Big data aplicado

05/10/2023

Introducción a Big Data

1. Big Data

Big data es el almacenamiento y procesamiento masivo de datos.

El termino big data también se refiere a las nuevas tecnologías que lo hacen posible.

Recopilar una gran cantidad de datos e intentar sacar conclusiones de este, que permita tomar decisiones rápidas y efectivas.

¿Cuándo considerar la utilización de Big Data?

* Gran volumen de datos
* Datos en tiempo real
* Transformación y análisis de datos no estructurados

1. Historia

El termino de Big data aparece en 1997

Aparece las 3Vs

* Volumen
* Velocidad
* Variedad

1. Componentes de Big Data

Fuentes de datos: bases de dtos, ficheros, csv…

Los datos se pueden capturar en tiempo real o acumulando. Se tratan en tiempo real (stream) o en batch. También existen hibiridos

Hay alamacenes de datos con los datos de todas las fuentes “filtrados” (data warehouses) para los analisisis.

Todo este proceso debe estar orquestado

Fuentes de datos

Clasificados por su origen

* Generados por maquinas
* Personas
* Externos

Según su estructura

* Estructurado: registro DB, CSV
* Semi-estructurado: JSON, XML
* Datos no estructurados: un comentario, un tweet, texto plano

Almacenamiento de Datos

* Sistemas de archivos distribuidos: Hadoop
* Almacenamiento en la nube: S3, Azure Blob

Ventajas que proporcionan

* Medidas de seguridad
* Copas de seguridad
* Políticas para el buen tratado. Este concepto se llama **gobernanza de datos**

La gobernanza de datos establece las responsabilidades para que los datos de proceses de manera segura y ética

Mensajes en tiempo real

Procesamiento por Lotes

Recibir datos y prepararlos para su procesamiento posterior

Procesamiento en tiempo real

Analizar los datos a medida que se recolectan. Ejemplos:

* Coches eléctricos

Procesamiento hibrido

Aquitectura típica en Big Data

Lambda

Todos los datos se procesan a trabes de una capa en batch y en una capa de stream (speed). La capa batch alamacena los datos en crudo

Kappa

Se olvida de la capa sequiencial (batch) y trata todo en tiempo real)

Almacén de Datos Analíticos (Data Warehouse)

Almacenes especifcos preparados para lamacenar gran cantidad de datos.

Ejemplo: “En lugar de guarda un registro con fecha y hora, se guarda solo el mes para hacer una tabla resumen rápido de los datos”

Sus características son:

* Consolidación de datos
* Estructura optimizada (para preparar informes)
* Historización
* Rendimiento
* Herramientas de consulta y reporting
* Seguridad y control de acceso

Análisis e informes

Tipos de análisis

* Predictivo
* Diagnostico
* Descriptivo
* Prescriptivo con Big Data: optimizar procesos

Visualizaciones o informes

Un cuadro de mandos permite visualizar los datos de forma comprensible para el usuario. Es importante story-telling

Orquestacion

Garantizar que el fujo de datos se ejecuta corectametne

Aprendizaje autmáticos

* Big data es el conjunto de datos en gran volumen
* Aprendizaje autmático, es una parte de la IA que se parovecha de este gran volumen de datos

Ambos son complementarios entre sí

Las 3 V del Big Data

* Volumen:
* Variedad:
* Velocidad:

Las 5V

* Veracidad
* Valor

Las 7 V

* Viabilidad
* Visualización

ESTO SE DEBERIA DAR EN SISTEMAS BID DATA

Infraestructura

* Hardware: servidores y dispositivos de alta velocidad
* Sofware: SO, herramientas

Ejemplos de infraestrucutra Big data

* Apache Hadoop:
* Apache Spark:
* Bases de datos NoSQL: datos no estructurados

La infraestructura debe permiten el proceso de una gran cantidad de datos

Ejemplo PowerBI

Cuadro de mandos ITACA

Se pueden descargar datos desde ITACA en formato XML. Tiene los siguientes (tablas): alumnos, grupos,

Credenciales

bigdata bigdata

Conceptos:

Clúster: Varios equipos trabajando en común