Data Analytics

Big data es un término que describe el gran volumen de datos – estructurados y no estructurados – que inundan una empresa, una institución, un ente gubernamental, etc. Sin embargo, no es la cantidad de datos lo importante, lo que importa es lo que las organizaciones hacen con los datos. El BigData puede ser analizado para obtener insights que conlleven a mejores decisiones y acciones de estratégicas.

Historia

El término "big data" se refiere a los datos que son tan grandes, rápidos o complejos y lo difícil o imposible de procesarlos con los métodos tradicionales.

El acto de acceder y almacenar grandes cantidades de información para la analítica ha existido desde hace mucho tiempo, pero el concepto de big data cobró impulso a principios de la década de 2000 cuando el analista de la industria, Doug Laney, articuló la definición actual de grandes datos como las tres V:

Volumen: Las organizaciones recopilan datos de diversas fuentes, como transacciones comerciales, dispositivos inteligentes (IO), equipo industrial, vídeos, medios sociales y más. En el pasado, su almacenamiento habría sido un problema, pero el almacenamiento más barato en plataformas como los data lakes y el Hadoop han aliviado la carga.

Velocidad: Con el crecimiento del Internet de las Cosas, los datos llegan a las empresas a una velocidad sin precedentes y deben ser manejados de manera oportuna. Las etiquetas RFID, los sensores y los medidores inteligentes están impulsando la necesidad de manejar estos torrentes de datos en tiempo casi real.

Variedad: Los datos se presentan en todo tipo de formatos: desde datos numéricos estructurados en bases de datos tradicionales hasta documentos de texto no estructurados, correos electrónicos, vídeos, audios, datos de teletipo y transacciones financieras.

La nube se refiere a un conjunto de recursos informáticos, como servidores, redes, almacenamiento y software, que están disponibles en línea y que pueden ser accedidos a través de Internet. En otras palabras, la nube es un modelo de entrega de servicios informáticos en el que los usuarios pueden acceder a recursos informáticos de alta calidad y a gran escala a través de la web sin tener que invertir en su propia infraestructura de TI.

El procesamiento distribuido se refiere al uso de múltiples sistemas informáticos para trabajar juntos en la realización de una tarea o procesamiento de datos. En lugar de tener un solo sistema que realice una tarea, el procesamiento distribuido divide la tarea en partes más pequeñas y las distribuye en varios sistemas para acelerar el proceso.

En el contexto de la nube, el procesamiento distribuido es una forma común de aprovechar los recursos disponibles en la nube. Al usar la nube, las organizaciones pueden acceder a recursos informáticos a gran escala que pueden ser compartidos por varios usuarios. Los sistemas de procesamiento distribuido pueden ser utilizados para aprovechar estos recursos de la nube y realizar tareas complejas de procesamiento de datos.

Hay varias ventajas en el uso de la nube y el procesamiento distribuido juntos. En primer lugar, la nube permite a las organizaciones acceder a una gran cantidad de recursos informáticos a un costo menor que la inversión en su propia infraestructura. En segundo lugar, el procesamiento distribuido permite a las organizaciones procesar grandes volúmenes de datos más rápido y con mayor eficiencia. Además, el procesamiento distribuido puede mejorar la capacidad de recuperación y la disponibilidad de los sistemas al permitir que las tareas se distribuyan en varios sistemas, en lugar de depender de un solo sistema.

Campias, Juan
Data Engineer

Un **Data Lake conserva todos los datos**, no sólo los que podrían utilizarse actualmente, sino también aquellos que podrían necesitarse en un futuro.
el **Data Warehouse estudia** muy bien **qué datos incluir**, cuáles son las fuentes de los datos. Además, se necesita dedicar tiempo para entender el negocio y así seleccionar y perfilar los datos necesarios. Por lo que el **Data Warehouse** al final, **contiene** un **modelo de datos altamente estructurado**, diseñado para la generación de informes.
Data Gobernance:

1. Permitir una mejor toma de decisiones
2. Reducir la fricción operativa
3. Proteger las necesidades del personal con interés en los datos
4. Capacitar a la dirección y al personal en general a adoptar enfoques comunes para problemas de datos
5. Construir procesos estándar repetibles
6. Reducir costes y aumentar la eficacia coordinando esfuerzos
7. Garantizar la transparencia de los procesos

Python es un lenguaje de programación popular en el análisis de datos debido a su flexibilidad y capacidad para procesar grandes cantidades de datos de manera eficiente. Pandas es una biblioteca de Python que proporciona estructuras de datos y herramientas para el análisis de datos. Se utiliza ampliamente en el análisis de datos debido a su facilidad de uso y su capacidad para manejar grandes cantidades de datos.
Los Data Engineers utilizan Python y Pandas en una variedad de formas para trabajar con datos. A continuación, se detallan algunos de los casos de uso más comunes:

1. Manipulación de datos: los Data Engineers utilizan Pandas para manipular y limpiar datos. Esto puede implicar la eliminación de valores atípicos, la eliminación de datos faltantes y la transformación de datos en un formato más adecuado para el análisis.
2. Automatización de procesos de ETL: Python se utiliza comúnmente para automatizar los procesos de ETL (extracción, transformación y carga) que son necesarios para integrar datos de múltiples fuentes en un sistema de datos centralizado. Los Data Engineers utilizan Pandas para procesar y transformar los datos y Python para automatizar el proceso de ETL.
3. Análisis exploratorio de datos: Pandas se utiliza para realizar un análisis exploratorio de datos, lo que implica la exploración de los datos para identificar patrones y tendencias. Los Data Engineers utilizan Pandas para calcular estadísticas descriptivas y para visualizar los datos de manera efectiva.
4. Análisis de datos en tiempo real: Python y Pandas también se pueden utilizar para analizar datos en tiempo real. Esto puede ser útil en situaciones en las que es necesario realizar análisis en tiempo real, como en el caso de datos de sensores o transacciones financieras en tiempo real.

Además de Pandas, los Data Engineers también pueden trabajar con otras herramientas de análisis de datos en Python, como Numpy, Scikit-learn y TensorFlow. Numpy se utiliza para trabajar con matrices y realizar operaciones matemáticas, Scikit-learn se utiliza para desarrollar modelos de aprendizaje automático y TensorFlow se utiliza para desarrollar modelos de aprendizaje automático complejos, como redes neuronales.
En conclusión, Python y Pandas son herramientas importantes para los Data Engineers en el análisis de datos. Se utilizan para manipular y limpiar datos, automatizar procesos de ETL, realizar análisis exploratorio de datos y analizar datos en tiempo real. Además de Pandas, los Data Engineers también pueden trabajar con otras herramientas de análisis de datos en Python, como Numpy, Scikit-learn y TensorFlow.

Campias, Juan
Data Engineer

Los Data Engineers utilizan SQL para administrar grandes volúmenes de datos y asegurar que estén disponibles y sean accesibles para otros miembros del equipo de análisis de datos.

1. Creación y administración de bases de datos: SQL se utiliza para crear y administrar bases de datos relacionales. Los Data Engineers utilizan SQL para definir esquemas de bases de datos, crear tablas y definir restricciones de integridad de datos. Además, SQL se utiliza para realizar tareas de mantenimiento y optimización de bases de datos, como la indexación y la limpieza de datos.
2. Extracción de datos: SQL se utiliza para extraer datos de bases de datos relacionales. Los Data Engineers pueden utilizar SQL para escribir consultas que recuperen datos específicos de una o varias tablas en una base de datos. Esto permite a los Data Engineers trabajar con grandes volúmenes de datos y extraer solo los datos relevantes para sus análisis.
3. Transformación de datos: SQL se utiliza para transformar datos en bases de datos relacionales. Los Data Engineers pueden utilizar SQL para escribir consultas que actualicen, inserten o eliminen datos en una tabla. Esto permite a los Data Engineers realizar tareas de limpieza y transformación de datos dentro de la base de datos.
4. Integración de datos: SQL se utiliza para integrar datos de múltiples fuentes. Los Data Engineers pueden utilizar SQL para unir datos de diferentes tablas o bases de datos en una única tabla o base de datos. Esto permite a los Data Engineers trabajar con datos de múltiples fuentes y realizar análisis más completos.
5. Consultas complejas: SQL se utiliza para escribir consultas complejas que involucren múltiples tablas y condiciones. Los Data Engineers pueden utilizar SQL para escribir consultas que realicen análisis complejos y agregaciones de datos.

En conclusión, SQL es una herramienta importante para los Data Engineers en la Ingeniería de Datos. Se utiliza para administrar y manipular bases de datos relacionales, extraer datos, transformar datos, integrar datos y realizar consultas complejas. SQL permite a los Data Engineers trabajar con grandes volúmenes de datos y asegurar que estén disponibles y sean accesibles para otros miembros del equipo de análisis de datos.

Importancia:

La importancia del big data no gira en torno a la cantidad de datos que tienes, sino en lo que haces con ellos. Puedes tomar datos de cualquier fuente y analizarlos para encontrar respuestas que permitan:

1. reducir los costos
2. reducir el tiempo
3. desarrollar nuevos productos y optimizar las ofertas
4. tomar decisiones inteligentes.

Cuando se combinan grandes datos con análisis de alta potencia, se pueden realizar tareas relacionadas con los negocios como:

- Determinar las causas de origen de fallos, problemas y defectos casi en tiempo real.
- Generar cupones en el punto de venta basados en los hábitos de compra del cliente.
- Recalcular portafolios de riesgo completos en minutos.
- Detectar el comportamiento fraudulento antes de que afecte a su organización.

Tres de los roles más comunes en este campo son:

Data Analyst: Es un profesional que se encarga de analizar datos para extraer información significativa que pueda ayudar a las empresas a tomar decisiones. Los analistas de datos a menudo trabajan con bases de datos, hojas de cálculo y herramientas de visualización de datos para realizar análisis exploratorios y generar informes. En términos de habilidades

técnicas, los Data Analysts deben ser competentes en SQL, ya que necesitan saber cómo extraer datos de una base de datos. Además, los analistas de datos también deben tener habilidades en Excel para trabajar con hojas de cálculo. Aunque no es tan crucial como SQL, Python también puede ser útil para los Data Analysts, especialmente si necesitan hacer análisis estadísticos complejos o trabajar con grandes cantidades de datos.
Data Engineer: Es responsable de la infraestructura necesaria para soportar el procesamiento de datos en una empresa. Los ingenieros de datos diseñan, construyen y mantienen sistemas de bases de datos y de procesamiento de datos que permiten a las empresas almacenar y acceder a grandes cantidades de información. En términos de habilidades técnicas, los Data Engineers deben ser expertos en SQL, ya que necesitan saber cómo diseñar y administrar bases de datos complejas. Además, los ingenieros de datos también deben tener habilidades en Python, ya que a menudo trabajan con herramientas de Big Data como Apache Hadoop y Spark, que requieren conocimientos de programación.
Data Scientist: Es un profesional que utiliza técnicas avanzadas de análisis de datos y herramientas de aprendizaje automático para resolver problemas empresariales complejos. Los científicos de datos se encargan de explorar grandes conjuntos de datos y de desarrollar modelos predictivos y algoritmos para obtener información valiosa. En términos de habilidades técnicas, los Data Scientists deben ser competentes en SQL y Python. SQL es esencial para extraer datos de bases de datos y Python es utilizado para la programación de algoritmos de aprendizaje automático y análisis estadísticos avanzados. Además, los Data Scientists deben tener un conocimiento profundo de estadísticas y matemáticas, así como una sólida comprensión de los conceptos de ciencia de datos.
En conclusión, el conocimiento de SQL es esencial para todos los roles en análisis de datos. Python, por otro lado, es más importante para los Data Engineers y Data Scientists, pero también puede ser útil para los Data Analysts. En general, cada uno de estos roles tiene habilidades y conocimientos únicos, pero todos son esenciales para ayudar a las empresas a tomar decisiones basadas en datos y mejorar su eficiencia y eficacia en el mercado.
Para ser un buen **Data Analyst**, se necesitan habilidades técnicas en:

1. SQL: es la habilidad más importante para un Data Analyst, ya que es la herramienta principal para extraer y manipular datos de una base de datos.
2. Excel: también es una habilidad clave, ya que es una herramienta muy utilizada para análisis de datos y visualización.
3. Herramientas de visualización de datos: Tableau, Power BI y Google Data Studio son algunas de las herramientas más utilizadas para visualizar datos de manera efectiva.

Para ser un buen **Data Engineer**, se necesitan habilidades técnicas en:

1. SQL: como en el caso del Data Analyst, es una habilidad fundamental para el Data Engineer, ya que es la herramienta principal para diseñar, administrar y manipular bases de datos complejas.
2. Python: se utiliza para automatizar procesos de ETL (extracción, transformación y carga de datos), así como para trabajar con herramientas de Big Data como Apache Hadoop y Spark.
3. Herramientas de Big Data: Apache Hadoop, Spark, Hive y Pig son algunas de las herramientas más utilizadas en la gestión de Big Data.

Para ser un buen **Data Scientist**, se necesitan habilidades técnicas en:

1. SQL: al igual que en los otros dos roles, es una habilidad esencial para extraer y manipular datos.
2. Python: es la herramienta principal para trabajar con algoritmos de aprendizaje automático y análisis estadísticos avanzados.
3. Herramientas de aprendizaje automático: Scikit-learn, TensorFlow y Keras son algunas de las herramientas más utilizadas en el desarrollo de modelos de aprendizaje automático.

En resumen, aunque SQL es una habilidad esencial en todos los roles, cada uno de ellos requiere habilidades y herramientas específicas para realizar su trabajo de manera efectiva. Los Data Analysts se enfocan en el análisis y visualización de datos, los Data Engineers en la gestión y procesamiento de datos, y los Data Scientists en el desarrollo de modelos de aprendizaje automático y análisis estadísticos avanzados. Por lo tanto, la elección de herramientas y habilidades dependerá del trabajo específico que se esté realizando en cada uno de estos roles.