UNIVERSIDAD DE QUINTANA ROO
DIVISIÓN DE CIENCIAS E INGENIERÍA

"Big Data: Conceptos Básicos, Tecnologías y Aplicaciones".

Trabajo monográfico
PARA OBTENER EL GRADO DE

INGENIERO EN REDES.

PRESENTA
Jacqueline Estefani Sansores Cuevas.

supervisores
Dr. Homero Toral Cruz.
Dr. José Antonio León Borges.
M.M. José Raúl García Segura.

supervisores suplentes
Dr. Freddy Ignacio Chan Puc.
Dr. Francisco Méndez Martínez.

CHETUMAL QUINTANA ROO, MÉXICO. DICIEMBRE DE 2020

UNIVERSIDAD DE QUINTANA ROO
DIVISIÓN DE CIENCIAS E INGENIERÍA

TRABAJO MONOGRÁFICO TITULADO
"Big Data: Conceptos Básicos, Tecnologías y Aplicaciones".

ELABORADO POR
Jacqueline Estefani Sansores Cuevas.

BAJO SUPERVISIÓN DEL COMITÉ DEL PROGRAMA DE LICENCIATURA Y APROBADO COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE:

INGENIERO EN REDES

COMITÉ SUPERVISOR

SUPERVISOR: _____
Dr. Homero Toral Cruz.

SUPERVISOR: _____
Dr. José Antonio León Borges.

SUPERVISOR: _____
M.M. José Raúl García Segura.

SUPLENTE: _____
Dr. Freddy Ignacio Chan Puc.

SUPLENTE: _____
Dr. Francisco Méndez Martínez.

CHETUMAL QUINTANA ROO, MÉXICO, DICIEMBRE DE 2020

DEDICATORIAS

Dedico este trabajo monográfico a mi familia, a mi abuelito Wilberth Sansores Góngora quien con su apoyo pude continuar con mis estudios, en especial a mi abuelita Isabel Reyna Canul QEPD, quien no pudo ver este logro y sé que se hubiese sentido muy orgullosa. A mi padre Wilberth A. Sansores Canul quien siempre con su amor y cariño estuvo ahí para acompañarme en cada actividad y logro escolar. De igual manera a la familia Loria Chulim por cobijarme y hacerme sentir como una hija más, por apoyarme con el cuidado de mi pequeño y por alentarme a seguir con mi formación académica.

También quiero dedicar este trabajo a mi Esposo, Alen Arturo Loria Chulim por su apoyo, amor y comprensión, a nuestro pequeño Allen Emmanuel Loría Sansores quien es mi motor y motivación, es mi anhelo que este logro sea también una motivación para él y lograr tener un título académico en el futuro.

AGRADECIMIENTOS

A todos mis profesores por su paciencia, dedicación y por ser parte fundamental durante mi formación académica. De manera especial a mi asesor Dr. Homero Toral Cruz por alentarme, apoyarme, por su tiempo, por su dedicación y sobre todo por la gran paciencia que me ha tenido para la realización de este trabajo monográfico.

También quiero agradecer por el apoyo brindado a mi tutora académica, M.T.I Melissa Blanqueto Estrada y finalmente a nuestra máxima casa de estudios “La Universidad de Quintana Roo”, que me brindó los mejores estándares de calidad de estudios y prepárame hacia el ámbito laboral.

RESUMEN

La sociedad en su vida cotidiana registra y almacena diversos tipos de información; por ejemplo, nuestros antepasados tales como los egipcios y mayas dejaban su evidencia plasmada en papiros, talladuras en piedras e ilustraciones. En la actualidad, debido a la gran cantidad de información que se maneja en diversos escenarios de la vida diaria, campos de la ciencia, empresas, etc., ha surgido la necesidad de realizar el análisis, administración y gestión de los datos. A este fenómeno se llama Big Data. En otras palabras, se puede decir que día con día manejamos datos tradicionales o comunes; sin embargo, estos se vuelven enormes y dan origen a grandes volúmenes de datos, convirtiéndose en datos no estructurados.

Es importante mencionar que este tipo de información o datos de gran tamaño no son fácil de procesar mediante las técnicas y plataformas de datos tradicionales. Muestran una capacidad de respuesta lenta y falta de escalabilidad, rendimiento y precisión. Sin embargo, para hacer frente a estos retos que demanda Big Data, se han desarrollado varios tipos de distribuciones y tecnologías.

El objetivo principal de esta monografía es presentar los conceptos básicos, aplicaciones y las tecnologías desarrolladas recientemente para la tecnología de Big Data.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPITULO 1

INTRODUCCIÓN

CAPÍTULO 1: INTRODUCCIÓN

En la actualidad, la sociedad está aprendiendo a vivir en un mundo digital que se ve envuelto en datos. Las empresas y organizaciones necesitan administrar y manejar el crecimiento de sus datos, los cuales son cada vez más grandes y exponencialmente más voluminosos. También deben aprender a manejar datos en nuevas y diferentes formas no estructuradas. A este fenómeno se llama *Big Data*.

Desde una perspectiva evolutiva, Big Data no es nuevo. El avance hacia Big Data es una continuación de la búsqueda de la humanidad antigua por medir, registrar y analizar el mundo.

El desarrollo de aplicaciones de Big Data se ha vuelto cada vez más importante en los últimos años. De hecho, varias organizaciones de diferentes sectores dependen cada vez más del conocimiento extraído de grandes volúmenes de datos. Sin embargo, en el contexto de Big Data, las técnicas y plataformas de datos tradicionales son menos eficientes. Muestran una capacidad de respuesta lenta y falta de escalabilidad, rendimiento y precisión.

Para afrontar los complejos retos del Big Data, se ha trabajado mucho en los últimos años. Como resultado, se han desarrollado varios tipos de distribuciones y tecnologías.

Esta monografía es una revisión que analiza los conceptos básicos, aplicaciones y las tecnologías desarrolladas recientemente para Big Data. Proporciona no solo una vista global de las principales tecnologías de Big Data, sino también comparaciones según las diferentes capas del sistema, como la capa de almacenamiento de datos, la capa de procesamiento de datos, la capa de consulta de datos, la capa de acceso a datos y la capa de gestión. Clasifica y analiza las principales características, ventajas, límites y usos de las tecnologías. También se presentan las principales aplicaciones desarrolladas en los últimos años, tales como: redes inteligentes, e-salud, Internet de las cosas (IoT), redes sociales, transporte y logística.

1.1 ANTECEDENTES

La sociedad en su vida cotidiana registra y almacena diversos tipos de información; por ejemplo, nuestros antepasados tales como los egipcios y mayas dejaban su evidencia plasmada en papiros, talladuras en piedras e ilustraciones. Durante la existencia de la humanidad, se han registrado datos de diversas formas; por ejemplo, en la época paleolítica se emplearon métodos rudimentarios para el almacenamiento de los datos con la ayuda de palos o muescas de huesos, lo cual se puede interpretar como una de las primeras muestras de que ya existía el interés por recopilar, contar y guardar los datos como información. Tiempo después surge el Abaco y las primeras bibliotecas que sirvieron para almacenar y consultar conocimiento, más adelante gracias al desarrollo de la primera computadora mecánica fue posible predecir posiciones astronómicas, después de siglos surge el análisis de datos estadísticos, así como el surgimiento del término *Bussines Intelligence*, surgen las bases de datos relacionales y el internet quien tuvo gran impacto causando la revolución de la recolección, almacenamiento y análisis de datos que se generaban.

En la actualidad los datos son gestionados a través de herramientas que facilitan el manejo, tratamiento, almacenamiento y uso de estos, el desarrollo de equipos electrónicos combinados con el uso de Internet propicia un gran volumen de datos que requieren ser gestionados.

Todo esto ha llevado al inicio de Big data, el cual día con día va evolucionando debido a las tecnologías y a la forma exponencial en la que se van generando grandes cantidades de datos.

1.2 Tecnología Big Data

Big data se define como un concepto abstracto [1]. Los grandes volúmenes de datos son una colección de conjuntos muy grandes de datos con una gran diversidad de tipos, razón por la cual resulta difícil procesarlos mediante el uso de métodos de procesamiento o plataformas de procesamiento de datos tradicionales [2].

Según IDC, Big Data es una nueva generación de tecnologías y arquitecturas diseñadas para extraer valor económico de grandes volúmenes de datos heterogéneos habilitando una captura, identificación y/o análisis a alta velocidad. Big Data se caracteriza por tener cinco dimensiones: volumen, variedad, velocidad, veracidad y valor [3].

La compañía de Cisco, Big Data significa un gran negocio debido a que los grandes volúmenes de datos presentan una poderosa herramienta para ayudar a las empresas a ser más conscientes, predictivas y ágiles. El entorno competitivo se hace más complejo como grandes datos crecientes [4]. Sin duda alguna Big Data cuenta con una gran cantidad de datos que se está generando y se está guardando tan rápido que está inundando a la sociedad, y por supuesto a las ciudades. Big Data se está convirtiendo en el próximo recurso natural que explotar y los retos que debe gestionar incluyen capturar, almacenar, buscar, compartir, transferir, analizar y visualizar [5].

Dado al aumento de los datos estructurados y no estructurados de los sistemas que se encargan de los grandes volúmenes; las organizaciones se ven en la necesidad de obtener ayuda de plataformas capaces de procesar, almacenar, extraer el valor de los tamaños y formas, dando así seguridad a los usuarios para la obtención de análisis de los datos, siendo este tan amplio y diverso.

Big Data se ha caracterizado principalmente por la gran cantidad de datos, que a su vez se ha visto en la necesidad de recurrir a herramientas para el almacenamiento de este, así como el uso de software para el procesamiento de los datos, de manera general podríamos decir que ha sido una tendencia para el avance en la tecnología misma que ha sido impulsada por la transformación

digital, pasando por una serie de procesos. Las nuevas tecnologías han permitido que el desarrollo sea a menor costo, así como el almacenamiento y procesamiento, permitiendo mayor facilidad para el manejo de los datos tanto como en aquellos que son realizados por lotes o en tiempo real.
Entonces podemos concluir que hoy en día Big Data se ha definido de diferentes maneras y en diversos lugares y se describe como un conjunto de recursos que permiten la gestión y análisis de cantidades masivas de datos, con un alcance y dimensiones en constante crecimiento [6].

Big Data es quien se encarga del proceso de encontrar la información para luego transformarla en conocimiento. Existen tres elementos en el proceso Big data que a continuación se explicara de manera breve.

- **Datos:** Información que es parte elemental, que carecen de relevancia, son datos que sin un fin, propósito o utilidad no son útiles. Podemos poner como ejemplo el nombre de una persona.
- **Información:** Es el conjunto de datos que han sido procesados y ordenados para comprenderlos, teniendo relevancia y/o propósito, dependiendo de la importancia de la información se pueden dar soluciones y tomar decisiones.
- **Conocimiento:** Es la adquisición de información de valor, con la finalidad de comprender por medio de la razón, teniendo como resultado procesos de aprendizaje a través de la experiencia en donde se involucra la toma de decisiones.

1.3 Importancia de Big Data

Con el desarrollo de la revolución tecnológica, miles y millones de personas están generando grandes cantidades de datos mediante el uso de diversos dispositivos; es decir, de manera conceptual Big Data hace referencia al gran volumen de datos generados por la humanidad a través de diferentes dispositivos; en donde, el procesamiento y análisis de estos datos son de gran valor para la toma de decisiones de una organización [7].

Los datos se han convertido en algo muy importante y valioso para las empresas. Dicho en otras palabras, la búsqueda de Big Data es directamente atribuible a la analítica, que ha evolucionado de ser una iniciativa empresarial a un imperativo comercial, Big Data se trata de mejores análisis en un espectro más amplio de datos [8]. Big Data tiene una naturaleza compleja que requiere potentes e innovadoras tecnologías y algoritmos avanzados. Por lo que las herramientas tradicionales se han convertido ineficientes para el procesamiento de grandes volúmenes de datos [9].

1.4 Beneficios de Big Data

Las nuevas tecnologías aplicadas a Big Data buscan aportar competitividad a las diferentes organizaciones, sean públicas o privadas, pues es la competitividad la que permite que perduren en el tiempo y puedan alcanzar sus objetivos [10]. Los datos al convertirse en información ayudan a tomar decisiones referentes a la estrategia que algunas empresas, organizaciones o los gobiernos implementan.

La transferencia de la información a una base de datos suele ser un gran desafío, los datos no estructurados se convierten en estructurados, y en ellos existe gran variedad de datos, diferentes formatos y fuentes de texto (documentos, mensajes, audios etc.) El beneficio es la respuesta de velocidad en tiempo real

considerando el alto volumen de datos que posteriormente son convertidos en información para toma de decisiones.

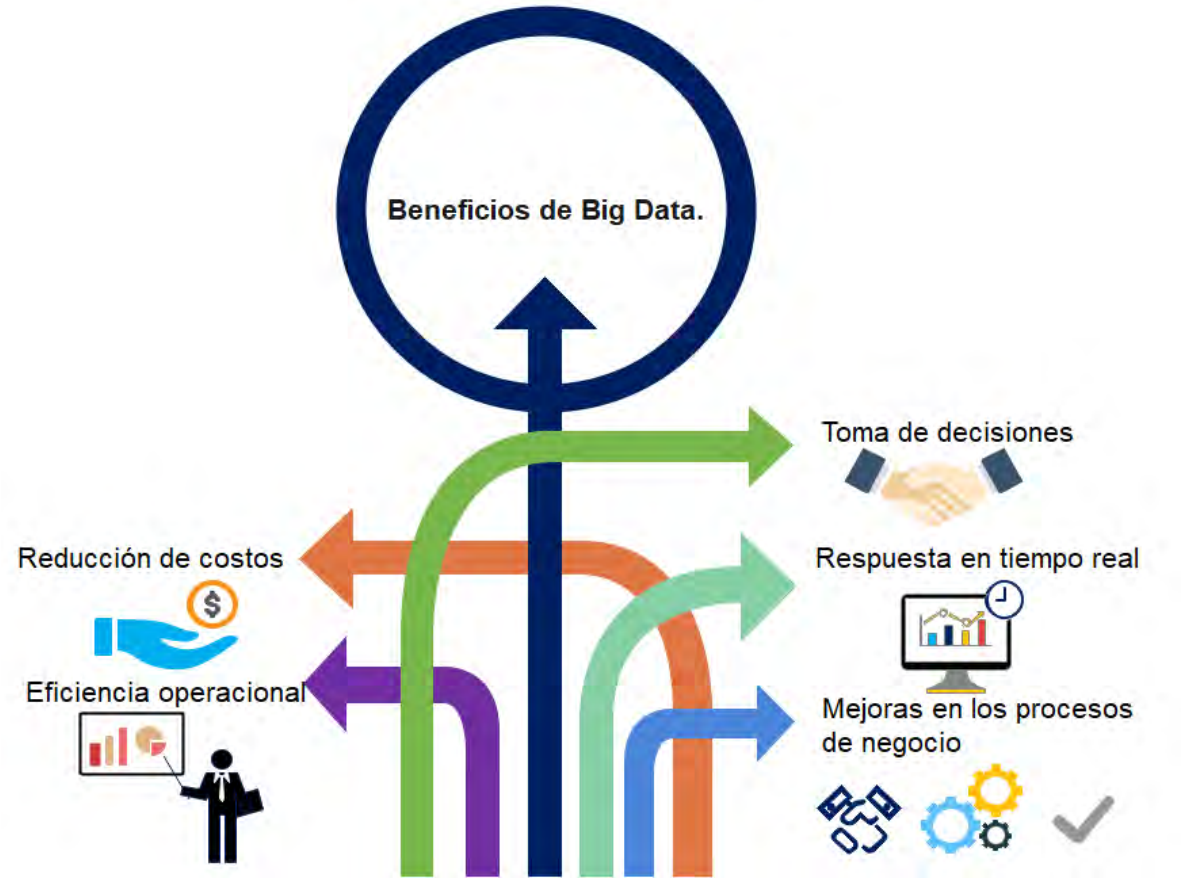


Figura 1.1 Representación de beneficios de Big Data

El uso del Big Data se encuentra basado en la capacidad en la que se procesa y se obtiene la información, por ello se puede decir que Big Data genera los siguientes beneficios:

- **Toma de decisiones:** mediante la creación de sistemas con la capacidad de procesar grandes volúmenes de datos estructurados, se facilita de forma dinámica la interpretación de la información para las empresas u organizaciones y de esta manera puedan tomar decisiones inteligentes.
- **Respuesta en tiempo real:** es el lado positivo y es uno de los beneficios que ha generado que el funcionamiento de los sistemas se mantenga en equilibrio, ya que esta tecnología no solo procesa y almacena información

sino también se encarga de recibir y procesar los datos en tiempo real obteniendo la información que necesitan de manera rápida, catalogando así a Big data como una tecnología ágil y veloz.

- **Reducción de costos de capital:** Tener el conocimiento de los datos a tratar, permite utilizar los datos que se tienen creando una reducción en el software, hardware y otros costes de infraestructura, bridando así una satisfacción en cuanto al uso y respuesta de las necesidades.
- **Eficiencia operacional:** Una reducción de los costes laborales debido a los métodos más eficientes para la integración de datos, la gestión, el análisis, y la entrega.
- **Mejoras en los procesos de negocio:** Un incremento en los ingresos o ganancias debido a las nuevas o mejores maneras de hacer negocios, incluyendo mejoras en las transacciones comerciales, la gestión sostenible de las comunidades, y la distribución apropiada de los servicios sociales, sanitarios, educativos y servicios [3].

1.5 Características de Big Data

Desde una perspectiva evolutiva, la generación de grande volumen de datos no es algo nuevo. El avance hacia Big Data es una continuación de la antigua búsqueda de la humanidad para medir, registrar y analizar el mundo. Diversas empresas han estado utilizando sus datos y realizando análisis durante décadas. Big Data tiene una naturaleza compleja que requiere potentes tecnologías y algoritmos avanzados. Las herramientas tradicionales estáticas ya no pueden ser eficientes para el análisis de grandes volúmenes de datos. La definición más común y generalizado de grandes volúmenes de datos está dada con base a las 3V [11]. La mayoría de los científicos de datos y expertos definen grandes volúmenes de datos por las siguientes características principales [9]: Volumen, Velocidad y Variedad.

1.5.1 Volumen (datos en reposo)

En Big Data es la capacidad de procesar grandes cantidades de información y es el principal atractivo de análisis de grandes volúmenes de datos [12]. El volumen se analiza en términos de Bytes [13].

En la actualidad, la conversación sobre los volúmenes de datos ha cambiado de terabytes a petabytes con un cambio inevitable a zettabytes, y todos estos datos no pueden almacenarse en sus sistemas tradicionales [14]. Un zettabytes es un billón de gigabytes (GB), o mil millones de terabytes [8].

Figura 1.2 Representa el volumen en términos de Bytes

1.5.2 Velocidad

En Big Data la velocidad es una de las características principales, esta tecnología se encarga de procesar y analizar de manera rápida con la finalidad de extraer la información útil y más relevante. La velocidad implica flujos de datos, estructurada creación de los documentos, y la disponibilidad para el acceso y la entrega. Se puede pensar que la velocidad es la forma en que se

mueven los datos y al ritmo que se utilizan, por otro lado, las aplicaciones para analizarlos demandan que la velocidad de respuesta sea de manera rápida, es decir, a raíz de que se crea, almacena datos de manera constante y se realizan los procesos en tiempo real, el resultado requiere de alta velocidad. Dato curioso respecto a la velocidad es que debido a que las empresas se encuentran tratando sus datos en petabytes en lugar de terabytes y el aumento de sensores y otros flujos de información ha generado que el flujo constante de datos tenga un ritmo que ha imposibilitado el manejo de sistemas tradicionales [14]. Los petabytes ahora son referenciados por el término de zettabytes [8]. La velocidad de datos que circulan a través de los sistemas varía de integración por lotes y carga de datos a intervalos predeterminados a transmisión en tiempo real de los datos. La integración por lotes en el almacenamiento de datos tradicional es hoy el principal método de procesamiento de datos utilizando *Hadoop*. Mientras que la carga de datos es el dominio de las tecnologías tales como el procesamiento de eventos complejos, motores de reglas, análisis de texto y buscar, inferencia, aprendizaje automático, y arquitecturas basadas en eventos en general [3].
Entonces podemos concluir que la clave para evaluar los requisitos de velocidad de datos grande es entender los procesos y requerimientos de los usuarios finales.

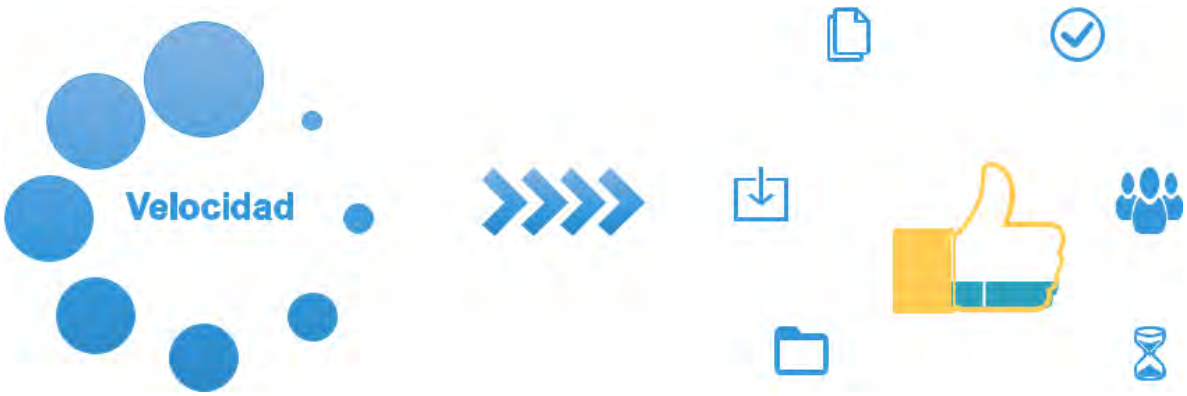


Figura 1.3 Representa la velocidad de procesos y análisis de datos

1.5.3 Variedad

Estos datos no tienen una estructura fija y rara vez se presentan en una forma perfectamente ordenada y lista para su procesamiento [12].

La mayor velocidad de los datos normalmente se transmite directamente a la memoria, en vez de escribirse en un disco. En otros términos, la variedad hace una representación a todos los tipos de datos [14] y la funcionalidad de la variedad como característica de Big Data es el hecho de capturar todos los datos que pertenecen al proceso de toma de decisiones [8].

Aplicaciones de grandes volúmenes de datos suelen combinar datos de una variedad de fuentes de datos distribuidas y en múltiples formatos (por ejemplo, vídeos, documentos, comentarios, registros), [9] tanto internos como externos a una organización y los datos de diferentes tipos como lo son estructuradas, semiestructuradas y no estructurados. Los datos pueden ser estructurados, estos son en donde se almacenan las bases de datos relacionales sabiendo que todo es predefinido como lo es la longitud, denominación y formato. Los datos no estructurados estos son los que no se encuentran definidos. Aquí es cuando podemos decir que Big Data tiene una gran riqueza, puesto que tiene una gran diversidad de tipos de datos. El surgimiento de nuevas fuentes de datos, comenzando por lo que más está en función, como lo son las redes sociales y la interacción con los dispositivos móviles, lo cual generan información en cada transacción, esto ha generado un aumento de en el grado de complejidad tanto en el almacenamiento como el procesamiento. Si se cuenta con un gran volumen de información, entonces sabemos que existe una variedad de datos, mismos que pueden representarse de diversas formas, alguno de los ejemplos son los medios electrónicos, es decir, audio, video, sistemas de GPS, sensores digitales utilizados en industrias, automóviles, medidores eléctricos, etc. Por tanto, estas aplicaciones que analizan los datos requieren que la velocidad de respuesta sea lo más rápida para así obtener la información correcta en el tiempo preciso [13].

IBM define la variedad como la representación de todos tipos de datos, dando así un gran cambio en los requisitos de análisis de los datos estructurados, semi

estructurados y no estructurados siendo parte de la toma de decisiones y de la visión del proceso [14].

1.5.4 Características que se suman para formar las 5V

Las tres V se encuentran en gran medida en diversas literaturas, sin embargo, otros autores e institutos como IEEE toman en cuenta las características de Valor y Veracidad, sumado así 5V. Estas dos últimas surgen a raíz de las preguntas ¿cuál es la veracidad de los datos y cuanto se puede confiar en ellos?; en la realidad y hoy en día Big Data han tomado gran importancia e incluir estas características cumplen con el propósito de obtener información pertinente, además de desarrollar una definición adecuada, la gran investigación de datos también debe centrarse en cómo extraer su valor, cómo usar los datos, y como transformar un conjunto de datos en grandes datos [1]. Como resultado, las soluciones Big Data se caracterizan por procesamiento complejo en tiempo real, relación de datos y capacidades avanzadas de analítica y búsqueda [15]. Es por ello por lo que se definirá estas dos características como parte importante de la funcionalidad de Big Data.

1.5.5 Veracidad (Datos en duda)

Caracterizado por contener la verdad o hecho, es decir, precisión y certeza. Poniendo en duda todo aquello que cause inconsistencia, engaños, fraudes, correos no deseados y latencia [8] [12], a medida que se multiplican los canales de interacción, la información de valor es cada vez más el resultado de la combinación de datos de múltiple origen y tipología que puede estar en forma estructurada, semiestructurada o no estructurada [16]. La veracidad en Big Data es el grado de confianza que se da o se establece en los datos.

Según un estudio realizado por IBM, las organizaciones tienen confianza suficiente en la calidad de los datos y análisis de los que disponen como para utilizarlos en sus procesos de toma de decisiones cotidianos.

1.5.6 Valor (Datos en resalte)

Diseñado para extraer económicamente valor a partir de volúmenes muy grandes de una amplia variedad de datos, habilitando de alta velocidad de captura, descubrimiento, y / o análisis [12], en el contexto de Big Data, valor hace referencia a los beneficios que se desprenden del uso de Big Data (reducción de costes, eficiencia operativa, mejoras de negocio) [16]. En la Figura 1.4 se da a conocer las características de las 3V para Big Data y las características complementarias.

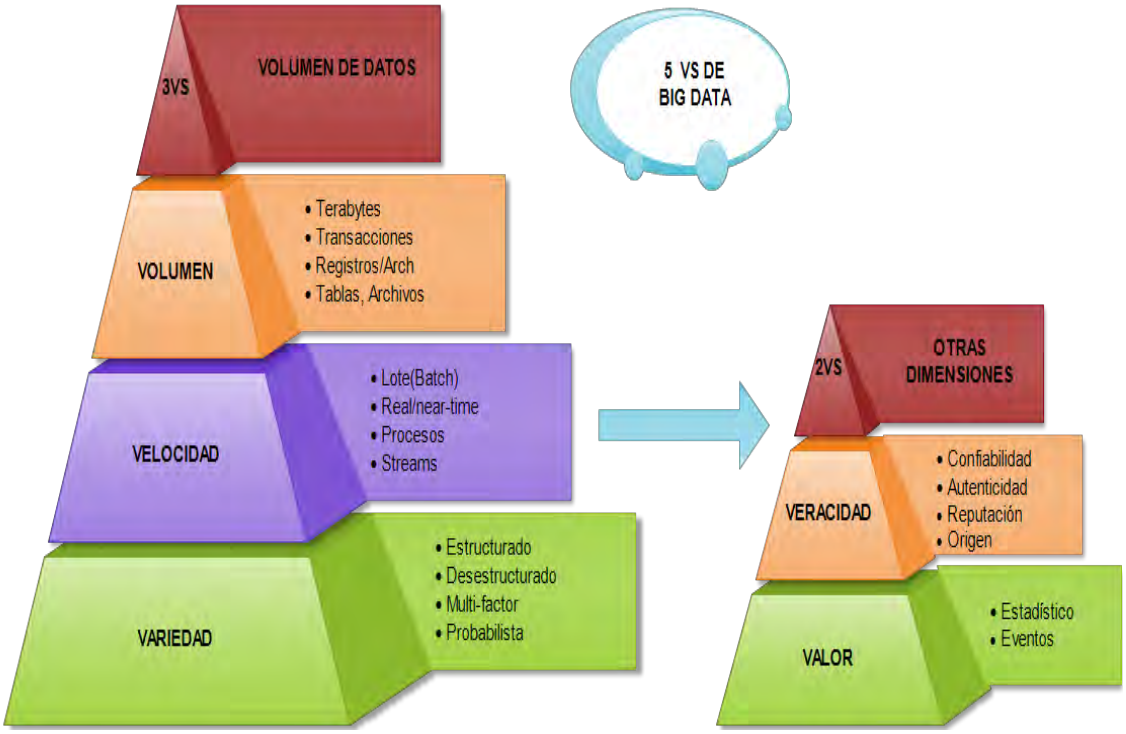


Figura 1.4 Características de Big Data

En resumen, grandes volúmenes de datos son enormes conjuntos de datos compuestas por datos estructurados y no estructurados, a menudo con la necesidad de un análisis en tiempo real y el uso de tecnologías y aplicaciones para almacenar, procesar, analizar y visualizar la información de múltiples fuentes complejas. Desempeña un papel primordial en el proceso de toma de decisiones en la cadena de valor dentro de las organizaciones [11].

1.6 Clasificación de Big Data

Se clasifican las tecnologías de Big Data en aquellas que dan soporte a la captura, transformación, procesamiento y análisis de los datos, pueden ser estructurados, semiestructurados o no estructurados [16]. Se encuentran clasificados dependiendo de la necesidad que enfrenten las organizaciones o empresas, tomando en cuenta de la información a analizar y de los problemas a resolver, así como la diversa variedad de tipo de datos [2] [3]. Se tiene el conocimiento de que existe una amplia variedad de tipos de datos a analizar, y se puede clasificar para poder entender mejor su representación [15]. IDC clasifica tecnologías Big Data en dos cubos distintos: uno para grandes volúmenes de datos en movimiento, el otro para grandes volúmenes de datos en reposo [3].

1.6.1 Tipos de datos en reposo

Los grandes volúmenes de datos en reposo son los que conocemos como datos estructurados y no estructurados [3][15]. Cabe mencionar que cuando hablamos de los datos que están semiestructurados o no estructurados es para dejar en claro que todos los datos tienen alguna estructura [8].

- **Estructurados:** Como su nombre lo indica son aquellos datos que están formados de manera estructurada definida, se pueden mencionar las hojas de Excel o una base de datos SQL. Se les puede dar un uso a los datos de manera fácil y pueden ser utilizados para futuros análisis y predicciones.
- **No estructurados:** Caso contrario del tipo de dato mencionado con anterioridad, estos son aquellos que no están formados de manera estructurada tampoco se encuentran definidos, es decir, cuando nos referimos a datos no estructurados, nos referimos a los subcomponentes que no tienen estructura. Comúnmente los podemos ver cuando redactamos el cuerpo de un email, o cuando hacemos uso de aplicaciones como chats en una conversación, los datos escritos en un fichero de Word

o bien en las bases de datos NoSQL. Este tipo de datos resulta complicado realizar informes y analizar, debido a la información contenida es valiosa, pero no se encuentra estructurada ni catalogada.

- **Datos semiestructurados:** son aquellos datos que se encuentran en el medio, es decir dentro de ellos existe los datos estructurados y parte de los datos no estructurados.

Figura 1.5 Tipos de datos

Datos estructurados	Datos semiestructurados	Datos no estructurados
Hojas de Excel con contenido de: • Nombre • Fechas de nacimiento **Bases de datos SQL**	**Correos electrónicos** • Destinatario, receptores son la parte estructurada • El cuerpo de un email es la parte no estructurada	• Redes sociales • Equipos móviles • Cámaras • Dispositivos GPS

Tabla 1.1 Tipos de datos

1.6.2. Tipos de datos en movimiento

Big Data cuenta con muchas fuentes, desde que se da un clic con el mouse en algún sitio web sabemos que en ese momento ya se puede capturarse en como un registro web.

- Contenido generador por el usuario y redes sociales como:
    - ❖ Facebook
    - ❖ Tweets
    - ❖ Datos de Stream (Amazon, Netflix, YouTube)

Estos datos se pueden capturar y se analizan para una mejor comprensión por que genera demasiada información mediante comentarios o reproducciones.

- Datos transaccionales generados por gran escala de registros como:
    - ❖ Sistemas web
    - ❖ Transacciones comerciales
    - ❖ Registros de facturación
    - ❖ Registros detallados de llamadas
- Conexión de maquina a máquina (M2M) es la tecnología en donde se conecta con otros dispositivos mismos que generan gran volumen de datos y requieren ser analizados y está dada por:
    - ❖ Sensores (Temperatura, presión)
    - ❖ Medidores (Velocidad)
    - ❖ Se caracterizan por la transmisión continua de los datos sobre el consumo de electricidad
- Ciencia de datos a partir de datos de experimentos intensivos como:
    - ❖ Datos o genomas celeste
    - ❖ Información biométrica (huellas digitales, escaneo de retinas).
- Datos generados por las personas:
    - ❖ Notas de voz
    - ❖ Correo electrónico
    - ❖ Documentos

La tecnología de Big Data tiene como objetivo minimizar la necesidad del hardware y reducir los costos de procesamiento [11].

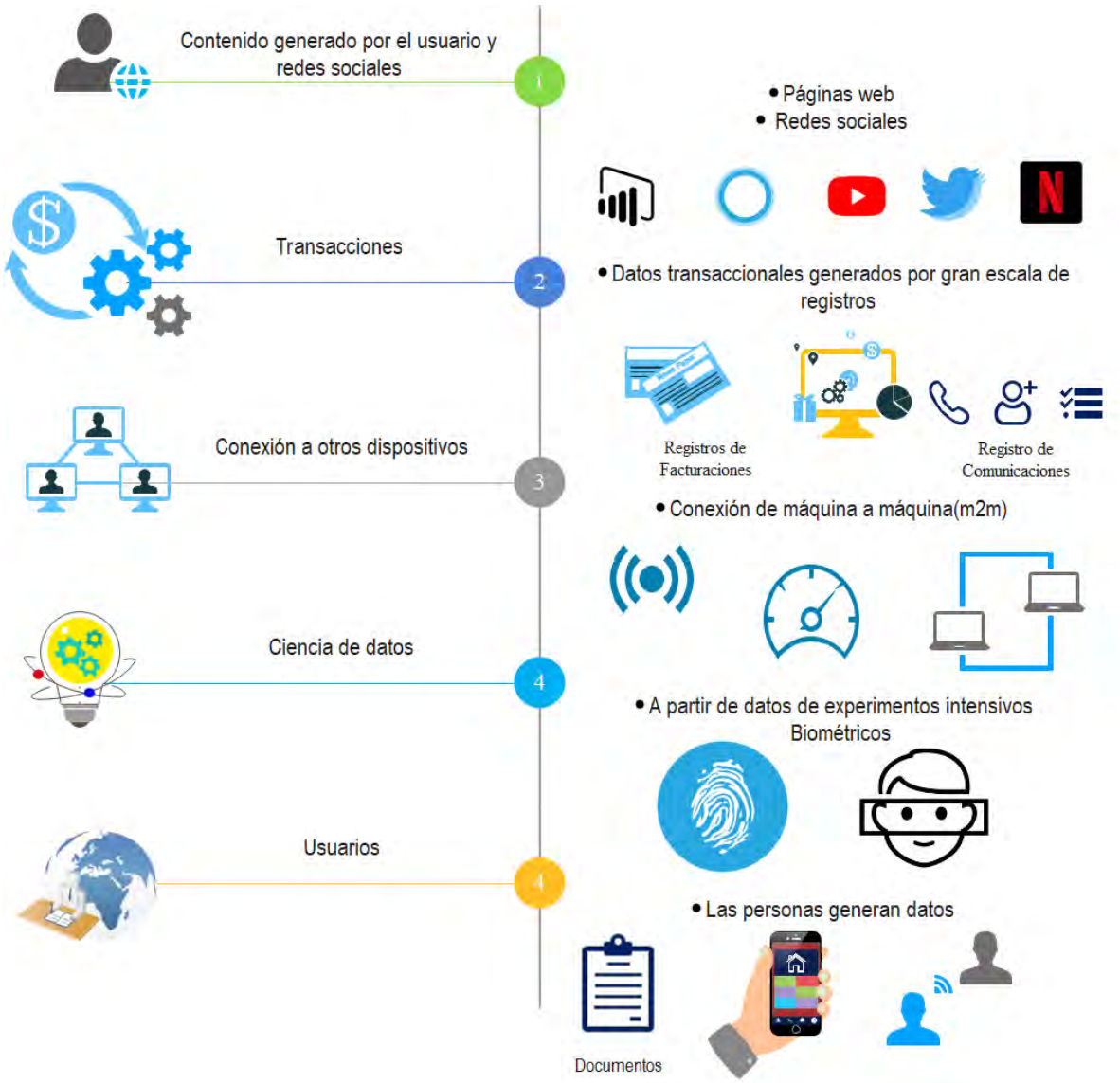


Figura 1.6 Representación de los tipos de datos en movimiento

1.6 Desafíos de los grandes volúmenes de datos

Algunas de las problemáticas para el desarrollo de las aplicaciones de los grandes datos se presentan como grades retos en la adquisición de datos, almacenamiento, gestión y análisis. Los sistemas de gestión y análisis

tradicionales toman como base el sistema de gestión de base de datos relacionales [1] permitiendo así crear, actualizar y administrar una base de datos relacional, esto surge a raíz de que los sistemas de gestión de bases de datos relacionales (RDBMS) no podían hacerse cargo de enorme volumen y la heterogeneidad de los datos grandes. A continuación, se muestra el proceso de análisis en la Figura 1.7, donde se descubre el conocimiento en minería de datos [2].

Figura 1.7 Desafíos en los análisis de datos

Desafíos en el Análisis de Big Data incluyen inconsistencias e incompletitud de datos, escalabilidad, oportunidad y seguridad de datos. A continuación, se menciona los desafíos previos y de gran importancia para el análisis de datos.

1.6.1 Representación de datos o almacenamiento

Muchos conjuntos de datos tienen ciertos niveles de heterogeneidad en el tipo, la estructura, la semántica, la organización, la granularidad y la accesibilidad [1]. Los conjuntos de datos crecen en tamaño debido a que cada vez más los recopilan dispositivos móviles ubicuos con detección de información [2].

1.6.2 Reducción de la redundancia y la compresión de datos

Es eficaz para reducir el costo indirecto de todo el sistema en la premisa de que los valores potenciales de los datos no se ven afectados, así como eliminar la ambigüedad y resolver los problemas [1].

1.6.3 Datos de gestión del ciclo de vida

Los valores ocultos en grandes volúmenes de datos dependen de la actualidad de los datos [1]; esto es con el fin de que sea eficiente y facilite la extracción de imagen fiable y optimizar recursos [9].

1.6.4 Mecanismo de análisis

El sistema de análisis de grandes volúmenes de datos que se deberá procesar grandes cantidades de datos heterogéneos en un tiempo limitado. Las RDBMS tradicionales tienen una gran falta de escalabilidad y capacidad de expansión es por ello por lo que tienen mayor ventaja en cuanto al manejo y tratamiento de datos estructurados [1].

1.6.5 Confidencialidad de los datos

El análisis de grandes volúmenes de datos puede contemplar los riesgos de seguridad para procesar y proteger datos sensibles, y así garantizar su seguridad [1], haciendo los datos fiables, de manera accesible y manejable [6].

1.6.6 Fiabilidad y escalabilidad.

El sistema de análisis de grandes volúmenes de datos debe ser compatible con los conjuntos de datos actuales y futuras. Esto para codificar los datos de seguridad y privacidad [9].

CAPITULO 2

INTELIGENCIA DE NEGOCIOS (BUSINESS INTELLINGENCE)

CAPITULO 2: INTELIGENCIA DE NEGOCIOS (BUSINESS INTELLINGENCE)

La inteligencia de negocios (BI por sus siglas conocidas en inglés: Business Intelligence) se refiere al uso de datos en una empresa para facilitar la toma de decisiones. Es un conjunto de estrategias y herramientas enfocadas al análisis de datos de una empresa mediante el análisis de datos existentes. Al combinar la tecnología, las herramientas y los procesos, se transforman los datos almacenados en información, esta a su vez en conocimiento que se dirige a un plan o estrategia comercial.

Todas las empresas pueden recopilar datos; datos relativos a ventas, a compras, a inversiones, a tiempos, etc. Miles de datos y variables pueden ser estudiados y utilizados para tomar nuevas estrategias, conocer las fortalezas propias, y por supuesto, las debilidades. Es por ello por lo que Business Intelligence debe ser parte de la estrategia empresarial, en virtud de permitir manejar de manera optimizada el manejo de recursos, realizar un monitoreo de los objetivos y contar con la capacidad de tomar buenas decisiones con la finalidad de obtener mejores resultados.

En términos generales, el Business Intelligence trata de extraer los datos de la empresa de distintas fuentes mediante las herramientas de Big Data. Todo este análisis, debería permitir incrementar el nivel financiero, administrativo, y con las decisiones mejorar las acciones de la empresa.

Se puede decir que cuando se tienen los datos, pero se carece de información, es necesario profundizar en los datos y tener la capacidad para encontrar patrones de comportamientos, así como también monitorear, rastrear, entender y administrar para obtener respuesta de aquellas preguntas que den pie a maximizar el rendimiento de la empresa.

Una de las problemáticas que enfrentan las empresas, es cuando existe fragmentaciones en el manejo de información generando diferentes versiones, lo que lleva a la creación de resultados diferentes del mismo informe, siendo complicado para las empresas, a causa de que el nivel de rendimiento puede

tornase bajo, por ello surge la necesidad de aplicar herramientas Business Intelligence. Dichas herramientas, beneficia a las empresas para un mejor manejo de crecimiento, en vista de ser para ellos un reto evolucionar ante los procesos que generan cambios.

Muchas empresas se preocupan por el manejo de los costos, siendo este un detonante para tomar en consideración la solución de Business Intelligence, la cual se encarga del control de costos, midiendo la capacidad de gastos, identificando que tipo de negocio, productos, entre otras fuentes de datos. Las empresas guardan información que les es muy valiosa, la problemática inicia cuando se desea realizar la transformación de información en conocimiento y dirigirlo a un giro comercial que beneficie y obtenga ganancia a la empresa. Con la ayuda de indicadores de gestión, es posible realizar análisis, monitoreo y administración.

Los beneficios que se pueden obtener mediante Business Intelligence, se puede catalogar de distintos tipos [18].

- **Beneficios tangibles:** reducción de costos, generación de ingresos y reducción de tiempos en la que se realizan las actividades

  .

- **Beneficios intangibles:** se tiene la información disponible para tomar una decisión, generando que esta información sea tomada en cuenta por otros usuarios para tomar decisiones y competencia.

- **Beneficios estratégicos:** buscando todo aquello que permita formular una estrategia con el fin de saber a qué giro dirigirse, es decir, productos, clientes o mercados. Con la ayuda de los indicadores de gestión para conocer el negocio.

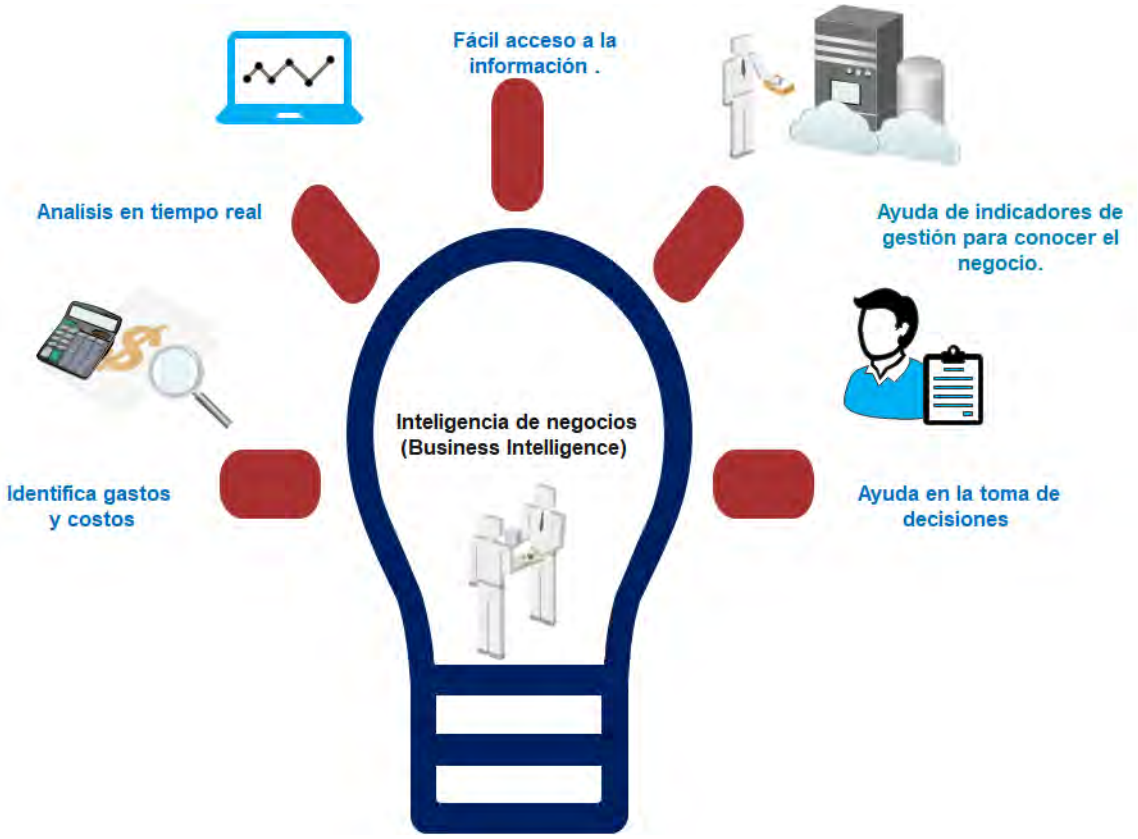


Figura 2.1. Beneficios que se obtienen mediante Business Intelligence

2.1 Componentes de Inteligencia de negocio

En esta sección abordaremos algunos componentes que se suman a la importancia de la inteligencia de negocios.

2.1.1Sistemas de información SQL o NoSQL

Es el almacén de datos que se encuentran en filas y columnas, el acceso es mediante consultas SQL. Es una base de datos relacional que se encuentra baja la escritura del lenguaje de consulta estructurado denominado SQL y se caracteriza por manejar con rapidez los datos dependientes del contexto [4]. Por otra parte “No solo SQL” por sus siglas conocido como NoSQL es una base de datos no relacional que se ha hecho popular debido a que puede trabajar con grandes volúmenes de información, NoSQL puede almacenar datos de cualquier estructura sin ser dependiente de SQL, estas estructuras permiten el

almacenamiento de información cuando se generan problemas de escalabilidad en las bases de datos relacionales, debido a las concurrencias de los usuarios y la demanda de millones de consultas [19].

2.1.2 Proceso ETL

Es el proceso de extracción, transformación y carga de datos, conocido por sus siglas en ingles Extract, Transfor, Load (ETL), dentro de sus funciones se encarga de filtrar, ordenar, combinar, limpiar, verificar y convertir los datos validos en tipos y formatos adecuados. Normalmente, se ejecutan en paralelo las tres fases del proceso ETL para ahorrar tiempo [8].

Los procesos ETL se pueden definir de acuerdo con lo que representan, de la siguiente manera:

- **Extracción:** Como su nombre lo indica se encarga de obtener y/o extraer los datos que se encuentran en el sistema de origen.

- **Transformación:** En este proceso se realizan varias operaciones, los diversos datos precedentes de repositorios digitales, normalmente no hay coincidencia en formato, es por ello por lo que resulta difícil integrarlos causando esto la necesidad de realizar la operación de transformación [20].

- **Carga:** Se introducen los datos en un almacén una vez que hayan sido extraídos y transformados al formato deseado para que así sean cargados a su destino. Hay casos en los que se sobre escribe la información que es vieja con la actual, así como también hay otros que se encargan de guardar el historial de los cambios realizados.

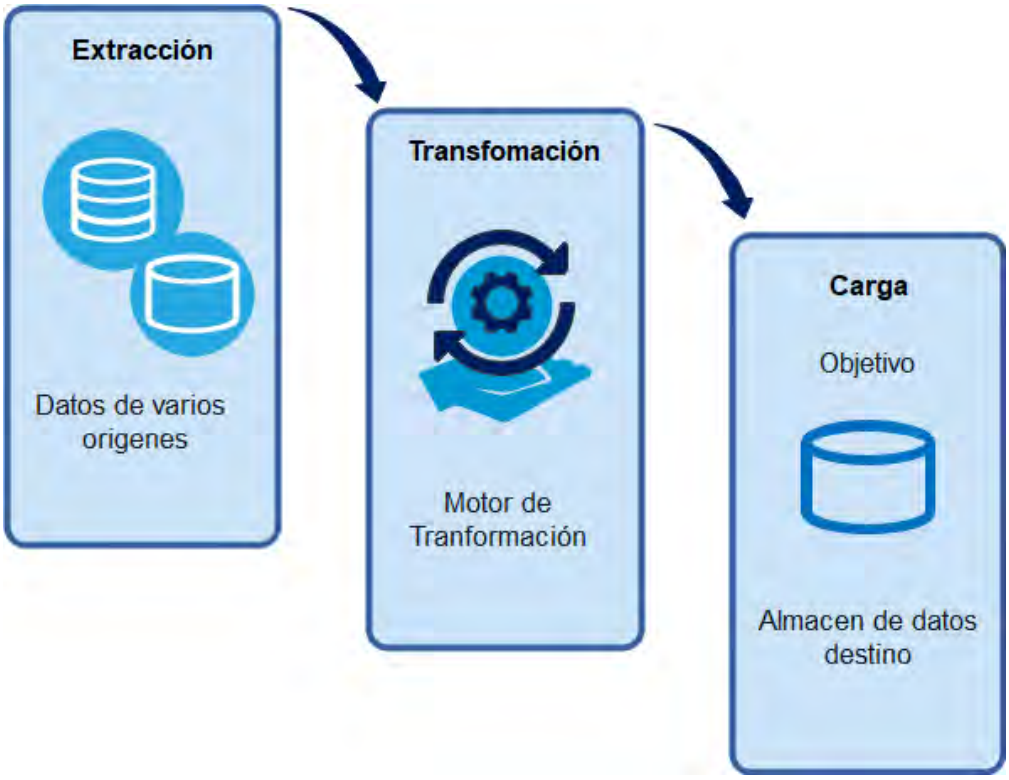


Figura 2.2 Fases del proceso ETL

2.1.3 Data Mart

Se dice que es una base de datos departamental, encargada del almacenamiento de los datos designada a un área específica. Una de sus características es contar con estructura óptima de datos con la finalidad de realizar un análisis de información detalladamente desde una perspectiva que cree una afectación a los procesos.

Forma parte del subconjunto de los datos del Data Warehouse y tiene como objetivo analizar la función o necesidad, con una cantidad especifica de usuarios. Los datos se encuentran estructurados en modelos de estrella o copo de nieve.

2.1.4 Data Warehouse o Almacén De Datos

Un Data Warehouse es una base de datos corporativa cuya característica principal es integrar y depurar información de una o más fuentes, para luego procesarla, lo que permite un análisis desde diversas perspectivas y con grandes velocidades de respuesta. La creación de un Data warehouse representa el

primer paso, para implantar una solución completa y fiable de LA Inteligencia de negocios [16]. Un Data warehouse es un medio de almacén de datos que proporciona una perspectiva más amplia, común e integrada de los datos de la organización.

Existen dos tipos de esquemas para estructurar los datos en un almacén de datos, las cueles son el esquema copo de nieve y estrella [21].

2.1.5 Procesamiento Analítico en Línea (OLAP)

Es una de las soluciones en inteligencia de negocios, permite extraer de manera fácil y selectiva los datos. Es una estructura multidimensional de base de datos encargada de organizar y relacionar los datos entre sí, se pueden visualizar en forma de cubos, estos cubos dentro de cubos de datos y cada una de los lados o caras del cubo es considerado una dimensión de los datos. Este procesamiento analítico en línea está constituido de manera compacta y es fácil de comprender, teniendo como finalidad visualizar y manipular los elementos de datos interrelacionados [21].

2.2 Modelo de Inteligencia de negocios

Está conformado por 4 fases en las cuales se considera un conjunto de herramientas, aplicaciones, infraestructura y mejor manejo de las practicas que permiten el análisis y acceso de los datos adquiridos y así optimizar el desempeño de las organizaciones.

- **Fuentes de información.** Son aquellas en las que se encuentran contenidas las bases de datos, sistemas de transaccionales asociadas a los sistemas de información SQL o NoSQL para almacenar la información el Data Warehouse.

- **Procesamiento de datos**. La adquisición de los datos está a cargo de las herramientas tecnológicas centradas al proceso de extracción, transformación y carga de los datos denominado por sus siglas ETL.

- **Almacenamiento.** Después del procesamiento de datos estos deben almacenarse y archivarse, para las grandes cantidades de datos existen bodegas de datos (Data Waterhouse, Data Mart) que se encargan de la extracción, transformación y almacenamiento, siento esta la parte central de la arquitectura del modelo de la Inteligencia de negocios (BI). El sentido de ellos es almacenar los datos de manera que sea mayor su administración, acceso y flexibilidad [16]. De igual manera, de acuerdo con Oracle, “una bodega de datos es una base de datos diseñada para permitir las actividades de inteligencia de negocios: existe para ayudar a los usuarios a comprender y mejorar el rendimiento de la organización” [17].

- **Agregación y aplicación de herramientas tecnológicas.** Aquí se analiza los volúmenes de datos mediante el uso de datos multidimensionales (OLAP) [16][24].

- **Visualización.** Permiten suministrar las cifras de interés a los usuarios e interesados finales, así como la navegación y el análisis [16][17]. Mediante reportes, análisis OLAP, cuadros de mando (Dashboard) y minería de datos [24]. La visualización de los datos, así como la interacción, forman parte esencial de las tecnologías de Big Data, esto ayuda a comprender mejor los datos, interpretar y hacer efectivo su uso [11].

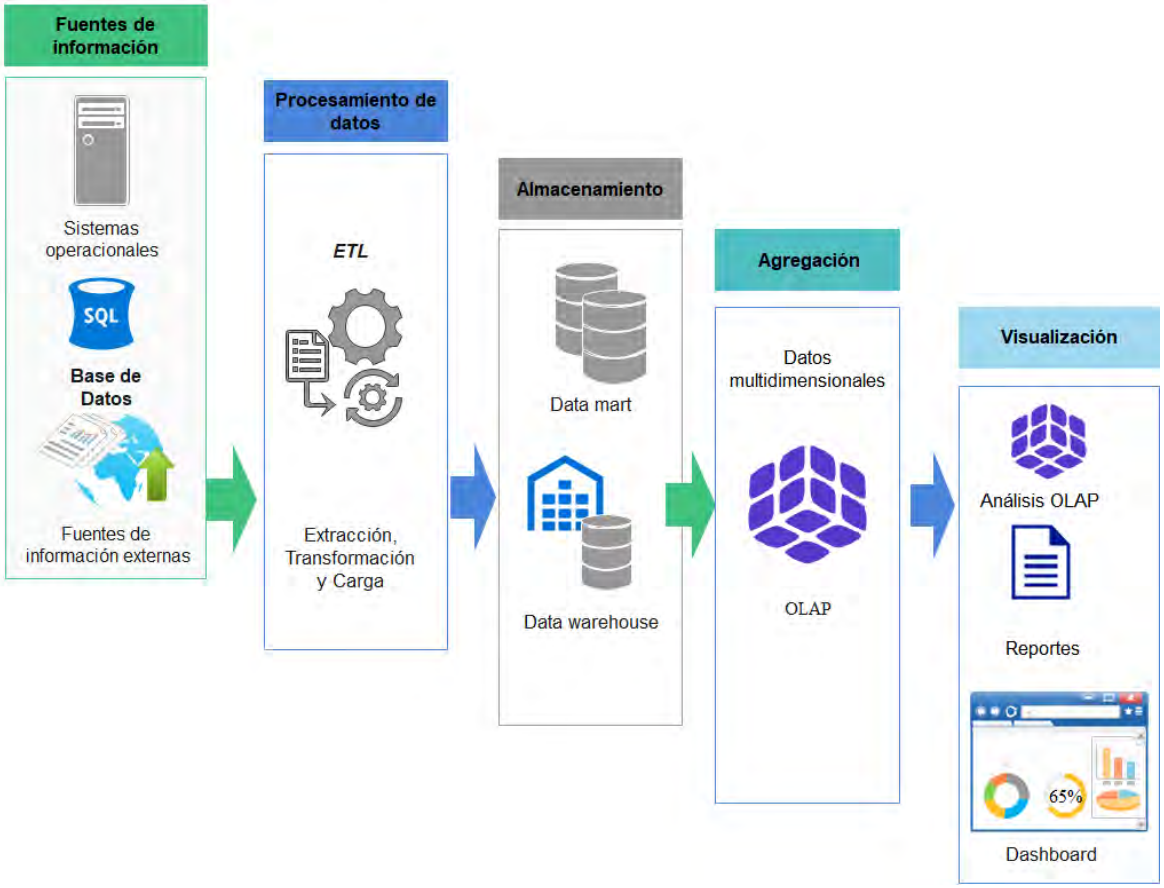


Figura 2.2 Representación del modelo de inteligencia de negocios (BI)

CAPITULO 3
ANALÍTICA BIG DATA

CAPITULO 3: ANALÍTICA BIG DATA

La primera impresión de Big Data es su volumen, por lo que el desafío más grande e importante es la escalabilidad cuando se trata de las tareas de análisis de Big Data [2].

Como se ha mencionado Big Data se caracteriza por la naturaleza compleja de los grandes volúmenes de datos que incluye las 5vs. Existen diversas técnicas analíticas, en donde se incluye la minería de datos, la visualización, el análisis estadístico y el aprendizaje automático [6].

- **Registro de datos / minería de datos:** La minería de datos se puede definir como el proceso en el cual se extrae el conocimiento partiendo de volúmenes de datos, es por ello que se le puede denominar como el proceso de descubrimiento del conocimiento. La minería de datos surge por la necesidad de conocer la información útil a partir de las bases de datos o Data Warehouse, con el crecimiento de los datos disponibles, la inteligencia de negocios usar opto la minería de datos para aplicar soluciones empresariales y comerciales, y así descubrir la información relevante [14]. Es de gran importancia para la minería de datos, los métodos para descubrir patrones interesantes y extraer valor oculto en tales enormes conjuntos de datos y flujos [6]. Big Data ha cambiado la forma de capturar y almacenar los datos, esto tomando en cuenta los dispositivos de almacenamiento, la arquitectura de almacenamiento de datos, así como los mecanismos de acceso. El requerimiento de más medios de almacenamiento ha creado la necesidad de tener una mayor velocidad en E/S para asumir el gran reto [2]. Una arquitectura de datos grandes tiene que adquirir datos de alta velocidad desde una variedad de fuentes (web, DBMS (OLTP), NoSQL, HDFS) y tiene que lidiar con diversos protocolos de acceso [4].

En la actualidad y lo más común es que los datos persistentes se almacenan en unidades de disco duro, mejor conocidos como HDD. Sabemos que el rendimiento de estos es más lento, sin embargo, estos están siendo reemplazados por las unidades de disco de estado sólido (SSD).

- **Análisis de estadístico:** El análisis del Big Datos se puede definir como el uso de técnicas analíticas avanzadas en grandes volúmenes de datos [4].

La clave para extraer valor de grandes volúmenes de datos es el uso de Analytics, que realizan la recolección y el almacenamiento a sí mismos agregando poco valor. Los datos tienen que ser analizados y sus resultados utilizados por los encargados de tomar decisiones y procesos de la organización [3].

- **Visualización:** El objetivo principal de la visualización de datos es representar el conocimiento de forma más intuitiva y eficaz mediante el uso de diferentes gráficos, es decir, se refieren principalmente a la visualización de forma múltiple, de múltiples fuentes y datos en tiempo real [4].

Una de las técnicas para la visualización de los datos más común es Big Data en la nube.

- **Aprendizaje automático (Machine learning):** Uno de los objetivos de aprendizaje automático es descubrir el conocimiento y tomar decisiones inteligentes. Se utiliza para muchas aplicaciones de palabra real, tales como motores de recomendación, sistemas de reconocimiento, la informática y la minería de datos y sistemas de control autónomo [6]. Los algoritmos de aprendizaje máquina se clasifican en supervisados y no supervisados [14].

3.1Tipos de analítica Big Data

Son utilizadas para Big Data y se pueden mostrar tres tipos de análisis en las diferentes aplicaciones en las tecnologías y las arquitecturas que se implementan tienen ciertas diferencias.

Podemos decir que la gran mayoría de los datos en bruto no dan a ofrecer demasiado valor, sin embargo al ser procesado mediante las herramientas y técnicas adecuadas se pueden extraer valiosas ideas y es por ello que gracias a la analítica avanzada, se puede tomar decisiones de forma más adecuada al utilizar los datos de manera concreta, permitiendo así automatizar cada proceso

reduciendo costos y tiempo en las tareas más recurrentes dando como resultado una mayor eficiencia en los procesos que son importantes en las organizaciones.

Figura 3.1 Tipos de Analítica Big Data

3.1.1 Analítica descriptiva

Se puede describir como el tipo de analítica más simple. Este proceso de análisis permite concentrar el Big Data en datos más pequeños, permitiendo así que las informaciones sean piezas más manejables.

Su trabajo es almacenar y agregar información a partir del análisis de los datos históricos permitiendo que se visualicen para que sea comprendido el estado en el que se encuentra una organización.

Si una organización trabaja con esta analítica podrá detectar las actividades de producción, si el giro fuese de ventas podría ver qué productos se están vendiendo. Es común ver como la tecnología ha evolucionado, así como el uso de redes sociales, a través de la analítica descriptiva se puede identificar de qué manera está repercutiendo en este ámbito, así como de la visualización de la actividad de las personas, es entonces ahí cuando se puede observar de qué manera va evolucionando y creando datos históricos [12].

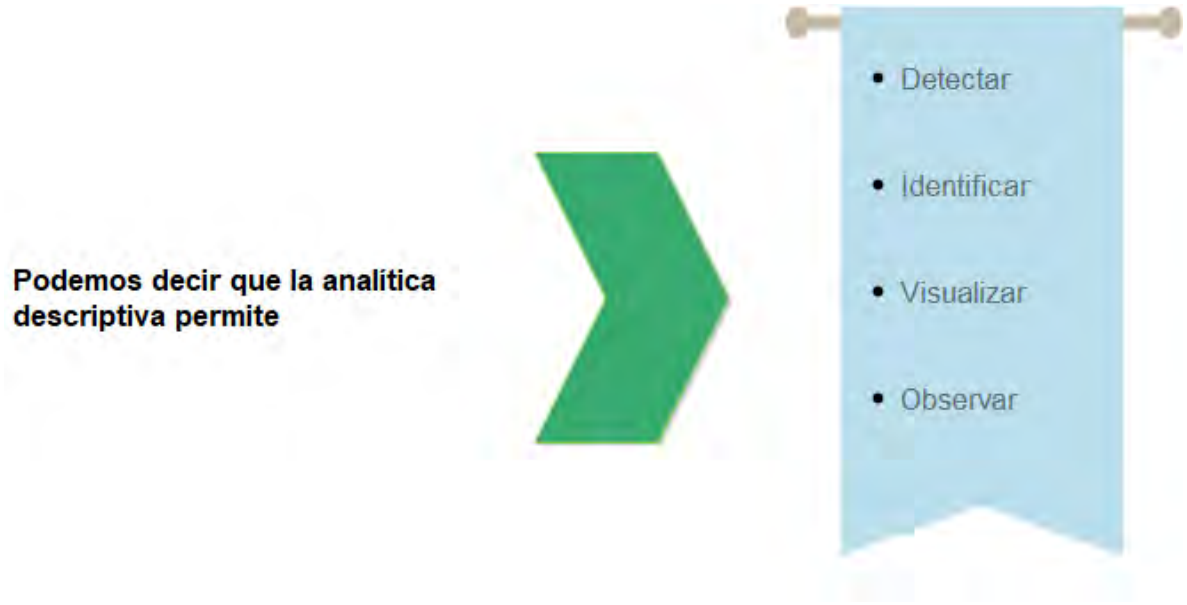


Figura 3.2 Actividades de Analítica descriptiva

3.1.2 Analítica predictiva

Es el encargado de predecir acerca del futuro o de sucesos por acontecer, por encargarse de agrupar y analizar los datos actuales e históricos, a través de la minería de datos se puede obtener la información y conocimiento de lo que se pretende predecir, así como los patrones de comportamiento respaldados por una serie de diferentes herramientas analíticas como son: algoritmos estadísticos, consultas y aprendizaje automático. El uso de ella se presenta cuando las organizaciones se encuentran en problemas difíciles de resolver, tales como:

- **Detección de fraudes.** Mediante el uso de herramientas de Big Data es posible almacenar grandes volúmenes de datos que permiten la detección de manera fácil y en tiempo real de la realización de algo indebido, así también permite realizar cruce de información, así también es posible el uso de algoritmos avanzados como machine learning. Uso en el marketing. Se ocupa para optimizar y determinar las respuestas o las compras que realizan los clientes, a través de este modelo ellos

pueden atraer a las empresas, mantener e ir incrementando la cantidad de clientes que le son viables.

- **Progresión en las operaciones**. Ciertas organizaciones lo utilizan para realizar un pronóstico en sus inventarios y gestión de recursos.

- **Probabilidades bajas de riesgo.** Comúnmente relacionado con la evaluación que se les realiza a los clientes en las organizaciones que tiene como objetivo realizar una valoración respecto a sus compras predeterminadas, un ejemplo es la evaluación de crédito en donde se ve relacionado la incorporación de todos los datos relevantes para así conocer que tan viable es la otorgación de crédito.

En conclusión, el análisis predictivo permite optimizar y admitir segmentación que es compleja para así automatizar los datos. En la actualidad muchas organizaciones se rigen de esta analítica para obtener un mejoramiento en sus operaciones y así adquirir ventaja sobre la competencia a través del uso de datos, algoritmos estadísticos y técnicas de machine learning [12].

3.1.3 Analítica prescriptiva

En esta etapa analítica se requiere primero haber aplicado la analítica descriptiva, como se mencionó antes se encarga de dar a conocer el estado actual en el que se encuentra la organización así después saltamos a la analítica predictiva para realizar una estimación de aquello que no se conoce, es decir de los sucesos a acontecer, a estos procesos se le integra como extensión la analítica prescriptiva, encargada de sugerir sobre toma de decisiones o acciones.

A través del análisis de los sistemas se ejecutan las técnicas de simulación y optimización para encargarse de realizar recomendaciones tales como reducción de costos, mejora de los beneficios y detección de alternativas

optimas tomando en cuenta las posibilidades y posibles soluciones que sean de conveniencia a seguir. La analítica prescriptiva está caracterizada por trabajar de manera inteligente y capaz de procesar planteamientos de propuestas, en base a la información que recopila hace una valoración de las posibles opciones a tomar para así poder al fin seleccionar mediante el proceso óptimo para favorecer el máximo rendimiento.

La analítica prescriptiva ofrece lo siguiente:

- Reducción de incapacidad, planificación de manera optimizada y decisiones de manera inteligente.
- Gestión de recursos costo - beneficio más eficiente.
- Riesgos e impacto sobre las decisiones.

La finalidad de las técnicas que se implementan para el análisis de datos es generalmente para permitirnos mejorar como gestionamos la información y datos que se obtienen.

CAPITULO 4

BIG DATA EN LA NUBE

(CLOUD COMPUTING)

CAPITULO 4: BIG DATA EN LA NUBE (CLOUD COMPUTING)

Figura 4.1 Big Data en la nube

Las nubes son entornos de TI que extraen, agrupan y comparten recursos escalables en una red. Dichos entornos, suelen crearse para habilitar lo que conocemos como Cloud Computing, que consiste en ejecutar cargas de trabajo dentro del sistema [10]. Esto es un componente clave para los grandes volúmenes de datos, no solo proporciona infraestructura y herramientas, sino que es un modelo de negocio. Cloud Computing, o computación en la nube logró virtualizar procesos que requerían de grandes inversiones en hardware, mismas que no siempre podían ser afrontadas por las organizaciones. Esto ha permitido también, que el crecimiento de los datos y su procesamiento se pueda escalar, así como se ha encargado de proporcionar infraestructura, servicios, plataformas y aplicaciones según se requiera en las redes [10], en relación con los grandes volúmenes de datos. Las cualidades principales del uso de la computación de la nube son caracterizadas por el enorme recurso de computación, almacenamiento y procesamiento de los datos, que tiene como finalidad brindar solución eficaz a la gestión de los datos. A medida que las aplicaciones se fueron incrementando, esto genero el impulso de su evolución de datos, siendo la demanda de aplicaciones y la computación en la nube

desarrollados a partir de las nuevas tecnologías, en conclusión, se puede decir que Cloud Computing no solo es computación si no también es un modo de servicio a través de la conectividad y gran escala de Internet. Existe una diferencia de esquemas de servidores tradicionales, el Cloud Computing da permiso a el tratamiento de los grandes volúmenes de información del Big Data caracterizándolo así por una palabra clave: la escalabilidad. Se puede mencionar los beneficios que se da al unir las tecnologías de Cloud Computing y el Big Data. Una de ellas es que Cloud Computing ofrece una manera económica de brindar soporte en las tecnologías Big Data, así como las aplicaciones analíticas a lo cual lleva a impulsar el sentido y valor empresarial [20]. Big data y Cloud Computing se complementan, trabajar en la nube ha sido sin duda alguna una tendencia en el desarrollo tecnológico, mientras que Cloud Computing ofrece a las empresas y los usuarios alta escalabilidad, alta disponibilidad y confiabilidad [11]. La computación en la nube ofrece todo esto a través de la virtualización de hardware. Por lo tanto, Big Data y computación en la nube son dos conceptos compatibles, la nube permite que el Big Data esté disponible, sea escalable y tolerante a fallas.

Ventajas

- Ofrece reducción de costos y escalabilidad.
- Acceso desdé cualquier dispositivo y agiliza el trabajo.
- Ahorras en equipamiento y tiempo de instalación.
- Acceso multiplataforma.
- Manejo de información en tiempo real.

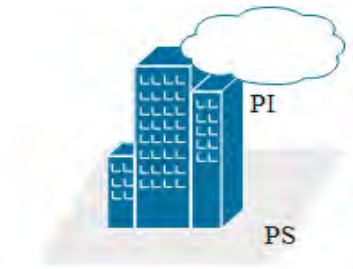
Desventajas

- Es dependiente de los proveedores, que se rija de buenas políticas.
- Conectividad nula o sin acceso a la conexión de Internet.
- Puede existir vulnerabilidad de la privacidad, los datos que son sensibles se encuentran fuera.
- Los riesgos de seguridad disminuyen cuando accedemos mediante protocolos https.

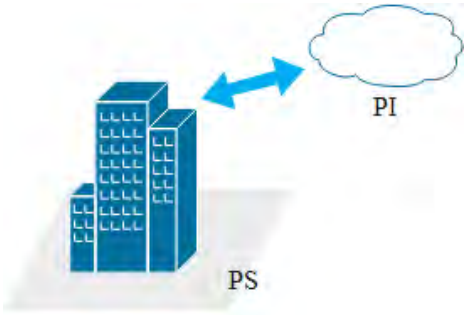
En los modelos de implementación existen algunos tipos de nubes en función de su privacidad [24].

- Nubes privadas.

Acceso y gestión desde una determinada organización.



PI=Provedor de Infraestructura
PS= Provedor de Servicios

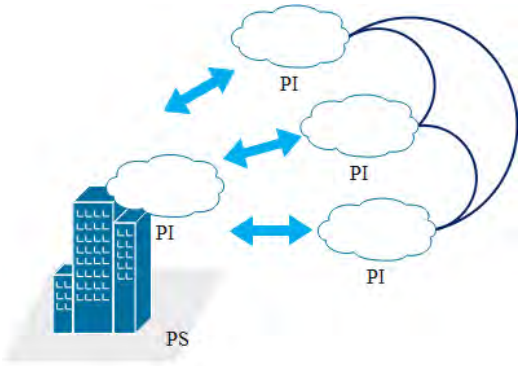


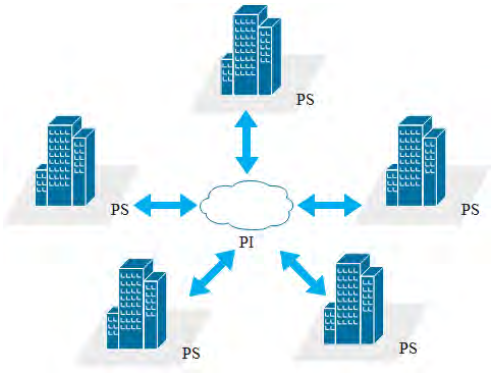
- Nubes públicas.

Pertenecen a un proveedor Cloud Computing y están abiertas al público.

- Nubes híbridas

En donde se compaginan las dos nubes anteriores.





- Nubes de comunidad

Comparte la infraestructura con varias organizaciones.

4.1 Servicios en la nube

Lo que se puede encontrar en los servicios basados en la nube vienen en diversas formas. Los proveedores externos son los encargados de brindar servicio en la nube publica mientras que nubes privadas se implementan dentro del firewall de una empresa.

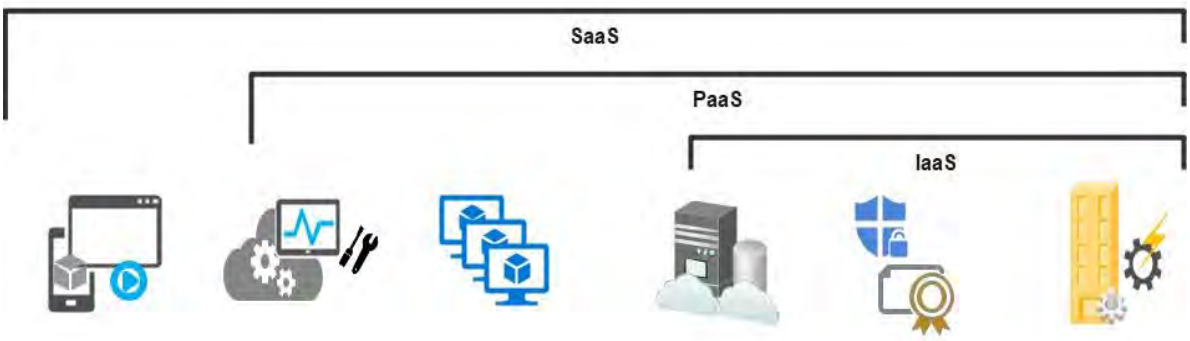


Figura 4.2 Servicios en la nube

Los servicios referidos en la nube pueden ser:

- Software como servicio (SaaS: Software as a Service).
- Plataforma como servicio (PaaS: Plataform as a Service).
- Infraestructura como servicio (IaaS: Infraestructure as a Service).

Estos son los encargados para poder acceder a una plataforma de Cloud Computing y estos son similares a los datos que se cargan de una empresa, así como el almacenamiento y análisis, los resultados se descargan a los usuarios y las aplicaciones.

4.1.1Software como servicio (SaaS)

De los tres niveles es el más conocido del Cloud Computing, brindando apoyo a través del software, mejorando y cubriendo los procesos de las empresas. Es una de las aplicaciones que se consumen de Internet y el acceso se da muchas

veces por el navegador. Para el uso, está condicionado el pago, esto es porque los datos se encuentran alojados en la plataforma del proveedor. Al adquirir este servicio, viene conjuntamente el hardware, el sistema operativo, el software de aplicación y del almacenamiento, todo esto bajo el cargo del proveedor, la tarea que el usuario tiene es cargar los datos y hacer uso del software de la aplicación o procesarlos. En la inteligencia de negocios (BI) y análisis existen muchos proveedores que se encargan de ofrecer sus servicios a través de la nube. Los que han destacado siendo los lideres son IBM Cognos y SAS.

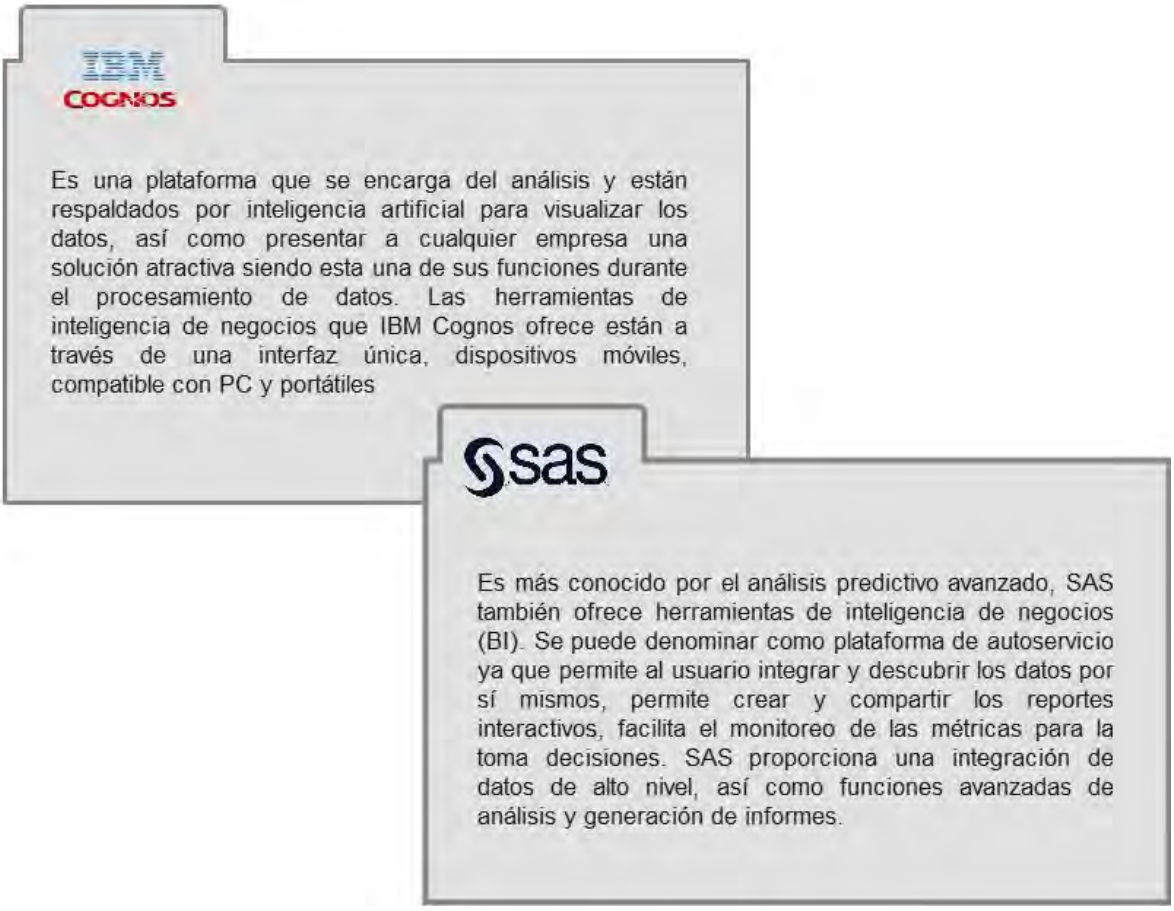


Figura 4.3 Líderes de servicios IBM Cognos y SAS

SaaS se ha convertido en una de las opciones más atractivas para las empresas que al no contar con los recursos financieros o humanos para implementar software y las aplicaciones de manera interna [19].

4.1.2 Plataforma como servicio (PaaS)

En esta plataforma el software para crear o ejecutar las aplicaciones no es proporcionado por el proveedor, únicamente le brinda la plataforma básica, es decir, le proporciona al cliente un entorno de desarrollo y herramientas para nuevas aplicaciones. Las empresas pueden crear y desarrollar sus propios servicios en una plataforma existente, he aquí PaaS se diferencia de SaaS. El servicio PaaS les brinda a las empresas una plataforma digital en la cual desarrollan e implementan sus propias aplicaciones y sus servicios, no tiene la necesidad de mantener espacio del servidor, del software de programación ni de los protocolos de seguridad que se encuentra internamente [20]. Se considera que se encuentra por plataformas compuesta por servidores, almacenamiento y sistemas operativos en conjunto con los sistemas que se encargan de la gestión de base de datos, cabe mencionar que no todas las plataformas tienen bases de datos, a esto se le añade herramientas de desarrollo y técnicas de contenedores. Las plataformas de servicio en la nube más conocidas son: Oracle Cloud Computing, Google App Engine (GAE) y Microsoft Windows Azure.

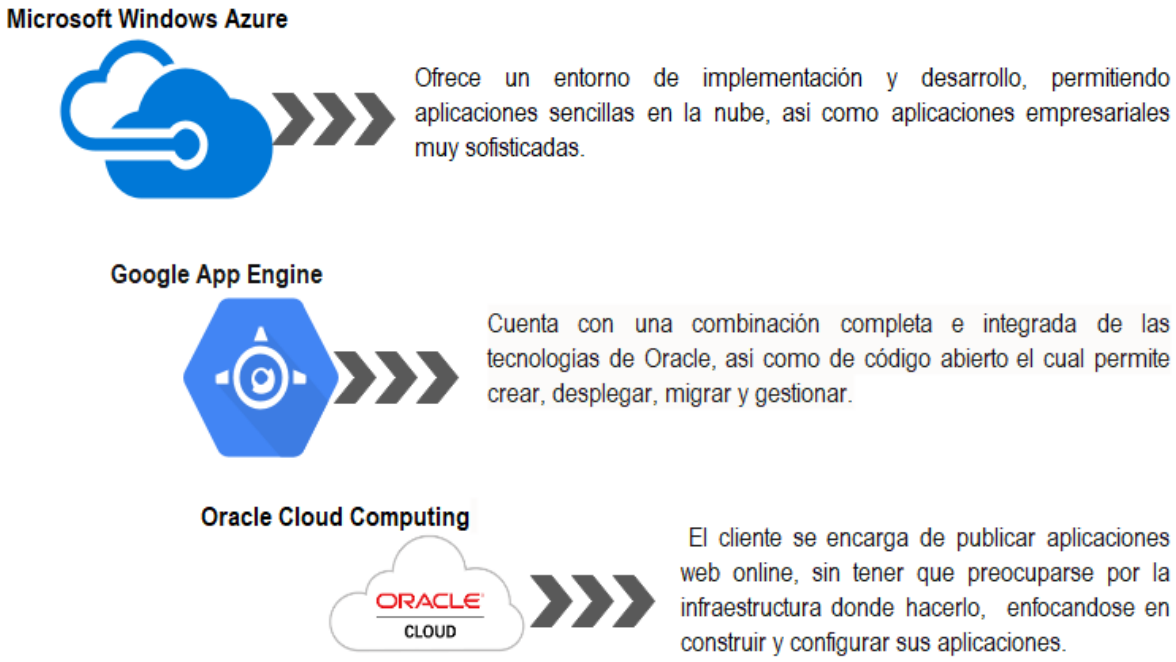


Figura 4.4 Plataformas de servicio en la nube

4.1.3 Infraestructura como servicio (IaaS)

A diferencia de los otros servicios, en este el proveedor proporciona el almacenamiento y la potencia de cálculo sin ser procesado, no integra el sistema operativo, así como tampoco el software de aplicación, quien se encarga de esto es el cliente, es decir, cargan una imagen que contiene la aplicación y el sistema operativo. En otros términos, este servicio ofrece infraestructura, lo cual significa almacenamiento, potencia de procesamiento y máquinas virtuales, todo esto el proveedor de la nube satisface las necesidades del cliente virtualizando los recursos de acuerdo con los acuerdos de nivel de servicio. Cloud Computing ofrece todo esto a través de la virtualización de hardware. Por lo tanto, Big Data y computación en la nube son compatible y congenian, por consiguiente, la nube permite que Big Data esté disponible, sea escalable y tolerante a fallas [18]. En Big Data se presentan interesantes servicios basados en la nube como Amazon EC2 el cual es parte de lo que oferta Amazon Web Services y Google Compute Engine [12]. Estos incluyen los servicios de virtualización como lo son las máquinas virtuales, así como los servicios de almacenamiento mediante disco conocido como almacenamiento no relacionado y las bases de datos denominado servicios de almacenamiento relacionado [19]. AWS (Amazone Web Services), es la plataforma informática en la nube más popular, ofrece un conjunto integral de servicios de informática el cual realiza analítica con Big Data. Este servicio es de paga, ofrece soluciones tecnológicas durante la recopilación de información, transmisión, almacenaje y análisis [22]. A raíz de que los servicios se han vuelto más sofisticados y están siendo mejor desarrollados, las empresas están optando por utilizar o implementar las plataformas de Cloud Computing, como se ha mencionado antes, la evolución de Big Data se caracteriza por el crecimiento del tráfico de datos y el volumen de datos, los formatos de datos que ahora son de múltiples fuentes y heterogéneos, y es por ello que deben ser tratados mediante un procesamiento de datos preciso y en tiempo real [11]. Entonces podemos concluir que las nubes son entornos donde se ejecutan las aplicaciones que a su vez al relacionarse con la parte

computacional se desarrolla el Cloud Computing, el cual se encarga de ejecutar todos los procesos de las aplicaciones de una nube. Cloud Computing es el entorno bajo el que se genera el almacenamiento de la información del Big Data y los modelos de servicio pueden ser administrados por los usuarios y los proveedores, siendo clasificados dependiendo del acceso que el cliente tenga [24].

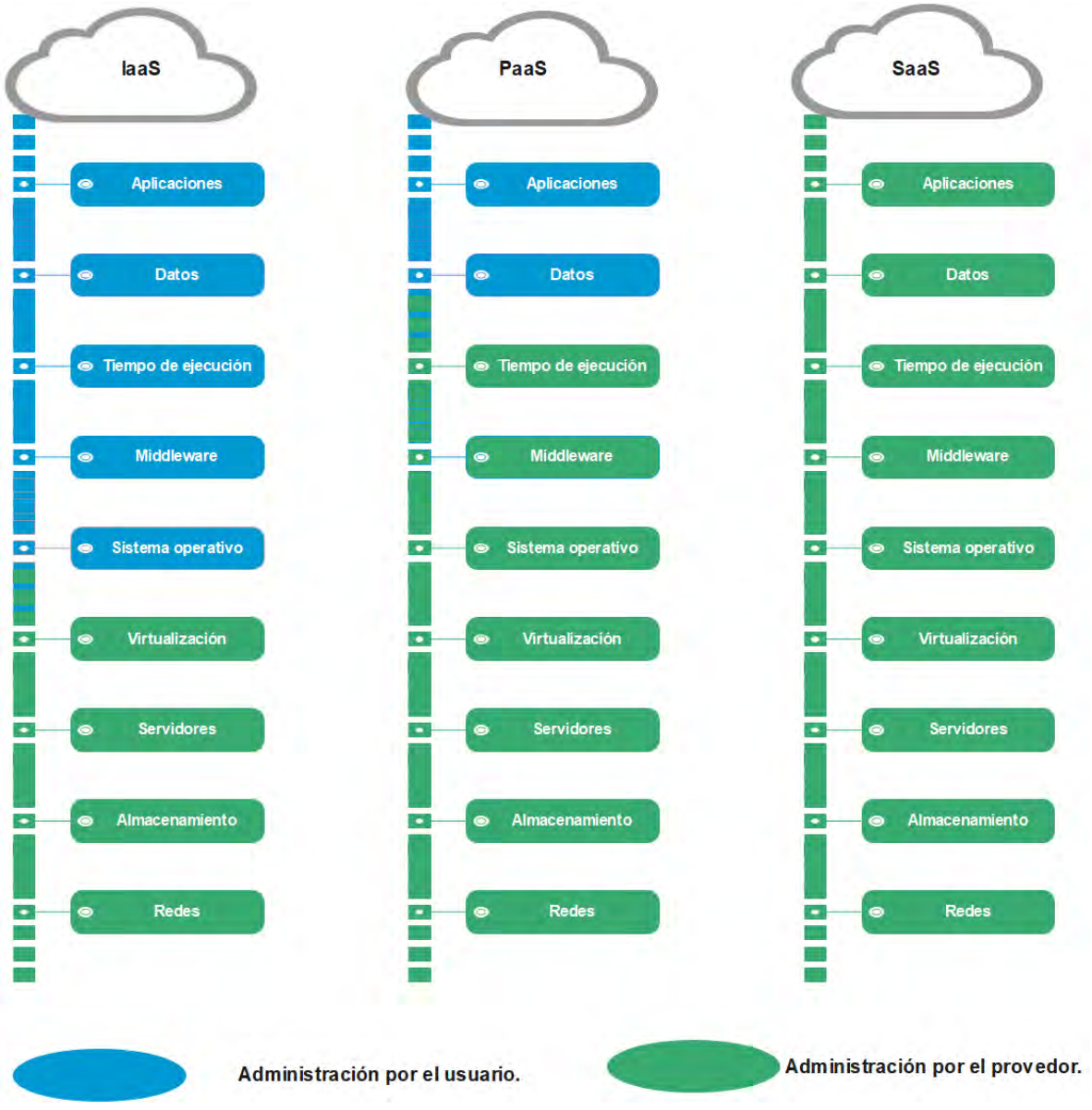


Figura 4.5 Representación de los modelos de servicios en la nube

CAPITULO 5
TECNOLOGÍAS DE BIG DATA

CAPITULO 5: TECNOLOGÍAS DE BIG DATA

La demanda masiva de datos en tiempo real ha causado que el procesamiento de los datos tradicionales no sea suficiente siendo un gran desafío [11]. Estos datos provienen de fuentes heterogéneas, estructurada de otra manera o bien sin estructurar**,** y se generan de manera rápida [14]. El conjunto de tecnologías brinda herramientas de procesamiento y almacenamiento de datos. Existen varias tecnologías fundamentales que están estrechamente relacionados con grandes volúmenes de datos también conocidas como herramientas [1].

Para capturar el valor de grandes volúmenes de datos, es necesario desarrollar nuevas técnicas y tecnologías para el análisis de esta. Las herramientas actuales se concentran en tres clases:

- **Herramientas basadas en el procesamiento por lotes:** La mayoría de las herramientas de procesamiento por lotes se basan en la infraestructura Apache Hadoop, como Mahout y Drya [2].
- **Herramientas basadas en el procesamiento de flujo:** Apache Storm y S4 son buenos ejemplos de plataformas de análisis de datos grandes de Streaming. El análisis interactivo procesa los datos en un entorno interactivo, permitiendo a los usuarios llevar a cabo su propio análisis de la información. El usuario se conecta directamente al ordenador y por lo tanto puede interactuar con él en tiempo real. Los datos se pueden revisar, comparar y analizar en formato tabular o gráfico, o ambos al mismo tiempo [2].
- **Herramientas de análisis interactivas:** Dremel y Apache Taladro de Google son las plataformas de grandes volúmenes de datos en base a un análisis interactivo [2].

5.1 Plataformas y software para tratamiento de Big Data

Se presentan tecnologías de software libre permitiendo así la generación de soluciones de Big Data tomando en cuenta aquellas necesidades que se presentan en un dominio de datos u organización, existen de igual forma tecnologías propietarias que soportan Big Data [14].

En nuestro entorno actual ya existen diferentes herramientas, la existencia de software que tiene como tarea el uso de la tecnología Big Data, la cual llamamos software de tratamiento, encargados de los grandes almacenes de datos, es entonces cuando se puede decir que MapReduce es la base de la programación de los diferentes herramientas y software. De igual manera le sigue Hadoop, el cual es el software más utilizado y actualmente tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto Hadoop [8].

Se ha creado una comunidad Hadoop misma que ha contribuido a enriquecer su ecosistema con varios módulos de código abierto. El poder de la plataforma Hadoop se basa en dos principales subcomponentes: el sistema de archivos distribuido Hadoop (HDFS) y el marco MapReduce, mismo que se mencionaran de manera descriptiva, seguido de Hbase y Apache Spark [6].

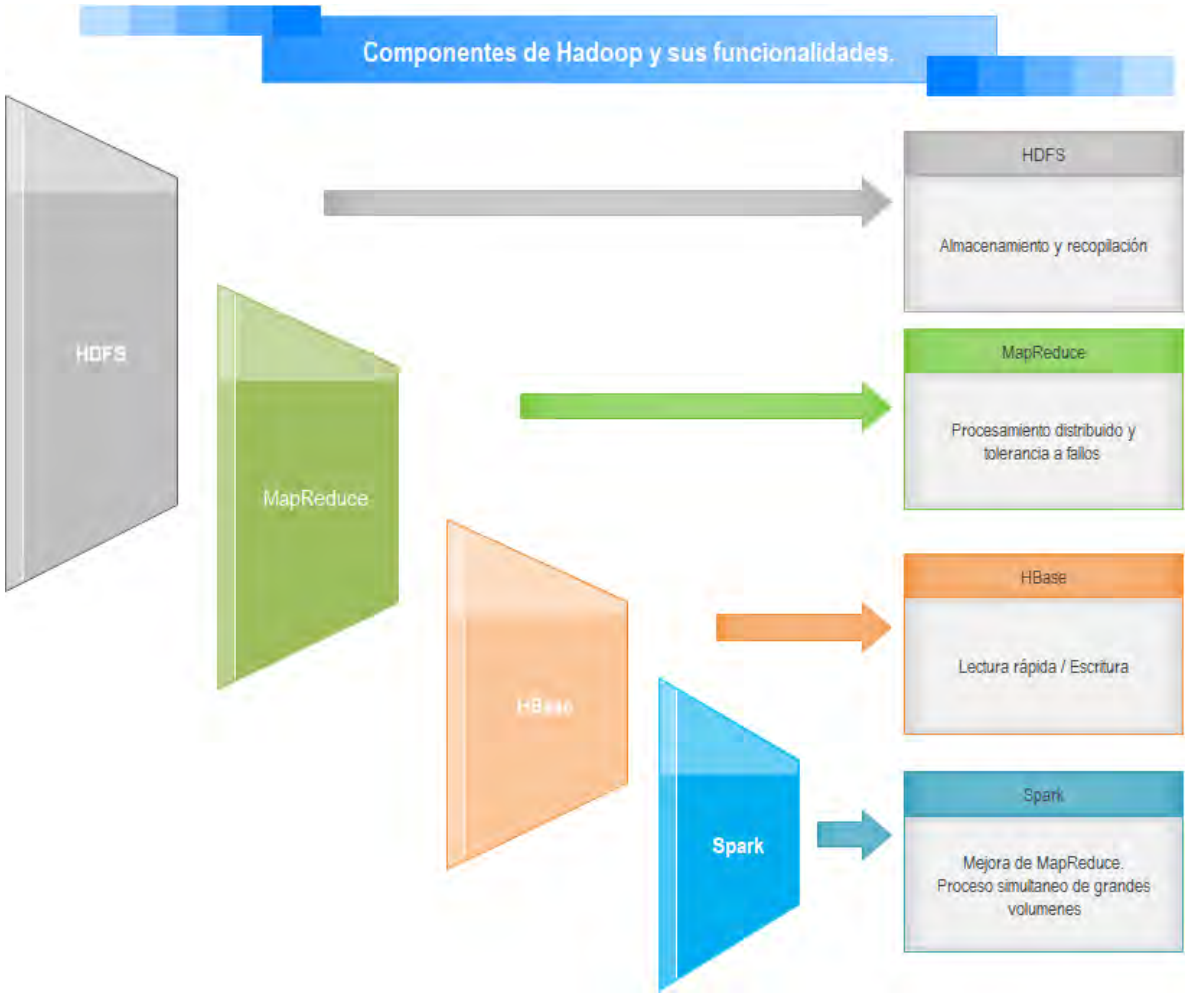


Figura 5.1 Componentes de Hadoop

5.2 Hadoop

La herramienta Hadoop es un entorno de desarrollo que permite almacenar, procesar y analizar grandes cantidades de datos, también definida como una librería de apache denominado framework. Fue creada con el propósito de responder a las necesidades de implementación de Big Data. Una de sus principales características es que es un software de código abierto. Es escalable, tolerable a fallas y es distribuido [16], desde un par de servidores alcanzando diversas maquinas o nodos, mismos que manejan almacenamiento y procesamiento de manera local [1]. Esta herramienta ha sido diseñada para evitar el bajo rendimiento y la complejidad encontrada cuando el procesamiento y el análisis de grandes volúmenes de datos se encuentran utilizando tecnología

tradicional [6]. La mayoría de estas herramientas son partes de apache y se construyen alrededor de la famosa Hadoop. Escrito en Java y creada por Doug Cutting. Hadoop trae la capacidad de procesar grandes cantidades de datos, independientemente de su estructura.

5.3 Características Hadoop

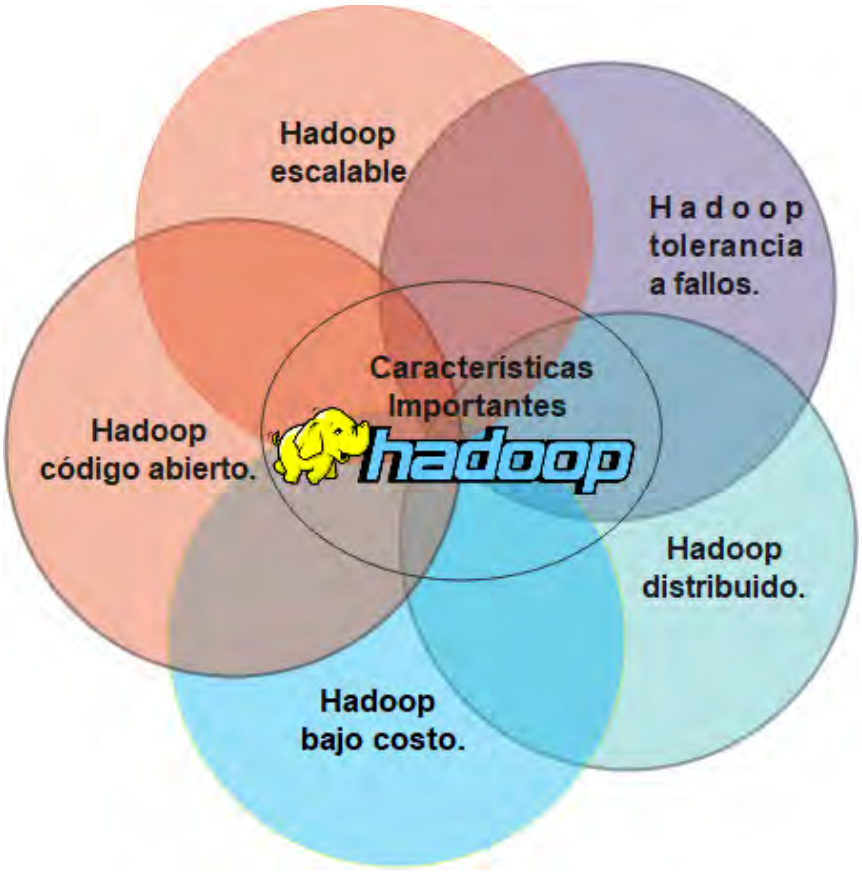


Figura 5.2 Características importantes Hadoop

5.3.1 Hadoop escalable

La escalabilidad masiva da como significado que no hay límites en la cantidad de datos, la capacidad, el volumen y los nodos en que se da el procesamiento [15]. Se encuentra diseñado con la finalidad de escanear a través de grandes conjuntos de datos para producir sus resultados a través de un lote distribuido altamente escalable [14]. Hadoop ha logrado cumplir con las expectativas sobre el procesamiento de los grandes volúmenes de datos por hacer posible lo que se creía imposible mediante la escalabilidad, y esto sucede tan solo añadiendo

nuevos nodos esclavos con la finalidad de almacenar más datos e incrementar la capacidad de procesamiento [16].

5.3.2 Hadoop tolerancia a fallos

La tolerancia a fallos sucede cuando uno de los nodos falla, este no afecta al sistema y continúa funcionando. El trabajo que realiza es la detección de fallo y a su vez tiene la capacidad de adaptarse a este teniendo como objetivo continuar con el servicio mientras se soluciona el problema del fallo reasignando el nodo maestro a otro, es decir se redistribuyen los trabajos. Se almacenan múltiples copias de todos los datos de manera automática [15.]El almacenamiento se da de manera redundante en el clúster, permitiendo así que la información siga disponible sin perdidas ni retrasos [16]. El procesamiento de datos y aplicaciones está protegido contra fallos del hardware.

5.3.3 Hadoop distribuido

Es el funcionamiento coordinado en un clúster mientras se ejecuta, es decir, puede trabajar los datos con gran rapidez, procesando de manera inmediata las enormes cantidades de tipos de datos [16]. Es un clúster de ordenadores que se encuentran conectados entre ellos de forma coordinada, permitiendo que el usuario no se preocupe por decidir en qué ordenador se ejecutarán, debido a que el clúster se encarga de las tareas de manera planificada y controlada por software [15].

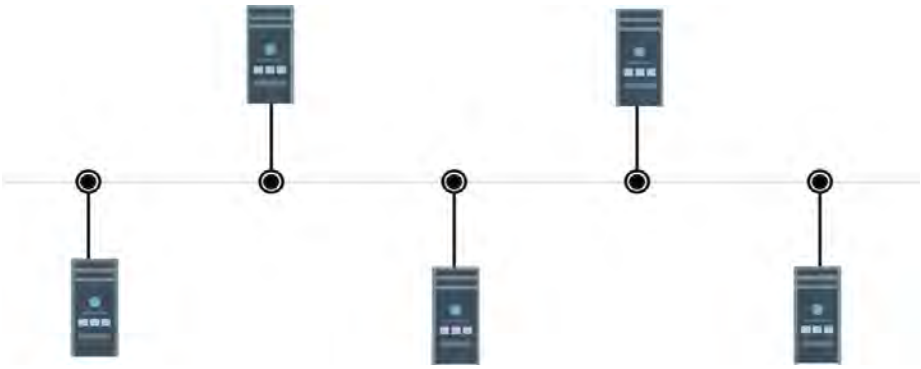


Figura 5.3 Clúster de Ordenadores

5.3.4 Hadoop código abierto

Hadoop *open source* tiene la ventaja que se encuentra disponible como código abierto permitiendo así crear modificaciones, utiliza modelos sencillos de programación. También es conocido como open source siendo un software de libre distribución, esta característica ha tenido mayor índice de crecimiento y esto ha logrado no solo que este código llegue a muchos usuarios si no también que la calidad sea mucho mejor que otros que no son de código abierto. Se pueden mencionar algunos de los más conocidos como desarrollos de software open source el sistema operativo para smartphones Android, el navegador de Internet Mozilla Firefox, etc.

Figura 5.4 Desarrollos de software Open Source

5.3.5 Hadoop bajo costo

Con la ventaja de que el código abierto es gratuito y el empleo en el hardware se da para el almacenamiento de las grandes cantidades de datos. Al ser un software de código abierto que habilita el procesamiento distribuido y el almacenamiento de grandes conjuntos de datos, Hadoop tiene la capacidad de aumentar la escala de un solo servidor a miles de servidores, siento esta una manera de adaptación a los cambios en base a la demanda [13].

Hadoop se toma como base y cuando se fusiona con otra tecnología o herramienta, se potencializan sus características permitiendo brindar un gran

poder de escalamiento [14]. Hadoop se compone de dos proyectos principales: Sistema de archivos distribuidos HDFS y MapReduce [4].

5.4 Hadoop Distributed File System (HDFS)

Hadoop Distributed File System conocido también por sus siglas HDFS el cual tiene la tarea de funcionar como un sistema de archivos que se encarga de recopilar toda la información posible, definiéndolo como un sistema de archivos distribuido, capaz de ser escalable y portátil, escrito en Java para el framework Hadoop. Este framework permite distribuir los archivos a diversas máquinas y es por eso que podemos decir que HDFS se aplica cuando la cantidad de datos es demasiada para una sola máquina [7].

La interfaz de HDFS sigue el modelo de sistemas de archivos basados en la UNIX, en donde se intercambia alguna Interfaz Portable del Sistema Operativo (POSIX)como requisitos para el desempeño [4] y es uno de los principales componentes de Hadoop, permite crear diferentes sistemas de archivos lo que permite tener replicas, mayor capacidad y rendimiento [16]. Una de las razones por la que fue diseñado es para el procesamiento de operaciones de alta latencia lote y así como manejar los datos estructurados y no estructurados y mantener grandes volúmenes, los archivos que se almacenan pueden ser más grande que un terabyte [6].

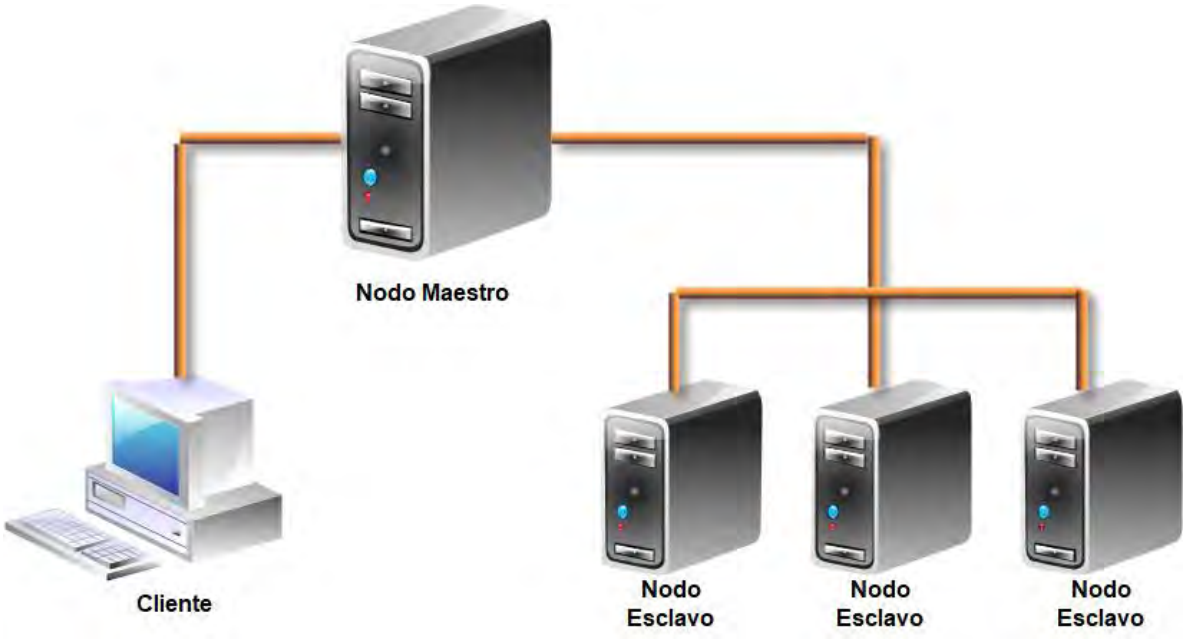


Figura 5.5 Representación del funcionamiento de la arquitectura de HDFS

5.4.1 Componentes que caracteriza a HDFS

HDFS se caracteriza por ser el más complejo que otros sistemas de archivos, debido a las complejidades e incertidumbres de las redes, se conforma por un Clúster y se divide en dos tipos de nodos:

- Nodo maestro: Se replican en múltiplos para facilitar el procesamiento en paralelo de grandes cantidades de datos.

- Nodo de datos: Este actúa como nodo esclavo y almacena los archivos en bloques, el tamaño de bloque por defecto de los cuales es de 64 MB [7].

HDFS se encuentra sostenido por tres componentes en su estructura que a continuación se describirá en la siguiente imagen:

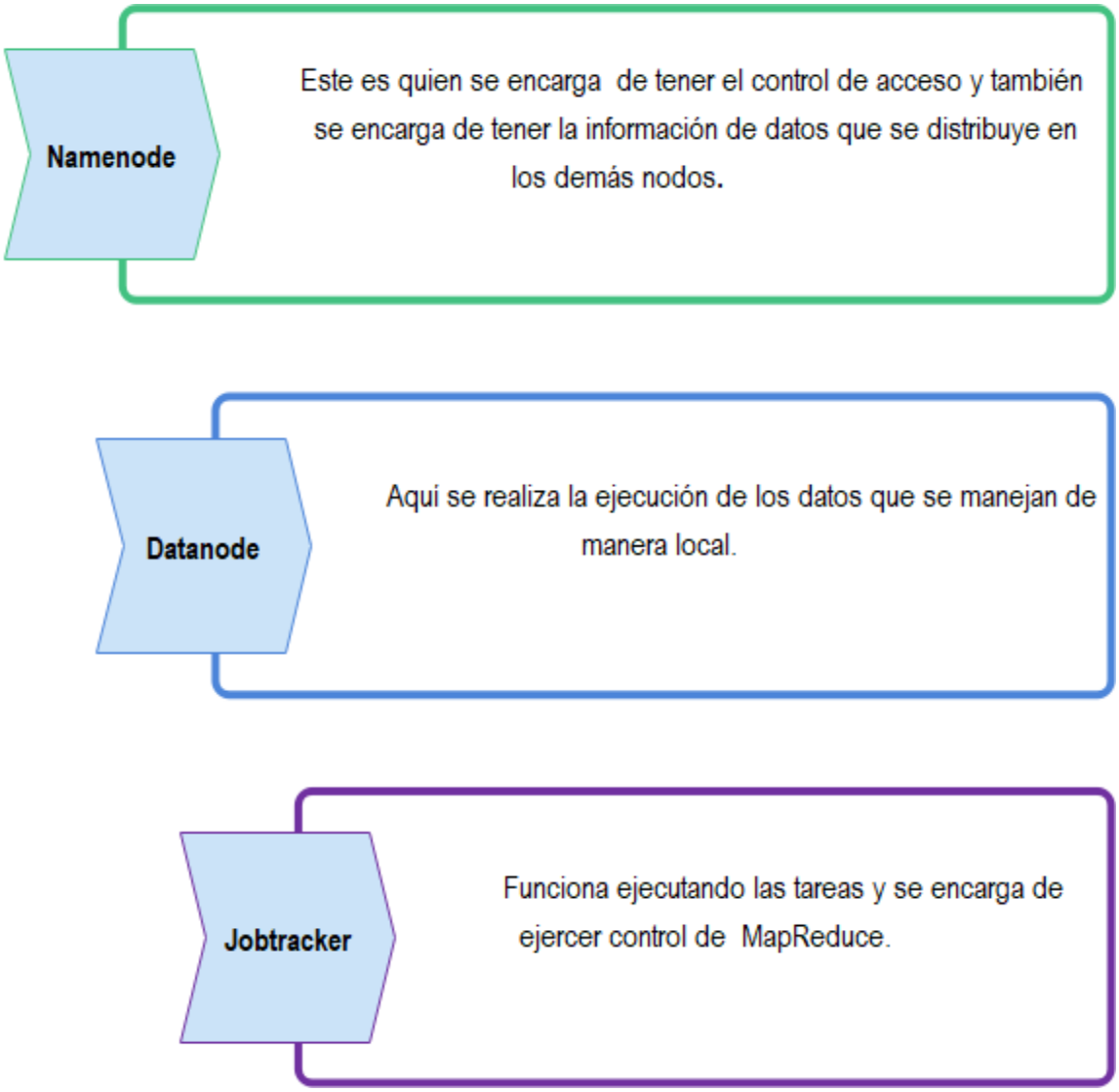


Figura 5.6 Componentes que conforman la estructura HDFS

Así mismo se puede mencionar que HDFS cuenta con algunas particularidades que lo destacan como características fundamentales en su funcionamiento:

- La tolerancia a fallos
- Permite el acceso a datos Streaming
- Facilita el trabajo
- Cuenta con un modelo sencillo coherente

HDFS se basa en la rapidez para recuperarse de las fallas que se puedan presentar en el hardware, sabiendo que una instancia de HDFS de acuerdo a

IBM puede tener miles de servidores y si alguno falla es algo que no se puede evitar, por ello HDFS se creó con la finalidad de detectar las fallas y así responder de manera rápida y eficaz, permite el acceso mientras se encuentra transmitiendo los datos, está encargado de procesar estos mismo por lotes, encargándose así del alojamiento de estos grandes volúmenes de datos para las aplicaciones que cuentan con una cantidad demasiado grande de datos estimados en de gigabytes a terabytes, HDFS se caracteriza por el ancho de banda de los datos que se agregan y por ser escalable de una gran cantidad de nodos en un solo clúster [16].

Su diseño le permite ser portable, es decir, le facilita moverse y transportarse en diversas plataformas de hardware permitiendo la compatibilidad con las variedades de sistemas operativos que se encuentran por debajo de este.

5.5 Mapa reducido (MapReduce)

Mapa reducido es hoy en día el principal modelo de programación y aplicación asociada para el procesamiento y la generación de grandes conjuntos de datos [4]. El paradigma MapReduce es un modelo de programación que permite el procesamiento de grandes volúmenes de datos en forma paralela, facilitando el manejo tolerable de errores en la manipulación de datos masivos lo que a su vez también permite de la forma más sencilla que diferentes procesos trabajen simultáneamente e interactúen entre sí y tiene la capacidad de dividir una petición por parte de un cliente en otros muchas partes y encargar el trabajo a múltiples nodos que funcionan en paralelo. Se puede decir entonces que, en la arquitectura de Hadoop, MapReduce es el principal encargado de gestionar los recursos y procesamientos de datos [16].

MapReduce tienen un enfoque que se muestra como base sólida de las soluciones Big Data, sobre todo desde el modelo de distribución de procesamiento donde se pueden hacer frente a los problemas de tratamiento de grandes volúmenes de datos que las herramientas tradicionales no son capaces de soportar [23].

5.5.1 Fases de MapReduce

MapReduce es un componente que tiene la capacidad de aislar al desarrollador de la programación paralela y consta de diversas fases que se describen a continuación [1]:

- Map: como su nombre lo indica es quien se encarga básicamente del mapeo de la información que entra, operando en un único bloque de HDFS ejecutándose siempre que le sea permitido que se almacene el bloque, minimizando el tráfico generado por la red.
- Reduce: se encarga de revisar que la información que se recibe sea procesada, misma que se mapea en primera instancia, es decir, se encarga de recolectar las respuestas de las subtareas en cada uno de los subnodos, combinando y agrupando así la respuesta final.

Estas fases tienen una etapa interna que permite la complementación para el funcionamiento y se le conoce como “shuffle and sort”, encargada de realizar un ordenamiento por clave de aquellos resultados que se emiten al mapear.

5.6 Apache Spark

Tiene un propósito general y se encarga del procesamiento de programación de datos distribuidos es un framework diseñado para ser rápido. Si bien como su nombre lo indica forma parte de los proyectos Apache, el cual se sabe que trabaja bajo la licencia de Open Source.

Apache Spark se encuentra formado por un clúster el cual es un motor de procesamiento distribuido llamado datos distribuidos resistentes (RDD) que se encarga de realizar operaciones como orquestar, distribuir y monitorizar sobre múltiples elementos de procesamiento de datos sobre varias máquinas de trabajo[25]. Es una versión mejorada de MapReduce, pues aprovecha el procesamiento simultaneo de grandes volúmenes de datos Spark, es un motor de cálculo rápido y general para datos de Hadoop. Spark proporciona un

modelo de programación simple y expresivo que admite una amplia gama de aplicaciones, que incluyen ETL, aprendizaje automático, procesamiento de secuencias y cálculo de gráficos [16]. El sistema base de computación de Apache Spark se encuentra basado en Hadoop MapReduce el cual tiene como función principal, dividir o trabajar de manera paralela, debido a que su instalación se da en un clúster de máquina.

Para entender un poco el funcionamiento de un clúster de maquina podemos dar como ejemplo que tenemos 20 máquinas trabajando con una versión de Apache Spark si se tiene un fichero muy grande en el cual se requiere procesar una gran cantidad de datos, este se puede dividir en 20 partes y cada una de las maquinas instanciadas con Apache Spark se encargara de una parte del fichero para tener como resultado final la unión, esto permite ganar velocidad y como sabemos para Big Data la velocidad es clave. Spark trabaja en conjunto con cualquier fichero que se encuentre alojado en HDFS o cualquier otro sistema que esté usando Hadoop, y así de esta manera permite tener ambos instalados a la vez [23].

Las principales características de Apache Spark son las siguientes:

- Su integración está a base de Apache Hadoop
- Mayor velocidad de procesamiento debido a que trabaja en la memoria
- Permite el trabajo en disco
- Brinda API para java, Scala y Python
- Procesos en tiempo real de los datos debido a que cuenta con un módulo llamado Spark Streaming trabajando en conjunto con Spark SQL.

5.6.1 Componentes de Spark

Se ha dado a conocer en Big Data gracias a su rendimiento y su gran variedad de librerías, por lo cual se ha caracterizado como una de las plataformas más reconocida, siendo prácticamente un todo en uno.

Este framework está conformado por los siguientes componentes principales:

- **Spark Core:** Parte central de Spark, conformado por el conjunto de librerías. Este se encarga de gestionar las funciones, es decir, del hecho de programar las tareas.
- **Spark SQL:** Está ubicado en la parte superior de la parte central de Spark. En este módulo se procesan los datos que se encuentran estructurados y semiestructurados, introducidos por los datos llamados RDD o dataframes.
- **Spark Streaming:** Es el procesamiento de los datos en tiempo real debido a la velocidad de la programación de la parte central de Spark.
- **Spark MLib:** Está compuesto por el módulo de librerías, demasiado completa que en ella se contienen gran cantidad de algoritmos machine learnig, este framework de aprendizaje rápido permite hacerlo práctico y escalable.
- **Spark Graph:** Está basado en un entorno de procesos gráficos denominado grafos DAG, permitiendo proporcionar API exclusivos para gráficos y de los cálculos de grafico en paralelo [23].

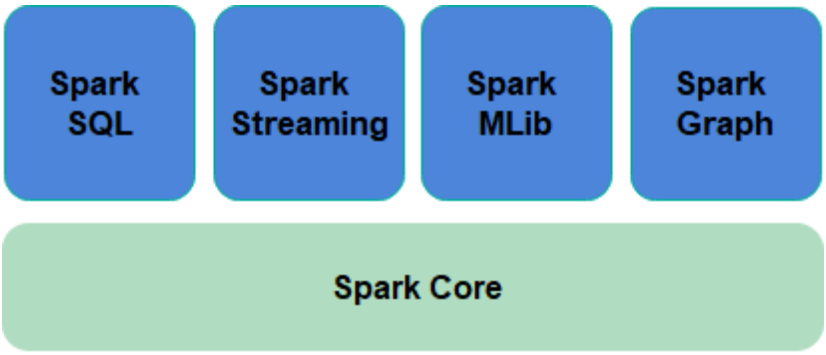


Figura 5.7 Componentes principales Spark

5.7 Apache Hbase

Es una base de datos no relacional distribuida. Es un proyecto de código abierto que se construye en la parte superior de HDFS. Está diseñado para operaciones de baja latencia. HBase se basa en el modelo de datos clave / valor, es un sistema de gestión de base de datos orientada a columnas distribuido [4].

HBase es más flexible que las bases de datos relacionales, tiene el potencial para soportar las altas tasas de actualización de tabla y para escalar horizontalmente en grupos distribuidos [6]. Un ejemplo de ello es cuando lee y escribe operaciones que implican todas las filas, pero tomando en cuenta solo un subconjunto de todas las columnas [7]. Es una base de datos distribuida escalable que admite el almacenamiento de datos estructurados para tablas grandes.

CAPITULO 6
APLICACIONES DE BIG DATA

CAPITULO 6: APLICACIONES DE BIG DATA

Big data trabaja con gran volumen de información, dicha información proviene de manera principal de las empresas en tanto la inteligencia de negocios y el procesamiento analítico en línea predecesores de las aplicaciones Big Data. Existen múltiples herramientas, algunas ya mencionadas con anterioridad como lo son, Hadoop, Spark y NoSql. Estos son utilizados en función de los datos analizados, como Big Data en marketing y ventas, seguridad, deportes, política, telecomunicaciones y salud.

6.1 Marketing y ventas

Las aplicaciones de Big Data en las empresas hablando particularmente en el área de marketing y con ayuda del análisis de correlación, Big Data permite predecir con una mayor precisión la manera en que sus consumidores se comportan, las técnicas utilizadas para ello son la minería de datos, el aprendizaje automático y el procesamiento de lenguaje natural; permite a los departamentos de marketing segmentar a los clientes de acuerdo con sus preferencias [9].

Como ejemplo se puede mencionar a Spotify el cual ofrece servicio de música, podcast y videos digitales en Streaming, permitiendo tener acceso a millones de canciones y del contenido de los artistas. El éxito que ha tenido esta empresa es gracias a la capacidad y destreza de la computación, es decir, que el principal negocio de Spotify es Big Data, gracias a los datos triangulados de los clientes. Este servicio trabaja aprendiendo de lo que el cliente va eligiendo como por ejemplo el tipo de música que escucha, genero, toma en cuenta el geoposicionamiento y la hora, como datos que le sirven para entender el comportamiento del cliente y así brindarle opciones y lo servicios que requiere [9].Otro caso muy común y utilizado en la actualidad es la plataforma de Streaming Netflix que permite ver series y películas en un dispositivo con conexión a internet el cual trabaja bajo el análisis predictivo utilizando Hadoop para recolectar lo que le gusta e interés a los usuarios [17].

6.2 Seguridad

Todo aquel dispositivo que se encuentre conectado o almacenado en la red son vulnerables a los ataques. Big Data se encarga del acceso a los datos, la organización y elegir los datos de más valor entregándolos y procesándolos. Las empresas se encuentran muchas veces con estos inconvenientes para aterrizar proyectos que precisen de análisis de los datos, también requieren encontrar una herramienta que se encargue de ello de manera segura y sea en tiempo real. En términos generales definimos que seguridad de datos hace referencia a la protección de la privacidad digital utilizado para evitar el acceso a intrusos o accesos no autorizados a los datos. La recolección de los datos y la utilización de herramientas analíticas que tienen la finalidad de extraer información plantea varias preocupaciones de privacidad [19]. Existen algunas herramientas que ayudan a mitigar las vulnerabilidades como por ejemplo Panda Adaptive Defense 360, esta es una empresa especializada en la creación de productos de seguridad, el cual se encarga de ofrecer soluciones que se basan en las técnicas de aprendizaje automático, teniendo como finalidad clasificar todo lo que ocurre en los sistemas de forma más efectiva, permitiendo detectar y bloquear procesos maliciosos, fugas de información, vulnerabilidades, así como solucionar el daño ocasionado por alguna brecha de seguridad [19].

6.3 Deportes

Tal vez surge la pregunta de cómo se relaciona Big Data con los deportes, es simple, existen aplicaciones que permiten definir las estrategias que se utilizaran en cada partido, prevenir futuras lesiones de los jugadores y así como también se puede llegar a conocer las preferencias de los aficionados.

La NFL es un ejemplo que emplean aplicaciones con ayuda de una plataforma, tiene una base de datos con gran cantidad de datos que manejan para la toma de decisiones, en el deporte americano la mentalidad permite adquirir

conocimiento en la recopilación y análisis de datos, distinguiéndose de los demás deportes, se toman a la tarea de crear una posibilidad de contar la inmensa cantidad de datos de cada uno de los jugadores y de los equipos con los que compiten [20].

6.4 Política

Big Data ha beneficiado a la comunicación política a través de la protección del comportamiento electoral, extracción, gustos, preferencias, intereses de la audiencia con la finalidad de realizar programas políticos que satisfagan las necesidades, organización de campañas electorales, así como también la interacción de partidos políticos y la población que vota [21].

6.5 Telecomunicaciones

Aquí entra la labor de la analítica de Big Data que como se mencionó anteriormente se encuentra los datos que se generan en la web, tecnología M2M en donde los dispositivos se conectan a otros dispositivos gracias al uso de sensores, las transacciones de los datos y los usuarios en base al operador generan una gran cantidad de datos estructurados.

6.6 Salud

El Big Data es útil en esta área para gestionar de manera más administrada y eficiente todo lo relacionado con la salud, específicamente en el campo de la medicina, el gran volumen de datos como los historiales médicos, las predicciones de los reingresos al hospital, las imágenes médicas, los datos de ensayos clínicos, así como el material genético. De manera exponencial han crecido. Los avances que se han tenido para gestionarlos se han visto involucrados la virtualización y el Cloud Computing, los cuales están dando la facilidad con el desarrollo de plataformas que permiten hacer más eficaz la captura, almacenamiento y manejo de los grandes volúmenes de datos [21].

Las aplicaciones de Big Data se encargan de aprovechar los datos para brindar mejoras en los negocios, así como también ayuda a tomar decisiones en esos datos para hacer que los clientes permanezcan y les sea rentable.

Se ha tenido en la actualidad un gran avance tecnológico así como el Internet de la cosas, aquellas fuentes que generan datos públicos y de los medios sociales brindando el acceso a una enorme cantidad de datos que se pueden explorar, he aquí es también donde se ve involucrado el análisis de los datos, si partimos de que el análisis de Big Data es capaz de proporcionar valores que son muy útiles mediante decisiones, sugerencias y/o apoyos, esto implica que para ello se requiere de una amplia gama de aplicaciones que son cambiantes y complejas.

Conclusión

Este trabajo dio a conocer los conceptos básicos, las tecnologías y las aplicaciones en Big Data. Sin duda alguna la nueva era y la transformación de los datos han llevado a buscar nuevas tecnologías para el manejo del gran volumen de datos que se generan día con día y con este aumento de datos, los sistemas de Big Data y, en particular, las herramientas analíticas, se han convertido en una fuerza importante de innovación que proporciona una forma de almacenar, procesar y obtener información sobre conjuntos de datos de petabytes. Los entornos en la nube aprovechan fuertemente las soluciones de Big Data al proporcionar entornos tolerantes a fallas, escalables y disponibles para los sistemas de Big Data. Las empresas adoptan trabajar con la inteligencia de negocios (BI) la toma de decisiones.

Big Data ha llegado para quedarse día a día se crean más y más información proveniente de diferentes lugares como las redes sociales, smartphones, las empresas las comunicaciones M2M, los sensores a través de las aplicaciones con ayuda de múltiples herramientas, los cuales se encargan de los grandes volúmenes de datos.

Se dio a conocer Big Data no únicamente como una tecnología que se encarga de obtener grandes cantidades de datos, sino también para analizar esos enormes volúmenes de datos y así conseguir información, darle valor y transformarla en conocimiento.

Referencias

[1] Yunhao, L., Chen, m., & Mao, S. (2014). *Big Data: A Survey.* Springer Science + Business Media Nueva York.

[2] Chen, C. P., & Zhang, C.-Y. (2014). Intensivas de datos de aplicaciones, retos, técnicas y tecnologías. (W. Pedrycz, Ed.) *ELSEVIER*.

[3] Ishikiriyama, C. S., & C. F. (2019). Big Data: Un panorama global. (A. E. Charles, Ed.) *Springer International Publishing AG*.

[4] Kacfah Emani, C., Cullot, N., & Nic, C. (2015). Comprensible Big Data: Una encuesta. (J. Nešetril, Ed.) *ELSEVIER*.

[5] (2020). *Oracle México*. Recuperado el Agosto de 2020, de https://www.oracle.com/mx/big-data/guide/what-is-big-data.html

[6] Ahmed Oussous, F.-Z. B. (October de 2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*.

[7] Fragoso, R. B. (2012). *IBM Developer*. Recuperado el Septiembre de 2020, de https://developer.ibm.com/es/articles/que-es-big-data/

[8] Khan, N., Yaqoob, I., Targio Hashem, I. A., Inayat, Z., Mahmoud Ali, W. K., Alam, M., . . . Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. (J.-R. Lee, Ed.) *The Scientific World Journal*.

[9] Andrea Sestino, M. I. (2020). *Internet of Things and Big Data as enablers for business digitalization strategies.* Elsevier. doi:https://doi.org/10.1016/j.technovation.2020.102173

[10] Zikopoulos, P. C., C. E., D. d., T. D., & G. L. (s.f.). *Comprensión de Big Data: análisis para Enterprise Class Hadoop y Streaming Data.* (S. Sit, Ed.) McGraw-Hill.

[11] WenhongTian, & YongZhao. (2015). Big Data Technologies and Cloud Computing. En W. Tian, *Optimized Cloud Resource Management and Scheduling* (págs. 17-49). Morgan Kaufmann.

[12] Watson, H. J. (2014). Tutorial: BigData Analytics: conceptos,tecnologías y aplicaciones. En *Comunicaciones de la Asociación de Sistemas de Información.* Fred Niederman.

[13] Hernández-Leal, E. J., Duque-Méndez, N. D., & Moreno-Cadavid, J. (2017). Big Data: an exploration of research, technologies and application cases. *TecnoLógicas, 20*.

[14] Hernández-Leal, E. J., Duque-Méndez, N. D., & Moreno-Cadavid, J. (2017). Big Data: an exploration of research, technologies and application cases. *TecnoLógicas, 20*.

[15] McCool, R. (s.f.). *Apache Hadoop*. Obtenido de https://hadoop.apache.org/

[16] Cano, J. L. (2007). *Business intelligence: competir con información.* (F. C. Banesto, Ed.) Obtenido de https://books.google.com.pe/books/about/Business_intelligence.html?id=g8A7QwAACAA

[17] School., E. B. (Octubre 2018). Apache Spark: Introducción, qué es y cómo funciona. *Spark*. Obtenido de https://www.esic.edu/rethink/tecnologia/apache-spark-introduccion-que-es-y-como-funciona

[18] Neves, P. C., B. S., J. B., & J. C. (2016). Big Data in Cloud Computing: features and issues. *Proceedings of the International Conference on Internet of Things and Big Data, Volumen 1*, pag. 307-314. doi:10.5220/0005846303070314

[19] C.L. Philip Chen, C.-Y. Z. (2014). *Data-intensive applications, challenges, techniques.* Elsevier. Obtenido de https://doi.org/10.1016/j.ins.2014.01.015

[20] Hrushikesha Mohanty, P. B. (2015). *Big Data a primer.* India: Springer. doi:DOI 10.1007/978-81-322-2494-5

[21] McCool, R. (2020). *Apache Hadoop*. Obtenido de Apache Hadoop project: https://hadoop.apache.org/

[22] School., E. B. (Octubre 2018). Apache Spark: Introducción, qué es y cómo funciona. *Spark*. Obtenido de https://www.esic.edu/rethink/tecnologia/apache-spark-introduccion-que-es-y-como-funciona

[23] Hrushikesha Mohanty, P. B. (2015). *Big Data a primer.* India: Springer. doi:DOI 10.1007/978-81-322-2494-5

[24] Min Chen, S. M. (s.f.). *Big Data Relacionado,Tecnologias,Desafios y Perpectivas del futuro.* Nueva York: Springer Cham Heidelberg.

[25] Argonza, J. S. (2016). Big data en la educación. *Revista Digital Universitaria*. Obtenido de http://www.revista.unam.mx/vol.17/num1/art06/

[26] Dean, J. (2004). MapReduce: simpli-fied data processing on large clusters. *Operating System Design and Implementation*.

[27] Pacheco, J. C., Garda, J. A., & Yañez, S. R. (2018). *Cloud Computing para PYMEs.* Ecuador: UTMACH. Obtenido de www.utmachala.edu.ec

[28] Cox, W. G. (2013). *Big Data Storage for DUMMIES* (EMC Isilon ed.). (L. John Wiley & Sons, Ed.) Inglaterra: Page Bros. Obtenido de www.customdummies.com