Universidad Tecnológica de Tula-Tepeji

PROCESADOR DE TEXTO

Docente: Cristian Emmanuel Pérez López

FUNDAMENTOS DE TI

Alumno: Gabriel Eduardo De Jesús Ortiz

17 de septiembre de 2022

INTRODUCCION

Big data hace referencia a las combinaciones de datos cuyo tamaño, complejidad y velocidad dificultan la captura, gestión, procesamiento y análisis por parte de las herramientas de software tradicionales. El estudio de Big data se hace complejo debido principalmente a que la mayoría de los datos generados por las tecnologías modernas, como los web logs, la maquinaria, los vehículos, las búsquedas en Internet, las redes sociales como Facebook, computadoras portátiles, teléfonos inteligentes entre muchas otras, son datos no estructurados. El objetivo de Big data es convertir el dato en información, para facilitar su análisis y así aprovechar su contenido para ayudar en la toma de decisiones, mejorar el marketing de las empresas y otras ventajas que se obtiene cuando se procesan los datos de la manera correcta; Actualmente el análisis de los datos se ha convertido en una ayuda para muchas empresas en la toma de decisiones. El big data es una herramienta moderna que requiere mucha atención, es por esto que se hace necesario elaborar un manual práctico para su aprendizaje y que este brinde un apoyo fundamental en el estudio de lo que es en realidad y de lo que se puede hacer con el Big data.

1. GENERALIDADES

1.2 Planteamiento Del Problema

Big data ha surgido como un área de estudio importante tanto para los profesionales como para los investigadores y es gracias a la tecnología moderna que su desarrollo e investigación se hace más importante y es de vital importancia que los ingenieros de sistemas o estudiantes interesados tengan conocimientos sobre él, ya que cada vez más las investigaciones tecnológicas combinan Big data y analistas con la nube, internet de las cosas, redes sociales, movilidad, etc….. Y es gracias a esto que muchas empresas han visto una oportunidad de negocio en el análisis de datos y han decidido implementarlo en sus operaciones de comercio; los documentos informativos sobre este tema son muy variados, confusos y la gran mayoría de artículos e investigaciones se encuentran en otros idiomas, las implementaciones que se han hecho son aisladas y no hay un documento claro que brinde la información de cómo se debe integrar el Big data utilizando las herramientas open source.

1.2 Objetivos

1.2.1 Objetivo General

Realizar un manual práctico que facilite el proceso de enseñanza-aprendizaje de lo que es Big data, de cómo debe ser implementado y de la importancia que tiene en la actualidad.

1.2.2 Objetivos específicos

- Identificar y analizar cuáles son las tecnologías requeridas, herramientas de software y los hardware necesarios para la correcta implementación de Big data.
- Configurar e integrar las herramientas para la elaboración del ambiente de Big data
- Elaborar el caso de estudio el cual se va a utilizar para realizar el testeo en el ambiente configurado.

1.3 Justificación

En la actualidad, con el auge de redes sociales, el internet de las cosas. El crecimiento de la población, entre otros muchos factores, han provocado gran crecimiento de los datos, por lo cual, las empresas y organizaciones han tenido que enfrentarse a nuevos retos que les permitan descubrir, analizar y sobretodo, entender el funcionamiento de herramientas no tan tradicionales. Es por esto, que resulta adecuado e interesante profundizar en este tema y proponer un manual práctico que permita entender lo que es Big Data, lo que significa y su importancia actualmente. la información que se genera cada segundo y que circula en las plataformas web puede ser vista por muchas empresas como una oportunidad de negocio, si esta información se traduce en cifras quien las utilice podrá detectar tendencias de mercadeo, orientar acciones que se van a llevar a cabo mediante la toma de decisiones entre muchas otras oportunidades que pueden ser provechosas si se aplica Big Data. La elaboración de este manual sobre Big Data brinda una ventaja en el desarrollo profesional a todas las personas que están involucradas en tecnologías de información, Ingenieros de Sistemas, científicos de datos, analistas, directores de TI que vean en Big Data un elemento competitivo y que deseen tener nuevas estrategias en sus negocios, ya que se podrá conocer más detalladamente de que se trata este tema y de cómo puede ser implementado

2. MARCO TEORICO

2.1 ¿Qué Es Big Data?

Desde su origen han existido diversas definiciones y explicaciones de lo que se significa Big Data; IBM que es una de las empresas más importantes a nivel mundial sobre tecnología define a Big Data como: “la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales”. Oracle compañía de software especializada en el desarrollo de aplicaciones locales o en la nube que aporta soluciones a muchas empresas da la siguiente definición: “Big Data describe una estrategia holística de gestión de la información que incluye e integra muchos nuevos tipos de datos y de gestión de datos junto con datos tradicionales”. Entonces, podemos denominar Big Data como el análisis y gestión de

grandes volúmenes de datos los cuales no pueden ser tratados de la manera convencional, y los cuales deben cumplir con la ley de las 4V´s del Big Data.

2.2 4v's De Big Data

Volumen: hace referencia a la cantidad de los datos que se generan por segundo, los datos son generados automáticamente por máquinas, redes e interacciones personales en sistemas como redes sociales; datos de valores desconocidos, como mensajes de Twitter, flujos de clics en páginas web, aplicaciones móviles, tráfico de red, equipos con sensores que capturan datos a la velocidad de la luz, etc.

El volumen delimita el concepto de datos masivos ya que no se pueden almacenar en ordenadores simples por el contrario requieren de una tecnología específica para el almacenamiento.

- Velocidad: se refiere al ritmo con el que los datos fluyen de fuentes como procesos de negocios, máquinas, redes e interacción humana con cosas como redes sociales, dispositivos móviles, etc. El flujo de datos es masivo y continuo. Estos datos en tiempo real pueden ayudar a los investigadores y las empresas a tomar decisiones valiosas que brindan ventajas competitivas estratégicas y retorno de la inversión si puede manejar la velocidad. Para aprovechar al máximo el significado de los datos su procesamiento debe realizase en tiempo real o en el menor tiempo posible. Para mejorar el análisis y la extracción de conclusiones se requiere una velocidad para acceder o visualizar los datos.

- Variedad: se refiere a las diferentes formas, fuentes y tipos que tienen los datos tanto estructurados como no estructurados o semiestructurados entre los que se incluye documentos de texto, audios, videos, emails, fotos, videos, sistemas de monitorización, PDF's, ficheros de sonido etc.

- Veracidad: La veracidad es cuán exacto o verdadero puede ser un conjunto de datos. Pero en el contexto de Big Data, la definición de veracidad adquiere un poco más de significado. Cuando se trata de la precisión del Big Data, no solo se trata de la calidad de los datos en sí, sino también de la confiabilidad de la fuente de datos, el tipo y el procesamiento. Eliminar aspectos como los prejuicios, las 11 anomalías o incoherencias, la duplicación y la volatilidad son solo algunos de los aspectos que contribuyen a mejorar la precisión de los grandes datos. La veracidad de los datos implica asegurar que el método de procesamiento de los datos reales tenga sentido en función de las necesidades del negocio y que el resultado sea pertinente a los objetivos. La interpretación de grandes datos de la manera correcta garantiza que los resultados sean relevantes y procesables.

2.3 Historia Y Evolución De Big Data

El nombre de Big Data es un nombre novedoso y el cual ha tenido un auge muy importante en esta era de la tecnología, pero su concepto ha sido implementado muchos años atrás. En 1880 se realiza un censo en los Estados Unidos de América, censo que tardo 8 años en tabularse, está sobre carga de información como fue denominada, fue fundamental para que se enfocaran en la importancia que tiene el tratamiento de la información y de la necesidad de desarrollar avances en la metodología para el tratamiento de los datos. Herman Hollerith desarrollo una maquina capaz de tomar la información depositada en tarjetas perforadas y analizarlos; la máquina de Hollerith como fue nombrada implemento un sistema que revoluciono el valor de los datos y disminuyo el tiempo de análisis de estos. La primera máquina de procesamiento de datos apareció en 1943 y fue desarrollada por los británicos para descifrar los códigos nazis durante la Segunda Guerra Mundial. Este dispositivo, llamado Colossus buscaba patrones en mensajes interceptados a una velocidad de 5.000 caracteres por segundo. De ese modo, se reduce la tarea de semanas enteras a solo unas pocas horas. En 1965 El gobierno de los Estados Unidos planea el primer centro de datos del mundo para almacenar 742 millones de declaraciones de impuestos y 175 millones de juegos de huellas dactilares en cinta magnética.

En la década de los 70 el análisis de los datos empieza hacer prioridad para las predicciones y la toma de decisiones, el modelo Black-Sholes que se crea en 1973 y su propósito era poder predecir el precio óptimo de las acciones en el futuro.

En el año 1999, aparece el término Big Data en Visually Exploring Gigabyte Datasets in Real Time, publicado por la Association for Computing Machinery. En esta publicación se describe el problema que se genera al almacenar grandes cantidades de datos sin una forma adecuada de analizarla.

En el año 2005 la web generada por los usuarios empieza a implementarse con mayor rapidez, la web 2.0 como fue denominada se logra implementando páginas web de estilo HTML con bases de datos basadas en SQL. En este año también es creada una herramienta de código abierto hadoop cuyo objetivo principal es el almacenamiento y el análisis de grandes datos.

La categorización de los datos es importante para analizarlos de la manera adecuada, en Big Data se definen tres tipos, datos estructurados, datos no estructurados y datos semi-estructurados los cuales son definidos teniendo en cuenta su precedencia y forma. A continuación, se definirán los tipos de datos:

- Datos Estructurados: se define datos estructurados a los datos que tienen definido su longitud, formato, números, fechas y que tienen esta información almacena en bases de datos relacionales como SQL. Los podríamos ver como si fuese un archivador perfectamente organizado donde todo está identificado, etiquetado y es de fácil acceso. Así que los datos estructurados son cualquier tipo de dato que se encuentre en un campo fijo dentro archivo o registro.

- Datos no Estructurados: Es aquella información que no está almacenada en tablas de bases de datos y no tiene definida una estructura interna. Estos datos son generados en su mayoría por los usuarios e incluyen 13 mensajes de correo electrónico mensajes de redes sociales, mensajes instantáneos y otras comunicaciones en tiempo real como documentos, imágenes, audio y vídeo.

• Datos Semi-Estructurados: Son aquellos datos que no residen de bases de datos relacionales, pero que presentan una organización interna que facilita su tratamiento, tales como documentos XML, CSV y datos almacenados en bases de datos NoSQL.

3. AREAS DE BIG DATA

3.1 RECOLECCIÓN

La recolección de datos hace referencia a una de las disciplinas de big data la cual en muy poco tiempo ha variado con mayor rapidez, esto es debido a que los datos son generados en grandes volúmenes, y son provenientes de muchas fuentes y de diversos dispositivos distribuidos por todo el mundo que trasmiten, procesan y recolectan los datos que son generados por las diversas actividades como la información generada por las redes sociales, plataformas digitales, datos de geolocalización, entre muchos otros.

3.2 Almacenamiento:

La información se ha convertido en una materia prima de gran valor. El almacenamiento masivo de datos y las nuevas fuentes de obtención de estos como las redes sociales, plataformas digitales, buscadores en internet entre otros no sólo afectan al mundo de los negocios, sino también al ámbito académico y a las Administraciones públicas. Es por esto por lo que se debe llevar acabo un almacenamiento escalable, es decir se debe implementar un sistema de almacenamiento que pueda variar su tamaño sin afectar el rendimiento general del sistema. Debido a esta necesidad han aparecido diferentes soluciones para tratar el almacenamiento masivo de datos, estas soluciones permiten establecer perfiles y hacer clasificaciones de datos, algunas de estas herramientas son:

Data Warehousing: Un data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso. En otras palabras, un data Warehouse es una arquitectura de almacenamiento de datos que le brinda a las empresas la capacidad de comprender y utilizar sus datos para tomar decisiones estratégicas.

Business Intelligence: hace referencia al uso de estrategias y herramientas que sirven para transformar información en conocimiento, con el objetivo de mejorar el proceso de toma de decisiones en una empresa. Las herramientas de Business Intelligence se han convertido en una potente herramienta, capas de analizar y procesar infinidad de datos, de infinidad de fuentes y así ayudar a las empresas a extraer conclusiones para mejorar sus cifras de negocio.

Cloud computing: ofrece servicios a través de la conectividad y gran escala de Internet. La computación en la nube democratiza el acceso a recursos de software de nivel internacional, pues es una aplicación de software que atiende a diversos clientes. La multilocación es lo que diferencia la computación en la nube de la simple tercerización y de modelos de proveedores de servicios de aplicaciones más antiguos. La computación en la nube ofrece a quienes adquieren su servicio la

capacidad de un pool de recursos de computación, mantenimiento, seguridad de los datos, y fácil acceso a la información.

3.3 Análisis

El análisis de datos es el proceso de examinar grandes cantidades de datos y así extraer información para descubrir patrones ocultos, correlaciones desconocidas y otra información útil, Los más importante del análisis de datos es procesar la información

de manera eficaz y en un tiempo razonable, de tal manera, que se puedan obtener resultados óptimos, Tal información puede proporcionar ventajas competitivas a través de organizaciones rivales y resultar en beneficios para el negocio.

3.4 VISUALIZACION

La visualización de datos permite representar cualquier tipo de información de una forma visual y sencilla, permite difundir el análisis previo de manera precisa y consistente, para ser visualizada al igual que proporciona la opción de comunicar el significado de los datos de una manera más entendible.

4. PARADIGMAS DE BIG DATA

El análisis de big data es diferente del análisis tradicional debido al gran aumento en el volumen de datos, debido a esto se hace imposible de manejar los datos usando los sistemas tradicionales de administración de bases de datos relacionales, para esto se necesitan paradigmas de programación que brindan una ayuda en el proceso y manejo de los datos de big data; en esta sesión se estudiaran los paradigmas que se centran en el desarrollo de aplicaciones y la gestión de grandes datos los cuales son MapReduce y MMP (Procesamiento Masivo en Paralelo)

4.1 Mapreduce

MapReduce es un framework que proporciona un sistema de procesamiento de datos paralelo y distribuido. Este paradigma se basa en enviar el proceso computacional al sitio donde residen los datos que se van a tratar, los cuales se coleccionan en un clúster Hadoop. MapReduce posee una arquitectura maestra / esclavo, la cual cuenta con un servidor maestro (JobTracker) y varios servidores esclavos (TaskTrackers), uno por cada nodo del clúster. Cuando se lanza un proceso de MapReduce se distribuyen las tareas entre los diferentes servidores del cluster y, es el propio framework Hadoop quien gestiona el envío y recepción de datos entre nodos. Una vez se han procesado todos los datos, el usuario recibe el resultado del clúster.

4.2 PROCESAMIENTO MASIVO EN PARALELO (MPP)

MPP es un paradigma que permite hacer cálculos para el procesamiento de consultas distribuidas, En MPP el procesamiento de datos se distribuye a través de un banco de nodos de cálculo, estos nodos están separados y procesan los datos en paralelo, los conjuntos de salida a nivel de nodo se ensamblan entre sí para producir un conjunto de resultados final.

5. PLATAFORMAS DE BIG DATA

La plataforma de Big Data es un tipo de solución de tecnologías de información que combina las funciones y capacidades de varias aplicaciones y utilidades de Big Data en una única solución. La plataforma de Big Data generalmente consiste en almacenamiento de Big Data, servidores, bases de datos, administración de Big Data, inteligencia comercial y otras utilidades de administración de Big Data. Una plataforma de análisis de datos grandes ayuda a extraer el valor de los datos. Los datos solo son útiles cuando se pueden derivar resultados comerciales beneficiosos, y para extraer los objetos valiosos de los datos, se deben adoptar las medidas adecuadas.

6. Instalación de un ambiente de Big Data y desarrollo de un caso practico

El objetivo de este capítulo es elaborar una serie de procedimientos paso a paso que describa el proceso de instalación de Hadoop, además de explicar el funcionamiento de cada uno de los servicios con los que cuenta esta herramienta para proporcionará un mejor entendimiento y un fácil aprendizaje de lo que es y lo que hace Hadoop.

6.1 Instalación De Un Ambiente De Big Data

Para la creación del ambiente de Big Data se utilizará la herramienta hadoop la cual permite el procesamiento distribuido de grandes volúmenes de datos mediante un cluster. Esta herramienta permite correr todas las aplicaciones dentro del mismo cluster, por lo tanto, toda la información quedará alojada allí y podrá ser utilizada sin ningún inconvenient.

CONTENIDO DEL CAPITULO	METODOLOGUIAS	RECURSOS	ESTÁNDARES DE EVALUACION
Conceptos • Definición de Big Data • 4V's de Big Data • Historia y evolución de Big Data • Tipos de datos **Procedimientos** • Leer el capitulo 2 de este manual practico • Practicar la lectura en libros de big data para aumentar los conocimientos, algunos de estos son: 1. Big Data: la revolución de los datos masivos (Kenneth Cukier y Viktor Mayer-Schonberger) 2. Analytics: el uso de big data en el mundo real	Ver conferencias acerca de Big Data y de cuál es su historia **Conceptos:** Realizar un resumen con la información más relevante del capítulo 2	• Capitulo 2 del presente manual practico • Conferencias sobre Big Data • Textos de apoyo • Reforzar los conceptos de Big Data con Exposiciones de los temas expuestos en el capítulo 2	Al finalizar el capítulo 2 se deberán tener claro los siguiente: • Definir en sus propias palabras que es Big Data • Saber qué hace Big Data • Inicios y evolución de Big Data

7. CONCLUCIONES

Big data es una tendencia que brinda una ayuda para el manejo de grandes volúmenes de información, principalmente es utilizado por grandes empresas, pero gracias a las nuevas tecnologías y su fácil acceso es posible ser utilizado por cualquier empresa o institución que desee vincularse con esta herramienta Las plataformas de Big Data al permitir el manejo de datos estructurados y no estructurados, presentan un gran beneficio para la toma de decisiones gracias a la facilidad de manejar todos esos tipos de datos, lo cual proporciona ventajas tanto para la vida profesional como para los diferentes campos de la ciencia La estructura de un ambiente Big Data ayuda a mejorar la manipulación de los datos, optimizando la gestión de la información respecto a tiempo y costo, logrando obtener mejores resultados en las estadísticas para una buena toma de decisiones.