UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

FACULTAD DE INGENIERÍA

HERRAMIENTAS, RETOS, OPORTUNIDADES, SEGURIDAD Y TENDENCIAS DEL BIG DATA

T E S I N A

PARA OBTENER EL TÍTULO DE

INGENIERO EN COMPUTACIÓN

PRESENTA

OMAR SÁNCHEZ VILLASEÑOR

DIRECTOR

DRA. ROSA MARÍA VALDOVINOS ROSAS

TOLUCA DE LERDO, ESTADO DE MÉXICO, ABRIL DE 2019

RESUMEN

La generación masiva de datos es un fenómeno que se encuentra en constante aumento, debido a que la adquisición y el almacenamiento de datos son cada vez más accesibles, sencillos y baratos, además gracias a la constante participación de las personas en el uso de redes sociales, transacciones bancarias, compras en línea, expedientes médicos, entre otras actividades. En consecuencia, se generan cada vez más datos lo que provocó el surgimiento del concepto Big Data.

El incremento en el volumen de adquisición de datos, la variedad de ellos y la velocidad a la cual estos se generan trajo como consecuencia que las herramientas tradicionales para el tratamiento de la información redujeran su efectividad para llevar a cabo este tipo de tareas. Ante esto, fue necesario buscar otro tipo de alternativas capaces de contener esa escala de información y lograr procesarla en tiempos admisibles. Actualmente, en las organizaciones uno de los activos más importantes que se tiene es la información, no obstante, para darle un valor agregado se requiere tener conocimiento de estrategias que favorezcan su análisis y contribuyan a la toma de decisiones.

Big Data se encuentra relacionada de manera importante con almacenar enormes cantidades de datos, aunque almacenar datos no es algo reciente, ahora se hace de manera desmesurada, ya que son cada vez más los dispositivos dotados de un sensor con capacidad de capturar información sobre las actividades diarias de los seres humanos, sus transacciones, registros médicos, sus preferencias de navegación, entre otros, es por lo que es importante conocer la importancia que tienen los datos en la actualidad, para poder identificar nuevas oportunidades, proveer de mejores servicios acorde con las necesidades reales de los clientes y tomar mejores decisiones.

Big Data crece de manera constante, aparecen nuevas herramientas que la integran, es posible encontrar diversos proyectos de software libre, así como organizaciones que lo están adoptando para aprovechar las enormes cantidades de datos que tienen almacenadas que por si solas no representan más que un problema, es la razón por la que es importante conocer sobre el tema y comprender sobre las ventajas que puede proporcionar en distintas áreas y conocer que posibilidades brindar para resolver diferentes tipos de problemas.

Al respecto, la presenta tesina brinda un panorama amplio con relación al Big Data, da a conocer su importancia, su ciclo de vida, algunas aplicaciones que se le están dando en México y el mundo, las herramientas de software libre sobresalientes en el mercado en el mercado actual, retos relacionados con sus características, proceso y adopción. También se presentan algunas oportunidades y beneficios que ofrece frente a otras tecnologías, beneficios para las organizaciones, beneficios para la educación, se habla sobre seguridad en Big Data, futuras tendencias para los datos, dispositivos, tendencias de inversión, hardware y software, así como algunas desventajas.

ÍNDICE

ÍNDICE DE FIGURAS **Página**

ÍNDICE DE TABLAS

INTRODUCCIÓN

Las Tecnologías de la Información y Comunicación (TIC) llegaron a revolucionar de manera importante la vida de los seres humanos y el mundo de la información *(Gil, 2016)*. En la actualidad, se generan millones de datos diariamente, tendencia que sigue creciendo de manera exponencial y, en consecuencia, ha dado un cambio de paradigma en las estrategias operativas dentro de las instituciones a nivel global *(Rama, 2016)*.

El crecimiento de la información digital es relativamente reciente, en el año 2000 sólo el 25% de la información mundial se encontraba almacenada de manera digital y el 75% se encontraba en papel. Actualmente más de un 98% de la información se encuentra digitalmente *(Gil, 2016)*, esto en gran parte a la utilización de "dispositivos móviles", sistemas de localización GPS, redes sociales, transacciones financieras, declaraciones de impuestos, registros médicos y el creciente número de dispositivos dentro del internet de las cosas *(Fragoso, 2012)*.

Actualmente se generan cantidades de datos sin precedentes, se cuentan con enormes repositorios de datos, sin embargo, el almacenarlos y el procesarlos ya no es el único problema, si no lograr una explotación adecuada de estos, conseguir un mejor aprovechamiento y volverlos valiosos. Para atender esta necesidad surge Big Data que permite:

- **Explorar grandes volúmenes de datos:** las compañías enfrentan el reto de dar valor a los datos para tomar mejores decisiones, mejorar las operaciones, reducir riesgos, llevar a cabo nuevas implementaciones y adoptar nuevas tecnologías.
- **Análisis en las operaciones:** permite analizar grandes cantidades de información en tiempo real acerca de operaciones importantes, lo cual es necesario obtener respuestas rápidas, lograr predecir futuras tendencias, e innovar en distintas áreas *(Balachandran, 2017)*.
- **Seguridad:** analiza inmensos repositorios de datos, para detectar nuevos patrones de conducta, predicción de acontecimientos y amenazas.

IMPORTANCIA DE LA TEMÁTICA

En las organizaciones, las grandes cantidades de datos están siendo utilizadas en su mayoría para propósitos comerciales, sin embargo, la implementación de Big Data representa un reto importante para las organizaciones ya que en primer instancia se deben dar a conocer las técnicas para el tratamiento de información, capacitación de personal, analizar información confiable, eliminar la intuición, crear nuevos roles y no oponerse a la innovación, por ello es importante conocer y prepararse para futuros cambios, nuevas necesidades del entorno, conocer los retos, las tendencias y oportunidades que ofrece esta tecnología, y de este modo ser capaces de competir en un mercado tan complejo.

Realizar el tratamiento de datos es un desafío aún cuando el almacén de datos es posible analizarlo con arquitecturas tradicionales como los sistemas gestores de bases de datos, lo que se convierte en un desafío particular cuando los métodos tradicionales pierden su efectividad debido al formato y al gran volumen de los datos. Por lo antes expuesto en la presente tesina plantea la investigación documental orientada a realizar una revisión bibliográfica y del estado del arte en el contexto de Big Data, para dar a conocer los retos tecnológicos, herramientas, futuras tendencias que se tienen en esta área del conocimiento para el Ingeniero en Computación y carreras afines.

OBJETIVOS

De lo anterior se propone el objetivo de realizar un análisis documental sobre los retos, tendencias, oportunidades, seguridad, proceso y aplicación que ofrece el Big Data al manejo y tratamiento de la información, así como a la toma de decisiones, mediante la consulta de artículos en revistas internacionales, investigaciones académicas, con el objetivo de conocer acerca de Big Data, sus características, aplicaciones, herramientas, beneficios, tendencias, así como las ventajas y desventajas frente a otras tecnologías.

De manera particular se proponen los siguientes objetivos:

- Realizar una reseña histórica sobre el Big Data, con la meta de conocer sus antecedentes y cómo logró ser una tecnología competente y aceptada en el área de las ciencias computacionales.
- Entender lo que significa Big Data, sus características, así como saber hasta qué punto se puede considerar o no como tal.
- Conocer el proceso que sigue Big Data para el análisis de volúmenes de información.
- Revisar casos de aplicación de Big Data y empresas donde ya están experimentando con esta tecnología, así como los resultados que han obtenido.
- Conocer el software y/o herramientas disponibles para Big Data, así como los requerimientos en hardware para su implementación.
- Lograr identificar las ventajas y desventajas que ofrece el Big Data frente a las tecnologías de procesamiento de información actuales o tradicionales.
- Aprender aspectos importantes sobre la seguridad respecto al Big Data.
- Reconocer los retos y oportunidades en el presente y a futuro del Big Data.

METODOLOGÍA Y TÉCNICA DE INVESTIGACIÓN

Con la intención de cumplir con los objetivos planteados, en la presente tesina se realizó una investigación exploratorio-descriptiva, ya que su objetivo central se orienta a analizar e investigar el impacto que tiene el Big Data, así como describir lo que el estado del arte establece en cuanto al grado de influencia que esta área tiene para las organizaciones hoy en día. Para ello, la estrategia metodológica considera las siguientes etapas:

- Revisión bibliográfica. Se acudió a la revisión de artículos y estudios publicados en fuentes formales de consulta, tales como: artículos científicos en revistas indizadas, artículos de congresos nacionales e internacionales, periódicos, investigaciones académicas y todas aquellas relacionadas al objeto de estudio.
- Selección de estudios. Se revisaron estudios realizados en español e inglés, de autores nacionales e internacionales, con las siguientes bases de datos como fuentes de consulta principal: Redalyc, ERIC, Scielo y Google académico.
- Operacionalización de las variables. En el estudio realizado se destacan las siguientes variables de estudio con sus respectivas definiciones operacionales:
    a. Big Data: Es una nueva generación de tecnologías y arquitecturas diseñadas para extraer valor económico de grandes volúmenes de una amplia variedad de datos *(Olofson, 2012)*.
    b. *Entornos de trabajo*: son los encargados de realizar la computación sobre datos en un sistema de datos *(Ellingwood, 2016)*.
    c. Revisión de los *entornos de trabajo* más importantes de Big Data, revisar sus características, ecosistema y herramientas con las que cuentan para llevar a cabo el procesamiento de datos.
    d. Beneficios del uso del Big Data. Todo aspecto de impacto favorable al utilizar el Big Data para la solución de problemas o realización de actividades.
    e. Revisión de futuras tendencias, retos, seguridad y oportunidades para Big Data.

f. Efectos negativos del Big Data. Consecuencias desfavorables del uso de Big Data en alguna práctica o algún medio en específico.

- Implicaciones éticas. A lo largo de la revisión bibliográfica se respetarán los derechos de autor de los artículos y de los diferentes escritos revisados, citándolos en el cuerpo del documento y colocando la referencia completa en la sección de Bibliografía.
- Desarrollo de la investigación. Para dar respuesta a la pregunta ¿Es necesario el Big Data para la toma de decisiones? Se brinda un resumen de la influencia del Big Data en las empresas, el futuro de la tecnología, los retos que se le presentan, herramientas y tendencias.
- Conclusiones y Discusión. Una vez analizada la información del estado del arte, se presentan las principales conclusiones del estudio, así como una breve discusión respecto a los supuestos planteados por los autores de las diferentes fuentes consultadas.

ESQUEMA DE CONTENIDO

La Tesina se encuentra organizada en siete capítulos los que se encuentran estructurados de la siguiente manera. Primeramente, se presenta una introducción general a Big data resaltando la importancia de la temática, objetivos generales y específicos, también se describe la estrategia metodológica que se siguió para realizar la investigación, la sección de referencias, así como las variables de estudio.

En el primer capítulo I se presenta una reseña histórica sobre Big Data con algunos de los hitos más importantes que hicieron posible Big Data, después la importancia y los costos de almacenamiento y procesamiento, posteriormente se define BD y se describen sus V o características, finalmente se dan ejemplos de algunas aplicaciones en México y el mundo.

Por otro lado, en el Capítulo II se aborda el ciclo de vida de Big Data, en el que se muestran las principales fuentes generadoras de datos, la ingesta de los datos a un sistema de almacenamiento distribuido, seguidamente se enlistan las diferentes formas de procesar los datos con Big Data y por ultimo se aborda el análisis de los datos.

El Capítulo III se introduce a los entornos de trabajo para Big Data, de los que se tomaron los provenientes del software libre entre ellos Apache Hadoop, Apache Spark y Apache Flink se enlistan sus características, ventajas, desventajas y en que aplicaciones tienen una ventaja competitiva sobre los demás.

También se destinan dos capítulos para hablar sobre los retos, oportunidades, seguridad y tendencias de Big Data respectivamente, posteriormente se agrega un capitulo extra para hablar sobre los beneficios que tiene BD vistos en dos vertientes: los beneficios de Big Data frente a las tecnologías tradicionales y los beneficios que puede traer a las organizaciones en a reducción de costos, toma de decisiones, nuevos productos, detección de fraudes. Por otro lado, se destina una sección para abordar las desventajas de Big Data visto de diferentes puntos de vista. Finalmente, se presentan conclusiones y recomendaciones junto con las referencias de los materiales empleados para la consulta de este trabajo.

CAPÍTULO I
BIG DATA Y APLICACIONES

1.1 RESEÑA HISTÓRICA

A finales de la década de los años cuarenta con la introducción de las computadoras, se dio inicio al procesamiento de datos, ellas no sólo eran capaces manejar los datos si no convertirlos en información *(Clegg, 2017)*, antes de la aparición de la computadora la información se analizaba de manera manual ya que esta era contable, pero en la actualidad las bases de datos ofrecen la posibilidad de preservar información de interés para una organización, realizar análisis, obtener conocimiento sobre algún tema en particular y automatizar procesos.

Hoy día, los volúmenes de información crecieron exponencialmente en los últimos años, y en consecuencia las herramientas con arquitecturas tradicionales como las bases de datos dejaron de contar con la capacidad de procesar tal cantidad de información debido a que ellas normalmente almacenaban sólo datos estructurados, pero con el tiempo surgió la necesidad de incorporar datos no estructurados como parte de la administración general *(Borkovich, 2014)*, situación que propició el desarrollo de nuevas tecnologías para conseguir mejores maneras de analizar datos y extraer información. En este sentido, Big Data es el resultado de una combinación de tecnologías que han evolucionado a través del tiempo, que permiten explorar, manipular y gestionar grandes cantidades de datos en un tiempo aceptable para obtener información que apoya la toma de decisiones *(Hurwitz, 2013)*.

No obstante, algunos autores toman los años ochenta como el inicio del Big Data puesto que, fue cuando hubo un auge mayor de las computadoras personales, aunque un estudio sobre la evolución de Big Data como tema de investigación muestra que estuvo presente desde la década de los setenta, siendo introducido por primera vez en la computación por Roger Mangoulas *(Lungu, 2012)*, para definir una gran cantidad de datos que no es posible gestionar con técnicas tradicionales de gestión de datos debido a la complejidad y tamaño de estos, pero no fue sino hasta 2008 que se incluyó por primera vez el término en publicaciones de manera formal.

La historia del Big Data está representada principalmente por eventos que llevaron a su impulso y principalmente por la necesidad de algunas empresas importantes de internet y TIC como Google, Facebook, Twitter, Yahoo! y Apple entre otras, además de la importancia que tuvieron algunos *entornos de trabajo* para que Big Data fuera posible. En la Figura 1 se muestra una línea de tiempo con importantes hitos de la historia de Big Data, la aparición de algunas empresas, el nacimiento de algunos entornos de trabajo como hadoop y Spark, la llegada de las bases de datos NOSQL para solucionar problemas encontrados con las bases de datos relacionales, partiendo de 1997 con importantes acontecimientos como la fundación de Google, el uso de la nube para almacenar información, la era del Zetabyte en 2015 y la liberación de TensorFlow una importante herramienta para Machine Learning entre otros.

1997, El problema de Big Data, investigadores de la nasa.

1998, Se fundo Google

1999, Nacimiento de Apache Software Foundation

2000, Doug Cutting lanzó su proyecto de búsqueda de indexación: Lucene

2000, L Page y S. Brin escribieron el artículo "La anatomía de un motor de búsqueda web hipertextual a gran escala"

2001, Las 3Vs, el documento de Doug Laney "Gestión de datos 3D: control de volumen, velocidad y variedad"

2002, Doug Cutting y Mike Caffarella comenzaron Nutch, un subproyecto de Lucene para sitios web de rastreo.

2003, Sanjay Ghemawat publicó "El sistema de archivos de Google" (GFS)

2003, Cutting y Caffarella adoptaron la idea de GFS y crearon Nutch Distribute File System (NDFS) más tarde, se convirtió en HDFS.

2004, Google comenzó a desarrollar Big Table.

2004, Yonik Seeley creó Solr para el motor de búsqueda de esquemas flexible, centrado en la lectura, orientado a documentos y flexible.

2004, Jeffrey Dean y Sanjay Ghemawat publicaron "Procesamiento de datos simplificado en un cluster" o MapReduce

2005, Nutch estableció Nutch MapReduce

2005, Damien Katz creó Apache CouchDB, anteriormente Lotus Notes

2006, Cutting y Cafarella comenzaron Hadoop o un subproyecto de Nutch.

2007, Yahoo desarrolló Apache Pig ejecutado en Hadoop.

2007, 10gen, una empresa de nueva creación trabajó en (PaaS). Más tarde, se convirtió en MongoDB.

2008, Apache Hive (extender SQL), HBase (Gestionar datos) y Cassandra (Sin esquema) para soportar Hadoop

2008, Mahout, un subproyecto de Lucene integrado en Taste.

2008, Hadoop se convierte en proyecto ASF de primer nivel.

2008, TUB y HPI iniciaron el Proyecto Stratosphere y luego se convirtieron en Apache Flink

2009, Hadoop combina de HDFS y MapReduce. Clasificación de un TB 62 segundos sobre 1,460 nodos

2010, Google tiene licencia para ASF Hadoop

2010, Apache Spark, una plataforma de computación en clúster se extiende desde MapReduce para primitivas en memoria

2011, Apache Storm se lanzó para un marco de cálculo distribuido para flujo de datos

2012, Apache Dill para el motor de consultas SQL sin esquema para Hadoop, NoSQL y almacenamiento en la nube

2012, Fase 3 de Hadoop - Aparición de "Yet Another Resource Negotiator"(YARN) or Hadoop 2.

2013, Mesos se convirtió en un proyecto de Apache de alto nivel.

2014, Spark tiene más de 465 colaboradores en 2014, el proyecto ASF más activo.

2015, Inicio en la era Zeta Byte.

2016, Se libera la version 1.0 de Apache Mesos

2017, Google libera TensorFlow su librería de Machine Learning.

2018, Se libera la version 1.0 de Apache Ranger

Figura 1 Corta historia de Big Data obtenido de (Buyya, 2016) y traducido a español agregando hitos 2016-2018

1.2 IMPORTANCIA DE LA CAPACIDAD DE ALMACENAMIENTO

Big Data se encuentra estrechamente relacionado con la capacidad de almacenar, administrar y procesar de manera eficiente cantidades de datos cada vez mayores. Actualmente, las computadoras ofrecen la posibilidad de analizar grandes cantidades de datos, la evolución tecnológica en cuanto a almacenamiento permite continuar almacenando cada vez más de manera automática y sencilla.

Todo comenzó entre los años setenta y ochenta cuando se tuvo el primer desafío "pasar del Megabyte al Gigabyte", debido a la necesidad de almacenar datos y ejecutar consultas relacionales, para realizar análisis e informes en grandes negocios. Después a finales de la década de los ochenta la popularización de la tecnología digital provocó que los volúmenes de datos

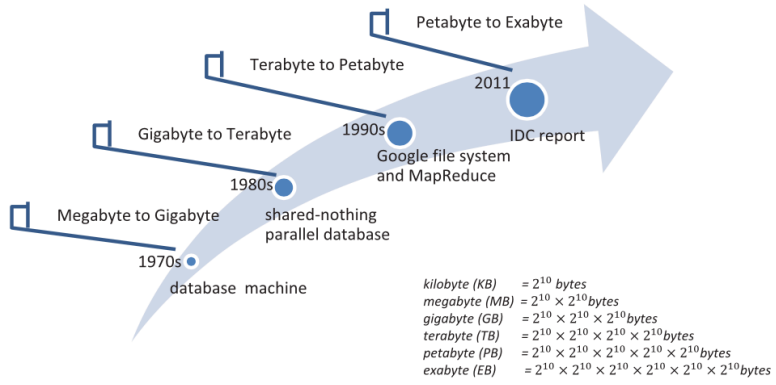


Figura 2 Historia breve con importantes hitos desde Megabyte hasta Exabyte obtenido de (HU, 2014)

se expandieran a varios Gigabytes e incluso Terabytes (Figura 2).

A finales de los años noventa con el rápido desarrollo de la web 1.0, se introdujo a todo el mundo a la era del internet junto con sistemas masivos estructurados y semiestructurados. A mediados del 2000 con la web 2.0 se aceleró la creación de datos mediante redes sociales, blogs y wikis *(Nugultham, 2012)*. Actualmente, de acuerdo con las tendencias de desarrollo los datos almacenados y generados por empresas, instituciones, redes sociales, información gubernamental y hospitales, excederán los PB *(HU, 2014)*.

Cada vez son más los dispositivos con un sensor capaces de conectarse a internet y generar datos. De acuerdo a la Figura 3 para el 2005 se contaban con aproximadamente 130 Exabytes de información en el mundo, presentando un aumento en el 2010 hasta un total de 1227 Exabytes y finalmente en el 2015 un total de 7910 Exabytes lo que da un panorama del crecimiento importante que se está dando en la generación de datos *(Balachandran, 2017)*.

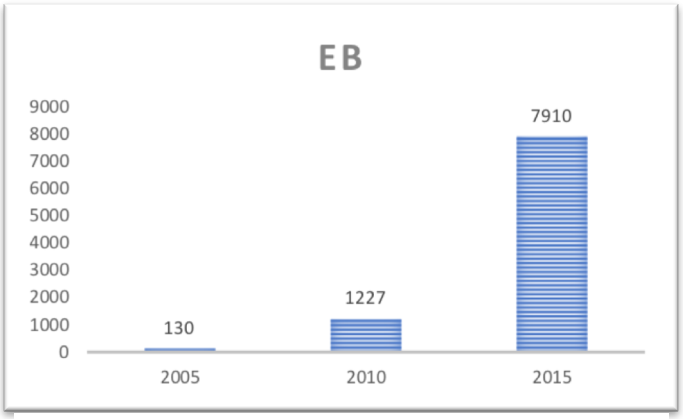


Figura 3 Incremento de los datos entre 2005 y 2015 en HB obtenido de (Olofson, 2012)

1.3 COSTOS DE ALMACENAMIENTO Y PROCESAMIENTO

Almacenar datos no es nuevo, sólo que ahora se almacena de manera astronómica, debido a la evolución del almacenamiento y a la disminución de costos de este, en consecuencia, debido a que el costo del almacenamiento sigue bajando, la cantidad de datos digitales continúa creciendo y a su vez permite la innovación impulsada por datos.

Para 1994 sólo el 3% de los datos de todo el mundo se encontraban almacenados de manera digital, sin embargo, para el año 2007, el 94% de la información ya se encontraba almacenada de manera digital *(Alliance, 2015)*.

En la Figura 4 se muestra como fue disminuyendo el precio del almacenamiento por Gigabyte a través de los años partiendo de 1992 era de $569 dólares mientras que para el 2012 disminuyo a $0.03, es posible apreciar en la gráfica el precio de almacenamiento disminuye un 38% aproximadamente cada año lo que hace que almacenar datos no sea el principal problema.

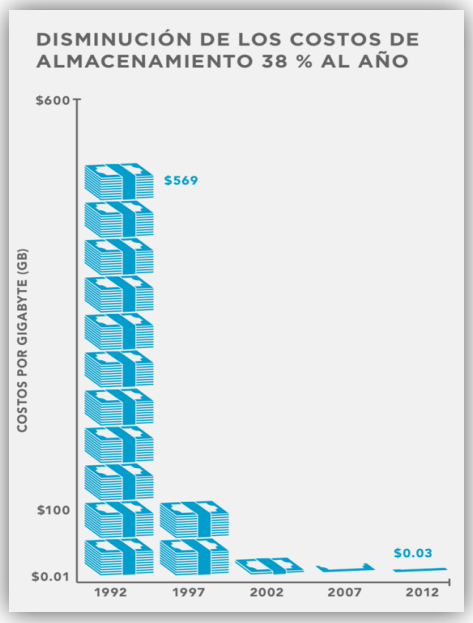


Figura 4 Disminución del precio del Gigabyte (Alliance, 2015)

Otro factor importante que contribuyó al almacenamiento fue la disminución del costo de la potencia de cómputo. De acuerdo con la Figura 5 en 1992 el costo por millón de transistores era de $222 dólares y para el 2012 disminuyó a $0.06 dólares por millón, ello hizo posible crear equipos con mejor rendimiento a bajo costo *(Hagel, 2013)*, eso más la reducción del precio del almacenamiento ha hecho que almacenar y procesar datos sea cada vez más accesible.

Figura 5 Disminución del costo de procesamiento obtenido de (Hagel, 2013)

1.4 BIG DATA

La definición para Big Data es diversa, difícil de precisar ya que no se tiene un consenso y no se encuentra establecida una sola, por ello a continuación se presentan algunas definiciones provenientes de diversos autores y organizaciones sobresalientes en el tema.

El McKinsey Global Institute *(Manyika, 2011)*, la revista RDU de la UNAM, Sadam Madden del MIT *(Madden, 2012)* y el Instituto Nacional de Estándares y Tecnología (NIST) *(HU, 2014)*, señalan que Big Data hace referencia al tratamiento y análisis de enormes repositorios de datos cuyo tamaño está más allá de capturar, almacenar, administrar, analizar y procesar con herramientas de bases de datos o analíticas convencionales *(Salazar, 2016)*.

IDC empresa pionera en el estudio de Big Data lo define como *"una nueva generación de tecnologías y arquitecturas diseñadas para extraer valor económico de grandes volúmenes de una amplia variedad de datos" (Olofson, 2012)*. De acuerdo a *(Contreras, 2016)* Big Data "es la ciencia de Datos que tiene por objetivo la obtención, limpieza, preparación y análisis de datos de cualquier fuente o naturaleza", se trata de un gran volumen, alta velocidad y variedad de información *(Gartner, 2018)*, y demandan formas rentables para procesar información que permitan la toma de decisiones y automatización de procesos.

Por último, *(Gil, 2016)* especifica que Big Data es el conjunto de tecnologías que permiten tratar cantidades masivas de datos provenientes de fuentes dispares, con el objetivo de poder otorgarles una utilidad que proporcione valor, *(O´Reilly Media, 2012)* y estos datos exceden la capacidad de procesamiento de los sistemas de bases de datos convencionales *(Villanustre, 2016)*.

Después de revisar las definiciones anteriores, es posible concluir que Big Data requiere enfoques analíticos avanzados y es útil para analizar, explorar, procesar y comprender fuentes de datos cada vez mayores. No obstante, esas definiciones son subjetivas a magnitudes de datos más grandes de lo acostumbrado, y hacer uso de arquitecturas tradicionales haría imposible gestionar la información en un lapso aceptable de tiempo *(Cuza, 2016)*, la complejidad de Big Data se describe desde los GB en adelante

(Kejariwal, 2012), y *(Cuza, 2016)* ha establecido una escala de acuerdo al volumen de datos que puede ser considerado Big Data en la actualidad (Figura 6*)*.

En la tabla 1 se muestra una comparación de datos tradicionales Vs Big Data tomando en cuenta criterios como el volumen, tiempo de generación, estructura, fuente, integración, almacenamiento y acceso de los datos.

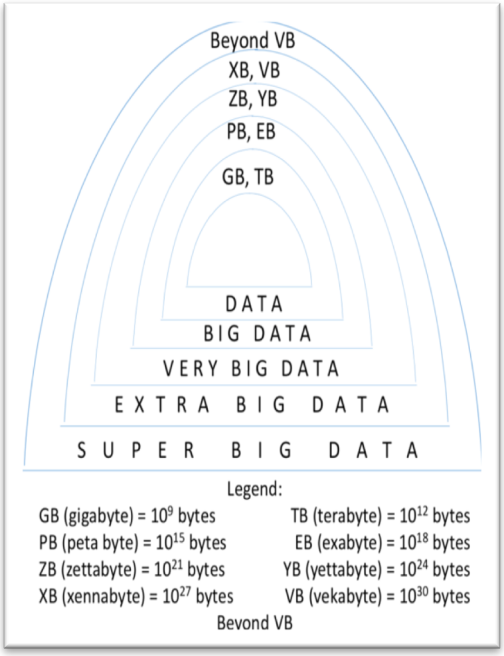


Figura 6 Big data de acuerdo con el volumen de los datos obtenido de (Cuza, 2016).

Tabla 1 Comparación entre datos tradicionales y Big Data obtenido de (Zanoon, 2017)

	Datos tradicionales	**Big Data**
Volumen	GB	TB y PB
Tasa de generación de datos	Por hora; por día	Más rápido
Estructura de datos	Estructurado	Semiestructurado o no estructurado
Fuente de datos	Centralizado	Totalmente distribuido
Integración de datos	Sencilla	Difícil
Almacén de datos	RDBMS	HDFS, NOSQL
Acceso a los datos	Interactive	Por lotes o en tiempo real

1.5 CARACTERÌSTICAS DE BIG DATA

Douglas Laney notó en el año 2001 aumento en los datos a lo largo de tres dimensiones las cuales son Volumen, Velocidad y Variedad, sin embargo, IBM agregó un nuevo aspecto, la veracidad *(Hurwitz & Associates, 2012)*, posteriormente Yuri Demchenko añadió un aspecto a los 4 de IBM el Valor y finalmente Microsoft agrego a los aspectos la Visibilidad dando un total de 6vs *(Buyya, 2016)* que caracterizan a Big Data (Figura 7).

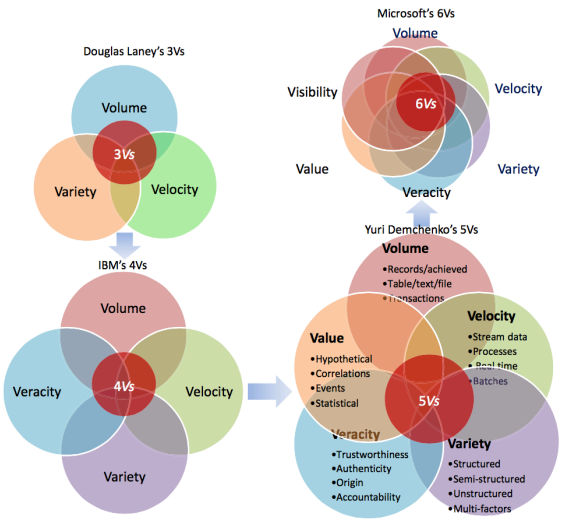


Figura 7 3Vs, 4Vs, 5Vs y 6Vs de Big Data obtenido de (Buyya, 2016)

- **Volumen:** El tamaño de los datos caracteriza a Big Data *(Rehman, 2016)*, representa el desafío principal, dado que los sistemas tradicionales no logran manejar grandes volúmenes de datos *(Mahmood, 2016)*, hablando de orden de TB a PB y estructuras como registros, transacciones, archivos y tablas obtenidos por una organización, para su tratamiento y generar nuevo conocimiento *(Villanustre, 2016)*.
- **Velocidad:** Se refiere a la rapidez para la adquisición de datos, ya que estos se pueden obtener de diferentes maneras y ritmos diferentes *(Mahmood, 2016)*. Los datos pueden almacenarse, procesarse y administrarse *(Villanustre, 2016)*, ahora la velocidad a la que se generan datos es muy elevada debido a las maneras de obtenerla también han aumentado de manera considerable *(Viñals, 2012)*.
- **Variedad:** Se refiere al tipo de datos que Big Data puede comprender, la información puede ser estructurada, semiestructurada y no estructurada *(Lungu, 2012)*. No sólo crecen los datos si no su patrón de crecimiento también lo hace, se puede encontrar la información en distintos tipos, pueden estar en forma de documentos de texto plano, con formato enriquecido, correos electrónicos, mensajes de texto, audio, imágenes, video, etc. *(Villanustre, 2016)*. De toda la variedad existente de datos generados de las diversas formas posibles los que más abundan son los tipos no estructurados *(Morales,*

2016), el 20% se encuentran de manera estructurada y el 80% de forma no estructurada *(Gil, 2016)*.

- **Veracidad:** Se refiere a la credibilidad y exactitud de las fuentes de datos *(Elragal, 2014)*, y cuán precisos son los datos recopilados *(Mahmood, 2016)*, así como la idoneidad para los fines de uso. También se refiere a la calidad y el grado que un líder utiliza la información para la toma de decisiones y de acuerdo a los beneficios obtenidos se puede dar una evaluación a la calidad del trabajo con Big Data *(Villanustre, 2016)*, la utilidad de los sistemas Big Data incrementa cuando los datos se recopilan de fuentes confiables y seguras *(Rehman, 2016)*.
- **Valor:** Define la utilidad y la usabilidad de los sistemas Big Data *(Rehman, 2016)*. Es la ganancia que ilustra la información después de algunas operaciones de procesamiento, esto debido a que estas operaciones pueden contribuir a descubrir patrones ocultos en los datos que afectan diferentes dominios y actividades *(Mahmood, 2016)*. Entonces en el contexto del Big Data se refiere tanto al costo de la tecnología como el valor del uso derivado de la misma. Por otro parte de acuerdo a IDC *(Olofson, 2012)*, el valor derivado depende de la reducción del costo de Hardware y Software, eficiencia de las operaciones, mejora de los procesos de negocio (aumento en los ingresos o beneficios del negocio), por lo general se le denomina análisis de Big Data *(Koseleva, 2017)*.
- **Visibilidad:** La visibilidad enfatiza en que es necesario tener una imagen completa de los datos para tomar una decisión informada *(Buyya, 2016)*, debido a que el análisis exploratorio puede ayudar a la detección, aislamiento y descubrimiento de fenómenos interesantes *(Landmarka, 2017)*.
- **Volatilidad:** La volatilidad se refiere al tiempo de almacenamiento de los datos después de procesarlos, ya que la volatilidad tiene impacto directo en los macro datos, como el volumen y la veracidad *(Mahmood, 2016)*, por ello en las organizaciones existen políticas de almacenamiento de datos para que la información no tenga interferencias ni daño.

1.6 APLICACIONES DEL BIG DATA

En México y el mundo se desarrollan múltiples aplicaciones mediante el uso de Big Data para resolver problemas, algunos ejemplos se describen a continuación iniciando por México y posteriormente en el mundo.

1.6.1 APLICACIONES DE BIG DATA EN MÉXICO

De acuerdo con EMC Digital Universe en México se tiene un crecimiento constante de dispositivos dentro del internet de las cosas, de igual modo en los últimos años se ha incrementado el uso del internet, la proliferación de los teléfonos inteligentes y el uso de las redes sociales y por otro lado con el paso del tiempo se han visto reflejadas disminuciones en el costo de tecnología de captura de datos y almacenamiento de información.

En el caso de México para el 2014 se contaban con alrededor de 99 EB de datos y se espera un crecimiento a 720 EB para el 2020. El tamaño, la diversidad de los datos en México y el crecimiento de ellos puede ser desalentador, las compañías enfrentan grandes desafíos al implementar sistemas de análisis predictivo, análisis de negocios y otras herramientas para el análisis de datos y toma de decisiones en tiempo real *(IDC , 2014)*. Algunas de las aplicaciones de Big Data que se han implementado en México son las siguientes:

Caso 1. Big Data y turismo en México *(Gonzalez, 2016)*

Con el uso de Big Data en México, ha permitido identificar tendencias en los destinos turísticos en México y fortalecer las políticas publicas dentro del sector. Es un proyecto pionero presentado por BBVA Bancomer en colaboración de la secretaria de turismo, con este estudio se busca fortalecer el diseño de políticas publicas y mejorar el crecimiento ordenado de la actividad turística. Bajo el nombre de Big Data y turismo se hizo un estudio inicial con 86 millones de usuarios de tarjetas bancarias nacionales y extranjeras durante un año. Se analizaron 111 pueblos mágicos y los principales corredores turísticos del país.

Todos los usuarios generan información importante en cuanto a gastos y movilidad. Uno de los objetivos fue registrar quienes son los usuarios que realizan más compras, turistas nacionales o de procedencia extrajera. Es un sector de gran importancia ya que gran parte del PIB en México proviene de este sector por lo que los resultados del estudio ayudarán a ofrecer mejor atención a clientes y anticipar movimientos de visitantes en pueblos mágicos y zonas turísticas.

Caso 2. Big Data la clave para la transformación en la industria en México *(Romero, 2019)*

La industria 4.0 está transformando a compañías locales como a transnacionales en México al incorporar nuevas tecnologías dentro de su vida empresarial, Big data lo esta haciendo de forma importante. Expertos en manufactura comentan que durante las etapas de producción existen problemas para obtener datos de manera útil y segura lo que arrastra con problemas hasta etapas finales, en consecuencia, se necesita volver a realizar en ocasiones nuevamente análisis de datos. Como respuesta a este problema a este problema Sadvik Coromant desarrollo una plataforma conformada con diversos sensores, herramientas de conectividad, algoritmos, sistemas en la nube y análisis de datos que permiten a los fabricantes utilizar Big Data para ganar precisión en las fases de producción.

Sadvik Coromant está invirtiendo en la manufactura 4.0 para brindarles a los fabricantes la posibilidad de mejorar sus procesos de producción en todas las fases, reducir desperdicios, optimizar los recursos y ganar precisión a través de su plataforma llamada CoroPlus.

1.6.2 APLICACIONES DE BIG DATA EN EL MUNDO

La industria de la salud genera grandes cantidades de datos a partir del tratamiento de registros médicos relacionados con los pacientes. Big Data tiene varios desafíos dentro de este sector entre ellos el desarrollar herramientas adecuadas para lograr el análisis adecuado de los datos recopilados de diversos sensores, información relacionada con tratamientos, pacientes y medicamentos *(Anita, 2015)*, llevar una observación adecuada

de los datos que colabore a la predicción de epidemias, curar importantes enfermedades, evitar muertes prevenibles y mejorar la calidad de vida *(Kim, 2016)*.

Caso 1: Beth Israel Deaconess Medical Center (BIDMC) *(Halamka, 2015)*

Beth Israel Deaconess Medical Center (BIDMC) es una institución en Boston que recibe aproximadamente 700 mil pacientes al año, actualmente utiliza Big Data como apoyo para la toma de decisiones clínicas inteligentes, por medio de los historiales médicos de pacientes.

En el BIDMC trabajan con alrededor de 4 PB de datos, los cuales son usados para crear aplicaciones reales que logren conducir a decisiones acertadas para los pacientes. Algunas personas utilizaban una aplicación llamada BIDMC@Home la cual es posible vincularla a través de un brazalete con un dispositivo móvil con lo que es posible obtener grandes de cantidades de números en bruto y descubrir con el tiempo resultados interesantes sobre la salud de la persona.

Se observó que los pacientes hacían uso de dispositivos móviles, por lo que decidieron llevar BIDMC@Home a nivel organización con el objetivo de obtener la mayor cantidad de datos de forma continua debido a que el monitoreo continuo puede producir resultados clínicos excelentes y beneficios financieros para ambas partes. Al ser presentada a la administración de BIDMC y se obtuvo el financiamiento para el proyecto, también se basaron en la regulación para la obtención de los datos de los pacientes de acuerdo con leyes federales relacionadas, además con ello buscan contribuir con otros hospitales al dar un resumen de cómo fue el desarrollo de un paciente, su tratamiento y recuperación.

Dentro del hospital hacen uso I2B2, una herramienta capaz de acceder a grandes bases de datos de varias instituciones. Primeramente, disponible en hospitales de Harvard, ahora disponible en más de 60 centros académicos en todo el mundo, ahora es una forma de encontrar una enorme cantidad de pacientes con un mismo problema, edad, nacionalidad similar, así como el tratamiento que fue eficiente con esos pacientes, y otros casos similares y con base en ese conocimiento previo aplicar el mejor tratamiento posible. Con el uso de este tipo de herramientas no se busca descartar a los médicos,

sino reducir su carga y ubicarse un paso adelante logrando decisiones clínicas acertadas para sus pacientes.

Caso 2: Apoyo en la atención médica en la India *(Anita, 2015)*

En la India están utilizando Big Data para capturar información sobre sus pacientes con el propósito de obtener un panorama más complejo sobre la gestión de la salud y la participación del paciente, y finalmente resolver lo siguiente:

- Brindar servicios centrados en el paciente: lograr proporcionar un alivio más rápido a los pacientes, lograr detección de enfermedades en etapas tempranas, basadas en historiales clínicos disponibles, minimizando medicamentos que provoquen efectos secundarios.
- Detectar propagación de enfermedades de manera oportuna: mediante el análisis de los registros de pacientes que viven en una ubicación geográfica en particular, con el objetivo de que los médicos puedan asesorar de mejor manera a las personas que residen en esa región.
- Monitorear la calidad de hospitales.
- Mejoras en los tratamientos: dando tratamiento personalizado a un determinado paciente.

Para llevar a cabo este proyecto, se utilizó Apache Hadoop, para la ingesta de datos se utilizo Apache Sqoop para importar los datos provenientes de bases de datos relacionales, por otro lado, se uso Apache Flume para los datos generados por dispositivos. Para almacenar los datos se utilizó HDFS, el procesamiento de datos se hizo mediante MapReduce y la utilización de algoritmos de aprendizaje automático con el objetivo de predecir una enfermedad lo antes posible.

Caso 3: Proyecto de ciencia ciudadana mediante eBird *(Kelling, 2015)*

eBird es un proyecto para la conservación de especies. Todo comienza con la compresión de patrones de distribución, abundancia y movimientos de individuos, los cuales se encuentran impulsados por una serie interactiva de procesos climatológicos, geológicos y ecológicos. Para ello utiliza técnicas de Big Data para acceder y analizar

datos, para recopilar datos sobre las aves y distribución en distintas regiones durante todo el año.

Los datos para este proyecto provienen de voluntarios que a través de internet y aplicaciones móviles recaudan observaciones de aves. Hoy día se cuenta con un aproximado de 250 mil participantes, se obtuvieron aproximadamente 260 millones de observaciones de aves en todo el mundo, registrando un 97% de las especies de aves conocidas en el mundo y se encuentra en un repositorio accesible.

Cada observación de eBird contiene 7 variables que identifican al observador, la ubicación de las observaciones, la duración de búsqueda de las aves, distancia recorrida durante la búsqueda, número de individuos de cada especie observada, número de personas en la búsqueda y el reporte enviado fue parcial o completo acerca de las aves observadas.

Para la distribución del hábitat se incluyen variables que describen la topografía, la cobertura del suelo, la latitud y longitud, estas se almacenan en la base de datos para cada ubicación de búsqueda, con el objetivo de en un futuro predecir la probabilidad de ocurrencia y abundancia de una especie en cualquier lugar que se desee hacerlo, principalmente Big Data fue útil para analizar los patrones de distribución de las aves en determinadas ubicaciones geográficas.

El entorno de trabajo utilizado no se especifica, sin embargo, es probable que se utilizara un híbrido con arquitectura Lambda, el cual es posible realizar procesamiento tipo Batch con la capa de lotes y en caso de llegar nuevos datos la capa de velocidad se puede hacer responsable de analizar datos nuevos mientras los datos de la capa de lotes están listos para finalmente combinar las vistas resultantes y tener la información deseada.

CAPÍTULO II
CICLO DE VIDA DE BIG DATA

Las industrias son atraídas por los beneficios que Big Data puede ofrecer, uno de los objetivos de una solución basada en Big Data es extraer respuestas a diferentes incógnitas mediante un análisis de enormes repositorios de datos para generar cambios a la manera en que se piensa y utiliza la información. El ciclo de vida del Big Data consta de cuatro fases principales Generación, Adquisición, almacenamiento y análisis de datos (Figura 8).

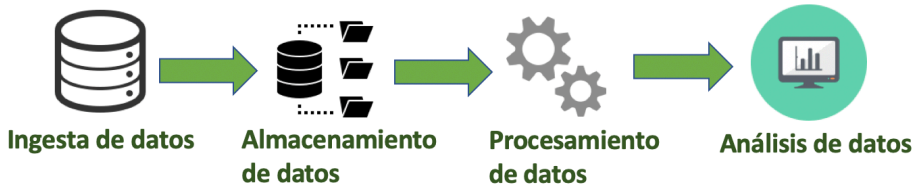


Figura 8 Ciclo de vida de Big Data obtenido de (Hassania, 2017) modificado agregando imágenes.

2.1 GENERACIÓN DE DATOS

Se refiere a como se generan los datos a partir de diversas fuentes, las cuales a través del tiempo se incrementaron, creando más repositorios de datos disponibles para realizar análisis sociales y económicos *(Hassania, 2017)*. Con la llegada de internet se ocasionó el Big Bang de los datos, creció de manera imparable y transformó la manera de interactuar en un marco económico y social posicionándolo como un medio básico; personas, empresas, organismos públicos en la actualidad generan grandes cantidades de información diariamente a través de internet *(Blazquez D. , 2018)*, mediante búsquedas, redes sociales, blogs, además de otras fuentes importantes como IOT(Internet de las cosas), registros médicos, móviles entre otros *(Fouada, 2015)* , a continuación se presenta una clasificación de las diferentes fuentes de datos:

- **Datos de empresas:** El uso de la tecnología ha sido fundamental en el crecimiento empresarial, se estima que el volumen de los datos comerciales se duplicará cada uno o dos años, debido a las transacciones comerciales que se

realizan a través de internet y sumando a ello los datos generados dentro de la empresa *[10]*.

- **Web y medios sociales:** Abarca toda la información de contenidos web e información de redes sociales *(Fragoso, 2012)*, como Facebook, Twitter, LinkedIn *(Soares, 2012)*. Twitter es uno de los más populares con un total de 332 millones de usuarios activos mensualmente y envían un promedio de 500 millones de tweets por día. Por otro lado, Facebook con 1650 millones de usuarios activos, representa una de las fuentes de datos más completas para analizar comportamientos sociales y económicos además de otros como LinkedIn, YouTube, Instagram, en este sentido se considera como una de las fuentes más importantes dando origen al término social Big Data *(Blazquez, 2018)*.
- ***Machine-to-Machine:*** son las tecnologías que permiten conectarse a otros dispositivos alámbrica o inalámbricamente *(Soares, 2012)*. Es una de las fuentes de datos de bajo costo que ayudan a recopilar información sobre actividades de la vida cotidiana o a capturar algún evento en particular *(Fragoso, 2012)*, lo logran mediante una conexión a internet y el uso de sensores para tareas especificas *(Blazquez, 2018)*. Este tipo de tecnologías hacen el "Internet de las cosas", presentes en el sector transporte, servicios públicos, comercios, hospitales, creciendo un 30% anualmente *(HU, 2014)*, en telecomunicaciones con los registros de llamadas *(Fragoso, 2012)*, se generan cada vez mayores cantidades de datos por ejemplo para el año 2010, 4000 mil millones (60%) de las personas en el planeta ya usaban un teléfono móvil *(HU, 2014)*.
- **Biometría:** Todo lo relacionado a biometría, huellas digitales, escaneo de retina, reconocimiento facial *(Fragoso, 2012)*, esto tipo de datos han crecido debido al avance de la tecnología y es muy usado por la policía, sistema legales y agencias de inteligencia *(Soares, 2012)*.
- **Datos científicos:** Cada vez existen más aplicaciones científicas que generan conjuntos de datos muy grandes, entre las principales áreas destacan la Biología computacional, la astronomía y la Física de alta energía *(HU, 2014)*.

La Figura 9 muestra ejemplos de fuentes importantes que diariamente generan cantidades importantes de datos, entre ellas destacan empresas tecnológicas, comercio en línea, salud, datos científicos, entre otras.

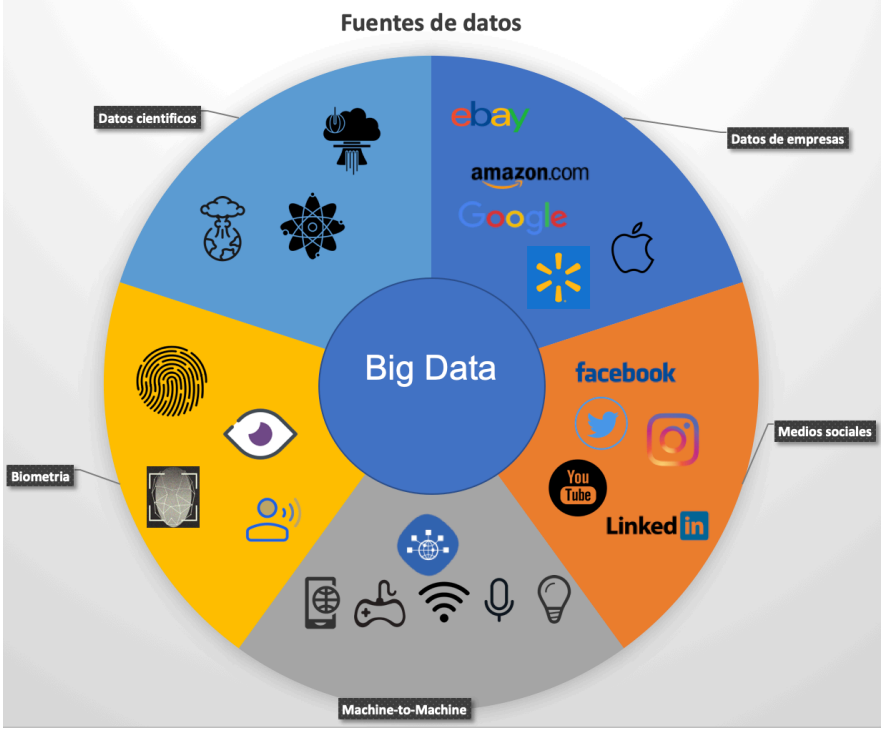


Figura 9 Resumen de fuentes típicas de datos en Big Data

2.2 ADQUISICIÓN O INGESTA DE DATOS AL SISTEMA

El ingreso de datos es el proceso de tomar datos sin procesar para agregarlos al sistema, sin embargo, la complejidad de ello depende de la calidad de los datos y de cuán lejos estén del formato deseado antes de iniciar el procesamiento, una de las maneras para agregarlos es mediante las herramientas de ingestión proporcionadas por los entornos de trabajo.

La elección de la tecnología a utilizar depende del tipo de procesamiento que se llevará a cabo, ya sea un procesamiento por lotes o tiempo real, además de la fuente de los datos. Existen varias herramientas que se han desarrollado para realizar esta tarea algunas de ellas se mencionan a continuación:

- **Apache Sqoop**. Herramienta perteneciente al ecosistema de Hadoop, orientada al procesamiento por lotes, útil para la gestión de datos provenientes de bases de datos relacionales. Entre sus características principales se encuentran las siguientes *(Friedman, 2015)*:
    a. Cuenta con una función Sqoop con la que se pueden crear archivos en gran variedad de formatos.
    b. Permite la importación de grandes cantidades de datos entre bases de datos relacionales como MySQL y Oracle a herramientas del ecosistema de Hadoop como HDFS, Apache Hive y Apache HBase *(Vohra, 2016)*.
    c. Tiene un alto rendimiento cuando se habla de importar bases de datos de tamaño masivo y exportación de tablas optimizadas.
- **Apache Flume.** Es una herramienta que tiene gran potencial para agregar e importar registros de aplicaciones y servidores que se encuentren generando datos en tiempo real *(ellingwood, 2016)* y está basado en flujos de datos de transmisión para recopilar, agregar y transferir grandes cantidades de datos. Los principales componentes de Flume son: Flume Channel y Flume Sink, admite diferentes tipos de fuentes, canales y sinks *(Vohra, 2016)*:
    a. Fuentes: admite diferentes fuentes como Http, Avro, Thrift, Exec.
    b. Canales: admite canales JDBC y canal de archivos.
    c. Sinks: los tipos de sinks que admite: Sistema de Archivos Distribuido de Hadoop (HDFS), MorphlineSolr Sink y HBase Sink.

Apache Flume cuenta disposiciones limitadas y complejas con el objetivo de asegurar alta disponibilidad y entrega garantizada de los datos, normalmente se basa en transmisión, pero a menudo se impone una orientación a lotes debido a que los datos se almacenan en HDFS *(Friedman, 2015)*. A continuación la Tabla 2 muestra un comparativo entre Apache Sqoop y Apache Flume.

Tabla 2 Comparación entre Sqoop y Flume obtenido de (Lakhe, 2016) y modificado agregando procesamiento, ventajas y desventajas y sistema operativo.

	Apache Sqoop	Apache Flume
Arquitectura	Tiene una arquitectura basada en conectores. Un conector es un código que es capaz de conectarse al origen de datos correspondiente y recuperar los datos que se escribirán en HDFS, o viceversa.	Tiene una arquitectura basada en agentes. Un agente es un código o programa que obtiene datos de transmisión desde la fuente.
Almacenamiento	HDFS es una fuente o un destino para datos que utilizan Sqoop.	Escribe datos en varios canales, y HDFS puede ser uno de los canales (o destinos).
Carga de datos	Las cargas de datos para Sqoop no son controladas por eventos.	Puede tener cargas de datos que son impulsadas por eventos.
Utilización	Se utiliza principalmente para la transferencia de datos desde (y hacia) fuentes de datos estructurados como RDBMS.	Se utiliza para mover datos de transmisión masiva a HDFS.
Procesamiento	Realiza grandes cargas de datos mediante procesamiento por lotes	En tiempo real
Ventajas	Es muy útil para la ingesta de grandes cantidades de datos provenientes de Bases de datos relacionales como Oracle, SQL Server, MySQL a bases de datos NOSQL y Sistemas de Archivos Distribuidos como HDFS.	Es muy útil para la carga de datos de transmisión en tiempo real, por ejemplo: los tweets generados en Twitter, los datos del flujo de clics de las aplicaciones web o los archivos de registro de un servidor web.
Desventajas	Cuando se requiere baja latencia Sqoop no es una opción ideal.	Aún y cuando la latencia es baja, la cantidad de datos que puede manejar es limitada.
Sistema Operativo	Linux	Linux

- **Apache Kafka**: proporciona un rendimiento alto con baja latencia para el manejo de datos en tiempo real, es capaz de manejar miles de lecturas y escrituras a nivel de segundos por parte de miles de clientes *(Vohra, 2016)*, proporciona una alta disponibilidad y escalabilidad horizontal debido a que los flujos de datos son divididos en particiones y distribuidos a través del clúster *(Buyya, 2016)*. Entre los usos más comunes de este son el procesamiento de flujos, el seguimiento de actividad de un sitio web y log de registros.

Durante el proceso de ingestión de datos regularmente se lleva a cabo algún nivel de análisis, clasificado y etiquetado, este proceso es llamado ETL (Extracción, Transformación y Carga), las operaciones típicas que se pueden incluir sobre este proceso son la modificación de datos entrantes para formatearlos, llevar a cabo una categorización de los datos, un etiquetado de los mismos, realizar un filtro de los datos innecesarios o incorrectos y aplicar algunos criterios de evaluación para que cumplan ciertos requisitos.

2.3 ALMACENAMIENTO DE DATOS

Después de la ingestión de datos al sistema, los datos son entregados a los componentes encargados de administrar el almacenamiento, para que ellos se puedan preservar de manera confiable en el disco. Pese a que el proceso parece simple, la cantidad de datos que ingresan al sistema, la disponibilidad y la computación distribuida hace que sea necesario hacer uso de un sistema de archivos más complejo, ello significa que es necesario hacer uso de un sistema de archivos distribuido para llevar acabo el almacenamiento se esos datos aún sin procesar.

Una de las soluciones principales en Big Data para resolver ese problema es el sistema de Archivos HDFS (Sistema de Archivos Distribuidos de Hadoop) el cual da la posibilidad de llevar a cabo la escritura de grandes cantidades de datos a través de un clúster, con la finalidad de que puedan ser accedidos por los recursos de cómputo y cargados en la RAM, para su posterior procesamiento en memoria.

2.3.1 SISTEMA DE ARCHIVOS DISTRIBUIDOS DE HADOOP (HDFS)

HDFS es un sistema de archivos distribuido diseñado para almacenamiento de archivos en nodos, tiene una arquitectura maestro esclavo *(Usama, 2017)*, es un proyecto de Apache, se encuentra escrito en Java y se usa para almacenar datos de entrada y salida de aplicaciones. Es altamente tolerante a fallas, usa replicación de datos y está diseñado para ejecutarse en hardware básico, proporciona acceso de alto rendimiento a los datos

de la aplicación y es adecuado para aplicaciones que tienen conjuntos de datos muy grandes.

Los archivos en HDFS se dividen por defecto en bloques de 128MB y cada uno de los bloques es replicado en los múltiples nodos normalmente 3 veces (3 equipos) de manera predeterminada, ello con el objetivo de lograr un procesamiento en paralelo y cada uno de los equipos sea capaz de procesar cada uno de los bloques, otra de las razones por las que el bloque es replicado es para tener una tolerancia a fallos (Figura 10).

HDFS adopta la arquitectura maestro-esclavo y se encuentra conformado por un NAMENODO en donde se almacenan los metadatos y los registros de cada una de las replicas, es capaz de saber en todo momento en donde se encuentran cada uno de los bloques que conforman a el archivo *(Zhang, 2012)*.

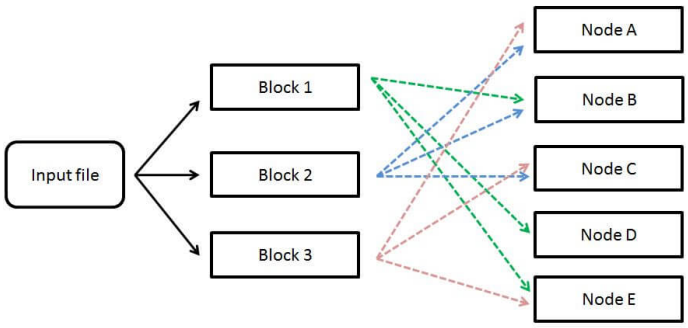


Figura 10 Sistema de Archivos Distribuidos de Hadoop HDFS obtenido de (Carvajal, 2016)

2.3.2 BASES DE DATOS NOSQL

Además del sistema de archivos HDFS de Hadoop existen otros sistemas distribuidos en donde es posible importar los datos y contar con un acceso más estructurado. Las bases de datos NOSQL son una opción adecuada ya que se encuentran diseñadas con las mismas consideraciones de tolerancia de fallas y se puede utilizar alguna de ellas dependiendo del criterio de organización y preservación de datos.

Las bases NOSQL en contraste con las Bases de Datos Relacionales (SQL) tienen algunas diferencias muy particulares, una de ellas es el escalamiento, en el caso de las BDR es sencillo escalar hacia arriba con hardware más rápido, sin embargo, se requiere

invertir cada vez más capital en hardware nuevo con mejores características, además de que en ocasiones en necesaria ingeniería adicional de soporte. En cambio las bases NOSQL están diseñadas en su mayoría para escalar horizontalmente utilizando clústeres de manera distribuida de bajo costo, con un incremento en el rendimiento sin aumentar el tiempo de respuesta, además de que este tipo de Bases de datos generalmente son de código abierto cosa que normalmente no pasa con las BDR *(MongoDB, 2018) (Amazon, 2018)*.

Las bases de datos NOSQL se desarrollaron para hacer frente a aplicaciones modernas, contando con características específicas como almacenar grandes cantidades de datos, y en caso de fallas evitar interrupciones, no cuentan con un esquema predefinido, por lo que el mantenimiento suele ser complejo en éstas *(Nath, 2016)***,** cuentan con los siguientes beneficios:

- **Esquemas dinámicos:** las Bases NOSQL se encuentran diseñadas para permitir inserciones sin ningún esquema predefinido cosa que las Bases de Datos Relacionales no es posible, lo cual hace que realizar pruebas o cambios significativos sea sencillo sin preocuparse por alguna interrupción en el servicio.
- **Replicación:** la gran mayoría admiten la replicación de manera automática de datos para mantener la disponibilidad en caso de interrupciones u otros eventos.
- **Distribución automática:** Con esta característica se logra que los datos se distribuyan de manera nativa a través de una cantidad de servidores y se evita una arquitectura muy costosa y monolítica *(MongoDB, 2018).*

De acuerdo con la Tabla 3 y la Figura 11 existen diversos tipos de Bases de Datos NOSQL *(MongoDB, 2018)* , *(Amazon, 2018)*, *(Jaramillo, 2014)* y a continuación se describe cada uno de los tipos:

Tabla 3 Tipos de Bases de datos (MongoDB, 2018) , (Amazon, 2018), (Jaramillo, 2014).

Tipo de Base de datos NOSQL	Descripción
Bases de datos de documentos	Se encuentran diseñadas para almacenar documentos semiestructurados como XML y JSON, permiten anidamiento profundo y se pueden lograr estructuras muy complejas, aunque también tienen sus desventajas y una de ellas por ejemplo recuperar el valor de un registro significa obtener todo el lote y lo mismo para las actualizaciones lo cual a su vez afecta al rendimiento.
Bases Clave-Valor	son las NOSQL más simples ya que cada uno de los elementos con un nombre de atributo (clave) junto con su valor.
Bases basadas en columnas	Este tipo de bases de datos se encuentran optimizadas para consultas sobre grandes conjuntos de datos, almacenan grandes columnas de datos en lugar de filas entre ellas destacan Casandra y Hbase.
Bases de datos de grafos	Son utilizadas para almacenar información sobre redes de datos, usa almacenamiento en memoria cache para mejorar el rendimiento de la aplicación, entre ellas existen NeoJ4.

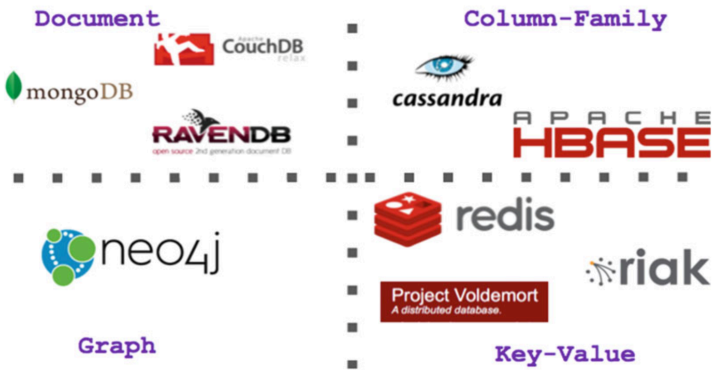


Figura 11 NOSQL clasificación de acuerdo con el modelo de datos usado obtenido de (Estrada, 2016)

Dentro de las Bases de datos NOSQL principales que pueden usar en Big Data destacan ***Cassandra*** la cual fue desarrollada por Facebook en 2008, posteriormente liberado como un proyecto abierto en 2010 se convirtió en el proyecto principal de fundación Apache *(Zaforas, 2016)*, puede ejecutarse desde un servidor básico hasta múltiples centros de datos, en la tabla 4 se muestran las características principales de Cassandra así como algunas de sus aplicaciones en entornos reales.

Tabla 4 Características y aplicaciones de Cassandra DB (instaclustr, 2019)

Característica / aplicación	**Descripción**
Escala lineal	Su crecimiento incrementa de manera lineal horizontal a medida que se añaden más nodos.
Tolerante a fallos	Los datos se replican de manera automática en múltiples nodos para tolerancia a fallas.
Elástico	El rendimiento de lectura y escritura aumenta linealmente cuando se integran nuevos nodos, sin inactividad e interrupción de aplicaciones.
Durable	Buena opción para aplicaciones donde no se quieren fallos ni perdida de datos.
Descentralizado	No hay puntos de falla ni cuellos de botella en la red ya que cada nodo es idéntico.
Aplicaciones en producción	• Apple: 75,000 nodos y almacena más de 10 PB de datos. • Netflix: 2500 nodos, 420 TB de datos y un billón de peticiones por día. • EBay: 100 nodos, 250 TB.

- **Cloud BigTable:** Es un servicio de base de datos NOSQL desarrollado por Google, es de alto rendimiento para cargas de trabajo analíticas y operativas, está diseñado para manejar cargas de trabajo masivas a baja latencia y alto rendimiento, es utilizada para aplicaciones operativas en IOT, análisis de usuarios y análisis de datos financieros, esta base impulsa a Google Maps y Gmail (Tabla 5). Además, se integra con herramientas de Big Data como lo es Hadoop y con la Api de HBASE *(Google, 2018)*, admite alto rendimiento de lectura y escritura y es una fuente ideal para operaciones con Map-Reduce *(Zhang, 2012)*.

Tabla 5 Características y aplicaciones de BigTable (Zhang, 2012), (Developers, 2018).

Característica / Aplicación	**Descripción**
Escalabilidad	Es escalar en proporción directa al número de equipos instalados en el clúster.
Simple administración	Actualizaciones y reinicios de manera transparente, alta durabilidad en los datos, el usuario sólo se encarga de diseñar sus esquemas de tablas.
Elástico	Cuando se realizan cambios en el clúster es cuestión de minutos para equilibrar el rendimiento en todos los nodos del clúster.
Aplicaciones principales	• Datos de marketing: historiales de compras, preferencias de clientes. • Datos financieros: historiales de transacciones, precios de acciones y cambio de divisa. • IOT: información de electrodomésticos u otros dispositivos dentro. • Datos de series de tiempo: Uso de CPU y memoria a lo largo del tiempo para múltiples servidores.

- **Amazon DynamoDB:** Es un servicio de base de datos NOSQL para aplicaciones que necesitan una latencia de milisegundos consistente. Es una base en la nube administrada y compatible con modelos de almacenamiento de documentos y valores clave, es ideal para aplicaciones IOT, web y móviles, cuenta con almacenamiento en cache con el objetivo de lograr que la lectura y escritura sean muy rápidas con una latencia baja, algunos de sus beneficios y aplicaciones se muestran en la Tabla 6.

Tabla 6 Características y aplicaciones Amazon DynamoDB obtenido de (Amazon, 2018).

Característica / aplicación	**Descripción**
Rendimiento rápido y consistente	Esta construido para ofrecer rendimiento constante y rápido en cualquier escala en todas las aplicaciones, su latencia es de milisegundos.
Escalable	Su escala es automática, su capacidad incrementa o viceversa a medida que los volúmenes aumentan o disminuyen.
Totalmente administrado	Sólo es necesaria la conexión y creación de tablas de datos las tareas de administración son transparentes para el usuario.
Flexible	Admite estructuras de datos de documentos y valor/clave.
Aplicaciones	• Soluciones IOT. • Almacenamiento de datos sobre juegos de dispositivos móviles, consolas y computadoras de escritorio.

2.4 CÓMPUTO DE DATOS

Una vez que se encuentran los datos disponibles en el sistema ya es posible iniciar con el procesamiento de datos, existen diferentes alternativas para atender los problemas de procesamiento de Big Data, las principales técnicas empleadas para este análisis son el procesamiento por lotes, en tiempo real y sistemas con un enfoque híbrido. En primera instancia el procesamiento por lotes es una excelente solución al problema de del volumen de los datos, en segundo lugar para atacar el problema de la velocidad existen entornos de trabajo que son una excelente opción para realizar procesamiento en tiempo real y finalmente existen *entornos de trabajo* con un enfoque híbrido que combinan ambas tecnologías (procesamiento por lotes y en tiempo real) estos tienen la capacidad de tratar el volumen y la velocidad cuando se requiere analizar cantidades grandes de datos ya sean estáticos o dinámicos *(Ellingwood, 2016)*.

2.4.1 SISTEMÁS DE PROCESAMIENTO BATCH (POR LOTES)

El procesamiento por lotes implica procesar grandes conjuntos de datos y devolver un resultado una vez que el cálculo sea completado. Las tareas que requieren grandes volúmenes de datos se manejan frecuentemente mediante este tipo de operaciones *(Ellingwood, 2016)*, en el proceso por lotes los datos primeramente son almacenados para su posterior análisis, se logra manejar grandes cantidades de datos, sin embargo, el tiempo de respuesta es largo, por lo cual no es recomendable en situaciones donde el tiempo de procesamiento es prioridad.

Entre las principales características del sistema de procesamiento por lotes son los siguientes:

- **Delimitado:** Los conjuntos de datos representan una colección finita de datos.
- **Persistente:** Los datos regularmente se encuentran respaldados por algún almacenamiento permanente.
- **Extenso:** El procesamiento por lotes es la única alternativa para procesar conjuntos de datos extremadamente grandes.

El procesamiento por lotes es la solución para procesar grandes volúmenes de datos estáticos, debido a que funciona con datos que ya se encuentran almacenados en el sistema. Este tipo de procesamiento no es adecuado para procesos analíticos en tiempo real, procesos analíticos a corto plazo, análisis de precios o análisis de clientes en tiempo real *(Ra, 2015)***.**

2.4.2 SISTEMÁS DE PROCESAMIENTO EN TIEMPO REAL

El término procesamiento en tiempo real significa que eventos se ejecutan dentro de un intervalo de tiempo específico, dentro de los sistemas de TI es típicamente en el orden de mili, micro o incluso nanosegundos dependiendo del sistema *(Buyya, 2016)*. El objetivo del procesamiento en tiempo real es mejorar la velocidad de Big Data para el procesamiento de los datos y con una baja latencia.

Este tipo singular de sistemas pueden manejar casi una cantidad ilimitada de datos sin embargo sólo procesan uno o muy pocos a la vez y es ideal para los datos en los que el

sistema debe responder a cambios o picos *(Ellingwood, 2016)*. Algunas de las características clave del procesamiento en tiempo real son baja latencia, alta disponibilidad y escalabilidad horizontal *(Buyya, 2016)*.

Este tipo de procesamiento logra proporcionar resultados de manera oportuna y con mayor precisión en comparación con el procesamiento por lotes *(Ra, 2015)*. En la Tabla 7 se muestra una comparación entre el procesamiento por lotes y el procesamiento en tiempo real mostrando las diferencias en cuanto a criterios de entrada de datos, almacenamiento, hardware, el procesamiento y aplicaciones**.**

Tabla 7 Comparación entre procesamiento en tiempo real y lotes (Ellingwood, 2016), (Buyya, 2016), (Ra, 2015)

	Procesamiento en tiempo real	**Procesamiento por lotes**
Entrada de datos	Nuevos datos o actualizaciones nuevas	Repositorios de datos
Tamaño de datos	Infinito o desconocido	Conocido y finito
Almacenamiento	Almacenamiento durante el procesamiento	Se almacena regularmente
Hardware	Cantidades de memoria limitadas	Múltiples CPU y memorias
Tiempo	Segundos o milisegundos	Mucho más tiempo
Aplicaciones	Monitoreo de trafico en sitios web, y minería en sitios web	Es adoptado en múltiples dominios

2.4.3 SISTEMÁS DE PROCESAMIENTO HÍBRIDO

Algunos *entornos de trabajo* son capaces de realizar procesamiento por lotes y en tiempo real, en diversos escenarios es necesario para la solución de problemas en donde son necesario ambos tipos de procesamiento, se consigue mediante un modelo híbrido que combina ambas tecnologías e intenta ofrecer una solución general para el procesamiento de datos, cuenta con una arquitectura Lambda conformada de los siguientes componentes *(Ellingwood, 2016)*:

- Capa de lotes: en esta capa la información es preparada para su gestión y los datos se encuentran normalmente en su estado original.
- Capa de servicio (resultados del lote): carga y expone las vistas de lote en un almacén para que puedan ser consultados.

- Capa de velocidad (Procesamiento en tiempo real): sólo se calcula con datos nuevos que requieren baja latencia, para consultarse los resultados completos, posteriormente las vistas de lotes y en tiempo real tienen que consultarse y fusionarse los resultados.

En el apartado 3.3 se explica más a fondo la arquitectura Lambda y como es su funcionamiento a detalle en cada una de las capas.

2.5 ANÁLISIS DE DATOS

El análisis de datos es el proceso que busca obtener respuestas a interrogantes y patrones ocultos en la información, es la etapa más importante ya que las respuestas servirán de apoyo para tomar decisiones dentro de la empresa o institución. Este análisis trata la información a través de la observación, medición o realizando experimentos de interés *(HU, 2014)*. Los métodos de análisis y aplicación difieren debido a la diversidad de los datos *(Hassania, 2017)* y se encuentra clasificado en tres niveles de acuerdo con las profundidades del análisis que se quieran llevar a cabo, se tiene el análisis descriptivo, el análisis predictivo y el análisis prescriptivo *(Pyne, 2016)*.

- **Análisis descriptivo**. El análisis descriptivo es el más simple dentro Big Data, implica una descripción y resumen de patrones, examina los datos para definir el estado actual de una determinada situación *(Sivarajah, 2016)*, profundiza en datos históricos hasta que patrones, variaciones y excepciones se vuelven evidentes *(Pyne, 2016)*. Este tipo de análisis usa métodos estadísticos simples como media, mediana, moda, desviación estándar, varianza, frecuencia de eventos y técnicas como correlación de variables y creación de gráficos para identificar tendencias en los datos y verlos representados de una mejor manera.
- **Análisis predictivo.** En un análisis predictivo se realiza una modelización estadística con el fin de identificar tendencias o eventos a futuro *(Sivarajah, 2016) (Wamba, 2016)*, combina datos de diversas fuentes *(HU, 2014)*, busca nuevas tendencias o eventos a futuro, no garantiza un 100% de efectividad sin embargo,

es útil a las empresas para tomar decisiones de manera más informada *(Pyne, 2016)*. Este tipo de análisis se basa en modelos de aprendizaje supervisado y no supervisado. Se clasifica en dos tipos *(Sivarajah, 2016)*: técnicas de regresión (modelos logit multinomiales) y aprendizaje automático

- **Análisis prescriptivo.** Los análisis anteriores evalúan el pasado para predecir el futuro, el objetivo del análisis prescriptivo investiga la relación causa-efecto y se trata de llevar a cabo una optimización mediante pruebas aleatorias, modelos numéricos y simulaciones para evaluar el impacto de estas y su vez apoya a los tomadores de decisiones *(Pyne, 2016), (Sivarajah, 2016)*. Existen varias fuentes de datos de las cuales se puede realizar un análisis detallado, puede ir desde texto simple, datos estructurados y no estructurados, audio, imágenes video y medios sociales.
    a. **Análisis de texto:** Se refiere al proceso de análisis de texto no estructurado para extraer información relevante *(Edison, 2016)* y conocimiento útil *(HU, 2014)*, este permite a las organizaciones convertir gran cantidad de textos generados en resúmenes significativos *(Hassania, 2017)* puede realizarse en correos electrónicos, blogs, foros en línea, tweets, noticias y registros de llamadas, entre las técnicas utilizadas para este análisis son: análisis estadístico y lingüística computacional, extracción de información y resumen de texto *(Villanustre, 2016)*.
    b. **Análisis de audio:** Es utilizado para extraer información de datos de audio no estructurados *(Hassania, 2017)*, las aplicaciones comunes se dan en los centros telefónicos y los servicios de salud con atención al cliente *(Edison, 2016)*, este tipo de datos son valiosos para analizar el comportamiento de compra de los consumidores *(Wamba, 2016)*.
    c. **Análisis de video:** El análisis de video es el uso de diversas técnicas para extraer información significativa, rastrear y analizar secuencias de video. Es una de las principales áreas de operación en la gestión de marketing y en aplicaciones de video vigilancia *(Villanustre, 2016), (Hassania, 2017)*, *(Edison, 2016)*. Este tipo de datos tienen el potencial de dar valor agregado a empresas de comercio electrónico, empresas como Netflix los usa para predecir los

hábitos de visualización y evaluar la calidad de experiencias *(Wamba, 2016)*.

d. **Análisis de medios sociales:** Es el análisis de datos estructurados y no estructurados de redes sociales como Facebook, Twitter, Instagram *(Villanustre, 2016), (Hassania, 2017)*, y este análisis es posible aplicar para predecir las preferencias y gustos de los clientes *(Wamba, 2016).*

La Figura 12 muestra a modo de resumen lo explicado en esta sección concerniente a las fases que conforman el ciclo de vida.

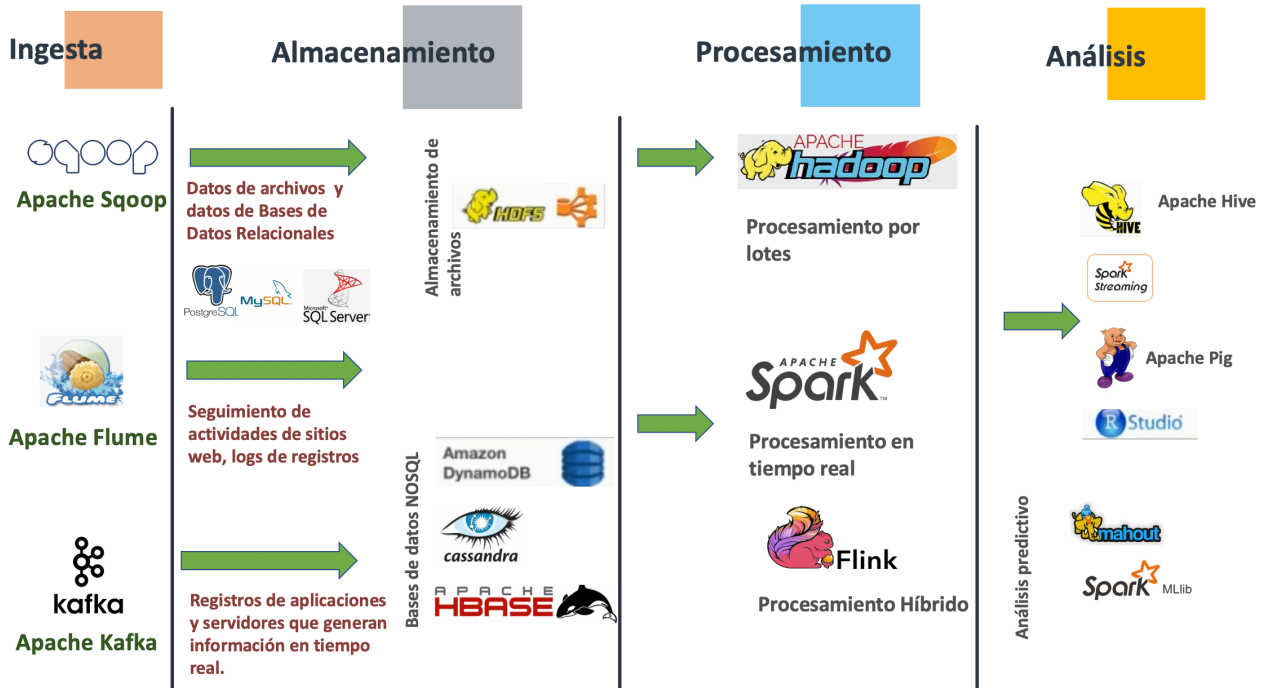


Figura 12 Resumen Ciclo de Vida Big Data

CAPÍTULO III

ENTORNOS DE TRABAJO PARA BIG DATA

Existen varios *entornos de trabajo* para procesamiento con Big Data algunos lo hacen por lotes (Apache Hadoop), otros realizan ese procesamiento en memoria (Apache Spark), y también existen otros que ofrecen ambas opciones siendo un híbrido (Apache Flink). La Figura 13 muestra la evolución de los *entornos de trabajo* iniciando con Hadoop el primer framework para procesar datos por lotes, Spark para procesar datos en tiempo real y ultimo Flink como un híbrido para procesar datos por lotes y en tiempo real *(Buyya, 2016)*. A continuación, se describen cada uno de ellos y las herramientas que los conforman.

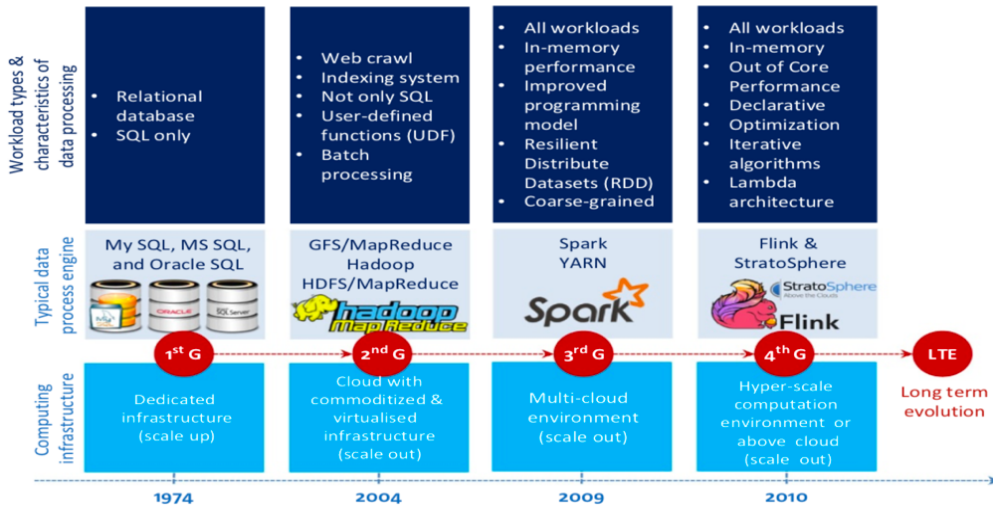


Figura 13 Evolución de los datos y motores de procesamiento Big Data obtenido de (Buyya, 2016)

3.1 APACHE HADOOP

Apache Hadoop es un entorno de trabajo de código abierto de Apache Software Fundation *(Konda, 2015)*, este se encuentra inspirado en el proyecto de Google File System (GFS) y en el paradigma de programación Map-Reduce *(Fragoso, 2012)*. Hadoop es un entorno *de trabajo* que procesa por lotes, cuenta con la capacidad de procesar grandes volúmenes de datos, utiliza la arquitectura Maestro-Esclavo y hace uso de su propio sistema de archivos para el almacenamiento de datos (HDFS), Hadoop ejecuta el código en donde se encuentran los datos, dando una mejora en el tiempo de

procesamiento, además de contar con una gran tolerancia a fallos debido a su sistema distribuido *(Venner, 2009)*.

La figura 14 muestra el ecosistema básico de Hadoop, que consta de dos componentes principales. Para la parte de almacenamiento de datos, se basa en un sistema de archivos distribuido HDFS y una base de datos NOSQL Apache Hbase. Por otro lado, cuenta con YARN/MapReduce para la parte del procesamiento, además de otras herramientas que lo complementan como HBase, Apache Sqoop, Apache Flume, Pig, Hive SQL, Mahout las cuales se describen posteriormente.

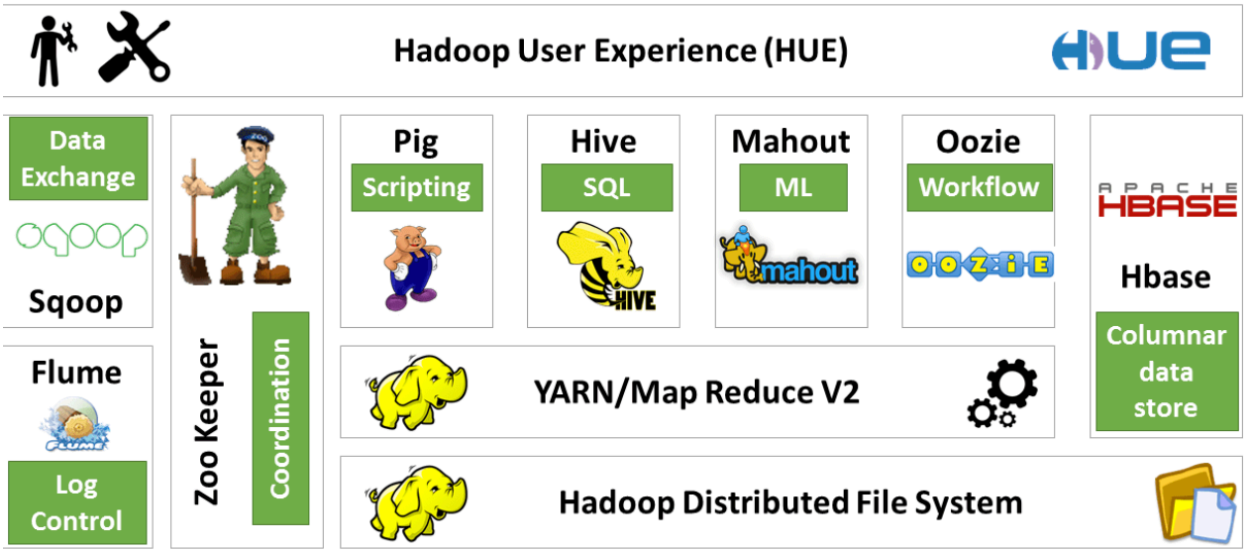


Figura 14 Ecosistema básico de Hadoop obtenido de (Atencio, 2016)

3.1.1 HBASE

Apache HBase es una base de datos NOSQL distribuida y escalable incluida dentro de las distribuciones de Hadoop, es de código abierto *(Vohra, 2016)*, diseñada para operaciones de baja latencia, tiene un modelo de datos clave-valor orientado a columnas, en donde las filas se definen mediante una clave y a ella se le asocian un número de columnas en donde se almacenan los valores asociados *(Presser, 2015)*. Normalmente HBase es utilizada en aplicaciones que pueden requerir filas dispersas, es decir que cada fila es capaz de utilizar solo algunas de las columnas, además de que suele ser rápido debido a que el acceso se realiza a través de la clave principal.

HBase al ser una base de datos NOSQL acepta datos estructurados y semiestructurados de manera natural, pero también puede guardar siempre y cuando estos no sean demasiado grandes, en otras palabras, soporta todos los tipos de datos, permite modelos de datos dinámicos y flexibles al no restringir los datos que almacena. Fue diseñado para ejecutarse en clúster y no en un sólo equipo, cada uno de los nodos en los que se encuentra contribuye con recursos de almacenamiento, procesamiento y caché *(Dimiduk, 2013)*.

A diferencia de las aplicaciones HDFS tradicionales, permite el acceso aleatorio a filas, evitando la lectura secuencial, sin embargo, no permite o no tiene la operación JOIN por lo que se debe de realizar a nivel aplicación *(Presser, 2015)*. Uno de los enfoques principales de HBase son las operaciones Crear, Leer, Actualizar y Eliminar (CRUD) en tablas amplias y dispersas, aunque no permite transacciones (proporciona bloqueo limitado a algunas operaciones atómicas) *(Lublinsky, 2013)*, HBase hace uso de HDFS para su almacenamiento de datos persistentes, lo que permite aprovechar las bondades del sistema de archivos como lo es la replicación, y la conmutación por error.

Por otro lado, ofrece funciones como consultas en tiempo real, búsquedas en lenguaje natural, actualmente se incluye en muchas soluciones de Big Data y sitios web controlados por datos un ejemplo de ello es Facebook.

3.1.2 MAPREDUCE Y YARN

MapReduce y YARN constituyen dos opciones importantes para realizar el procesamiento de datos en Hadoop, se encuentran diseñados para administrar la programación de tareas, los recursos y el clúster.

Hadoop MapReduce es un modelo de programación de procesamiento por lotes que se utiliza principalmente para el análisis de una gran cantidad de datos estáticos *(Ra, 2015)*. Para ello se descompone la tarea de procesamiento en múltiples tareas y a través de dos pasos (Asignar y reducir), para realizar la programación y la asignación en el nodo de gran escala, además de ser un sistema de programación en paralelo tiene una gran tolerancia de errores, la distribución y el balance de una carga de datos *(Venner, 2009)*.

La Figura 15 muestra como funciona MapReduce, cada uno de los pasos que realiza para el procesamiento de datos *(Venner, 2009) (Oussous, 2017)* a continuación se describe cada uno de los pasos:

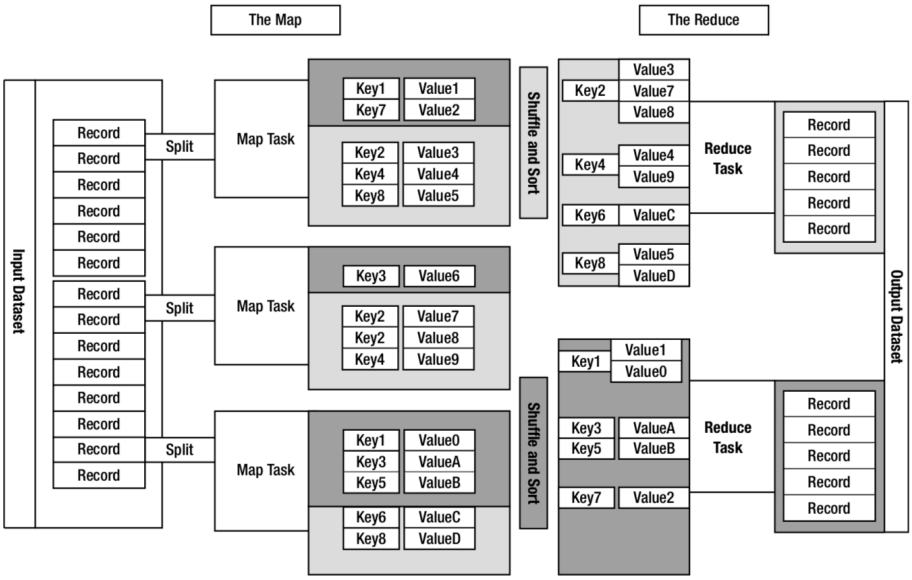


Figura 15 Modelo MapReduce obtenido de (Venner, 2009)

- Splitting: primero se tiene un archivo de entrada que se desea procesar con un formato especifico, se realiza varios Split del archivo y cada uno de los trozos del archivo son el número de tareas que se ejecutaran en paralelo, estos generalmente se almacenan en el sistema de archivos de Hadoop HDFS.
- Mapping: después de ello actúa el Record Reader toma cada una de las partes en las que se dividió el archivo y los convierte en pares clave-valor, posteriormente el Mapper procesa cada uno de ellos de manera individual a través de varias tareas map en todo el clúster de manera paralela.
- Shuffling and sorting: Seguidamente la fase Suffle and Sort se va a encargar de tomar los datos, hacer la ordenación y agrupación de las mismas llaves, para ser enviadas al mismo reduce.
- Reducing: La función Reduce se utiliza para procesar los datos de salida intermedio, hace un conteo de claves de acuerdo con el programa definido.

- Aggregating: finalmente, MapReduce Store genera un archivo de salida por cada uno de los reduce en el mismo formato de entrada (en caso de no especificar uno diferente), también es posible agruparlo en un sólo archivo.

En la figura 16 se muestra un ejemplo de conteo de palabras aplicando el modelo MapReduce. En la entrada se tienen archivos muy grandes, los cuales son divididos en pequeños archivos y a su vez distribuidos en diferentes equipos. Posteriormente, cada uno de los archivos es procesado mediante una indexacion clave-valor que en el Shuffling and sorting son odenados de acuerdo a su valor en diferentes nodos para después pasar al *reducing* para hacer un conteo de claves y finalmente se escriben en disco los resultados.

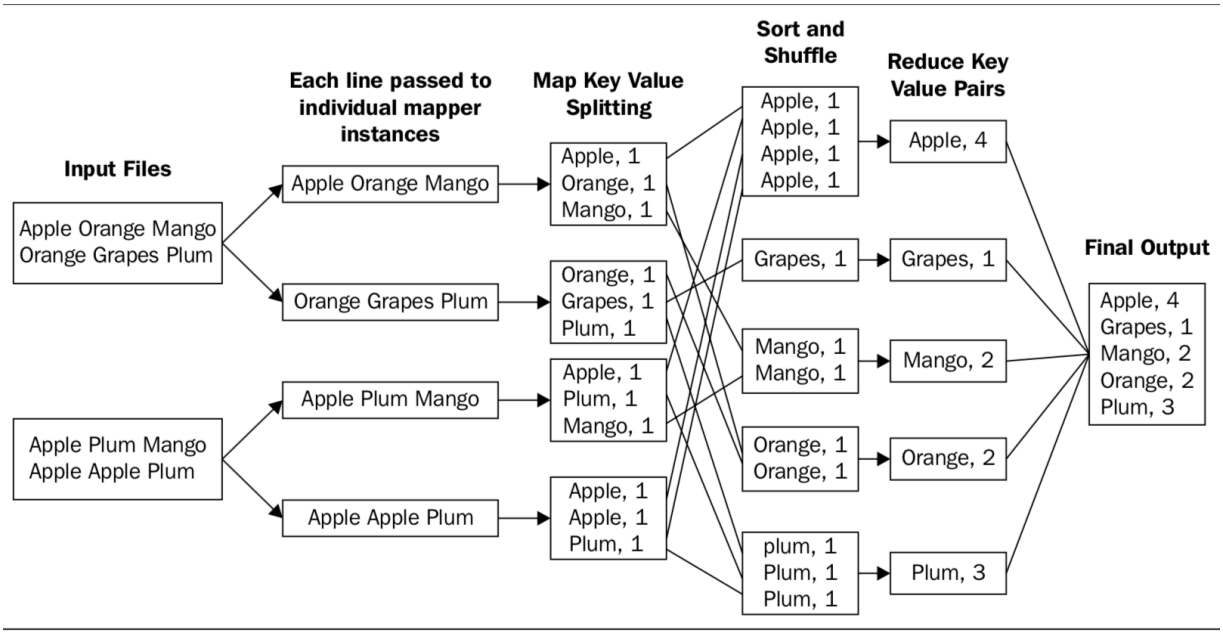


Figura 16 Ejemplo de conteo de palabras con MapReduce obtenido de (Achari, 2015)

A pesar de que MapReduce es eficiente, tiene la desventaja de estar limitado totalmente en lotes, por lo que no permite el procesamiento de datos en tiempo real. Con la llegada de YARN este aspecto cambia debido a que es un programador distribuido y realiza las siguientes actividades *(Holmes, 2015)*:

- Atender peticiones para la creación de contenedores, en donde un contenedor es un proceso al cual se le fijan recursos de manera muy estricta.

- Monitorear contenedores o procesos en ejecución.

YARN es un framework responsable de programar, monitorear recursos y aplicaciones que se ejecutan dentro del clúster, trabaja en la parte superior de HDFS, se ocupa de distribuir datos durante el análisis *(Alaka, 2018)*, das más posibilidades a escalabilidad, cuenta con mejor paralelismo en comparación con MapReduce, es posible procesar datos por lotes y en tiempo real *(Oussous, 2017)*. Dentro de YARN existen dos componentes principales *(Holmes, 2015)*:

- ResourceManager: es el proceso maestro de YARN donde su función es designar los recursos dentro del clúster, responde a solicitudes de clientes para la creación de contenedores.
- NodeManager: es un proceso esclavo que se ejecuta en cada nodo de un clúster, sus funciones son crear, supervisar y eliminar contenedores, además de informar de manera continua al ResourceManager acerca del estado de cada uno de los contenedores.

En la figura 17 se muestran los componentes de una aplicación YARN, en donde una aplicación inicia con el cliente de YARN, el cual se comunica con el ResourceManager para la creación de una nueva instancia de YARN ApplicationMáster, el cual a su vez se involucra nuevamente al cliente ya que se encarga de informar de los requisitos físicos que se requieren al ApplicationMáster. Por otro lado, el ApplicationMáster no realiza trabajos directamente relacionados con la aplicación ya que estos son delegados a los contenedores, sin embargo, sí es responsable de administrar los contenedores específicos de la aplicación, como por ejemplo preguntarle al ResourceManager cuando

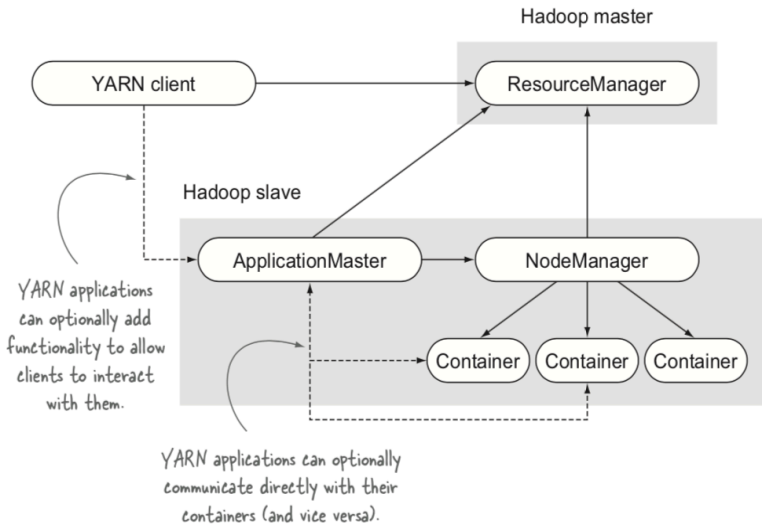


Figura 17 Interacciones típicas de una aplicación YARN obtenido de (Holmes, 2015).

se desea crear un nuevo contenedor y posteriormente ponerse en contacto con el NodeManager y delegarle la función de la creación del contenedor, donde el ApplicationMáster debe especificar los recursos en cuanto a memoria y procesador.

Además, el ApplicationMáster es el encargado de la tolerancia a fallos, recibe mensajes del ResourceManager cuando algún contenedor falla y este decide si crear alguno nuevo o simplemente ignorar el evento. Cuando se pasa de MapReduce a YARN, este es compatible con la API de MapReduce lo cual hace que los usuarios sólo tengan que volver a compilar sus programas de MapReduce para poder ejecutarlos en YARN *(Oussous, 2017)*.

3.1.3 APACHE SQOOP Y FLUME

Apache Sqoop es un software de código abierto, cuenta con una interfaz de línea de comandos (CLI), Sqoop garantiza una ingesta de datos masivos de forma eficiente entre bases de datos relacionales y Hadoop. Entre sus ventajas proporciona un rápido rendimiento, tolerancia a fallas y utilización óptima para reducir las cargas de procesamiento a sistemas externos *(Oussous, 2017)*. Permite una fácil integración con Hbase, Hive y Oozie y cuando los archivos son importados desde HDFS la salida puede ser en varios archivos, los cuales pueden ser archivos de texto delimitados, archivos binarios. En el proceso de exportación de Sqoop toma archivos de texto limitado de HDFS en paralelo para analizarlos e insertarlos en una base de datos destino.

Apache Flume está diseñado para recopilar, agregar y transferir datos externos a HDFS, es capaz de importar datos de bases relaciones, maneja una arquitectura flexible, además de ofrecer funciones importantes como tolerancia a fallas y servicio de recuperación de fallos

3.1.4 APACHE OOZIE

Apache Oozie es una herramienta para gestionar los flujos de trabajo en Hadoop, se encuentra implementado con la arquitectura cliente-servidor, cuenta con los siguientes servicios *(Lublinsky, 2013)*:

- Oozie Flujo de trabajo: es un componente que proporciona soporte para definir y ejecutar una secuencia controlada de trabajos MapReduce, Hive y Pig.
- Paquetes de Oozie: permite definir y ejecutar un paquete de aplicaciones, lo que permite agruparlas en un conjunto y administrarlas juntas.
- Acuerdo de nivel de servicio (SLA) de Oozie: provee asistencia para el durante la ejecución de las aplicaciones en un flujo de trabajo.

En Apache Oozie el cliente puede mandar un flujo de trabajo y las bibliotecas necesarias al servidor Oozie, para después el servidor Oozie envié de manera individual los trabajos a Hadoop, mientras eso sucede el cliente puede obtener el estado del trabajo y administrarlo a través de este servicio. Para llevar a cabo un flujo de trabajo, Oozie utiliza flujos de control en el cual se pueden ejecutar una o más opciones:

- Ejecución de Job MapReduce
- Ejecución de script Pig o Hive
- Ejecución de programas Shell o Java.
- Manipulación de datos vía comandos HDFS
- Comandos remotos con SSH
- Enviar e-mails.

Para llevar a cabo un flujo de trabajo Oozie utiliza XML, es posible validar cada flujo de trabajo debido a que Oozie tiene una utilidad mediante línea de comandos para verificar si éste tiene errores o no y con base en ello poder ejecutarlo sin problemas.

La ventaja de esta herramienta es que se puede programar un flujo de trabajo el cual se puede ejecutar en ciertas fechas establecidas por el administrador y al final de cada proceso mandar una notificación que muestre los resultados obtenidos.

3.1.5 HIVE

Es una herramienta que fue creada por Facebook ahora un proyecto *open source* de apache, proporciona un alto nivel de abstracción sobre MapReduce sirve para realizar consultas tipo SQL sobre los datos utilizando un lenguaje llamado HiveQL *(Achari, 2015)*. Las consultas lanzadas son traducidas a lenguaje MapReduce y evita que analistas de

datos que no tienen conocimientos en programación Java tengan que llevar a cabo esta tarea *(Jain, 2017)*. Hive se ejecuta en la máquina cliente en donde se crean las sentencias HiveQL, posteriormente el interpretador Hive se encarga de hacer la conversión a código Java, optimizarlo, generar un archivo.jar, posteriormente se envía a un clúster Hadoop para su ejecución y finalmente se observa el resultado sin necesidad de programar en Java.

Hive consta de dos partes un Metastore que es el lugar donde éste guarda la estructura de las tablas detalladamente, mientras que los datos de manera física se guardan en el sistema de archivos HDFS, en otras palabras, las consultas operan sobre tablas de manera similar a una RDBMS, en donde una tabla es un directorio de HDFS que contiene uno o más archivos *(Achari, 2015)*. Cabe mencionar que los datos se almacenan en archivos de texto plano, y en el Metastore se crea la estructura que tiene la tabla de donde se cargarán los archivos, la manera de crear una tabla es muy similar a crear una tabla en una BDR, la extensión de un archivo de HiveQL.hql, acepta los tipos de datos primitivos típicos de una Base de datos relacional y otros más complejos como Arreglos y Mapas.

3.1.6 APACHE PIG

Apache Pig es una plataforma para realizar un análisis y procesamiento de datos en Hadoop, fue desarrollado originalmente por Yahoo, ahora proyecto *open source* de Apache, entre sus objetivos es lograr flexibilidad, productividad y mantenimiento. Los principales componentes de Pig son los siguientes: un lenguaje de flujo de datos, un Shell interactivo para ejecutar sentencias (grunt), su intérprete de Pig y un motor de ejecución *(Jain, 2017)*.

Pig maneja diferentes tipos de datos, los datos primitivos existentes en varios lenguajes de programación, datos tipo Map, Tuplas de manera similar a SQL y datos tipo Bag que es un contenedor de tuplas *(Achari, 2015)*. Pig es muy útil cuando es necesario realizar muchas combinaciones o existen muchas tablas intermedias, debido a que su sintaxis es muy simple y da a elegir al desarrollador entre usar Hive o este lenguaje.

3.1.7 APACHE MAHOUT

Apache Mahout es una librería de aprendizaje automático, se ejecuta en la parte superior de Hadoop, admite diferentes algoritmos de clasificación, agrupación, y filtrado colaborativo, además de proporcionar API de desarrollo para Java *(Vohra, 2016)*.

3.2 APACHE SPARK

Apache Spark es un motor de análisis unificado para el procesamiento de datos a gran escala *(Apache Software Fundation, 2018)*, es un framework de código abierto de uso general para computación distribuida en Big Data utilizando primitivas en memoria *(Ahmeda, 2016)*, puede ejecutarse en grupos de Hadoop a través del gestor de tareas YARN, sobre Apache Mesos, Kubernates de manera independiente o en la nube y busca atender las limitaciones de Hadoop *(Buyya, 2016)*. Apache Spark cuenta con escalabilidad lineal y tolerancia a fallas, y admite la integración de herramientas del ecosistema de Hadoop, tiene la capacidad de leer y escribir datos en formatos compatibles con MapReduce y de trabajar con bases de datos NOSQL como lo son Hbase y Cassandra *(Buyya, 2016)*.

A diferencia de MapReduce, Spark realiza el procesamiento de los datos en la memoria principal, cuando los datos no caben en memoria pasan a almacenarse en disco, lo que le permite funcionar bien con bloques de cualquier tamaño, de igual forma cuando se satura la memoria Cache y por algún motivo no hay suficiente espacio en ella y no es posible almacenarlos en la memoria principal se van directo a disco y desde allí son accedidos para su uso. Spark se puede ejecutar hasta 100 veces más rápido cuando se ejecuta en memoria y puede ser hasta 10 veces más rápido cuando se ejecuta en disco en comparación como modelos Map-reduce *(Achari, 2015)*, *(JayaLakshmi, 2018)*. Para ejecutar Apache Spark en un clúster es necesario que tenga un sistema de archivos compartido como HDFS o NFS, también es necesario que Java se encuentre instalado

en cada uno de los nodos o uno de los administradores de clústeres como YARN o Mesos *(Foundation, 2018)*.

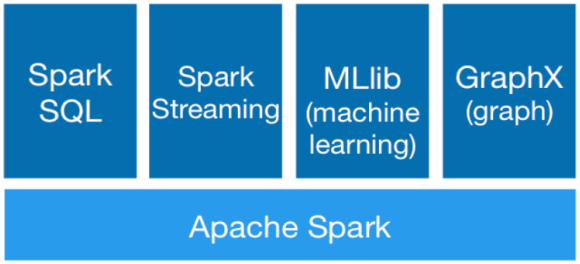


Figura 18 Apache Spark entorno de trabajo obtenido de (Apache Software Fundation, 2018).

Apache Spark tiene múltiples componentes integrados (Figura 18) entre ellos se tiene en primera instancia Apache Core, es la base o núcleo y en el apoyan las demás librerías y herramientas que hacen posible el Framework *(Apache Software Fundation, 2018)*. Apache Core es el encargado de la distribución de tareas, programación, lectura y escritura de datos todo ello a través de una interfaz ya sea Java, Python, Scala o R *(Ellingwood, 2016)*. La interfaz se centra en la abstracción de las RDDs, las cuales invocan operaciones como Map, Filter o Reduces además de que estas proporcionan implementaciones como la recopilación de datos distribuidos, la tolerancia a fallos de los nodos, la capacidad de utilizar varias fuentes de datos y el paralelismo *(JayaLakshmi, 2018)*.

3.2.1 SPARK SQL

Es un paquete de Apache Spark diseñado con el objetivo de trabajar con datos estructurados, este proporciona una interfaz del tipo SQL para realizar tareas con esos datos. Se puede escribir un Spark Query y la sentencia será muy similar a SQL, de este modo proporciona una abstracción de estilo SQL que simplifica el trabajo con conjuntos de datos estructurados. Entre las principales características de Spark SQL destacan las siguientes *(Apache Software Fundation, 2018)*:

- Es utilizable en Java, Python, Scala o R
- Acceso uniforme a datos: permite conectarse a cualquier base de datos de la misma manera e incluso es capaz de unir datos entre ellas.
- Integración: Permite ejecutar consultas SQL O HiveQL en almacenes ya existentes debido a que Spark SQL admite la sintaxis de HiveQL.
- Conectividad Estándar: permite la conectividad a través de JDBC o ODBC.

- Puede cargar datos de diversas fuentes como JSON, Hive y Parquet *(Karau, 2015)*.

3.2.2 SPARK STREAMING

Es una extensión de Spark que puede ejecutar tareas a lo largo de un intervalo de tiempo (un intervalo de micro bloques), el flujo de datos que llega a tiempo se divide en muchas partes para su procesamiento en paralelo *(Petrov, 2018)*, además es tolerante al flujo de datos en vivo lo que lo hace ideal para el procesamiento y análisis en tiempo real *(Buyya, 2016)*. Es posible ejecutar Spark Streaming en un clúster independientemente del administrador de recursos, es compatible con HDFS (normalmente usado en producción por alta disponibilidad), FLUME, KAFCA y ZeroMQ, es posible desarrollar aplicaciones analíticas e interactivas y cuenta con un entorno local de desarrollo.

3.2.3 MLLIB (MACHINE LEARNING) Y GRAPHX

Gracias a que el núcleo de Apache Spark es rápido y de propósitos generales es ideal con múltiples componentes especializados de alto nivel para varias cargas de trabajo como lo es el aprendizaje automático. MLlib es la librería de aprendizaje automático de Apache Spark que tiene algoritmos de clasificación, regresión, agrupación, filtrado colaborativo *(Apache Software Fundation, 2018)* y otras primitivas de bajo nivel como el algoritmo de optimización de descenso del gradiente genérico, algunos de ellos desarrollados por la misma comunidad de Apache Software Fundation *(Karau, 2015)*.

Los algoritmos de MLlib son para funcionamiento en paralelo que se ejecutan bien en clústeres, con la finalidad de aprovechar al máximo las capacidades de un sistema distribuido, aspecto por el que viejos algoritmos que no están diseñados para este tipo de plataformas son desechados al momento del desarrollo de MLlib *(Karau, 2015)*.

Por otro lado, GraphX es la Api de Spark para trabajar con algoritmos basados en grafos, esta cuenta con una amplia gama de algoritmos ya implementados *(Achari, 2015)*, es mejorada constantemente ya que con cada nueva versión de Spark también esta es actualizada *(JayaLakshmi, 2018)*.

3.3 APACHE FLINK

Apache Flink es un entorno de trabajo distribuido para análisis de datos a través de flujos de datos limitados e ilimitados *(The Apache Software Foundation, 2019)*, es capaz de manejar tareas por lotes, en donde los lotes son considerados como flujos de datos finitos y es por ello por lo que trata al procesamiento como un subconjunto de procesamiento de flujo *(Ellingwood, 2016)*.

Flink tiene la capacidad de ejecutarse en entornos de clústeres comunes, realizar cálculos en memoria y a cualquier escala. Cuenta con una arquitectura maestro-esclavo, y ésta se encuentra compuesta por un Job Manager y uno o varios administradores de tareas. El Job mánager (o nodo maestro) tiene como función coordinar los cálculos en el sistema Flink (Figura 19), mientras que los administradores de tareas se encargan de ejecutar los programas de manera distribuida *(Sonia Bergamaschi, 2017)*.

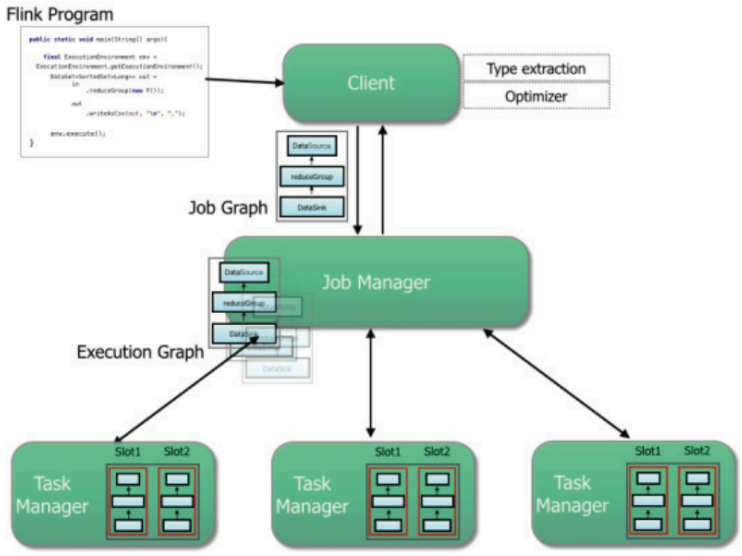


Figura 19 Ciclo de vida de la ejecución de una aplicación en Flink obtenido de (Mahmood, 2016)

Flink admite diferentes lapsos de tiempo (tiempo de evento, tiempo de ingestión, tiempo de procesamiento), esto ayuda a los programadores a controlar como se correlacionan los eventos. Una ventaja de Flink es que su arquitectura es transparente a los programadores, por lo que con sólo hacer uso de su API se llevan a cabo los programas. Apache Flink puede procesar los datos como flujos ilimitados (*Streaming*) o acotados (por lotes) *(The Apache Software Foundation, 2019)*.

- Flujos de datos sin límites: este tipo de flujos tienen un inicio, pero no un final definido, los datos son proporcionados a medida que estos son generados, se procesan continuamente, es decir se gestionan en lapsos de tiempos definidos y

para garantizar la integridad de resultados en el procesamiento es necesario que los datos entren en un orden especifico al momento de la ingesta de estos.

- Flujos de datos acotados: en los datos acotados, la ingesta de datos puede darse antes de que estos sean procesados (como en Apache Hadoop), debido a que el tamaño se encuentra definido, en este caso como el conjunto de datos es finito el procesamiento generalmente se hace por lotes.

Apache Flink tiene la capacidad de integrarse con todos los administradores de clústeres más comunes como Hadoop YARN, Mesos y Kubernates *(The Apache Software Foundation, 2019)* y se encuentra diseñado para trabajar de manera correcta con cualquiera de ellos, aun así, tiene la opción de configurarse como un clúster independiente.

Apache Flink se encuentra diseñado para ejecutar aplicaciones en cualquier escala, mediante la división de cargas grandes en tareas más ligeras, que son distribuidas a través del clúster, con la ventaja de aprovechar al máximo toda la memoria principal, recursos CPU y disco. Es por lo que Flink se basa en una arquitectura Lambda, donde la idea principal es construir arquitecturas Big Data con base en diferentes capas, y cada una de estas es basada en la funcionalidad que se encuentra debajo de ellas.

En la figura 20 se observa el funcionamiento de la arquitectura Lambda en primer término la nueva información adquirida por el sistema es enviada a la capa de lotes y a la capa de velocidad. En la capa de lotes la información se prepara para su gestión los datos se encuentran en su estado original, a la nueva información se le aplican funciones específicas mediante un proceso Batch, esta capa emite vistas de lotes (Batch Views) como resultado de esas funciones.

Posteriormente es necesario cargar las vistas de lotes a un lugar donde sea posible consultarlas, es allí en donde la capa de servicio (Serving Layer) entra en función, ya que se encarga de realizar un indexado de las vistas de lotes para que puedan ser accedidas

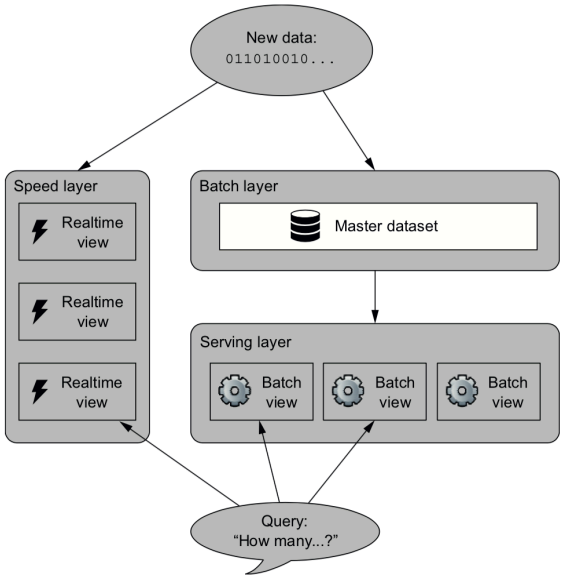


Figura 20 Diagrama de la arquitectura Lambda obtenido de (JamesWarren, 2015)

con la menor latencia posible, permite realizar lecturas aleatorias en ella, cuando se cargan nuevas vistas de lotes hace un intercambio automático para que así se muestren resultados más actualizados.

La capa de servicio se actualiza de manera constante cada vez que la capa de lotes termina un pre procesamiento de una vista por lotes, por lo que existen datos que no se encuentran disponibles los cuales ingresan cuando se encuentra ejecutando una función de pre procesado de datos *(JamesWarren, 2015)*. Es por lo que se cuenta con un sistema de procesamiento de datos en tiempo real (capa de velocidad) toma los datos que se ingresan en las últimas horas con el fin de compensar las latencias que se tienen la capa de lotes y de servicio este es su propósito principal.

A diferencia de la capa de Lotes la capa de velocidad logra el mismo trabajo en menor tiempo y actualiza las vistas a medida que llegan nuevos datos en lugar de volver a calcular todos ellos nuevamente, por último, es posible resolver consultas fusionando los resultados de las vistas por lotes y en tiempo real *(JamesWarren, 2015).*

Entre las ventajas de esta arquitectura es que en cada una de las capas es posible implementar diferentes herramientas de acuerdo con la necesidad del problema a resolver, por ejemplo en la parte de almacenamiento de la capa de servicio no es necesario el utilizar una base de datos tan compleja que permita escrituras aleatorias mientras que en la capa de velocidad si lo es, existen varias bases de datos para resolver esos problemas para la primera vertiente es posible utilizar ELEPHANT DB para la primer vertiente y Cassandra para la segunda *(Buyya, 2016)*.

La Figura 21 muestra de forma concentrada las características de los entornos de trabajo explicados anteriormente.

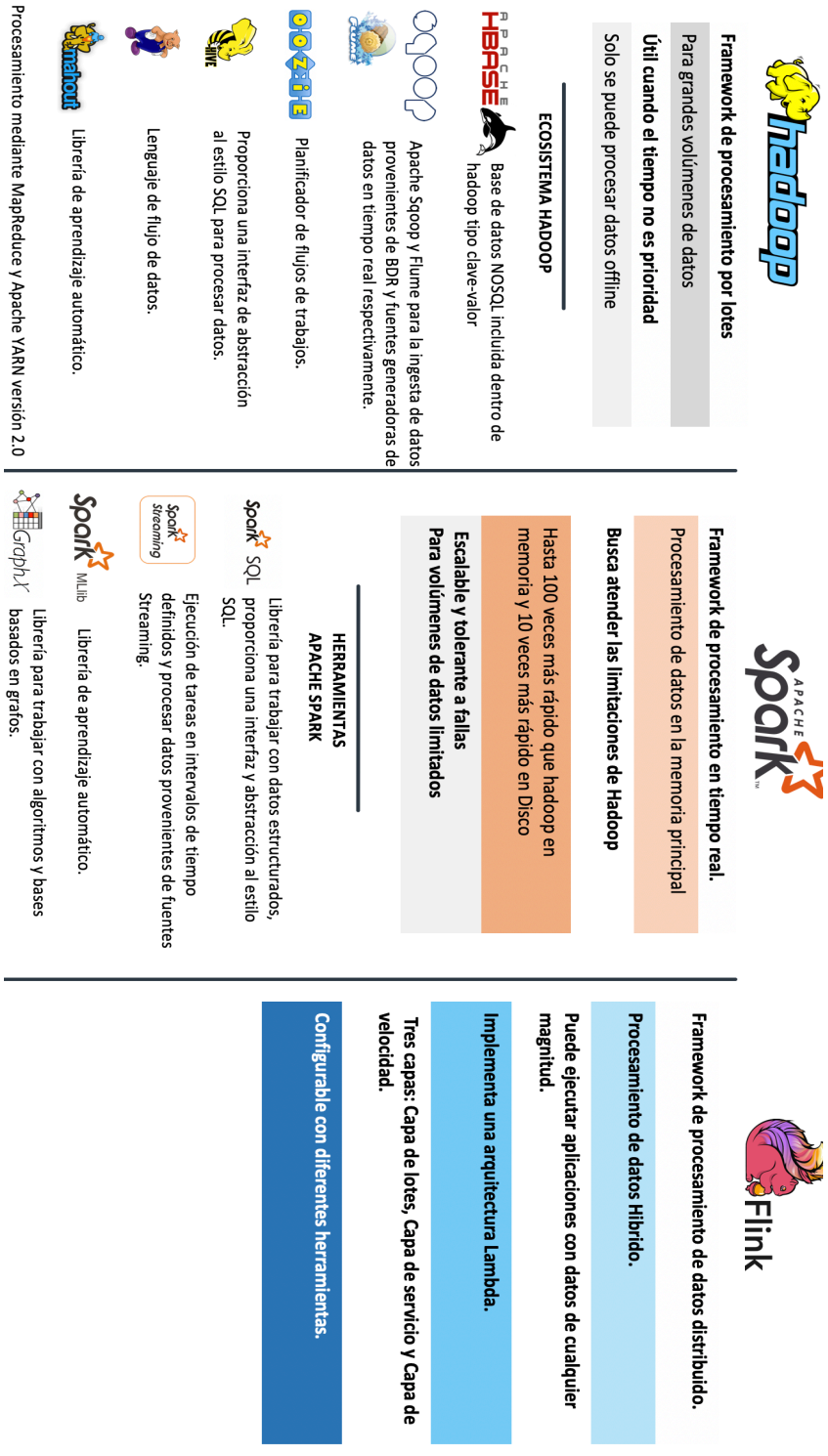


Figura 21 Resumen de entornos de trabajo para Big Data

CAPÍTULO IV
RETOS Y OPORTUNIDADES DEL BIG DATA

4.1 RETOS

Big Data proporciona grandes beneficios en cuanto análisis, desarrollo e implementación de la tecnología, no obstante, el analizar datos con ella tiene varios desafíos antes de lograr los resultados esperados, uno de los primeros retos al que se enfrenta en el mundo de BD es el administrar enormes repositorios de datos y no necesariamente a almacenarla como tal si no también a la extracción y el comprender como utilizarla.

Existen varios retos para Big Data de los cuales se pueden clasificar de la siguiente manera:

- Retos relacionados con los datos.
  - **a)** Retos relacionados con las características de los datos.
  - **b)** Retos relacionados con el proceso de los datos
  - **c)** Retos relacionados con la gestión de los datos.
- Retos relacionados con la adopción de la tecnología: no todas las empresas están dispuestas a adoptar esta nueva tecnología, debido a los gastos y desafíos que ésta representa.
- Retos relacionados con la gobernanza de datos.

4.1.1 RETOS RELACIONADOS CON LAS CARACTERÍSTICAS DE LOS DATOS

Cuando se habla de los retos relacionados con los datos se refiere al volumen, la velocidad, la variedad, veracidad y calidad de los datos (Figura 22) *(L'HEUREUX, 2017)*.

- **Retos para el volumen:** Por parte del volumen las escalas de datos que se manejan con Big Data van desde Terabytes a Exabytes, lo que hace que sea un

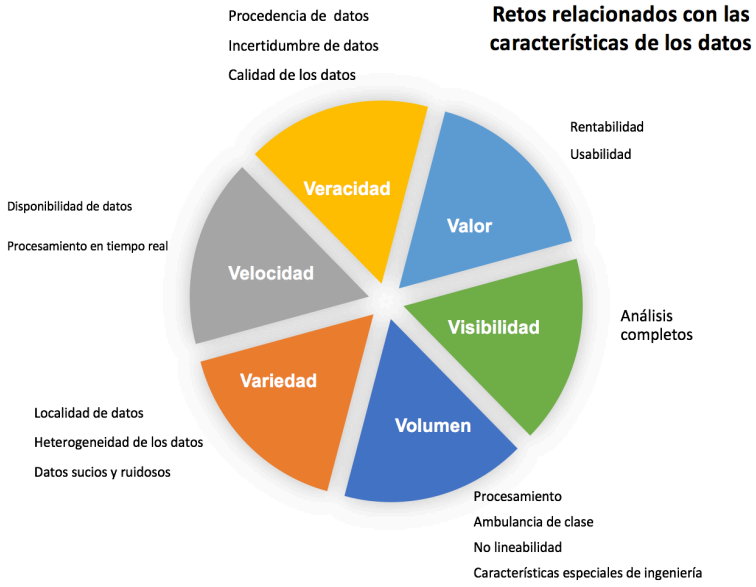


Figura 22 Retos relacionados con las características de los datos obtenido de (L'HEUREUX, 2017) y modificado agregando dos V (valor y visibilidad)

desafío obtener la información deseada de esos conjuntos de datos en tiempos aceptables *(MustafaKamal, 2017)*, debido a que esa escala aumenta de manera importante la complejidad computacional, lo que hace que incluso operaciones triviales se vuelvan costosas, por lo tanto, el aumento en tamaño de datos incrementara drásticamente el tiempo empleado para procesar los datos y en ocasiones es posible que algunos algoritmos de análisis de datos queden inútiles ante este problema. A medida que los datos aumenten se debe realizar un análisis de la arquitectura que se esta utilizando ya que el rendimiento de los algoritmos dependerá en gran parte de ello *(L'HEUREUX, 2017)*.

En ocasiones muchos algoritmos en el momento de ser procesados los datos requieren que los cálculos se vayan guardando en la memoria principal o en un sólo archivo en disco, pero cuando el tamaño de datos es demasiado grande y la memoria principal no es suficiente el acceder a disco en ocasiones puede conducir al fracaso ya que el tiempo se expande de lectura y escritura.

- **Retos para la variedad:** es un desafío muy importante ya que los datos se encuentran en diferentes formatos y en la mayor parte de los casos no siguen una plantilla o formato *(MustafaKamal, 2017)*. Otro aspecto importante es la ubicación de los datos, puesto que en ocasiones se asume que los datos se encuentran concentrados en una sola base de datos o un conjunto de archivos, sin embargo, la mayoría de las veces no es así por lo general se encuentran en una gran cantidad de archivos los cuales se ubican en diferentes ubicaciones físicas.

Otro desafío es la heterogeneidad sintáctica de los datos ya que el análisis de Big Data involucra integración de fuentes diversas, y los datos generalmente son diferentes en términos de formato, modelo de datos, tipos de datos, formatos de archivo y codificación de datos. Por otro lado, se tiene la heterogeneidad semántica y va más con las diferencias en los significados e interpretaciones, la dificultad aumenta cuando se integran conjuntos de datos desarrollados por diferentes partes.

Por último, un desafío importante está relacionado con la calidad de los datos ya que estos pueden estar *sucios* y causar ruido, lo cual dificulta y entorpece la disponibilidad de los datos para el análisis lo cual hace que se incrementen los recursos a cuanto tiempo para prepararlos para el análisis *(L'HEUREUX, 2017)*.

- **Retos para la velocidad:** el procesar muchas cantidades datos e implementar algoritmos dependen mucho en la disponibilidad de los datos, en si ya se encuentran concentrados al momento o estarán entrando a medida que se realiza el procesamiento de ellos se tiene que tomar en cuenta la transmisión de datos y revisar si los datos llegan a tiempo para continuar con el trabajo *(L'HEUREUX, 2017)*. Las arquitecturas tradicionales no se encuentran diseñadas para el manejo de flujos constantes, lo que lleva a la necesidad de procesamiento en tiempo real ya que le valor comercial de un sistema de este tipo reside en la capacidad de respuesta instantánea.
- **Retos para la veracidad:** la confiabilidad de los datos depende la procedencia de los datos, la calidad de las fuentes entre otras *(L'HEUREUX, 2017)*. Es importante identificar el origen de los datos ya que con ello es posible identificar los pasos, las transacciones y procesos a los cuales fueron sometidos los datos que se detecten como no validos, lo que hace tener una información contextual que ayude a un mejor procesamiento y a la implementación de algoritmos de aprendizaje automático.

  Actualmente se reúne la mayor cantidad de datos posible sin embargo los medios utilizados para la recopilación de datos en ocasiones pueden producir incertidumbre y por tanto impactar a la veracidad de los datos *(MustafaKamal, 2017)*.

- **Retos para el valor:** el principal reto dentro del valor es que el costo del sistema no sobrepase el valor derivado del uso de esta, que la usabilidad sea y las ganancias sean mayores después de realizar el procesamiento de los datos, exista una mejora de procesos y eficiencia de operaciones.
- **Retos para la visibilidad:** es complejo obtener una imagen completa de todos los datos almacenados, además en ocasiones es complicado entender los resultados obtenidos de los análisis exploratorios lo que entorpece el descubrimiento de nuevos hallazgos interesantes.

4.1.2 RETOS RELACIONADOS CON EL PROCESO DE LOS DATOS

Durante el procesamiento y análisis se presentan un grupo de desafíos, los cuales comienzan desde el momento de la obtención de datos hasta la interpretación y resultados *(MEHMOOD, 2016)*.

- Retos en la adquisición e integración de datos: este reto se encuentra relacionado con la obtención de datos de diversas fuentes de almacenamiento, una de las principales barreras se da con la procedencia de los datos y la discrepancia en la procedencia de los datos. En consecuencia, ello afecta la capacidad de extraer información procesable, por lo que es necesario aplicar filtros inteligentes para mejorar la calidad de los datos, acotar opciones de captura para eliminar captura de información no deseada y descartar inconsistencias. Por otro lado, el integrar enormes cantidades de datos representa un reto debido a que en muchas ocasiones ellos carecen de información que se vincule.
- Retos para el análisis de datos: dentro del análisis de datos uno de los principales retos es entender las salidas obtenidas después del procesamiento de datos, y usar la herramienta adecuada para su análisis.
- Retos en la interpretación de datos: los resultados deben de ser comprensibles y útiles para la toma de decisiones.

4.1.3 RETOS RELACIONADOS CON LA GESTIÓN DE LOS DATOS

Big Data tiene un enorme desafío para lograr una administración y cuidar esos datos de accesos no deseados, hoy día la cantidad de repositorios de datos son extremadamente grandes, muchos de ellos son de carácter confidencial como datos relacionados con el seguro social, declaraciones de impuestos, procedimientos médicos, transacciones financieras y otros datos personales importantes *(MustafaKamal, 2017)*. Es por lo que para una organización representa un desafío cuidar de ellos, contar con una estructura solida, manejo de políticas dentro de la organización que sean capaces de lograr un régimen correcto de la información dentro y fuera de las instalaciones, procedimientos correctos, capacitación de los empleados para el manejo correcto de los datos *(MEHMOOD, 2016)*.

4.1.4 RETOS RELACIONADOS CON ADOPCIÓN DE LA TECNOLOGÍA

Aun y cuando Big Data proporciona grandes beneficios y muchas empresas están apostando por esta tecnología existen desafíos hacen que el adoptar esta tecnología sea difícil entre ellos los siguientes:

- Gastos operacionales: los datos están en constante aumento, lo que conlleva costos de almacenamiento, costos de mantenimiento y costos de rendimiento. Big Data mantiene los datos de manera distribuida con el fin de tener redundancia, además de que ello conlleva nuevos gastos de adquisición de hardware y procesamiento.
- Capacitación de personal: actualmente el personal dentro de las empresas en su mayoría no cuenta con la suficiente capacitación para hacer uso de Big Data, en consecuencia, el capacitar al personal puede generar importantes gastos y tiempo.
- Cambio de paradigma: no todos están dispuestos a cambiar las soluciones implementadas actuales e intentar con nuevas tecnologías.

4.1.5 RETOS RELACIONADOS CON LA GOBERNANZA DE LOS DATOS

La gobernanza de los datos es tomada por las organizaciones como un punto de partida para obtener enfoques potenciales para mejorar la calidad de sus datos, por lo que los datos deben estar alineados con las políticas de gobierno corporativo con el objetivo de aprovechar al máximo la información y mantenerla como su activo clave para la organización y respaldar el logro de las operaciones *(Ghavami, 2016)*.

Mediante la formulación de políticas y el conjunto de procesos bien aplicado se busca que garantice que los activos de datos importantes se gestionen de manera correcta, formal y sistemática en toda la empresa *(Ghavami, 2016)*. Sin embargo, además de enfrentar retos relacionados con el crecimiento de datos, la infraestructura y la escalabilidad existen algunos otros referentes a la categorización, el modelado y el mapeo de datos a medida que estos son capturados y almacenados, esto es debido a la naturaleza desestructurada de los datos y a su vez compleja. Por ello la gobernanza juega un papel importante debido a que si esta es eficaz se puede garantizar que los datos que se utilizarán para posteriores análisis serán de una calidad superior lo que puede agilizar procesos posteriores.

4.2 OPORTUNIDADES

Big Data tiene una extensa gama de oportunidades en las organizaciones, su éxito depende de descubrir tendencias y generar nuevas ideas con base al análisis de datos, por lo que es necesario tomar medidas, anticipar nuevas decisiones e ir de la mano con la innovación. Muchas empresas adoptan esta tecnología para resolver problemas en áreas particulares, encontrar un enriquecimiento agregado al analizar sus repositorios de datos, de igual forma buscar economizar en cuanto a infraestructura y resolver problemas al tener un almacén enorme de datos además de realizar acciones productivas con ellos para obtener ventajas adicionales.

Desde grandes empresas hasta pequeños comercios de distintos sectores buscan tener un mayor conocimiento de su mercado ya que puede mejorar la interacción organización-

persona, por lo que, si son capaces de recopilar información acerca del comportamiento, transacciones, preferencias de búsqueda, representa un conjunto de oportunidades al negocio para entregar servicios de mejor calidad y personalizados. Actualmente con el aumento de las ventas en línea, la proliferación de dispositivos móviles ayuda a aprender sobre el comportamiento de los clientes, su perfil, sus preferencias, lo que ayuda a favorecer a empresas para así enfocarse en ciertos productos o servicios, mejorando sus ventas e ingresos *(Portela, 2016)*.

En base a IDC se tiene una predicción para el 2021 un 10% de los gastos a nivel empresarial serán destinados a software, datos e implementación de nuevos algoritmos, con el objetivo de mejorar la productividad y automatizar procesos *(IDC, 2018)*. El tamaño de las empresas influirá en el tamaño de los ingresos que estas adquirirán, se estima que aquellas empresas con un capital humano mayor a los 1000 empleados serán las responsables de dos terceras partes de las oportunidades y beneficios con Big Data, sin embargo, a su vez también serán las que mayor inversión harán más de un 47%, mientras que las pequeñas y medianas empresas serán muy importantes con una cuarta parte de los ingresos mundiales *(IDC, 2018)*, el resto será para otros sectores.

La Figura 23 resume las oportunidades que se pueden obtener con el uso de Big data va desde el valor agregado a la información, servicios de mayor calidad y áreas nuevas de oportunidad.

Figura 23 Resumen de oportunidades con Big Data

CAPÍTULO V
SEGURIDAD Y TENDENCIAS DEL BIG DATA

5.1 SEGURIDAD EN BIG DATA

Uno de los mayores desafíos para Big Data es la protección de la privacidad de los datos, así como la de los individuos que proporcionan esta información. La privacidad se ve afectada por el desarrollo de la tecnología, ya que la mayoría de población a nivel mundial actualmente cuenta con herramientas disponibles para realizar búsquedas de internet, teléfonos inteligentes y se encuentran registrados en una o más redes sociales donde comparten información personal, videos, fotografías y audios de manera continua *(YU, 2016)*.

En muchas áreas se recopilan datos provenientes de diferentes fuentes, cada vez crecen más y las organizaciones enfrentan desafíos en cuanto a la privacidad. Empresas como Google y Amazon son capaces de conocer nuestras preferencias en cuanto a hábitos de navegación y compras electrónicas. Redes sociales como Facebook almacenan información sobre nuestras relaciones sociales y vida personal, Instagram guarda toda nuestra experiencia de diversos lugares, YouTube hace recomendaciones multimedia basadas en nuestro historial de búsqueda, todo ello con el impulso de Big data para la recopilación, almacenamiento y reutilización de la información con la finalidad de generar beneficios comerciales, que a su vez ponen en duda la privacidad y seguridad de los usuarios *(MEHMOOD, 2016)*.

No obstante, la privacidad de los usuarios puede ser violada en las siguientes circunstancias *(MEHMOOD, 2016)*:

- Cuando en una base de datos se encuentra información personal almacenada y se combina con otros conjuntos de datos externos a ella, se produce una violación a la privacidad debido a que esta mezcla puede inferir en hechos de los usuarios, por lo que estos datos deben ser secretos y no ser revelados a otros.

- Cuando la información de un usuario es recopilada con el fin de darle valor agregado a un negocio, (los hábitos de compra de una persona pueden revelar mucho sobre su estilo de vida e información personal).
- Cuando información confidencial es almacenada y procesada en ubicaciones que no cuenten con suficiente seguridad, lo que en consecuencia puede ocasionar fugas de datos en fases de almacenamiento o procesamiento.

Son cada vez más los dispositivos dotados de un sensor capaces de reunir datos relacionados con las actividades diarias de las personas, la mayor parte de las aplicaciones son realizadas por los proveedores de servicios para sus clientes, se basan normalmente en la ubicación del usuario, la información puede ser entrelazada con lugares específicos como el hogar, el trabajo entre otros, por lo que estos datos deben de ser protegidos y realizar grandes esfuerzos ya que en ocasiones pueden ser accedidos por agencias de seguridad y gubernamentales o usuarios maliciosos.

Derivado de lo anterior expuesto, la seguridad es un tema que debe ser abordado de la mejor manera posible para evitar consecuencias criticas tanto para los usuarios como la organización, no obstante, el desafío crece de forma significativa cuando los datos no se encuentran omnipresentes, cuando no se tienen controles de seguridad adecuados que garanticen que los datos solo serán accedidos por personas autorizadas, para combatir ello en muchos lugares se llevan análisis de registros, revisiones constantes de los flujos a través de la red, detección de eventos anormales en los sistemas y detección de intrusos.

Para proteger la privacidad de los datos durante el ciclo de vida de Big Data se están desarrollando mecanismos que buscan realizar esta tarea, tal como se describe a continuación.

5.1.1 PRIVACIDAD EN LA GENERACIÓN DE DATOS

Existe la generación de datos activos y pasivos, cuando el propietario proporciona su información a terceros y acepta sus políticas en el caso de los activos, mientras que los datos pasivos se generan por la actividad concurrente en línea por parte del usuario,

durante este proceso se busca evitar la falsificación y restringir el acceso a ellos a agentes externos.

El usuario se encuentra en su derecho de no compartir datos cuando estos revelen demasiada información o el juzgue que no debe compartirse y en caso de proporcionarlos, debe poner en practica métodos de control efectivos para que un tercero no pueda hurtar sus datos, poner en funcionamiento herramientas de cifrado, o utilidades que distorsionen información verdadera para evitar acceso a datos confidenciales, el uso de estas medidas no garantiza la protección total de los datos, sin embargo, puede mejorar considerablemente la privacidad.

5.1.2 PRIVACIDAD EN EL ALMACENAMIENTO DE DATOS

Muchas instituciones almacenan inmensas cantidades de datos y si estas se ven expuestas puede ser perjudicial ya que la información respecto a los usuarios puede ser divulgada, por esa razón las instituciones deben implementar controles de seguridad ya sea a nivel archivo, seguridad a nivel base de datos o a nivel aplicación con la meta de cumplir con las normas de seguridad de los datos y generar confianza para los usuarios.

5.1.3 PRIVACIDAD EN EL PROCESAMIENTO DE LOS DATOS

La privacidad también se debe cuidar en el procesamiento de los datos y se puede dividir en dos partes, entre tanto la primera la finalidad es proteger la información de la no divulgación debido a la información delicada que puede contener sobre el propietario de los datos, mientras que la segunda involucra extraer información importante sin violar la privacidad.

Previo al procesamiento se pueden modificar los datos mediante técnicas de anonimizacion para evitar revelar información personal del propietario, eliminado identificadores antes de almacenarlos para su procesamiento. Una vez realizado ese proceso los valores confidenciales se ocultan, y entre las técnicas de anonimizacion se encuentran las siguientes:

- Generalización: busca reemplazar el valor de atributos específicos por una descripción menos especifica.

- Supresión: se refiere al reemplazo de valores con otro tipo de caracteres con la finalidad de proteger la información un ejemplo de ello es el número de seguro social en donde algunos valores son ocultados con xxxx xx 1923, con el fin de proteger la identidad del dueño
- Permutación: la relación cuasi-identificador y el atributo sensible se anula al dividir un conjunto de registros en grupos.
- Perturbación: Los valores originales son modificados por valores sintéticos en los cuales la modificación no difiere de manera significativa la información estadística calculada a partir de los datos originales.

La implementación de diversas técnicas colabora para mantener la privacidad, sin embargo, estudios recientes muestran que conjuntos de datos pueden ser atacados. En un estudio se realizó una recopilación de movilidad durante 15 meses de 1.5 millones de personas a los que se les aplicó operaciones de anonimizacion para eliminar datos significativos como el nombre, dirección, numero telefónico. Cada hora se obtenía nueva información de cada individuo, posterior a que se llevó a cabo el procesamiento del conjunto de datos se obtuvieron resultados importantes arrojando que es posible identificar a una persona con un 95% de efectividad con sólo cuatro puntos espacio temporales *(YU, 2016)*.

5.2 TENDENCIAS DEL BIG DATA

Con Big Data nuevos modelos de negocio se aproximan, los servicios serán cada vez mejores y personalizados de acuerdo con la necesidad de los usuarios. El poder de las empresas se definirá por el software que ejecuten, por que tan informada sea su toma de decisiones en todos los niveles, el buen manejo de los costos operativos mediante la búsqueda de reducción de gastos y generar mayores ganancias. Algunas de las principales tendencias se describen a continuación.

5.2.1 TENDENCIAS DE BIG DATA PARA DATOS Y DISPOSITIVOS

Cada vez crece el número de dispositivos que cuentan con sensores y son capaces de generar datos y conectarse entre sí mediante una red, actualmente el universo digital

crece alrededor de un 40% anualmente, debido a que por lo menos 2000 mil millones de personas realizan su trabajo en línea y por los millones de dispositivos con sensores que envían y reciben datos *(IDC, 2014)*.

De acuerdo con un estudio realizado por *EMC Digital Universe* se prevé para el año 2020 un incremento importante en los datos y se pase de 4.4 billones de GB de datos existentes en el mundo en el año 2014 a 44 Billones de datos almacenados, dando un incremento 10 veces su valor en tan solo 6 años, el incremento es drástico. Pero el crecimiento de datos no viene sólo, el número de dispositivos en el planeta se incrementará, para el 2014 se contaba con aproximadamente 200 mil millones, de los cuales solamente un 7% (14 mil millones) se encontraban interconectados entre si, pero para el 2020 se pronostica un incremento a 32 billones de dispositivos conectados y generando datos de manera continua *(IDC, 2014)*.

IOT anuncia una nueva era en la informática, para el año 2014 este tipo de tecnologías ya abarcaba un 2% y se espera que para el 2020 represente un 10% de todo el universo digital. Los dispositivos móviles también juegan un papel muy importante ya que se pronostica un crecimiento de generación de datos mediante este tipo de dispositivos de un 17% en el 2013 a un 27% para el 2020 *(IDC, 2014)*.

5.2.2 TENDENCIAS DE INVERSIÓN CON BIG DATA

A medida que pasa el tiempo son más las organizaciones que están implementando soluciones relacionadas con Big Data. Las industrias que en la actualidad realizan las mayores inversiones y soluciones para el análisis de negocios son las siguientes: la banca, la manufactura discreta, manufactura de procesos, servicios profesionales y gobiernos federal/central los cuales representan casi la mitad y se prevé que para el 2022 estas sean las industrias con mayores oportunidades de crecimiento y su inversión será de 129 mil millones de dólares *(IDC, 2018)*.

Estados Unidos es el mercado de mayor tamaño en soluciones Big Data con más 88 mil millones de dólares de ingresos, lo que representa aproximadamente un 50% del total en el mundo, tendencia que se sigue repitiendo. Lo sigue Europa occidental debido a que en el 2018 alcanzo alrededor de 35 mil millones, posteriormente países de Asia/

Pacifico sin incluir a Japón alcanzaron 23.9 mil millones, con un crecimiento a 27 mil millones para el 2022 con un crecimiento concurrente del 15.1% entre el 2017-22 *(IDC, 2018)*.

Japón promete ser el segundo país con más inversiones respecto a Big Data entre 2018 y 2019 seguido por países como Reino Unido, Alemania y China. Mientras que los países con mayor crecimiento en soluciones relacionadas con Big Data son Argentina con un 20.8%, Vietnam con un 19%, Filipinas con 19.5% e indonesia con un 19.4% *(IDC, 2018)*.

En el caso de Europa Central y Oriental se prevé que los ingresos por soluciones basadas en Big Data sean de 5400 millones de dólares en 2022, con un crecimiento anual del 11.3% en un periodo de 2017-22. Entre 2017 y principios de 2018 Rusia presento la mayor cantidad de ingresos un 40% con 1400 millones de dólares, seguida de Polonia con 850 millones y otros como Eslovaquia, Ucrania, Croacia con un 15% y República Checa con un 12.7% *(IDC, 2018)*.

5.2.3 TENDENCIAS DE INGRESOS CON BIG DATA

A nivel mundial empresas de diversos sectores actualmente ya aplican soluciones basadas en Big Data. Empresas como la banca, manufactura, servicios profesionales y gobiernos federales/centralizados los cuales combinados representaron un 49% (81 mil millones de dólares) de los 166 mil millones del total los ingresos en el 2018. Para el 2022 se espera que los ingresos mundiales mediante soluciones Big Data y Business Analytics alcancen los $ 260 mil millones de dólares con una tasa de crecimiento de un 11.9% en el periodo 2017-2022, mientras que las industrias con mayor crecimiento en cuanto a ingresos de Big Data y Business Analytics son las industrias minoristas con un 13.5%, la banca con un 13.2% y servicios profesionales con un 12.9% *(IDC, 2018)*.

5.2.4 TENDENCIAS DE SOFTWARE BIG DATA

Más del 50% de los ingresos obtenidos con Big Data se destinarán a TI y servicios empresariales *(IDC, 2018)*, de igual manera los ingresos relacionados con los servicios se encontrarán con un rápido crecimiento con un crecimiento anual compuesto de 13.2%. Se prevé un aumento significativo en nuevas inversiones para software a unos 90 mil

millones en 2022, en herramientas de consulta, informes, análisis de usuarios finales, herramientas de administración de datos *(IDC, 2018)*.

Se pronostica un gran crecimiento en dos categorías importantes en cuanto a tecnología por un lado las plataformas de software cognitivo/IA con un 36.5% aproximadamente y un 30.5% para los almacenes analíticos no relaciones (Bases de Datos NOSQL). De igual manera se estima un incremento en servidores, equipo de computo relacionado, almacenamiento del 7.3% y 27 mil millones en 2022 con respecto al 2018.

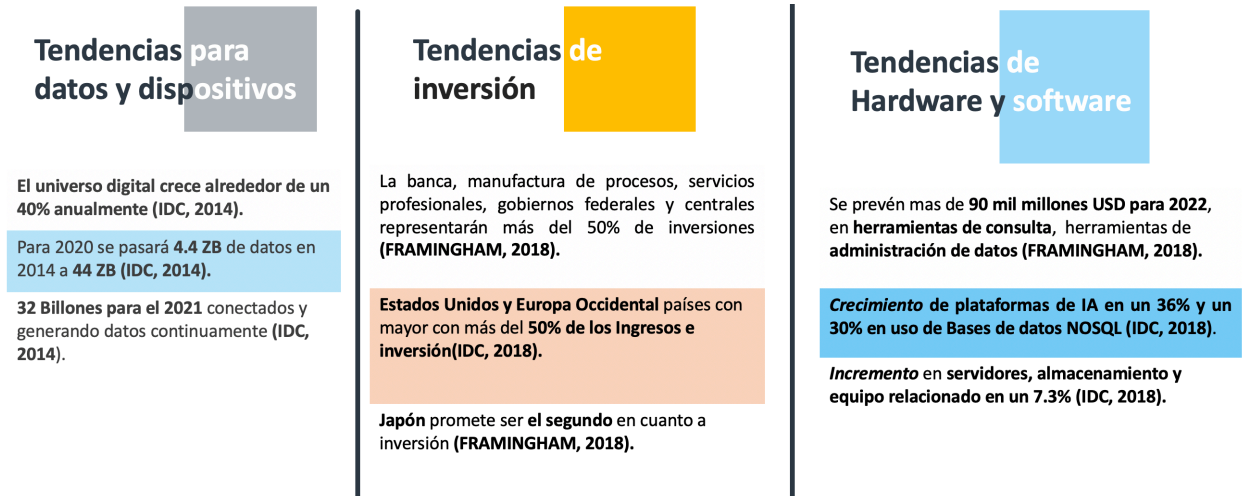


Figura 24 Resumen de Tendencias para Big Data

CAPÍTULO VI
BENEFICIOS Y DESVENTAJAS DE BIG DATA

Los beneficios del Big Data pueden expresarse desde dos perspectivas: respecto a las tecnologías tradicionales y en la mejora de las organizaciones.

6.1 BENEFICIOS DE BIG DATA FRENTE A TECNOLOGÍAS TRADICIONALES

Big Data tiene importantes beneficios en contraste con herramientas de procesamiento de datos tradicionales como las bases de datos y es posible separarlas de la siguiente manera *(LIANG, 2018)*:

- **Integridad:** esta no sólo se limita a capturar datos principales o los más significativos, si no que recopila datos relacionados con detalles más específicos de una determinada actividad para un futuro análisis. Cuando se hace uso de herramientas tradicionales, regularmente los datos capturados se limitan a resolver un problema, ello debido a que tienen que cuidar detalles como el almacenamiento y el rendimiento de la base de datos, mientras que con Big Data se persigue registrar la mayor cantidad de información posible, integrar mayor cantidad de detalles para más tarde analizarlos y con ello entender comportamientos y poder proporcionar servicios personalizados de acuerdo a los gustos de los usuarios finales.
- **Persistencia:** con Big Data es posible proceder a la captura de información constante, en el tiempo actual existe gran cantidad de dispositivos con sensores que son capaces de realizar lecturas de diferentes variables, lo cual produce una cantidad significativa de datos que son de gran utilidad. Un ejemplo de ello son dispositivos inteligentes que puedes acompañar con ropa o como un accesorio más, con la capacidad de medir presión arterial, temperatura corporal entre otras.
- **Multiplicidad:** la mayor parte de las herramientas tradicionales organizan los datos de una manera estructurada, sin embargo, en Big Data la mayor parte de esos datos son semiestructurados y no estructurados aun cuando estos sean más

difíciles de controlar debido a que son almacenados datos como video, audio o archivos de texto.

6.2 BENEFICIOS DE BIG DATA PARA LAS ORGANIZACIONES

La información pasó a ser uno de los activos más importantes para organizaciones, instituciones y otros sectores donde es posible implementar una solución mediante Big Data, es un tema fundamental ya que se debe evaluar el valor comercial de esos datos, para incrementar aun más su valor, ya que una implementación adecuada trae beneficios potenciales y es posible desbloquear un valor significativo así como obtener algunos beneficios como los que se enlistan posteriormente *(Balachandran, 2017)*:

- **Reducción de costos:** tecnologías como Hadoop y otros *entornos de trabajo* para Big Data, representan ventajas sobre los precios debido a la facilidad de implementación en hardware, ya que no es necesario adquirir servidores con grandes capacidades o supercomputadoras. Para las empresas es más accesible la adquisición de un conjunto de ordenadores con arquitecturas X86 conocido como “commodity hardware” o hardware básico y montar un clúster con múltiples nodos o múltiples equipos, supone un ahorro ya que no es necesario el comprar una supercomputadora.
- **Toma de decisiones:** la analítica siempre ha involucrado intenciones de mejorar la toma de decisiones en las organizaciones y lograr negociaciones más inteligentes con información que respalde una decisión.
- **Nuevos productos y servicios:** es uno de los objetivos más interesantes con el análisis del Big Data el crear productos y servicios para clientes, desde hace una década lo han hecho varias empresas que venden online y actualmente también las que no.
- **Recomendación de un nuevo producto:** La adopción de Big Data y su análisis ha demostrado ser una herramienta muy poderosa para las empresas en línea. Almacenar y trabajar con grandes cantidades de información ha sido un desafío para cualquier operación y es allí donde Big Data ha construido un camino importante para administrar negocios enormes.

- **Detección de fraudes.** Las pérdidas por fraude se estima que ascienden a $9000 US por cada millón en ganancias. Esta situación se pretende reducir mediante la identificación de ideas relevantes a través del uso de Big Data y analizar datos en un nivel agregado para identificar fraudes relacionados con tarjetas de crédito, devoluciones de productos y robo de identidad *(Wamba, 2016)*.
- **Utilización eficiente de los recursos:** Con el tiempo muchos recursos se vuelven escasos o muy costosos por lo que integrar soluciones que hagan mejor la utilización de esos recursos es una de las áreas de oportunidad del Big Data *(Nuaimi, 2015)*.

Estos beneficios deben alcanzar altos niveles de sofisticación e implicación en términos de aplicaciones, recursos y personas involucradas *(Nuaimi, 2015)*, establecer políticas para garantizar la precisión de los datos, alta calidad, alta seguridad, privacidad y control de los datos, uso de estándares en documentación para proporcionar sobre contenido y usos de esos conjuntos de datos.

En la actualidad las empresas que utilizan Big Data son capaces de ofrecer mejores productos, tomar mejores decisiones, establecer relaciones con sus clientes, posicionarse adecuadamente en el mercado y transformarse en más ágiles y obtener ventaja competitiva frente a sus rivales *(Vega, 2015)*.

6.4 BENEFICIOS DE BIG DATA EN LA EDUCACIÓN

Hacer uso de Big Data en la educación es de gran utilidad para recoger y analizar datos sobre los estudiantes, ver sus deficiencias y mejorar sus habilidades en algún área de conocimiento *(Argonza, 2019)*. Big Data tiene importantes beneficios dentro de la educación, ya que mediante el análisis adecuado es posible mejorar los procesos de enseñanza y aprendizaje y dependiendo del tipo de análisis es posible averiguar lo siguiente *(IDD, 2019)*:

- Análisis descriptivo: mediante este tipo de análisis es posible encontrar patrones encontrar patrones dentro de los datos y averiguar que es lo que ocurre en el

área académica, para posteriormente mediante diagnósticos indagar en el porque de esos patrones de información.

- Análisis predictivo: mediante un análisis predictivo busca predecir posibles sucesos para con ello tomar decisiones.
- Análisis prescriptivo: busca el como se puede mejorar la calidad de la enseñanza y el aprendizaje por parte de los estudiantes.

Dentro la educación Big Data puede proporcionar diversos beneficios, de acuerdo con el enfoque con el que se realice en análisis. A continuación, se enlistan algunos de los beneficios que Big Data puede proporcionar a la educación:

- **Mejora de resultados académicos:** dentro de la educación buenas notas no significa un buen aprendizaje, el análisis de las trayectorias del alumno puede ser de gran ayuda para él para mejorar los resultados académicos detectando en primer plano cuales son las temáticas en las que debe reforzar, así como las habilidades que necesita mejorar para alcanzar buenas notas que le permitan acceso a nuevas oportunidades escolares e incluso laborales.
- **Satisfacer las necesidades de los estudiantes:** la información proporcionada por parte de los estudiantes es de gran utilidad, ya que ello puede dar un panorama general de lo que los alumnos necesitan del centro educativo, en otras palabras, recalcar cuales son sus debilidades y fortalezas. Partiendo del análisis de la información de los estudiantes tanto institución y como docentes pueden ser capaces de buscar estrategias para lograr satisfacer las necesidades de los alumnos partiendo del análisis de los métodos empleados en la enseñanza, los materiales utilizados, el contenido temático, etc.
- **Educación personalizada acorde a su necesidad:** no todos los estudiantes aprenden de la misma manera, es por lo que se tienen diferentes necesidades a otros con Big Data mediante el análisis adecuado se busca que se puedan llevar a cabo planes personalizados para los estudiantes, mejorar su desempeño, así como su experiencia durante el aprendizaje.
- **Incrementar las posibilidades de adquirir mejores empleos:** lograr adquirir nuevos conocimientos es de gran importancia, no obstante, también lo es poder

aplicarlos y con ello proporcione mayores oportunidades competitivas dentro del mercado laboral. Con el uso de Big Data en la educación permite a las universidades analizar la situación de los alumnos mientras se encuentran en el aula y también cuando ya se encuentran en el campo laboral, con ello es posible analizar la situación laboral de los egresados y verificar mediante un análisis ver las necesidades en cuanto a formación, lo que se busca en un escenario real y mejorar las oportunidades de empleo para los estudiantes.

Instituciones que utilizan Big Data buscan ser capaces de solucionar problemas con los estudiantes, evitar la deserción escolar, identificar las necesidades reales de aprendizaje, conocer el comportamiento de los estudiantes a lo largo de su formación, y optimizar los recursos para la formación empleando de manera efectiva los contenidos y herramientas del curso para la asignatura *(Gende, 2019)*.

6.4 DESVENTAJAS DE BIG DATA

En la actualidad muchas empresas buscan invertir en Big Data, sin embargo, ello supone un coste para su implementación, mantenimiento y capacitación del capital humano, elementos que deben de ser vistos antes de comenzar una implementación dentro de una organización. Algunas barreras relacionadas a capital humano son las siguientes:

- Existen perfiles Big Data, pero no son los requeridos: actualmente algunas universidades están comenzando a formar a estudiantes con perfiles Big Data, sin embargo, se requieren perfiles con mayor especialización sobre todo en el área de científico de datos, perfiles matemáticos, expertos en Machine Learning y con grandes habilidades en programación.
- Perfiles Big Data son talentos complicados: encontrar perfiles Big Data es complicado en el mercado actual y son muy demandados debido a que varias empresas están buscando incorporarlos.

Por las razones anteriores empresas buscan incorporar nuevos talentos a las compañías, formar nuevos perfiles con el personal calificado que ya tienen dentro de las

organizaciones, y así a los nuevos talentos son introducidos al mundo de Big Data. Entre las habilidades más demandadas por las empresas son las siguientes: Razonamiento analítico, computación en la nube, inteligencia artificial, desarrollo de aplicaciones móviles, traducción, procesamiento de lenguajes naturales, computación científica, marketing de redes sociales, animación, ciencia de datos *(Pontaza, 2019)*.

Big Data buscar almacenar la mayor cantidad de datos posible, no obstante, el tener grandes cantidades de datos no es equivalente a calidad, el propósito de almacenar datos es convertirlos en información valiosa, por lo que tener exceso de ellos puede complicar la gobernanza, mantenimiento y procesarlos. Es por lo que se deben tener bien definidos y claros los objetivos y beneficios que la organización pretende conseguir a través del análisis de datos, al hacerlo busca evitar el almacenamiento de datos innecesarios, mejor control de la calidad de los datos, logrando omitir datos con una calidad no apta para posteriores análisis que puedan afectarlo o distorsionarlo.

Actualmente se cuenta con una gran capacidad para almacenar, pero, aún y cuando los costos de almacenamiento y procesamiento han disminuido con el tiempo y también es posible encontrar cada vez más herramientas para análisis de software libre, muchas organizaciones apenas se encuentran en crecimiento por lo que están lejos de lograr implementar Big Data dentro de ellas.

Las empresas disponen de grandes cantidades de datos para poder realizar sus análisis de Big Data, lo cual es bueno ya que se pueden obtener grandes beneficios, pero al final del día, también se tiene desventajas al respecto, una de la más importante es la vulnerabilidad de los datos que puedan ser robados, lo cual puede traer consecuencias graves para la organización que pueden impactar directamente en el desprestigio de la marca o derivar en implicaciones legales *(Martín, 2018)*.

Añadido a lo anterior, el robo de datos trae otras consecuencias como la venta de datos de los consumidores, que se ser combinados con bases de datos exteriores, lo que puede generar grandes problemas para los usuarios ya que su privacidad se ve fuertemente comprometida, en consecuencia, muchos usuarios no están dispuestos a

proporcionar sus datos, ya que con ellos es posible conocer más sobre sus preferencias por lo que algunos usuarios lo consideran como si estuvieran vulnerando su privacidad.

Big Data puede ser muy efectivo sin embargo ello no significa que no se pueda engañar, esto cuando las fuentes de datos han sido alteradas de manera significativa, ocasionando un ruido significativo y que el análisis los resultados se vean alterados siendo diferentes a los esperados *(Arrieta, 2017)*. Por último, a medida que el número de dispositivos que contribuye a la generación de datos se incrementa, cada vez será más complicado el mantener la privacidad de los usuarios finales.

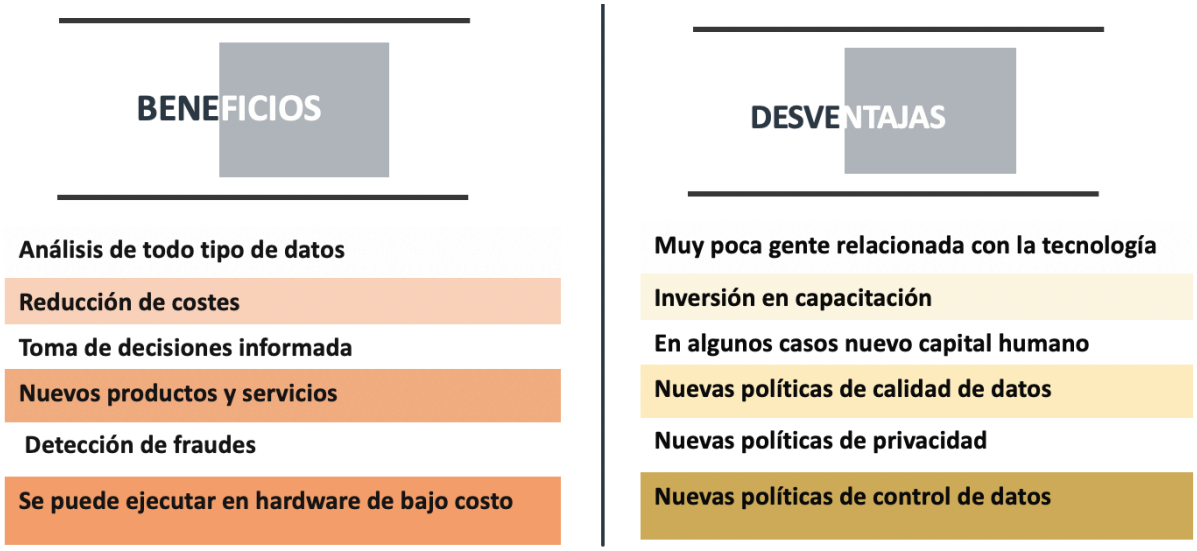


Figura 25 Resumen de los Beneficios y desventajas de Big Data

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

Después de realizar la presente investigación es posible verificar que no solo es posible analizar datos estructurados provenientes de archivos como hojas de calculo o de Bases de Datos Relacionales, si no que es viable el análisis de cualquier fuente de datos, crear diferentes aplicaciones, aplicarlo en diferentes sectores y explotar datos que anteriormente no se aprovechaban o eran desechados debido a que no se tenían las herramientas adecuadas para su análisis y todo ello gracias a las herramientas que se encuentran dentro de Big Data.

Aún y cuando es posible encontrar diversos entornos de trabajo y los precios de almacenamiento y procesamiento han disminuido con el paso del tiempo, existen muchos retos relacionados con Big Data, debido a que no es sencillo poner en práctica las nuevas tecnologías, ya que en primera instancia las empresas suelen oponerse a nuevos cambios, puesto que conlleva nuevas inversiones en tecnología y gastos en capacitación de personal. Sin embargo, empresas que lo han implementado han conseguido buenos resultados, mayores crecimientos, ganancias y lo más importante el verdadero valor de los datos.

Respecto a mi formación profesional, Big Data me parece una tecnología muy importante por la que se puede apostar e inclinar ya sea para lograr un perfil de Ingeniero de Datos o Arquitecto de Datos; el primero para trabajar con grandes volúmenes de datos y programar diferentes rutinas para la solución de problemas y, el segundo para elegir entre las tecnologías existentes las mejores para implementar una arquitectura y solucionar algún problema en especifico, además de que BD promete ser una de una de las mejores profesiones para el siglo XXI.

Otra razón por la que se puede tomar partido hacia Big Data es que en la actualidad si bien existen personas con perfiles Big Data no son suficientes para cubrir el mercado actual, o si bien algunas otras tienen conocimiento sobre herramientas relacionadas con BD ellas no cumplen con los perfiles y/o requerimientos solicitados por parte de las

empresas, lo que significa que son talentos complicados de encontrar, no obstante, es importante mencionar que algunas organizaciones están buscando incorporarlos para su posterior capacitación.

Existen diversos perfiles de graduados que actualmente buscan inclinarse por perfiles Big Data entre ellos el Ingeniero en Computación, sin embargo, no cuenta con una formación en tecnologías de Big Data, incluso los ingenieros con perfiles de desarrollo de Software sólo cuentan con una materia de análisis de datos lo cual no es suficiente, el temario se limita a instruir el desarrollo de software con bases de datos SQL que se vienen utilizando desde hace muchos años, no obstante, para trabajar con las tecnologías dentro de Big Data es necesario introducirse a bases de datos NOSQL, lenguajes de programación como Python, R, Scala además de aprender a utilizar herramientas relacionadas con cada una de las partes del ciclo de Big Data.

Por último, para adentrarse en el mundo de Big Data se requiere mucho esfuerzo, ya que requiere una gran demanda de habilidades y conocimientos, además de un largo camino de aprendizaje para poder llegar a él, es necesario conocer las diferentes herramientas disponibles, soluciones completas en el mercado, áreas de aplicación para emplearlas correctamente y poder obtener el verdadero valor de los datos.

RECOMENDACIONES

Una vez concluida la investigación documental presentada en esta tesina, algunas recomendaciones que puedo realizar para futuras investigaciones en cuanto al tema de Big Data son las siguientes:

- Actualmente cuando se habla de Big Data se aborda un tema muy extenso lo que complica abarcar todos los entornos de trabajo, herramientas, aplicaciones y APIs disponibles, por lo que solo se abarcaron las tecnologías más importantes relacionadas con el software libre (ver anexo 1).
- Aún y cuando dentro de las herramientas de software libre encontramos una amplia gama de herramientas para cada una de las fases de del ciclo de vida de

Big Data, diferentes entornos para solucionar distintos problemas, es muy importante estudiar y conocer otro tipo de tecnologías en el mercado, entre ellas las soluciones completas Big Data que ofrecen diferentes empresas mediante licencias de pago por uso entre ellas las del anexo 2.

- De igual forma también es importante indagar sobre las diferentes empresas que actualmente están implementado soluciones mediante Big Data, como son las aplicaciones que están desarrollando y tomarlas de referencia para resolver problemas con los datos (ver anexo 3).
- Por último, es importante mencionar que cada año aparecen nuevas soluciones Big Data por lo que es necesario la revisión continua de ellas, con la finalidad de encontrar diferentes alternativas para llevar a cabo una solución Big Data y así elegir entre ellas la más conveniente (ver anexo 4).

GLOSARIO DE ACRONIMOS Y METODOLOGÍA TÉCNICA

Acrónimo	Significado
BD	Big Data
KB	Kilobyte
MB	Megabyte
GB	Gigabyte
TB	Terabyte
PB	Petabyte
EB	Exabyte
ZB	Zetabyte
HDFS	Sistema de Archivos Distribuidos de Hadoop
YARN	Yet Aother Resource Negotiator (Negociador de Recursos)
RAM	Memoria de Acceso Aleatorio
IOT	Internet de las Cosas
BDR	Bases de Datos Relacionales

Datos estructurados: Se refiere a los datos que tienen una longitud y formato definidos *(Hurwitz, 2013)*, “tienen una mayor facilidad para accederse ya que cuentan con una estructura muy especificada *(Vega, 2015)*”, como hojas de cálculo y archivos, pueden procesarse fácilmente con herramientas de procesamiento tradicional como lo son las bases de datos SQL o mediante el manejo de archivos *(Fouada, 2015)*.

Datos semiestructurados: No cuentan con un formato definido, para separar un dato de otro se realiza mediante etiquetas**,** para ser leído son necesarias un conjunto de reglas, un ejemplo de estos son las bases de datos relacionales *(Vega, 2015)*.

Datos no estructurados: Son aquellos datos que no siguen un formato especifico *(Hurwitz, 2013)*, que no pueden ser normalizados, no se encuentran bajo ningún patrón, al almacenarse no se hace de forma relacional y no tienen una jerarquía de datos. Estos datos aun cuando no siguen un formato específico se pueden almacenar, modificar, actualizar y eliminar *(Vega, 2015)*, entre ellos Audio, video, SMS, artículos, imágenes satelitales *(Morales, 2016)*, etc.

Latencia: Se refiere al tiempo entre un evento que ocurre en el ambiente del sistema y el inicio de su procesamiento.

Alta disponibilidad: Disponibilidad se refiere a la capacidad de un sistema de realizar su función cuando sea necesario. En el caso de los entornos de trabajo algunas estrategias que se siguen para lograr este requisito son la distribución del procesamiento en múltiples nodos, es decir si un equipo falla otro puedo reemplazarlo. En cuanto a la información para lograr preservarla se realiza una replicación de datos en varios servidores.

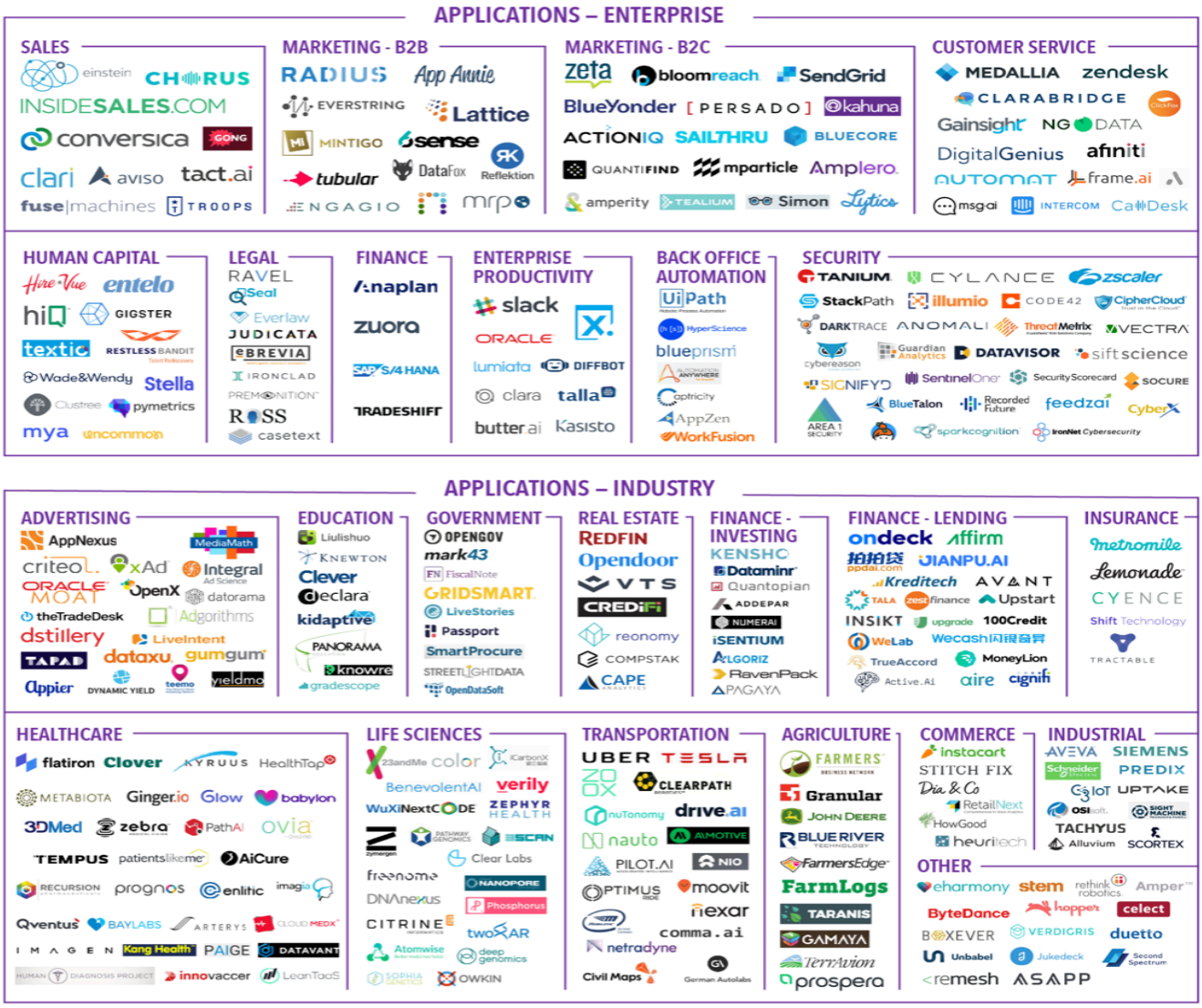
Escalabilidad horizontal: Esta característica se refiere a la capacidad de agregar más servidores aun grupo existente para incrementar el rendimiento o aumentar la capacidad.

ANEXOS

Anexo 1. Tecnologías de software libre para soluciones Big Data *(Turck, 2019).*

Anexo 2. Soluciones Big Data por parte de algunas empresas *(Turck, 2019)*

Anexo 3. Aplicaciones Big Data *(Turck, 2019)*



Anexo 4. Landscape completo *(Turck, 2019)*.

REFERENCIAS

Fragoso, R. B. (18 de 06 de 2012). IBM DeveloperWorks*. (IBM) Recuperado el 11 de 01 de 2018, de ¿Qué es Big Data?: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/*

O´Reilly Media, I. (2012). Big Data Now *(2012 Edition ed.). (O. REILLY, Ed.)*

Villanustre, B. F. (2016). Big Data Technologies and Applications. *(Springer, Ed.) Boca Raton, Florida, USA.*

Clegg, B. (2017). Big Data How the information Revolution is transforming our lives . *(Hotscience, Ed.) UK.*

Viñals, J. T. (2012). Del cloud computing al Big Data. *UPC Barcelona. Barcelona: Eureca Media, SL .*

Mahmood, Z. (2016). Data Science and Big Data Computing- Frameworks and Methodologies. *(D. o. Zaigham Mahmood, Ed.) Derby, UK: University of Derby.*

Olofson, C. W. (2012). Big Data: Trends, Strategies, and SAP Technology. *IDC Analyze the Future, SAP. SAP.*

Cuza, A. l. (2016). Big Questions on Big Data. Revista de cercetare si interventie sociala, 55*, 112-126.*

Hurwitz & Associates, F. H. (01 de 2012). Big Data a la velocidad de los negocios*. (IBM, Productor, & IBM) Recuperado el 19 de 01 de 2017, de IBM Software : https://www-01.ibm.com/software/mx/data/bigdata/*

Zaforas, M. (17 de marzo de 2016). Paragidma- Blog de tecnología y desarrollo*. Recuperado el 22 de enero de 2018, de Cassandra, dama de las bases de datos: https://www.paradigmadigital.com/dev/cassandra-la-dama-de-las-bases-de-datos-nosql/*

Nath, K. G. (Agosto de 2016). NoSQL Database: An Advanced Way to Store, Analyze and Extract Results From Big Data. International Journal of Advance Research in Computer Science and Management Studies, 4*(8), 206-207.*

Elragal, A. (2014). ERP and Big Data: The Inept Couple. Procedia Technology , 16*, 242 – 249.*

Soares, S. (3 de June de 2012). (DATAVERSITY) Recuperado el 2 de 2 de 2018, de Not Your Type? Big Data Matchmaker On Five Data Types You Need To Explore Today

Zhu Yan-li, Z. J. (2012). Research on Data Preprocessing In Credit Card Consuming Behavior Mining. Energy Procedia, 17*, 638-643.*

Venketesh Palanisamy, R. T. (in progress). Implications of big data analytics in developing healthcare frameworks – A review . Computer and Information Sciencie*, --.*

Anita, J. a. (2015). A Survey Of Big Data Analytics in Healthcare and Government. Procedia Computer Science, 50*, 408-413.*

Nugultham, K. (2012). Using Web 2.0 for Innovation and Information Technology in Education Course. Procedia - Social and Behavioral Sciences, 46*, 4607-4610.*

Halamka, J. D. (14 de 12 de 2015). Using Big Data to Make Wiser Medical Decisions*. (Harvard Business Review) Recuperado el 11 de 10 de 2018, de https://hbr.org/2015/12/using-big-data-to-make-wiser-medical-decisions*

YU, S. (2016). Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. IEEE Access, 4*, 2751-2763.*

JayaLakshmi, A. (2018). Performance evaluation of DNN with other machine learning techniques in a cluster using Apache Spark and MLlib. ScienceDirect , Articulo en Progreso*, 1-9.*

JamesWarren, N. M. (2015). Big Data Principles and best practices of scalable real-time data systems. *Shelter Island, NY 11964: Manning Publications Co.*

Jain, V. (2017). Big Data and Hadoop. *Khanna Book Publishing Co.*

Holmes, A. (2015). Hadoop IN PRACTICE. *Shelter Island, NY: Manning Publications Co.*

Friedman, T. D. (2015). Real-World Hadoop. *United States of America: O'Reilly Media, Inc.*

Vohra, D. (2016). Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools. *New York: Apress.*

Lakhe, B. (2016). Practical Hadoop Migration How to Integrate Your RDBMS with the Hadoop Ecosystem and Re-Architect Relational Applications to NoSQL. *Darien, Illinois USA: Apress.*

Presser, K. S. (2015). Field Guide to Hadoop An Introduction to Hadoop, Its Ecosystem, and Aligned Technologies. *United States of America.: O'Reilly Media, Inc.*

Achari, S. (2015). Hadoop Essentials. *Birmingham B3 2PB, UK.: Packt Publishing.*

Ghavami, P. K. (2016). BIG DATA GOVERNANCE Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics . *Washington, D.C.*

Pyne, . (2016). Big Data Analytics Methods and Applications. *India: Spinger.*

Madden, S. (-- de May-June de 2012). From Databases to Big Data. IEEE Internet Computing*, 4-6.*

Buyya, R. (2016). Big Data Principles and Paradigms. *Cambridge: ELSEVIER INC.*

Balachandran, B. M. (2017). Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence Big Data Analytics in the Cloud for Business Intelligence. Procedia Computer Science, 112*, 1112–1122.*

Ellingwood, J. (28 de 10 de 2016). Hadoop, Storm, Samza, Spark, and Flink: Big Data Frameworks Compared*. (Digital Ocean) Recuperado el 03 de 01 de 2019, de https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared*

Kim, S. M. (2016). Big data applications for healthcare: preface to special issue. *New York: Springer.*

Sonia Bergamaschi, L. G. (2017). BigBench workload executed by using Apache Flink. Procedia Manufacturing, 11*, 695-702.*

Ra, S. (2015). Apache Spark a Big Data Analytics Platform for Smart Grid. SMART GRID Technologies, 21*, 171-178.*

Wamba, S. A. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electron Markets, 26*, 173–194.*

Amazon. (- de - de 2018). What is NoSQL? *(Amazon) Recuperado el 05 de 01 de 2019, de https://aws.amazon.com/nosql/?nc1=f_cc*

Developers, G. (25 de 01 de 2018). Overview of Cloud Bigtable*. (Google) Recuperado el 4 de 02 de 2019, de https://cloud.google.com/bigtable/docs/overview?hl=es*

Borkovich, P. D. (2014). Big Data in the Information Age: Exploring the Intellectual Foundation of Communication Theory. Information Systems Education Journal, 12*, 15-26.*

Niño, M. (02 de 09 de 2015). Industria 4.0, Big Data Analytics, emprendimiento digital y nuevos modelos de negocio*. Recuperado el 21 de 01 de 2019, de Analizando las "V" (Volumen, Velocidad, Variedad) del Big Data: http://www.mikelnino.com/2015/09/volumen-velocidad-variedad-big-data.html*

MongoDB. (2018). Para las ideas gigantes*. (MongoDB) Recuperado el 22 de 01 de 2019, de Bases de datos NOSQL explicadas: https://www.mongodb.com/nosql-explained*

Gil, E. (2016). Big Data, Privacidad y Protección de datos . *Madrid, España: AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS.*

ellingwood, J. (28 de 09 de 2016). Una introducción a los conceptos y terminología de Big Data*. (DigitalOcean) Recuperado el 18 de 12 de 2018, de https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology*

Alliance, T. s. (20 de 10 de 2015). ¿Por que son importantes los datos? *(bsa.org) Recuperado el 12 de 10 de 2018, de https://www.bsa.org/search-result?keyword=porque%20son%20importantes%20los%20datos*

Pontaza, D. (21 de 02 de 2019). Talento, barrera en la implementación de Big Data en México...y el mundo. EXPANSIÓN*, págs. -.*

Gonzalez, C. (01 de 12 de 2016). El big data, al servicio del turismo en México. eldiario.es*, págs. -.*

Romero, D. (31 de 01 de 2019). Agencia EFE*. Obtenido de Big Data, la Clave para la Transformación de la Industria en México: https://www.efe.com/efe/america/comunicados/big-data-la-clave-para-transformacion-de-industria-en-mexico/20004010-MULTIMEDIAE_3884173#*

Data Scientist: The Sexiest Job of the 21st Century*. (s.f.). Obtenido de https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*

Davenport, T. (01 de 10 de 2012). Harvard Bussines Review*. Obtenido de Data Scientist: The Sexiest Job of the 21st Century*

Carvajal, J. (26 de 10 de 2016). Tecnologías Big Data Hadoop: Introducción,Componentes y Ecosistema*. Recuperado el 11 de 02 de 2019, de http://blog.jacagudelo.com/tag/tencologias-big-data/*

Jaramillo, S. (2014). SISTEMAS PARA ALMACENAR GRANDES VOLUMENES DE DATOS - BIG DATA STORES. Revista Gerencia Tecnológica Informática, 13*, 17-28.*

Estrada, R. (2016). Big Data Smack. *México City: APRESS.*

Zhang, S. (2012). Research on Key Technologies of Cloud Computing. En E. B.V. (Ed.), 2012 International Conference on Medical Physics and Biomedical Engineering. 33*, págs. 1791-1797. Hebei Province, China: Physics Procedia.*

Atencio, L. (30 de 06 de 2016). Big Data: El ecosistema básico en las empresas. *Recuperado el 10 de 01 de 2019, de http://blog.leonelatencio.com/big-data-ecosistema-basico-las-empresas/*

Venner, J. (2009). Pro Hadoop Build scalable, distributed applications in the cloud. *(Apress, Ed.) Verlag, New York, USA: Springer.*

Apache Software Fundation. (2018). (http://spark.apache.org/) Recuperado el 12 de 01 de 2019, de Apache Spark Motor de analisis unificado ultrarrapido.

L'HEUREUX, A. (2017). Machine Learning With Big Data: Challenges and Approaches. IEEEAccess, 5*, 7776 - 7797.*

Hurwitz, J. (2013). Big Data FOR DUMMIES *(Vol. 1). New Jersey, New Jersey, USA: John Wiley and Sons, Inc.*

Lungu, L. (- de 4 de 2012). Perspectives on Big Data and Big Data Analytics. Database Systems Journal, 3*.*

HU, H. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *Singapore: IEEE.*

Hagel, J. (4 de 10 de 2013). From exponential technologies to exponential innovation*. Recuperado el 12 de 01 de 2019, de https://www2.deloitte.com/insights/us/en/industry/technology/from-exponential-technologies-to-exponential-innovation.html*

Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. *Intitute, McKinsey Global.*

Salazar, J. (01 de 01 de 2016). Big Data en la Educación. Revista Digital Universitaria, 17*, 3-5.*

Contreras, D. C. (2016). La revolución del Big Data. *UNAM, Ciencias de la Administración, México.*

Gartner. (11 de 11 de 2018). Gartner IT Glossary*. Recuperado el 11 de 01 de 2017, de https://www.gartner.com/it-glossary/big-data*

Kejariwal, A. (2012). Big Data Challenges: A Program Optimization Perspective. Cloud and Green Computing (CGC), 2012 Second International Conference on. *Xiangtan, China: IEEE.*

Rehman, M. H. (2016). Big Data Reduction Methods: A Survey. Data Sci. Eng, 1*, 265-284.*

Vega, J. C. (Enero-Junio de 2015). Conociendo Big Data. redalyc.org, 24*, 63-77.*

Fouada, M. M. (2015). Data Mining and Fusion Techniques for WSNs as a Source of the Big Data. Procedia Computer Science, 65*, 778-786.*

Morales, M. D. (1 de Enero-julio de 2016). Los desafios del marketing en la era del Big Data. revista.ebci@ucr.ac.cr, 6*(1).*

Koseleva, N. (2017). Big data in building energy efficiency: understanding of big data and main challenges. Procedia Engineering, 172*, 544 – 549.*

Landmarka, A. D. (2017). Visualisation of Train Punctuality – Illustrations and Cases. Transportation Research Procedia, 27*, 1227–1234.*

IDC . (2014). MEXICO COUNTRY BRIEF THE DIGITAL UNIVERSE OF OPPORTUNITIES. *México: EMC2.*

Kelling, S. (2015). Taking a Big Data approach to data quality in a citizen science project. Ambio, 44*, 601-611.*

Hassania, A. (2017). A framework for Business Process Data Management based on Big A framework for Business Process Data Management based on Big Data Approach Data Approach. Procedia Computer Science, 121*, 740-747.*

Blazquez. (2018). Big Data sources and methods for social and economic analyses. Technological Forecasting & Social Change*, Articules in progrees.*

Usama, M. (2017). Job schedulers for Big data processing in Hadoop environment: testing real-life schedulers using benchmark programs. Digital Communications and Networks, 3*, 260–273.*

Google. (10 de 11 de 2018). Google Cloud Platform*. (G. Developers, Productor, & Google) Recuperado el 22 de 01 de 2018, de CLOUD BIGTABLE: https://cloud.google.com/bigtable/*

Sivarajah, U. (2016). Critical analysis of Big Data challenges and analytical methods. Elsevier Inc, 70*, 263-286.*

Edison, M. (2016). Concepts and Methods of Sentiment Analysis on Big Data. International Journal of Innovative Research in Science, Engineering and Technology, 5*(9), 16288-16296.*

Konda, S. (09 de DIC de 2015). Balancing & Coordination of Big Data in HDFS with Zookeeper and Flume. International Research Journal of Engineering and Technology (IRJET), 02*(09), 869-874.*

Dimiduk, N. (2013). HBase IN ACTION. *Shelter Island, NY: Manning Publications Co.*

Lublinsky, B. (2013). Professional Hadoop Solutions. *Indianapolis, IN: John Wiley & Sons, Inc.*

Oussous, A. (2017). Big Data technologies: A survey. www.sciencedirect, 30*, 431-448.*

Alaka, H. (2018). A framework for big data analytics approach to failure prediction of construction firms. Procedia Computer Science*, 2-9 -- articulo en progreso.*

Ahmeda, H. (2016). Performance Comparison of Spark Clusters Configured Conventionally and a Cloud Service. Procedia Computer Science, 26*, 99-106.*

Foundation, A. S. (2018). Apache Spark FAQ*. Recuperado el 10 de 01 de 2019, de http://spark.apache.org/faq.html*

Petrov, M. (2018). Adaptive performance model for dynamic scaling Apache Spark Streaming a, Streaaming. Procedia Computer Science, 136*, 109-117.*

Karau, H. (2015). Learning Spark. *CA United States of America.: O'Reilly Media, Inc.*

The Apache Software Foundation. (2019). Apache Flink® - Stateful Computations over Data Streams*. Recuperado el 16 de 01 de 2019, de https://flink.apache.org/*

MustafaKamal, M. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70*, 263-286.*

MEHMOOD, A. (2016). Protection of Big Data Privacy. IEEEAccess, 4*, 1821 - 1834.*

Portela, F. (2016). Why Big Data? Towards a project assessment framework. Procedia Computer Science, 98*, 604 – 609.*

IDC. (05 de 09 de 2018). Big Data and Business Analytics Solutions Revenues in Asia/Pacific (excluding Japan) Forecast to Value at USD 27 Billion in 2022, According to Latest IDC Spending Guide*. (IDC) Recuperado el 05 de 02 de 2019, de https://www.idc.com/getdoc.jsp?containerId=prAP44260318*

IDC. (15 de 08 de 2018). Revenues for Big Data and Business Analytics Solutions Forecast to Reach $260 Billion in 2022, Led by the Banking and Manufacturing Industries, According to IDC*. (IDC) Recuperado el 05 de 02 de 2019, de https://www.idc.com/getdoc.jsp?containerId=prUS44215218*

IDC. (April de 2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. (EMC2) Recuperado el 29 de 01 de 2019, de https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm*

IDC. (30 de 08 de 2018). Big Data and Analytics Spending in Central and Eastern Europe Moves from Preparation Stage to Deployment, According to IDC*. (IDC Analize the future) Recuperado el 05 de 02 de 2019, de https://www.idc.com/getdoc.jsp?containerId=prCEMA44243018*

LIANG, F. (2018). A Survey on Big Data Market: Pricing, Trading and Protection. IEEEAccess, 6*, 15132-15154.*

Nuaimi, E. A. (2015). Applications of big data to smart cities. SpringerOpen Journal*, 2-15.*

Rama, J. (2016). The implications of Big Data analytics on Business Intelligence: A qualitative study in China. Procedia Computer Science, 87*, 221 – 226.*

Turck, M. (01 de 04 de 2019). Big Data & IA Landscape*. Obtenido de http://mattturck.com/wp-content/uploads/2018/07/Matt_Turck_FirstMark_Big_Data_Landscape_2018_Final.png*

Martín, I. (23 de 10 de 2018). PublicaTIC*. Obtenido de Riesgos del Big Data: https://blogs.deusto.es/master-informatica/riesgos-del-big-data/*

Arrieta, E. (22 de 07 de 2017). Expansión Economia Digital*. Obtenido de Los peligros del 'big data': ¿estamos creando un mundo más injusto y desigual?: http://www.expansion.com/economia-digital/innovacion/2017/07/22/596f8c54ca47413b118b45df.html*

instaclustr. (14 de 02 de 2019). Hosted & Managed Apache Cassandra as a Service*. Obtenido de https://www.instaclustr.com/solutions/managed-apache-cassandra/?utm_campaign=All-SN-Cassandra&utm_medium=ppc&utm_source=adwords&utm_term=%2Bcassandra%20%2Bdb&hsa_mt=b&hsa_grp=41829854299&hsa_kw=%2Bcassandra%20%2Bdb&hsa_acc=1467100120&hsa_src=g&hsa_cam=384*

Zanoon, N. (2017). Cloud Computing and Big Data is there a Relation between the Two: A Study. International Journal of Applied Engineering *, 6970-6982.*

IDD, I. y. (20 de 03 de 2019). Innovación y desarrollo docente*. Obtenido de VENTAJAS Y RIESGOS DEL BIG DATA EN EDUCACIÓN: https://iddocente.com/big-data-educacion/*

Gende, I. M. (20 de 03 de 2019). unirrevista*. Obtenido de Big Data en Educación: Analítica de Aprendizaje y Aprendizaje Adaptativo: https://www.unir.net/educacion/revista/noticias/big-data-en-educacion-analitica-de-aprendizaje-y-aprendizaje-adaptativo/549203628743/*

Argonza, J. S. (20 de 03 de 2019). rdu revista digital universitaria unam*. Obtenido de Big Data en la educación: http://www.revista.unam.mx/vol.17/num1/art06/*

Blazquez, D. (2018). Big Data sources and methods for social and economic analyses. Technological Forecasting & Social Change*, Article in progress.*