Cabe mencionar que estos precios incluyen soporte técnico.

Tabla VIII. **Opciones de Licenciamiento de Productos Cloudera**

Producto de Cloudera	Licenciamiento por nodo ($)
Data Engineering	4 000,00
Data Science	4 000,00
Operational DB	6 000,00
Analytic DB	8 000,00
Enterprise Data Hub	10 000,00

Fuente: *Flexible pricing and licensing options*. https://www.cloudera.com/products/pricing.html. Consulta: 13 de abril de 2018.

- MapR

Conformada en 2014 por miembros procedentes de varias compañías, como Google, EMC Corporation, Veoh entre otros. Contribuye constantemente al desarrollo de los proyectos HBase, Pig, Hive y ZooKeeper[17]. Su objetivo principal es crear una plataforma unificada para Big Data, procesamiento y análisis predictivo en tiempo real. Además, ha implementado su propia versión de sistema de archivos distribuido llamado MapR-FS, que es una alternativa a HDFS.

Dispone de una base de datos NoSQL denominada MapR-DB, la cual es compatible con Apache Hadoop y está diseñada para manejar grandes volúmenes de datos. También, cuenta con MapR Streams, consiste en una herramienta para análisis de flujos de datos en tiempo real por medio de Hadoop y Spark.

[17] HARRIS, Derrick. *Why MapR Is Right to Give Back to Apache Hadoop*. https://gigaom.com/2011/06/01/why-mapr-is-right-to-give-back-to-apache-hadoop/. Consulta: 13 de abril de 2018.

A pesar de su corto tiempo en la industria, ha logrado valiosas alianzas con grandes empresas. Una de ellas es Amazon, que ofrece como servicio ediciones de MapR como parte de Elastic MapReduce de Amazon Web Services. Otro socio es Google, que por medio de MapR se logró procesar 1,5 trillones de bytes en un minuto empleando 2 103 máquinas virtuales en Google Compute Engine[18].

Los precios de los productos y soporte de MapR no son públicos. En cuanto a este último se refiere, ofrece acceso a documentación respectivo de cada producto. Una base de datos con notas y procesos para solventar errores y una plataforma muy activa de foros denominada MapR Converge Community[19].

En la tabla IX se presenta un resumen de las características de cada uno de los tres tipos de soporte ofrecidos por MapR. Los tiempos de respuesta o prioridad son acorde a acuerdos ajustables.

[18] METZ, Cade. *Google teams with prodigal son to bust data sort record*. https://www.wired.com/2013/02/google-mapr-data-sort-record/. Consulta: 13 de abril de 2018.

[19] Converge Community. https://community.mapr.com/. Consulta: 13 de abril de 2018.

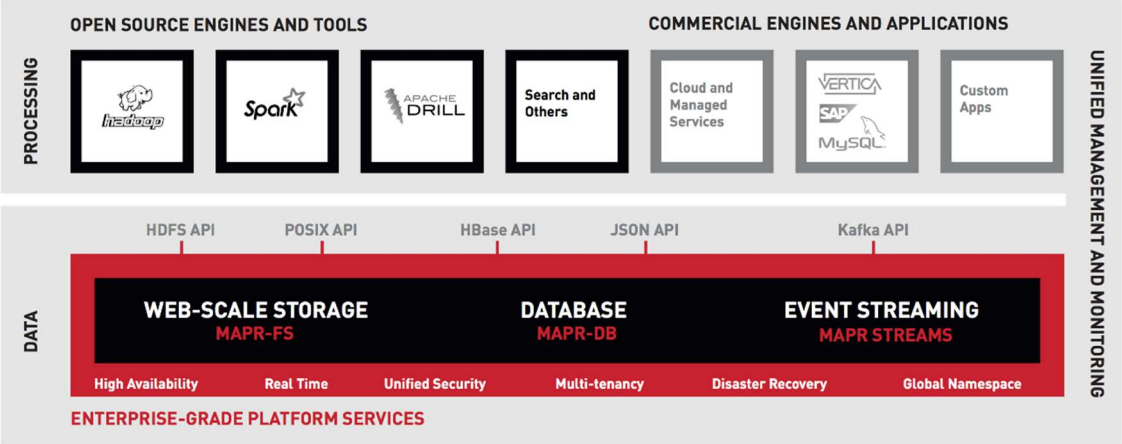
Tabla IX. **Resumen de soporte de MapR**

Característica	Soporte comunitario	Soporte estándar	Soporte TAM (Technical account management)
Actualización de productos	Sí	Sí	Sí
Entrenamiento a demanda	Sí	Sí	Sí
Documentación	Sí	Sí	Sí
Acceso y recursos comunitarios	Sí	Sí	Sí
Base de datos de conocimiento	Sí	Sí	Sí
Soporte comunitario	Sí	Sí	Sí
Creación de casos		Sí	Sí
Inicio de sesión al portal		Sí	Sí
Corrección urgente de bugs		Sí	Sí
Soporte 24 horas		Sí	Sí
Soporte telefónico 24x7		Sí	Sí
Designación de personal de soporte			Sí
Soporte predictivo y proactivo			Sí
Optimización de clúster y guía en proceso de actualización			Sí

Fuente: *MapR Support Overview*.
https://mapr.com/support/s/supportoverview?language=en_US. Consulta: 14 de abril de 2018.

MapR ofrece dos ediciones de su plataforma: Converged Community Edition y Converged Enterprise Edition. La primera de ellas es gratuita y de la segunda, como ya se mencionó no hay acceso a precios o costos.

Figura 15. **MapR Converged Data Platform**



Fuente: *A platform engineered for next-generation applications*. https://mapr.com/datasheets/mapr-converged-data-platform/. Consulta: 14 de abril de 2018.

Como se aprecia en la figura 15, las plataformas integran soluciones de código abierto como Apache Hadoop, Spark, Apache Drill, entre otros. Y es posible complementar con herramientas o aplicaciones comerciales como SAP, MySQL, etc. Además, del acceso a las implementaciones propias de MapR como lo son MapR-FS, MapR-DB y MapR-Streams.

No se dará detalle técnico, pero las APIs cobran especial relevancia al facilitar la interconexión entre las diferentes herramientas y el desarrollo de aplicativos para acceso, consumo o análisis de los datos o recursos.

Es importante hacer mención que la edición gratuita contempla ciertas restricciones según *end user license agreement* (EULA) de MapR[20]. Podría estar sujeto a un tiempo límite, a un periodo de prueba o bien a la prohibición de reventa de alguna implementación con base a este software gratuito.

- Hortonworks

Nace en 2011, integrada por un grupo de 24 ingenieros que pertenecieron al equipo original de Hadoop en sus inicios en Yahoo!. Ha establecido alianzas importantes con Microsoft, RedHat, SAP, entre otras compañías con la intención de habilitar una plataforma Hadoop sencilla de utilizar e implementar[21].

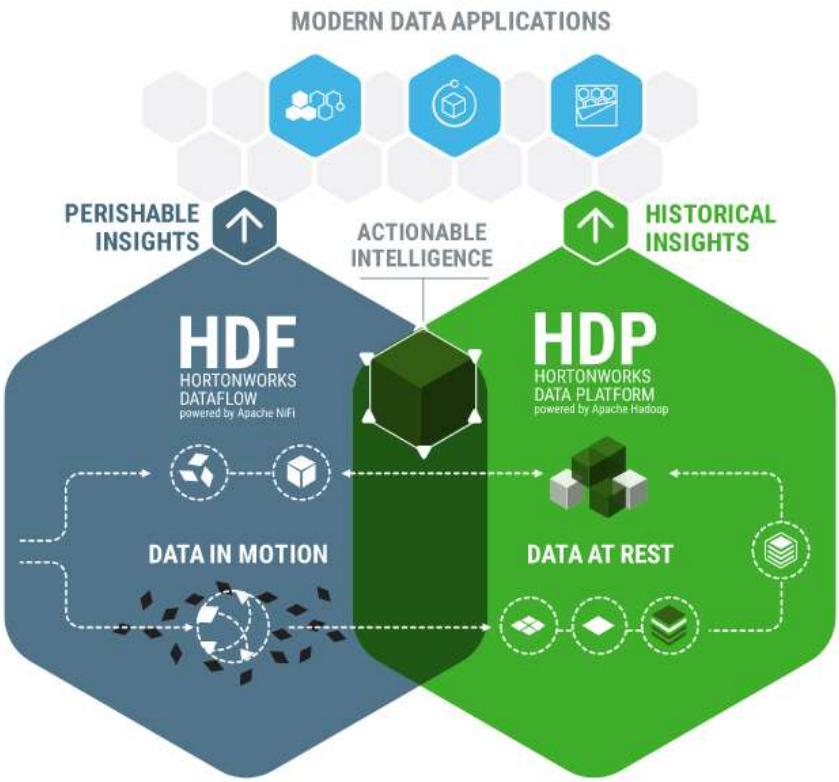
La idea de Hortonworks es promover el uso de software de código abierto, por tal razón no existen acuerdos de licenciamiento para los productos que ofrece. Dejando como opcional el contrato de servicios de soporte, consultoría o capacitación.

La propuesta de esta compañía gira en torno a herramientas para administración de flujo de datos a través de DataFlow y de datos estáticos por medio de Data Platform tal como se observa en la figura 16.

[20] *Mapr end user license agreement*. https://mapr.com/resources/eula/. Consulta: 17 de abril de 2018.

[21] *Quick facts about the business, technology and business teams*. https://hortonworks.com/about-us/quick-facts/. Consulta: 17 de abril de 2018.

Figura 16. **Hortonworks Connected Data Platforms**



Fuente: *Hortonworks Connected Data Platforms.* https://hortonworks.com/products/data-platforms/. Consulta: 18 de abril de 2018.

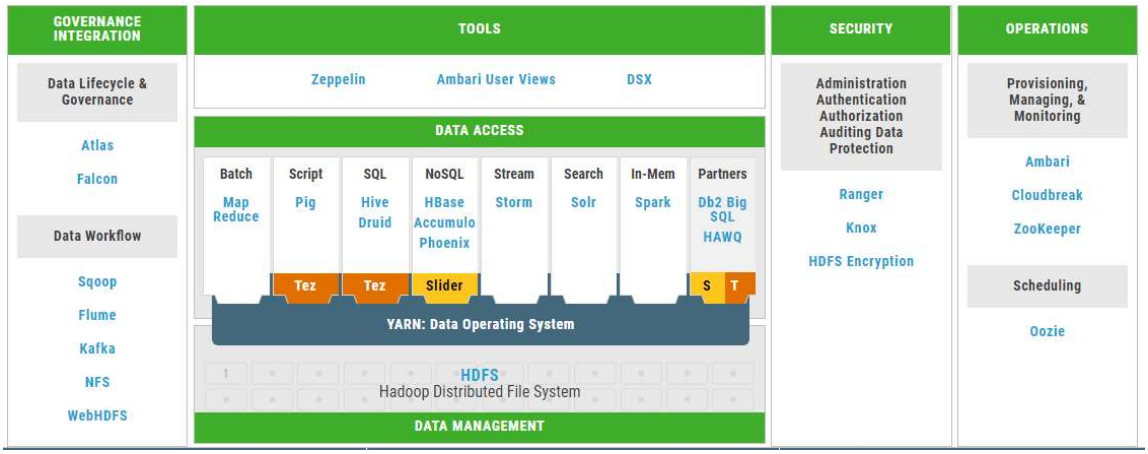
DataFlow provee mecanismos para recolectar, refinar y analizar datos en tiempo real. Esta solución integra Apache Nifi, Apache Kafka, Apache Storm y Druid[22]. Puede ser implementado en servidores en sitio o en la nube, además ofrece una interfaz gráfica que facilita todas estas tareas de administración de datos.

[22] *Hortonworks DataFlow (HDF).* https://hortonworks.com/products/data-platforms/hdf/. Consulta: 18 de abril de 2018.

En tanto, Hortonworks Data Platform (HDP) es una distribución de Apache Hadoop basada en una arquitectura centralizada por medio de YARN. Esta plataforma es capaz de interactuar con múltiples herramientas del ecosistema Hadoop. Por ejemplo, Sqoop y Flume para manejo de flujos de datos; Zeppelin y Ambari como herramientas de gestión de clústeres; uso de MapReduce, Pig, Hive, Hbase, Solr para procesamiento y almacenamiento de grandes volúmenes de datos.

Siguiendo la filosofía de código abierto, emplea como base YARN para administración de recursos y HDFS como sistema de archivos distribuido. Además, ofrece mecanismos para garantizar la seguridad de los datos almacenados y procesados, por ejemplo, por medio de HDFS Encryption. En la figura 17 se tiene una visión general de HDP.

Figura 17. **Hortonworks Data Platform (HDP®)**



Fuente: *Hortonworks Data Platform (HDP®).* https://hortonworks.com/products/data-platforms/hdp/. Consulta: 20 de abril de 2018.

Hortonworks ofrece opciones de soporte Enterprise[23] para Data Platform y DataFlow. No se tiene acceso público a precios o costos de las subscripciones, que en este caso son de renovación anual. Y se dispone de tres tipos de soporte: Enterprise, Enterprise Plus y Flex. La asistencia es 24x7 y podría ser en línea o por teléfono, el tiempo de respuesta va desde una hora hasta un día, esto depende de la severidad del caso.

Cada una de las subscripciones de soporte mencionadas garantiza solución de incidentes en las plataformas y cada uno de los componentes que las conforman. Además de la asistencia en la puesta en marcha de soluciones o pruebas de concepto sobre estas herramientas.

3.1.3. Capacitación de personal

Al implementar una nueva tecnología es de suma importancia capacitar al personal responsable de implementar o administrar las plataformas o soluciones, en este caso sobre Big Data, que al estar compuesto por una serie de mecanismos y herramientas para el almacenamiento y procesamiento de volúmenes masivos de datos la complejidad de aprendizaje se incrementa.

En el caso de Apache Hadoop y su ecosistema de software de código abierto, aunque existe basta documentación en línea de cada herramienta suele ser muy complejo iniciar con el aprendizaje y dominio de los productos. De tal forma, empresas como Cloudera, MapR y Hortonwork compilan y afinan el material para hacerlo más sencillo y digerible de una manera paulatina y lógica. Esto también es aplicable para el caso de Oracle y toda su gama de soluciones.

[23] *Enterprise Support Subscription*. https://hortonworks.com/services/support/enterprise/. Consulta: 20 de abril de 2018.

El principal propósito de dichas empresas es impartir el material de tal modo que prepare a los estudiantes para tomar exámenes de certificación, lo cual avala un conocimiento mínimo del manejo de las herramientas. Es probable que se dispongan de varias rutas o ramas de certificación, es decir, no es necesario dominar todas las soluciones.

De tal modo se dispone de cursos en línea, la mayoría de las veces son sesiones en tiempo real impartidas por instructores certificados. Otra modalidad, denominada entrenamiento a demanda, brinda sesiones grabadas lo cual permite a los alumnos avanzar a su propio ritmo u horario. Y dependiendo del país o región es probable que Oracle, Cloudera, MapR o Hortonworks cuenten con partners, empresas asociadas o distribuidores locales, avalados para impartir de forma presencial el material de los cursos.

Oracle cuenta con una extensa colección de cursos, que no se mencionaran excepto aquellos considerados como básicos para la administración de la base de datos y el manejo de Big Data. En la tabla X se brindará precios relacionados a cada modalidad, costo de certificación y duración de los cursos correspondientes a los temas o áreas mencionadas.

Tabla X. **Cursos de Oracle Database y Big Data**

Curso	Training on Demand	Live Virtual Class ($)	Classroom Training ($)	Examen de Certificación ($)	Duración
Oracle Database SQL Workshop I	NA	612,00	660,00	125,00	3 días
Oracle Database 12c Administration Workshop	935,00	1 020,00	1 100,00	150,00	5 días
Oracle Database 12c Backup and Recovery	935,00	1 020,00	1 100,00	150,00	5 días
Oracle Database 12c Managing Multitenant	NA	408,00	440,00		2 días
Oracle Big Data Fundamentals	935,00	1 020,00	1 100,00	150,00	5 días
Oracle NoSQL Database for Administrators	374,00	408,00	440,00		2 días

Fuente: *All Certifications*. https://education.oracle.com/pls/web_prod-plq-dad/db_pages.getpage?page_id=632. Consulta: 23 de abril de 2018.

Las certificaciones de Hortonworks tienen un costo de $250,00 cada una[24]. En cuanto a los cursos se refiere, estos pueden ser presenciales, online y a demanda a través de los denominados *Self-paced*. Dispone de una modalidad mixta o *blended*, esta consiste en acceso por subscripción anual a material de los cursos para avanzar a propio ritmo y sesiones cortas programadas con

[24] Hortonworks Apache Hadoop and Big Data Certifications. https://hortonworks.com/services/training/certification/. Consulta: 23 de abril de 2018.

instructores, ambos casos, en línea. La distribución de los cursos las ha colocado en tres rutas o ramas: desarrolladores, administradores o analistas.

Tabla XI. **Cursos disponibles en Hortonworks University**

Curso	Modalidad	Precio ($)
HDP Overview: Apache Hadoop Essentials	Live Training	700,00
	Live Training	2 800,00
HDP Operations: Administration Foundations	Blended	1 800,00
HDP Operations: Security	Live Training	2 100,00
HDP Operations: Security	Blended	1 350,00
HDP Developer: Quick Start	Live Training	2 800,00
HDP Analyst: Data Science	Live Training	2 100,00
HDF: NiFi Flow Management	Live Training	2 100,00
HDF: NiFi Flow Management	Blended	1 350,00

Fuente: *Hortonworks university*. http://learn.hortonworks.com/#. Consulta: 24 de abril de 2018.

MapR también cuenta con una serie de cursos y certificaciones. Por medio de Academy Essentials se distribuye material gratuito y por medio de subscripciones se tiene acceso a contenido en Academy Pro[25]. Los cursos están relacionados por medio de un rol y se pueden llevar series correspondientes a una herramienta en particular.

En el caso de Cloudera, también imparte cursos de forma presencial o en línea. Y sus cursos a demanda, a través de una subscripción, por lo general ofrecen acceso por 180 días a todo el contenido del curso. Es importante mencionar que ha separado por roles los cursos para hacer más sencilla la ruta de aprendizaje a tomar. Por lo general, el costo de cada curso en Cloudera es de $2 235,00 por estudiante o alumno. En la tabla XII un extracto de cursos impartidos por instructores de Cloudera.

[25] *On-demand Training Options*. https://mapr.com/training/on-demand/. Consulta: 24 de abril de 2018.

Tabla XII. **Extracto de cursos de Cloudera**

Curso	Role	Días
Data Scientist Training	Desarrollador, Analista	4
Data Analyst Training	Analista	4
Cloudera Administrator Training	Administrador	4
Big Data Architecture Workshop	Desarrollador	3

Fuente: elaboración propia.

De tal forma, a través de los cursos diseñados por cada una de estas empresas se facilita la especialización en cada uno de los productos que ofrecen. Lo cual sumado a la experiencia va cimentando los conocimientos adquiridos, lo cual beneficia sobremanera en mayor éxito en la implementación de nuevos proyectos o ajuste de plataformas existentes.

3.2. Costos de infraestructura

En cuanto a infraestructura se refiere, los costos son difíciles de estimar dado que no existe un lineamiento estricto que defina las características de los servidores o hardware. Se añade un grado de complejidad cuando se trata de una implementación nueva, donde no se tiene datos que sirvan como métrica para calcular los recursos necesarios.

Por tal razón surgen propuestas como la de Oracle, que por medio de su sistema de ingeniería conocido como *big data appliance* ofrece una plataforma completa para ejecutar cargas y procesos en Hadoop, Kafka e incluso NoSQL[26]. Entre otras características, ofrece acceso a una opción propia de Oracle

[26] *Oracle Big Data Appliance X7-2.* http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdataappliance-datasheet-1883358.pdf. Consulta: 01 de mayo de 2018.

llamada Big Data SQL, esto permite en una sola consulta SQL combinar datos provenientes de Hadoop, NoSQL, Kafka y base de datos Oracle. Lo anterior facilita sobremanera la integración de las diferentes plataformas de administración de datos o información.

En definitiva, Big Data Appliance reduce la complejidad que representa seleccionar cada uno de los componentes de hardware que conformaría la infraestructura para una solución de Big Data. Y dado que cuenta con una plataforma integrada de software, provoca que el tiempo se pueda aprovechar para desarrollar aplicaciones para procesamiento y análisis de datos en beneficio del negocio.

A continuación, se presentan detalles de software incluido en Big Data Appliance X7-2:

- Sistema operativo: Oracle Linux 5, Oracle Linux 6 u Oracle Linux 7

- Cloudera Enterprise 5 – Data Hub Edition con Soporte para:

    - Distribución de Cloudera de Apache Hadoop (CDH)
    - Cloudera Impala
    - Cloudera Search
    - Apache Hbase y Apache Accumulo
    - Apache Spark
    - Apache Kafka
    - Cloudera Manager con soporte para:
        - Cloudera Navigator
        - Cloudera Back-up y Disaster Recovery (BDR)

- Oracle Perfect Balance
- Oracle Table Access para Hadoop
- Oracle Java JDK 8
- MySQL Database Enterprise Server – Advanced Edition
- Oracle Big Data Appliance Enterprise Manager Plug-In
- Oracle R Distribution
- Oracle NoSQL Database Community Edition
- Opcional Oracle Big Data Connector
    - Oracle SQL Connector para Hadoop
    - Oracle Loader para Hadoop
    - Oracle XQuery para Hadoop
    - Oracle R Advanced Analytics para Hadoop
    - Oracle Data Integrator

En la tabla XIII se detalla el costo de Big Data SQL y de Big Data Connectors mencionados con anterioridad; cabe resaltar que ambos componentes son opcionales en el *appliance*.

Tabla XIII. **Precios de software opcional Big Data**

Software	Precio de licenciamiento ($)	Actualización de licencia y soporte ($)	Métrica de licenciamiento
Big Data Connectors	2 000,00	440,00	Procesador
Big Data SQL	4 000,00	880,00	Disco

Fuente: *Oracle Engineered Systems Price List*. http://www.oracle.com/us/corporate/pricing/exadata-pricelist-070598.pdf. Consulta: 02 de mayo de 2018.

El Big Data Appliance puede ser adquirido en su versión completa o inicial. La diferencia radica en la cantidad de nodos de cómputo, uno cuenta con 18 y el otro con 6, respectivamente. Las características del hardware de cada nodo no varían entre las versiones del rack; en la tabla XIV se presenta pormenores al respecto.

Tabla XIV. **Específicaciones de hadware de Big Data Appliance X7-2**

Full Rack	Starter Rack
18 nodos de cómputo/almacenamiento	6 nodos de cómputo/almacenamiento
Por Nodo: • 2 x 24 Core (2.1 GHz) Intel Xeon 8160 • 8 x 32 GB RAM DDR4-2666 MHz expandibles a 1.5 TB • 12 x 10 TB 7,200 RPM High Capacity SAS Drives • 2 x 150GB M.2 SATA SSD Drives • 2 x QDR 40 Gb/seg InfiniBand Ports • 1 x Dual-port InfiniBand QDR CX3 (40 Gb/sec) PCIe HCA • 1 x Built-in RJ45 1 Gigabit Ethernet port • 2 x 32 Port QDR InfiniBand Leaf Switch o 32 x InfiniBand ports o 8 x 10 Gb Ethernet Ports • 1 x 36 Port QDR InfiniBand Spine Switch	

Continuación de la tabla XIV.

○ 36 x InfiniBand Ports • Additional Hardware Components included ○ Ethernet Administration Switch ○ 2 x Redundant Power Distributions Units (PDUs) ○ 42U rack packaging • Spares Kit included ○ 1 x 10 TB High Capacity SAS disk ○ InfiniBand cables

Fuente: *Oracle Big Data Appliance X7-2.* http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdataappliance-datasheet-1883358.pdf. Consulta: 02 de mayo de 2018.

Una plataforma de este tipo, que provee el hardware y software de forma integral, permite desarrollar de forma rápida soluciones de Big Data, pues no se invierte tiempo en calcular los recursos necesarios; esto aumenta las probabilidades de que los proyectos sean exitosos y de máximo beneficio para las empresas. En la tabla XV se encuentra el precio de cada versión del *rack*, también el costo de soporte premier anual del hardware y del sistema operativo que se debe incluir dentro del presupuesto.

Tabla XV. **Precios de Oracle Big Data Appliance**

Sistema de ingeniería	Precio ($)	Soporte Premier Oracle para sistemas de ingeniería ($ anual)	Soporte Premier Oracle para sistema operativo ($ anual)
Big Data Appliance X7-2 Full Rack	679 750,00	81 570,00	54 380,00
Big Data Appliance X7-2 Starter Rack	249 750,00	29 970,00	19 980,00

Fuente: *Oracle Engineered Systems Price List*. http://www.oracle.com/us/corporate/pricing/exadata-pricelist-070598.pdf. Consulta: 03 de mayo de 2018.

El costo de instalación del Big Data Appliance ronda los $14 000,00. De tal manera, la versión inicial del *rack* representa $ 313 700,00 de inversión. Esta no es una cifra viable para muchas empresas y en muchos casos desalienta e impide realizar pruebas de concepto, en este caso concreto, de Apache Hadoop.

Por ello, una alternativa es construir un sistema propio para tareas con grandes volúmenes de datos y dada la visión de Hadoop, esta es una vía muy razonable, aunque pudiese representar una mayor cantidad de tiempo a invertir para ajustar una plataforma de este tipo. En la tabla XVI se tiene las especificaciones y precio de un servidor con similar capacidad de procesamiento que un nodo del Big Data Appliance.

Tabla XVI. **Especificaciones de servidor HP**

Concepto	Descripción
Producto	Smart Buy Proliant DL380 Gen10 6126
Procesador	Intel Xeon 8160 (24 core, 2.10 GHz, 24 MB, 33 MB, 115 W)
Memoria Estándar	32 GB
Discos duros instalados	Ninguno, suporta hasta 8 SFF
Fuente de poder	2 x 500 W
Costo	$ 7 747,00

Fuente: elaboración propia.

En tal sentido, se supondrá construir una plataforma de 6 nodos para Hadoop, que vendría a ser la versión inicial del Big Data Appliance. Al precio del servidor se debe sumar otras componentes de hardware y soporte de sistema operativo; la tabla XVII expone los detalles correspondientes.

El monto de la inversión inicial en hardware para una plataforma propia es de $ 179 133,00. Habría que considerar el costo de la instalación y de mantenimiento de los componentes de la infraestructura. Y a esto habría que sumarle, de ser el caso, el precio de soporte ofrecido por MapR, Hortonworks, Cloudera u otra alternativa.

Tabla XVII. **Costo de clúster**

Componente	Cantidad	Precio unitario ($)	Precio total ($)
Smart Buy Proliant DL380 Gen10 6126	1	7 747,00	7 747,00
32 GB RAM adicional	4	756,00	3 024,00
HP 3TB 12G SAS 7.2K rpm LFF (3.5 inch) Midline 1-year warranty hard drive	12	859,00	10 308,00
Costo por Nodo			**21 079,00**
Costo por server	6	21 079,00	126 474,00
Voltaire InfiniBand 4X QDR 36-port Reversed Air Flow Managed Switch	3	12 999,00	38 997,00
Costo total hardware para clúster			**165 471,00**
Red Had Enterprise Linux Server Standard (up to 1 guest) per socket pair (3 years)	6	2 277,00	13 662,00
Costo total por Hardware y soporte de SO			**179 133,00**

Fuente: PLUNKETT, Tom. *Oracle Big Data Handbook*. p 72.

Aunque el precio disminuye, como se expone en algunos artículos[27], a largo plazo la opción de un sistema de ingeniería como el Oracle Big Data Appliance es más conveniente, no solo en términos de costo monetario sino también de tiempo y recursos humanos.

Es de resaltar, que la inversión inicial sigue siendo demasiado alta para empresas pequeñas. Dejando la opción de Oracle Big Data Appliance, en gran cantidad de casos, para corporaciones o instituciones financieras multinacionales. Y aun cuando se construya una plataforma propia con menor capacidad a la expuesta haciéndola económicamente factible, en la mayoría de los casos no se tiene la experiencia suficiente para desarrollar proyectos de éxito en corto plazo.

[27] DIJCKS, Jean-Pierre. *Price comparison for big data appliance and hadoop*. https://blogs.oracle.com/datawarehousing/updated:-price-comparison-for-big-data-appliance-and-hadoop. Consulta: 07 de mayo de 2018.

Una solución a lo anterior, para grandes y pequeñas empresas, es optar por contratar servicios es la nube. Empresas como Oracle, Amazon, Microsoft, entre otras, han construido plataformas muy completas y estables, sobre las cuales ofrecen diversos servicios, entre ellos para Big Data. Dichos servicios se suelen pagar conforme al uso, típicamente por hora y en algunos casos por mes.

Tabla XVIII. **Servicios de Oracle Cloud Big Data**

Producto	Pago por consumo ($)	Mensual flexible ($)	Métrica	Detalles adicionales
Oracle Big Data Cloud Service – Pack inicial – 3 nodos	29,0322	19,3548	Entorno alojado por hora	3 nodos con 32 OCPU, 256 GB de RAM y 48 TB de almacenamiento por nodo
Oracle Big Data Cloud Service – Nodos adicionales	9,6774	6,4516	Nodo alojado por hora	1 nodo con 32 OCPU, 256 GB de RAM y 48 TB de almacenamiento por nodo
Oracle Big Data Cloud Service – OCPU adicionales	0,2400	0,1600	OCPU por hora	OCPU adicionales activadas. Se requiere una compra mínima de 32 OCPU, con un máximo de 480 OCPU por instancia de clúster. Las OCPU añadidas deben formar múltiplos de 32

Fuente: *Oracle Cloud Big Data*. https://cloud.oracle.com/es_ES/big-data/big-data/pricing. Consulta: 08 de mayo de 2018.

Al optar por la nube de Oracle, en etapa de prueba de concepto lo más conveniente es el pago por consumo. A modo de ejemplo, si se asume 4 horas diarias de pruebas, durante 20 días sería un monto de $ 2 322,576 por servicios. Una vez superada la etapa de evaluación debe realizarse un análisis

para determinar la tarifa apropiada, pero en definitiva representa una suma menor a la de un Appliance o plataforma propia.

En el caso de Microsoft Azure, Hadoop y herramientas de su ecosistema (Spark, Interactive Query, Kafka, Storm, HBase) están disponibles por medio de su servicio denominado HDInsight. Dentro de su gama de servidores dispone de nodos optimizados para memoria y de nodos para propósito general.

En la tabla XIX se presenta los precios para los actuales equipos optimizados para memoria, resta mencionar que, únicamente, se detalla precio por hora. El precio total está conformado por costo de licencia/soporte de sistema operativo y el costo de emplear HDInsight.

Tabla XIX. **Precio de HDInsight por nodos optimizados para memoria**

Instancia	CPU	RAM (GB)	OS ($/hora)	HDInsight ($/hora)	Total ($/hora)
D11 v2	2	14	0,149	0,038	0,187
D12 v2	4	28	0,299	0,075	0,374
D13 v2	8	56	0,598	0,150	0,748
D14 v2	16	112	1,196	0,300	1,496

Fuente: *Azure HDInsight pricing*. https://azure.microsoft.com/en-us/pricing/details/hdinsight/. Consulta: 11 de mayo de 2018.

A modo de plantear un ejemplo con nodos optimizados para memoria, si se construye un clúster Hadoop de 6 nodos con instancias D14 v2, en un periodo de 4 horas diarias por 20 días, el costo resultante es de $ 718,08. Y en caso se tuviera activo el clúster de 6 nodos durante un mes completo, el costo mensual sería de $ 6 549,00. Esta tarifa es muy viable sobre todo para pequeñas empresas o pruebas de concepto.

Algunos de los nodos de propósito general se presentan en la tabla XX. De igual manera, se despliega el precio por hora de cada tipo de instancia disponible. Como se observa, el precio por hora es menor a los nodos optimizados para memoria, lo cual significa que se incurre en menos gastos de operación.

Tabla XX. **Precio de HDInsight por nodos de propósito general**

Instancia	CPU	RAM (GB)	OS ($/hora)	HDInsight ($/hora)	Total ($/hora)
A1	1	1.75	0,06	0,017	0,077
A2	2	3.5	0,12	0,033	0,153
A3	4	7	0,24	0,066	0,306
A4	8	14	0,48	0,132	0,612
A5	2	14	0,22	0,033	0,253
A6	4	28	0,44	0,066	0,506
A7	8	56	0,88	0,132	1,012

Fuente: *Azure HDInsight pricing*. https://azure.microsoft.com/en-us/pricing/details/hdinsight/. Consulta: 11 de mayo de 2018.

Ambos tipos de nodos brindan suficiente cantidad de recursos de CPU y RAM no solo para pruebas de concepto sino para ambientes de producción o críticos. Y en el caso de optar por Microsoft Azure Insight se tiene el respaldo de Hortonworks en cuanto a la implementación de Hadoop y sus demás componentes, dada la sociedad entre estas dos compañías de software.

Amazon Web Services, o AWS según sus siglas en inglés, implementa, con base en la propuesta Hadoop de MapR el servicio llamado Amazon EMR (*elastic map reduce*). El precio del servicio varía respecto al tipo de instancia EC2 (*elastic compute cloud*) seleccionada a lo cual se le suma el precio por

Amazon EMR. Cabe mencionar, EC2 se refiere a servicios que proveen capacidad de procesamiento o cómputo en la nube de Amazon.

Al momento se cuenta con tres ediciones de MapR[28]: M7, M5 y M3. No se ahondará en detalles, pero la diferencia radica en cantidad de características presentes en cada edición. En cada una de ellas se encuentra implementada la distribución completa de Apache Hadoop. Siendo M3 una edición gratuita y por ende útil al iniciar con pruebas de procesamiento de Big Data.

Dado que existe una amplia gama de instancias EC2[29] de las cuales podemos seleccionar dependiendo de cantidad de RAM o CPU necesarios, en la Tabla XXI se dan algunos precios, pues se toma como base aquellas instancias que cuentan con 32GB o 64 GB de RAM y 8 o 16 CPU.

Los tipos de instancia denotan algún tipo de optimización, es decir CPU más veloces, mayor cantidad de memoria RAM por instancia o bien disponibilidad de discos más grandes y rápidos para tareas intensivas de almacenamiento.

En el siguiente ejemplo se toma la instancia m5.4xlarge pues brinda un buen balance de capacidad de cómputo (procesador 2,5 GHz Intel Xeon Platinum 8175), memoria RAM y recursos de red. De tal modo, se incurre en un costo de $ 76,80 para un periodo de 4 horas diarias durante 20 días. Y si se necesitara almacenamiento adicional, el costo en S3 por 10 TB de datos es de $ 235,40 mensual.

[28] *Amazon EMR with the MapR Distribution for Hadoop*. https://aws.amazon.com/emr/mapr/. Consulta: 12 de mayo de 2018.

[29] *Amazon EC2 Instance Types*. https://aws.amazon.com/ec2/instance-types/. Consulta: 12 de mayo de 2018.

Tabla XXI. **Precio de Amazon EMR y Amazon EC2**

Tipo de instancia	Instancia	Precio de Amazon EC2 ($/hora)	Precio de Amazon EMR ($/hora)	Precio total ($/hora)
Propósito general	m5.4xlarge	0,768	0,192	0,960
Propósito general	m4.4xlarge	0,800	0,240	1,040
Optimizado para cómputo	c5.4xlarge	0,680	0,170	0,850
Optimizado para cómputo	c4.4xlarge	0,796	0,210	1,006
Optimizado para memoria	r4.2xlarge	0,532	0,133	0,665
Optimizado para memoria	r3.3xlarge	0,665	0,180	0,845
Optimizado para almacenamiento	i3.2xlarge	0,624	0,156	0,780
Optimizado para almacenamiento	d2.2xlarge	1,380	0,270	1,650

Fuente: *Amazon EMR Pricing*. https://aws.amazon.com/emr/pricing/. Consulta: 14 de mayo de 2018.

A través de los servicios en la nube se facilita la implementación de proyectos de procesamiento de datos por medio de Apache Hadoop, pues las empresas pagan, únicamente, por el tiempo que consumen los recursos para ejecutar una o varias tareas. Además, permite hacer pruebas y mediciones rápidas para estimar la cantidad de recursos necesaria. Y la opción de un Appliance o plataforma propia no se descarta, sino es para casos que por políticas no se pueda alojar datos en la nube o algo por el estilo.

En definitiva, uno de los beneficios de los servicios en la nube es la rapidez que otorga para hacer cambios en la infraestructura y la flexibilidad para incrementar o disminuir la cantidad de los servidores para acoplarse a la demanda o carga del sistema o aplicativo. Es sin lugar a duda una opción a considerar si lo que se busca es evaluar, en este caso, el funcionamiento de Apache Hadoop.

4. DESEMPEÑO DEL ESCENARIO HADOOP

En los capítulos anteriores se plantearon las características y ventajas de Apache Hadoop y tomando en cuenta que se trata de una solución de Big Data asequible económicamente, se convierte en una opción al alcance de cualquier tipo de empresa para el procesamiento de grandes volúmenes de datos. En este capítulo se detallan las pruebas realizadas para la medición de rendimiento de Apache Hadoop al ser integrado a base de datos relacionales.

4.1. Escenario de pruebas

A modo de obtener resultados fehacientes las pruebas se realizarán en ambientes de igual características de hardware y sistema operativo. Esto permitirá una evaluación imparcial del procesamiento de datos brindada por Apache Hadoop o bien de Oracle Database como sistema de base de datos relacional.

Los distintos escenarios, que se detallarán más adelante, se desarrollarán contando con las siguientes especificaciones técnicas de hardware y sistema operativo:

- Procesador Intel® Core™ i7
- Memoria RAM de 16 GB
- Disco de estado sólido de 237 GB
- Oracle Linux 7.4

En cuanto a las especificaciones de la base de datos relacional que se empleará para las pruebas, en la tabla XXII se resumen los aspectos más importantes.

Tabla XXII. **Especificaciones técnicas de la base de datos Oracle**

Característica	Detalle
Versión	Enterprise Edition 12.1.0.2
Tipo	Relacional
Mínimo de RAM	4 GB
Mínimo de almacenamiento necesario para binarios	10 GB
ACID	Si
Tamaño mínimo de bloque de datos	2 KB
Tamaño máximo de bloque de datos	32 KB
Cantidad máxima de archivos de datos	65533
Tamaño máximo de un Bigfile Tablespace	128 TB con bloques de 32 KB 32 TB con bloques de 8 KB
Tamaño máximo de un Tablespace tradicional	128 GB con bloques de 32 KB 32 GB con bloques de 8 KB
Sistema Operativo soportado	Linux x86-64, HP-UX Itanium, Microsoft Windows 64-bit, Solaris 64-bit, AIX

Fuente: elaboración propia.

En los escenarios propuestos se hará uso de bloques de 8 KB de tamaño, que suele ser el valor recomendado para sistemas OLTP. También, se tiene contemplado verificar el procesamiento de datos con bloques de 32 KB, que por lo regular es empleado en sistemas OLAP.

En tanto, las especificaciones técnicas de Apache Hadoop se detallan en la tabla XXIII.

Tabla XXIII. **Especificaciones técnicas de Apache Hadoop**

Característica	Detalle
Versión	2.8.3
Tipo	Big Data
Mínimo de RAM	4 GB
Mínimo de almacenamiento necesario para binarios	2 GB
ACID	No
Tamaño mínimo de bloque de datos	1 MB
Tamaño predeterminado de bloque de datos	128 MB
Sistema Operativo soportado	Linux x86-64

Fuente: elaboración propia.

Como se aprecia en la tabla XXIII, no existe un valor máximo para el tamaño de bloque de datos y tampoco se coloca una cantidad máxima de nodos en el clúster Hadoop. Ambos factores son adaptables y/o dependientes del tipo de tareas a ejecutar y se pueden ir optimizando conforme se vayan obteniendo estadísticas de rendimiento. Para el caso de las pruebas propuestas se hará uso del tamaño de bloque predeterminado.

4.1.1. Descripción

Tomando como base las especificaciones de hardware y el sistema operativo detallados con anterioridad, se plantean los siguientes escenarios de prueba:

- Escenario A: base de datos Oracle Single Instance.
- Escenario B: base de datos Oracle Single Instance con opción de compresión avanzada.

- Escenario C: base de datos Oracle RAC de dos nodos con opción de compresión avanzada.
- Escenario D: Apache Hadoop en modo pseudodistribuido.
- Escenario E: Clúster Apache Hadoop de un nodo maestro y dos nodos de datos.

En cada uno de los escenarios se utilizarán 56 GB de datos para las tareas de procesamiento. Los mismos son un subconjunto del modelo presentado por TPC para la evaluación de rendimiento de los denominados sistemas de toma de decisiones[30]. Sistemas que por lo regular administran y procesan volúmenes grandes de datos.

TPC cuenta con un procedimiento oficial para la generación de datos[31], pero dado que se empleará un segmento del total de datos se ha optado por generarlos siguiendo las instrucciones presentadas en el tpcds-kit, disponible en un repositorio público de github[32]. Los 56 GB mencionados contiene datos de dos tablas, su distribución se observa en la tabla XXIV:

Tabla XXIV. **Especificación de tablas de pruebas**

Nombre	Número de columnas	Total de filas	Tamaño (MB)
STORE_SALES	23	432 003 263	57 701
DATE_DIM	28	73 049	10

Fuente: elaboración propia.

[30] *TPC-DS is a Decision Support Benchmark*. http://www.tpc.org/tpcds/. Consulta: 06 de junio de 2018.

[31] *TPC - Current Specifications.* http://www.tpc.org/tpc_documents_current_versions/current_specifications.asp. Consulta: 06 de junio de 2018.

[32] *TPC-DS benchmark kit*. https://github.com/gregrahn/tpcds-kit#tpcds-kit. Consulta: 06 de junio de 2018.

La estructura de los datos de prueba se describe en la tabla XXV. La tabla STORE_SALES la conforman ocho archivos y DATE_DIM está compuesta por un archivo.

Tabla XXV. **Estructura de datos de prueba**

STORE_SALES	DATE_DIM
ss_sold_date_sk (int)	d_date_sk (int)
ss_sold_time_sk (int)	d_date_id (char)
ss_item_sk (int)	d_date (date)
ss_customer_sk (int)	d_month_seq (int)
ss_cdemo_sk (int)	d_week_seq (int)
ss_hdemo_sk (int)	d_quarter_seg (int)
ss_addr_sk (int)	d_year (int)
ss_store_sk (int)	d_dow (int)
ss_promo_sk (int)	d_moy (int)
ss_ticket_number (int)	d_dom (int)
ss_quantity (int)	d_qoy (int)
ss_wholesale_cost (numeric)	d_fy_year (int)
ss_list_price (numeric)	d_fy_quarter_seq (int)
ss_sales_price (numeric)	d_fy_week_seq (int)
ss_ext_discount_amt (numeric)	d_day_name (char)
ss_ext_sales_price (numeric)	d_quarter_name (char)
ss_ext_wholesale_cost (numeric)	d_holiday (char)
ss_ext_list_price (numeric)	d_weekend (char)
ss_ext_tax (numeric)	d_following_holiday (char)
ss_coupon_amt (numeric)	d_first_dom (int)
ss_net_paid (numeric)	d_last_dom (int)
ss_net_paid_inc_tax (numeric)	d_same_day_1y (int)
ss_net_profit (numeric)	d_same_day_1q (int)
	d_current_day (char)
	d_current_week (char)
	d_current_month (char)
	d_current_quarter (char)
	d_current_year (char)

Fuente: elaboración propia.

4.1.2. Métricas a evaluar

Estas variables representan aspectos o métricas que se utilizarán para evaluar el desempeño en el procesamiento de datos en cada uno de los escenarios propuestos con anterioridad.

Tabla XXVI. **Métricas de evaluación**

Tipo	Métrica
Entrada	• Tiempo total de inserciones
Salida	• Tiempo de respuesta • Consumo de recursos

Fuente: elaboración propia.

A continuación, una breve descripción de las métricas mencionadas en la tabla XXVI:

- Tiempo total de inserciones: es el tiempo que conlleva ingresar o cargar el total de datos en la base de datos relacional o clúster de Hadoop.

- Tiempo de respuesta: es el tiempo que le toma a la base de datos o clúster Hadoop responder a consultas ejecutadas.

- Consumo de recursos: es el porcentaje de uso de CPU, memoria RAM y disco utilizado durante la carga y consulta de los datos, respectivamente.

4.2. Comparación de rendimiento

En las siguientes secciones se presentan, comparan y analizan los resultados obtenidos tras procesar los datos de prueba en cada uno de los escenarios mencionados con anterioridad.

4.2.1. Tiempo total de inserciones

En este caso, se detalla en la tabla XXVII el tiempo consumido para la carga de los 56 GB de datos en cada uno de los cinco escenarios de prueba propuestos.

Tabla XXVII. **Tiempos de inserción o carga de datos**

Escenario	DATE_DIM (seg)	STORE_SALES (seg)	Total (seg)	Total (min)
A	180,00	14 340,00	14 520,00	242,00
B	6,00	31 500,00	31 506,00	525,10
C	3,00	47 820,00	47 823,00	797,10
D	0,60	9 000,00	9 001,00	150,00
E	0,60	10 200,00	10 201,00	170,00

Fuente: elaboración propia.

Conforme a los resultados obtenidos, en los escenarios con base de datos relacional (A, B, C) el tiempo de carga aumentó al implementar compresión. Y también aumenta el tiempo al pasar a un ambiente RAC, en este caso debido a que dos nodos acceden de forma concurrente el almacenamiento compartido.

Respecto a los ambientes con Apache Hadoop, al no evaluar la estructura de los datos al insertarlos, se obtiene una mejora significativa en tiempo de

carga. El mejor tiempo de base de datos (A) supera en 92 minutos el mejor tiempo del ambiente con Hadoop (D).

Como se observa, al contar con dos nodos Hadoop se presenta un leve incremento de tiempo de procesamiento de datos, que como se verá más adelante se compensa en tiempo de búsqueda o consulta.

Derivado de lo anterior y de la cantidad total de registros, en la tabla XXVIII se presenta la cantidad de registros procesados por segundo por cada uno de los cinco escenarios.

Tabla XXVIII. **Inserciones por segundo**

Escenario	Registros insertados por segundo
A	29 757
B	13 714
C	9 035
D	48 005
E	42 352

Fuente: elaboración propia.

Con esto se aprecia de mejor forma la capacidad de procesamiento de herramientas de Big Data al no evaluar estructura o esquematización de los datos ingeridos. Se nota el mismo efecto al no hacer uso de compresión en las bases de datos relacionales, aunque conlleva emplear mayor espacio de almacenamiento.

4.2.2. Tiempos de respuesta

Esta medición representa el tiempo que conlleva obtener resultados de cada una de las consultas evaluadas, lo cual se observa en la tabla XXIX. Cabe mencionar que la primera de las consultas consiste en un conteo de registros de la tabla más grande y la segunda se trata de un *join* con algunas condiciones entre los campos de las dos tablas.

Tabla XXIX. **Tiempos de respuesta en consultas (minutos)**

Escenario	Consulta 1	Consulta 2
A	5,08	12,27
B	4,45	7,20
C	2,23	6,33
D	72,40	115,60
E	43,23	76,80

Fuente: elaboración propia.

De la tabla anterior, en todos los escenarios la segunda consulta consume más tiempo para retornar los registros consultados, en promedio se requiere el doble de tiempo. Esto debido a las diferencias comparaciones que debe hacer entre los campos para filtrar los registros de salida.

También es notable el tiempo utilizado en los clústeres de Apache Hadoop. Si se toma como referencia los tiempos obtenidos en el escenario A, se observa en promedio 13 veces más uso de tiempo en el escenario D y 7 veces mayor consumo de tiempo en el escenario E. Este comportamiento es esperado pues en Hadoop los datos son estructurados al consultarlos, mientras que en las bases de datos relacionales ya no es necesario ese paso.

En la tabla XXX se hace un recopilatorio de los tiempos que intervienen para realizar la primera de las consultas. Se aprecia de forma clara que los tiempos son altos al insertar en base de datos relacionales pero bajos al consultar. Lo inverso se presenta en clústeres de Hadoop.

Tabla XXX. **Comparativa de tiempos consulta 1 (minutos)**

Escenario	Tiempo de carga STORE_SALES	Consulta 1	Tiempo resultante
A	239,00	5,08	244,08
B	525,00	4,45	529,45
C	797,00	2,23	799,23
D	150,00	72,40	222,40
E	170,00	43.23	213,23

Fuente: elaboración propia.

En términos generales, los tiempos resultantes son mejores en los ambientes con Apache Hadoop. Y como se denota en la tabla anterior y en la tabla XXXI, los tiempos de consulta disminuyen al contar con mayor cantidad de nodos en el clúster de Hadoop.

Tabla XXXI. **Comparativa de tiempos consulta 2 (minutos)**

Escenario	Tiempo de carga ambas tablas	Consulta 2	Tiempo resultante
A	242,00	12,27	254,27
B	525,10	7,20	532,30
C	797,05	6,33	803,38
D	150,01	115,60	265,61
E	170,01	76,80	246,81

Fuente: elaboración propia.

Lo cual es un buen indicador de que si se trata de consultas para reportes de análisis estadístico o predictivo encontramos en Hadoop una solución factible.

4.2.3. Consumo de recursos

En cada una de las pruebas de los distintos escenarios se estuvo en constante observación del estado de los recursos de cada computadora, en la cual se encontraba operando ya sea la base de datos Oracle o nodo de Apache Hadoop. A continuación, se muestran los resultados obtenidos respecto del consumo de recursos en cada computadora utilizada.

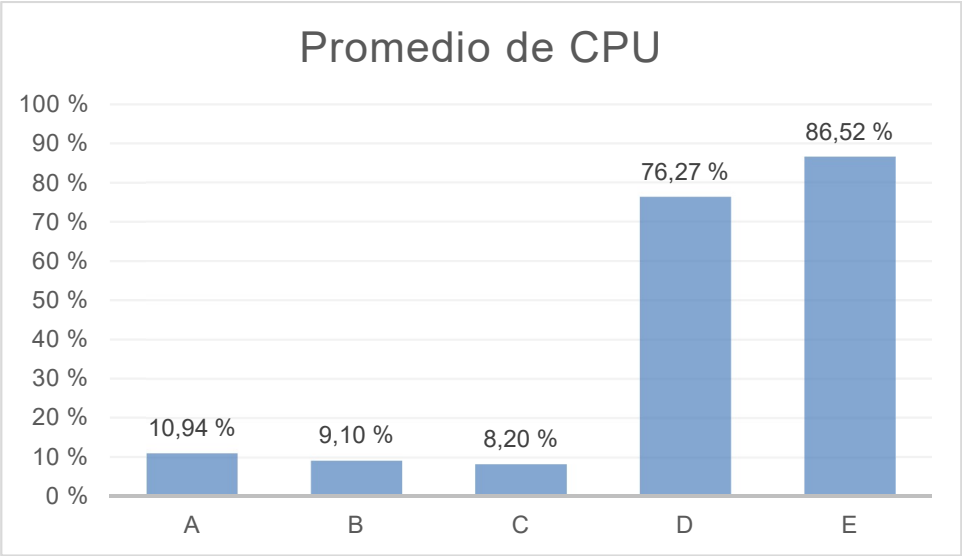
Tabla XXXII. **Porcentaje de recursos consumidos al insertar datos**

Escenario	CPU		RAM		Disco	
	Máximo	Promedio	Máximo	Promedio	Máximo	Promedio
A	20,00	10,94	6,29	5,70	50,50	18,17
B	27,00	9,10	9,31	6,90	90,67	38,50
C	38,00	8,20	7,01	6,21	41,00	17,33
D	99,95	76,27	21,73	18,83	34,67	21,00
E	99,7	87,38	15,53	12,36	2,29	1,29

Fuente: elaboración propia.

En la tabla XXXII se exponen el porcentaje máximo y el promedio de recursos consumidos en cada uno de los escenarios de prueba. La intención de contar con ambos valores es proveer una base para estimar los recursos necesarios para ambientes productivos.

Figura 18. **Consumo promedio de CPU al insertar datos**



Fuente: elaboración propia.

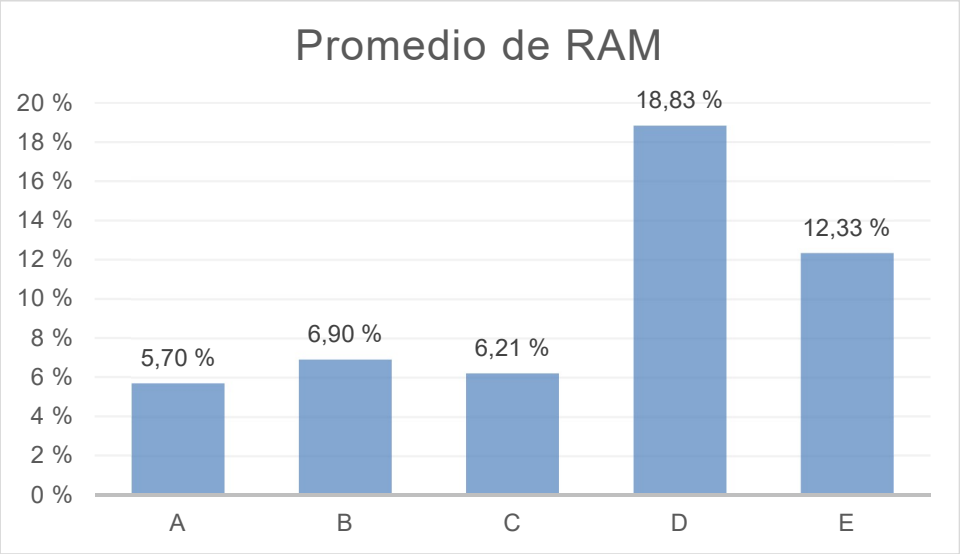
En la figura 18 se observa que en los ambientes con base de datos relacional aumenta el consumo de recursos de CPU y RAM cuando se emplea compresión de datos. En tanto, en los escenarios con Hadoop se mantiene un uso promedio por sobre el 76 % de CPU, esto se debe a la distribución de los archivos en forma de bloques.

Lo mismo sucede con el uso de RAM, pues se observa que en los entornos de Apache Hadoop se duplica e incluso triplica el uso promedio de recursos de memoria en comparativa con las bases de datos relacionales. Esto se aprecia de mejor manera en la figura 19.

Cabe destacar que el uso intenso o constante de CPU y RAM no es indicador que haya algún error o inconveniente en el escenario de prueba.

Desde un punto de vista práctico, tanto el CPU como la RAM son mucho más veloces y baratos que los medios de almacenamiento.

Figura 19. **Consumo promedio de RAM al insertar datos**

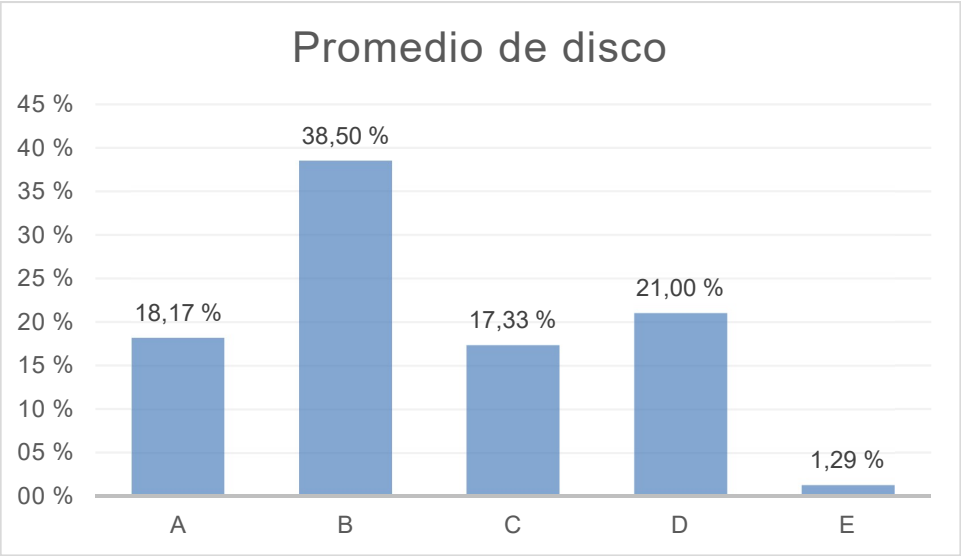


Fuente: elaboración propia.

Y en muchos casos, incluso en ambientes de producción resulta sencillo aumentar cantidad de CPU si se trata de máquinas virtualizadas, práctica común en muchas empresas. Si se tratara de equipos físicos, aumentar memoria RAM ya no representa un reto complejo como en años anteriores.

La nube ofrece muchos de estos beneficios, los cuales cobran mucha importancia al momento de hacer nuevas implementaciones y es muy sencillo cambiar o restructuras los recursos asignados para los diversos aplicativos o servicios.

Figura 20. **Consumo promedio de disco al insertar datos**



Fuente: elaboración propia.

Respecto a los escenarios con Apache Hadoop, como ya se mencionó se observa un uso elevado de CPU y RAM, pero poca actividad en el medio de almacenamiento, sobre todo si se trata de un clúster Hadoop con varios nodos como es el caso del escenario E. Esto debido a que los datos se distribuyen y cada nodo administra una porción del total de datos de forma local.

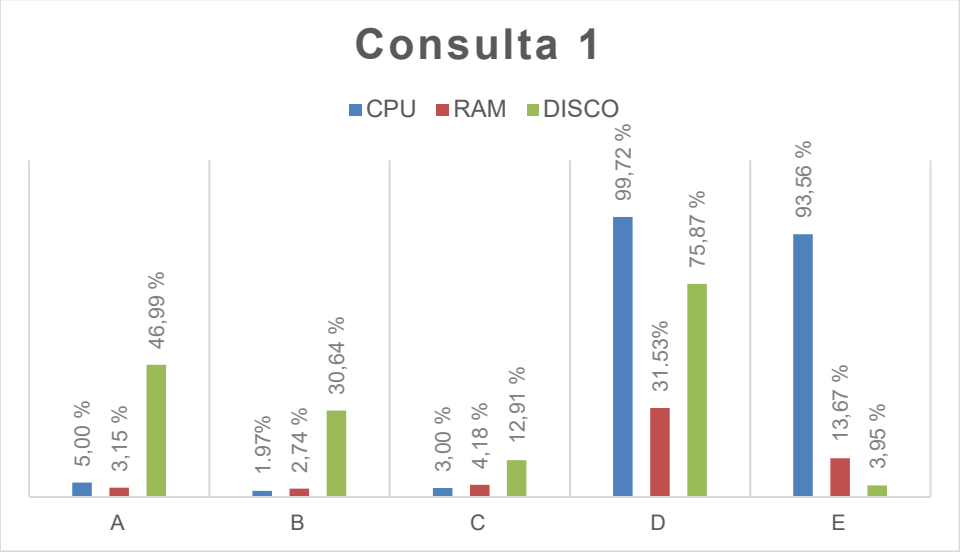
Tabla XXXIII. **Porcentaje de recursos consumidos en consulta 1**

Escenario	CPU		RAM		Disco	
	Máximo	Promedio	Máximo	Promedio	Máximo	Promedio
A	7,00	5,00	3,76	3,15	74,04	46,99
B	5,20	1,97	3,48	2,74	81,51	30,64
C	4,00	3,00	4,83	4,18	22,94	12,91
D	99,95	99,72	53,12	31,53	82,83	75,87
E	99,40	93,56	19,48	13,67	15,59	3,95

Fuente: elaboración propia.

En la tabla XXXIII se presenta los datos recopilados sobre el consumo de recursos de CPU, RAM y disco al momento de ejecutar la primera de las consultas en cada uno de los cinco escenarios de prueba. Y en la figura 21 se ilustra como en los ambientes con base de datos el medio de almacenamiento es el recurso con mayor demanda. En tanto en los escenarios con Apache Hadoop, es el CPU el recurso más utilizado.

Figura 21. **Consumo promedio de recursos en consulta 1**



Fuente: elaboración propia.

Al comparar los dos escenarios con Apache Hadoop se observa alta cantidad de lecturas en disco en el clúster con un solo nodo o en modo pseudodistribuido. Eso se debe a que en el clúster con dos nodos, cada uno de ellos administra de forma local una parte de los datos. El mismo efecto se produce en la RAM. En cuanto al uso de CPU es similar la demanda de este recurso en ambas modalidades de clúster.

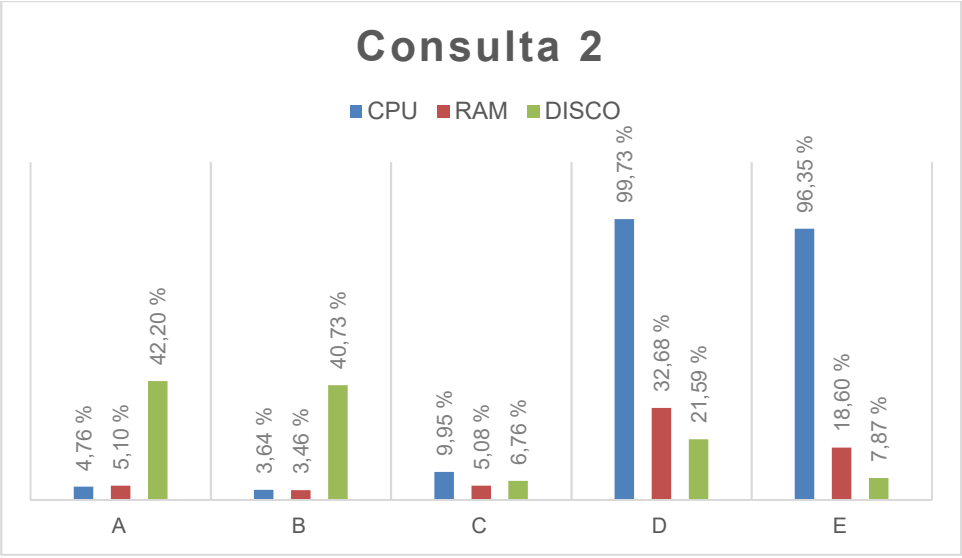
Tabla XXXIV. **Porcentaje de recursos consumidos en consulta 2**

Escenario	CPU		RAM		Disco	
	Máximo	Promedio	Máximo	Promedio	Máximo	Promedio
A	7,40	4,76	6,79	5,10	52,80	42,20
B	9,75	3,64	5,57	3,46	62,60	40,73
C	13,4	9,95	6,15	5,08	17,07	6,76
D	99,98	99,73	53,25	32,68	49,56	21,59
E	99,4	96,35	21,86	18,60	12,51	7,87

Fuente: elaboración propia.

Para la segunda consulta se presenta el mismo comportamiento en el rendimiento y consumo de recursos de CPU, RAM y disco como se presenta en la tabla XXXIV y se observa en la figura 22.

Figura 22. **Consumo promedio de recursos en consulta 2**



Fuente: elaboración propia.

4.3. Análisis de resultados de rendimiento

A raíz de los resultados obtenidos de las distintas pruebas realizadas en cada uno de los escenarios propuestos, se concluye que es factible la integración de Apache Hadoop a entornos con bases de datos relacionales para procesamiento de registros con propósito de análisis estadístico o predictivo.

En la tabla XXXV se presenta un resumen de los aspectos evaluados en los escenarios de prueba, se coloca una calificación de 1 a 5 a cada métrica analizada, siendo 1 el mejor y 5 el peor en desempeño.

Tabla XXXV. **Resumen de pruebas realizadas**

Métrica	Escenario A	Escenario B	Escenario C	Escenario D	Escenario E
Tiempo total de inserción	3	4	5	1	2
Tiempo de respuesta	3	2	1	5	4
Consumo de CPU	3	1	2	5	4
Consumo de RAM	1	2	3	5	4
Consumo de disco	5	3	2	4	1

Fuente: elaboración propia.

En términos generales, el tiempo de inserción o carga de datos es mejor en los escenarios con Apache Hadoop. Este tiempo puede ser mejorado conforme se aumenta la cantidad de nodos en el clúster Hadoop. La complejidad de administrar varios nodos de Hadoop disminuye al implementar soluciones como Apache Ambari o bien recurrir a ambientes en la nube como EMR de AWS o HDInsight de Microsoft Azure.

Aunque en los escenarios de prueba, los ambientes con base de datos superaron los tiempos de los entornos con Hadoop, lo cual es esperado dado que no requiere estructuración de los datos o registros, la ventaja que ofrece Hadoop es la generación de un archivo de salida y remover carga de trabajo de la base de datos, siendo esta última usualmente empleada para operaciones de tipo OLTP.

De tal manera es factible la integración de Apache Hadoop a bases de datos relacionales brindando versatilidad para la administración y procesamiento de datos de cualquier tipo y volumen. Esto es lo que ha captado el interés de empresas como Facebook, Yahoo!, Microsoft, Cisco, entre otras, para analizar a profundidad sus diversas fuentes de datos.

Es importante denotar que también existe desventajas al momento de implementar Apache Hadoop, como el utilizar múltiples herramientas de su ecosistema y por su naturaleza no brinda los beneficios de una base de datos. Por tal razón se complementa y puede coexistir con base de datos relacionales o incluso no relacionales denominadas NoSQL.

CONCLUSIONES

1. Es posible la integración de clúster de Apache Hadoop a base de datos relacionales para el procesamiento y transformación de datos.

2. La naturaleza de los aplicativos que han implementado el uso de Apache Hadoop y que han obtenido beneficios sustanciales son el procesamiento y la transformación de volúmenes amplios de datos sin estructura o para mostrar información en tiempo real.

3. Existen varios *framework* para procesamiento de datos en Apache Hadoop, tales como Apache Spark, Pig, Hive que en la mayoría de los casos MapReduce es la base de estos.

4. El grado de complejidad en la implementación de un clúster de Apache Hadoop, es relativa a los conocimientos de sistema operativo y de programación que posea el personal encargado de integrarla al entorno o arquitectura actual.

5. La implementación de Apache Hadoop no representa costos de licenciamiento, lo cual si aplica para algunas opciones como compresión y particionamiento de algunas bases de datos relacionales.

6. El desempeño de Apache Hadoop mejora conforme a la cantidad de nodos que integran el clúster.

RECOMENDACIONES

1. Efectuar un análisis de viabilidad para no tener ningún inconveniente con la integración de esta tecnología a los entornos o implementaciones existentes.

2. Explorar otras herramientas del entorno de Apache Hadoop, como Apache Ambari, pues a través de esta se facilita la administración de los clústeres; Apache Spark, Hive, Pig para el procesamiento de datos como alternativa a MapReduce.

3. Capacitar al personal con poca experiencia en sistema operativo LINUX y en programación de java o SQL, esto con la finalidad de hacer más cómoda la transición a esta tecnología.

4. Analizar la naturaleza de los datos o el caso de uso para asegurar que la implementación de Apache Hadoop es apta para solventar las diversas problemáticas con datos de diversas estructuras, fuentes o volúmenes.

BIBLIOGRAFÍA

1. Apache Hadoop [En línea]. <http://hadoop.apache.org/>. [Consulta: 05 de marzo de 2018].

2. AVEN, Jeffrey. *Sams teach yourself hadoop in 24 hours*. 1. Estados Unidos de América: Pearson Education Inc, 2018. 496 p.

3. DATE, C.J. *Introducción a los Sistemas de bases de datos*. 7. México: Pearson Educación, 2001. 920 p.

4. EADLINE, Douglas. *Hadoop 2 quick-start guide: Learn the essentials of big data computing in the Apache Hadoop 2 Ecosystem*. 1. Estados Unidos de América: Pearson Education Inc, 2016. 304 p.

5. PLUNKETT, Tom; MACDONALD, Brian; NELSON, Bruce; SUN, Helen; HORNICK, Mark F; LAKER, Keith; MOHIUDDIN, Khader; HARDING, Debra L; SEGLEAU, David; MISHRA, Gokula; STACKOWIAK, Robert. *Oracle Big Data handbook plan and implement an enterprise big data infraestructure*. Estados Unidos: McGraw-Hill Education, 2014. 464 p.

6. RUNGTA, Krishna. *Learn hadoop in 1 day*. India: Publicación Independiente, 2017. 105 p.