Modulo tres

¿Cómo debe iniciarse las empresas en la ciencia de datos?

Asegúrese de que haya una coherencia para que alguien que intente comprender esos datos 20 años después pueda hacerlo.

Basura entra, basura sale es una regla en cualquier tipo de análisis. Si algo no se mide, es muy difícil mejorarlo o cambiarlo. Así que el primer paso es la medición. Si las empresas tienen datos existentes, entonces deberían comenzar a analizarlos y limpiarlos. Si no disponen de datos existentes, entonces necesitan empezar a recopilarlos.

Todas las organizaciones utilizan la ciencia de datos por la misma razón: descubrir soluciones óptimas a los problemas existentes. Pero para descubrir estas soluciones, tu organización debe identificar el problema y establecer una comprensión clara del mismo.

La medición es el primer paso para que una organización resuelva su problema utilizando datos. Necesita capturar y recopilar sus datos. Si algo no se mide, es difícil mejorarlo o cambiarlo. Si su organización no está capturando los datos, ayúdelos a descubrir cómo capturarlos. Nunca sobrescriba datos antiguos. Siempre es relevante y nunca pasa de moda. Una vez que tengas los datos, puedes empezar a mirarlos y limpiarlos.

Como científico de datos, su trabajo es ayudar a su organización a identificar herramientas y desarrollar una estrategia de análisis. Tenga en cuenta estudios de casos al personalizar una solución potencial. Identifique sus herramientas de análisis y luego desarrolle sus modelos estadísticos y de aprendizaje automático.

Término	Definición
Modelos aritméticos	La ciencia de datos a menudo utiliza modelos matemáticos para analizar datos y predecir resultados.
Estudio de caso	Análisis en profundidad de una instancia de un tema elegido para extraer conocimientos que informen la teoría, la práctica o la toma de decisiones.
Minería de datos	Extraer información de datos sin procesar, como tomar decisiones, predecir tendencias o comprender fenómenos.
Ciencia de datos	El campo implica recopilar, analizar e interpretar datos para extraer información valiosa y tomar decisiones informadas.

Estrategia de datos	Un plan que describe cómo una organización recopilará, gestionará y utilizará datos para lograr sus objetivos.
Análisis predictivo	Utilizando datos, algoritmos, modelos y aprendizaje automático para realizar predicciones.

Modulo cuatro

Comprender los datos

Los datos son información no organizada que se procesa para hacerla significativa. Generalmente, los datos comprenden hechos, observaciones, percepciones, números, caracteres, símbolos e imágenes que pueden interpretarse para derivar significado. Una de las formas en que se pueden categorizar los datos es por su estructura. Los datos pueden ser estructurados, semiestructurados o no estructurados.

Los datos estructurados tienen una estructura bien definida o se adhieren a un modelo de datos específico, se pueden almacenar en esquemas bien definidos, como bases de datos, y en muchos casos se pueden representar de manera tabular con filas y columnas.

Algunas de las fuentes de datos estructurados podrían incluir bases de datos SQL y sistemas de procesamiento de transacciones en línea, u OLTP, que se centran en transacciones comerciales, hojas de cálculo, como Excel y Google Sheets

Los datos semiestructurados son datos que tienen algunas propiedades organizativas, pero carecen de un esquema fijo o rígido. Los datos semiestructurados no se pueden almacenar en forma de filas y columnas, como en las bases de datos. Contiene etiquetas o elementos, o metadatos, que se utilizan para agrupar datos y organizarlos en una jerarquía.

Algunas de las fuentes de datos semiestructurados podrían incluir correos electrónicos, XML y otros lenguajes de marcado. XML y JSON permiten a los usuarios definir etiquetas y atributos para almacenar datos en forma jerárquica, se utilizan ampliamente para almacenar e intercambiar datos estructurados.

Los Datos no estructurados son datos que no tienen una estructura clara y no se pueden organizar en una Base de datos relacional, en filas y columnas. No tiene ningún formato, secuencia, semántica, o reglas particulares. Los datos no estructurados pueden abordar la heterogeneidad de fuentes y tienen una variedad de aplicaciones de inteligencia empresarial y análisis.

Formatos de archivos como JPEG, GIF y PNG. archivos de video y Archivo de audio, documentos y archivos PDF, presentaciones de PowerPoint, registros de medios, y encuestas.

Fuentes de datos

Normalmente, las organizaciones cuentan con aplicaciones internas que las ayudan a gestionar sus actividades comerciales diarias, transacciones con clientes, actividades de recursos humanos, y sus flujos de trabajo. Estos sistemas utilizan Base de datos relacional, como SQL Server, Oracle, MySQL, e IBM DB2, para almacenar datos de manera estructurada.

Los datos almacenados en bases de datos y almacenes de datos pueden utilizarse como fuente para el análisis. Fuera de la organización, existen otros conjuntos de datos disponibles públicamente y en forma privada. Por ejemplo, organizaciones gubernamentales que publican conjuntos de datos demográficos y económicos de forma continua. Luego están las empresas que venden datos específicos. Por ejemplo, datos de puntos de venta, datos financieros o datos meteorológicos.

Los archivos planos almacenan datos en formato de texto sin formato, con un registro o fila por línea, y cada valor separado por delimitadores como comas, punto y coma o tabulaciones. Los datos de un archivo plano se asignan a una sola tabla, a diferencia de las bases de datos relacionales que contienen varias tablas. Uno de los formatos de archivos planos más comunes es CSV en el que los valores están separados por comas.

Los archivos de hojas de cálculo son un tipo especial de archivos planos que también reconocen datos en formato tabular, filas y columnas. Pero una hoja de cálculo puede contener varias hojas de trabajo, y cada hoja de trabajo puede asignarse a una tabla diferente. Aunque los datos en las hojas de cálculo están en texto simple, los archivos se pueden almacenar en formatos personalizados e incluir información adicional como formato, fórmulas, etc. Microsoft Excel, que almacena datos en formato XLS o XLSX, es probablemente la hoja de cálculo más común.

Los archivos XML contienen valores de datos que se identifican o marcan mediante etiquetas. Mientras que los datos en archivos planos son planos o se asignan a una sola tabla, los archivos XML pueden admitir estructuras de datos más complejas como las jerárquicas. Algunos usos comunes de XML incluyen datos de encuestas en línea, extractos bancarios, y otros conjuntos de datos no estructurados.

Las API y los servicios web generalmente escuchan las solicitudes entrantes, que pueden tener la forma de solicitudes web de los usuarios o solicitudes de red de las aplicaciones y devuelven datos en texto sin formato, XML, HTML, JSON, o archivos multimedia.

El web scraping se utiliza para extraer datos relevantes de fuentes no estructuradas. ganancias por acción, y precios históricos para comercio y análisis. También conocido como raspado de pantalla, recolección web, y extracción de datos web, el raspado web permite APIs de búsqueda y validación de datos, que pueden ser muy útiles para los Analista de datos en la limpieza y descargar datos específicos de páginas web en función de parámetros definidos. preparación de datos, así como para correlacionar datos.

Las API también se utilizan para extraer datos de fuentes de bases de datos, dentro y fuera de los Algunos usos populares del web scraping incluyen la recolección de detalles de productos de minoristas, fabricantes y sitios web de comercio electrónico para proporcionar comparaciones de precios, generar oportunidades de venta a través de datos públicos

Algunas aplicaciones populares utilizadas para procesar flujos de datos incluyen Apache Kafka, Apache Spark Streaming, y Apache Storm.

Término	Definición
Valores separados por comas (CSV)	Archivos de texto delimitados donde el separador es una coma. Se utilizan para almacenar datos estructurados.
Formatos de archivos de texto delimitados	Los archivos de texto se utilizan para almacenar datos donde cada línea o fila tiene valores separados por un delimitador. Un delimitador es una secuencia de uno o más caracteres que especifica el límite entre valores. Los delimitadores comunes incluyen la coma, el tabulador, los dos puntos, la barra vertical y el espacio.
bases de datos NoSQL	Las bases de datos están diseñadas para almacenar y gestionar datos no estructurados y proporcionar herramientas de análisis para examinar este tipo de datos.
Sistemas de procesamiento de transacciones en línea (OLTP)	Sistemas que se centran en el manejo de transacciones comerciales y el almacenamiento de datos estructurados.
bases de datos relacionales	Las bases de datos están diseñadas para almacenar datos estructurados con esquemas bien definidos y admitir métodos y herramientas de análisis de datos estándar.

Sensores	Dispositivos como los sistemas de posicionamiento global (GPS) y las etiquetas de identificación por radiofrecuencia (RFID) generan datos estructurados.
Hojas de cálculo	Se utilizan aplicaciones de software como Excel y Hojas de cálculo de Google para organizar y analizar datos estructurados.
Bases de datos SQL	Bases de datos que utilizan lenguaje de consulta estructurado (SQL) para definir, manipular y consultar datos en formatos estructurados.
Valores separados por tabulaciones (TSV)	Archivos de texto delimitados donde el delimitador es una tabulación. Se utilizan como alternativa a CSV cuando hay comas literales en los datos de texto.

Término	**Definición**
Adobe Spark	Un conjunto de herramientas de software que permite a los usuarios crear y compartir contenido visual como gráficos, páginas web y videos.
Habilidades analíticas	La capacidad de analizar información de manera sistemática, lógica y organizada.
Director de información (CIO)	Un ejecutivo de negocios responsable de los sistemas de tecnología de la información de una organización y de las iniciativas relacionadas con la tecnología.
Pensamiento computacional	Dividir problemas en partes más pequeñas y usar algoritmos, lógica y abstracción para desarrollar soluciones. A menudo utilizado, pero no limitado a la informática.

Clústeres de datos	Un grupo de puntos de datos similares y relacionados, distintos de otros clústeres.
Resumen ejecutivo	Generalmente ubicado al comienzo de un documento de investigación, esta sección resume las partes importantes del documento, incluidas sus conclusiones clave.
Clúster de computación de alto rendimiento (HPC)	Una tecnología de computación que utiliza un sistema de computadoras en red diseñado para resolver problemas complejos y computacionalmente intensivos en entornos tradicionales.
Computación matemática	El uso de computadoras para calcular, simular y modelar problemas matemáticos.
Matrices	Plural de matriz, matrices son un arreglo rectangular (tabular) de números que a menudo se utilizan en matemáticas, estadística e informática.
Stata	Un paquete de software utilizado para análisis estadístico.
Distribuciones estadísticas	Una forma de describir la probabilidad de diferentes resultados basados en un conjunto de datos. La “curva de campana” es una distribución estadística común.
Lenguaje de Consulta Estructurado (SQL)	Un lenguaje utilizado para gestionar datos en una base de datos relacional.

Red TCP/IP	Una red que utiliza el protocolo TCP/IP para comunicarse entre dispositivos conectados en esa red. Internet utiliza TCP/IP.

Recogida y organización de los datos

Un repositorio de datos es un término general utilizado para referirse a datos que han sido recopilados, organizados, y aislados para que puedan usarse en operaciones comerciales o extraerse para informes y análisis de datos. Puede ser una infraestructura de base de datos pequeña o grande con una o más bases de datos que recopilan, administran, y almacenan conjuntos de datos.

Una base de datos es una colección de datos, o información diseñada para la entrada, almacenamiento, búsqueda recuperación y modificación de datos. Y un Sistema de administración de bases de datos, o DBMS, es un conjunto de programas que crea y mantiene la base de datos. Permite almacenar, modificar, y extraer información de la base de datos mediante una función llamada consulta.

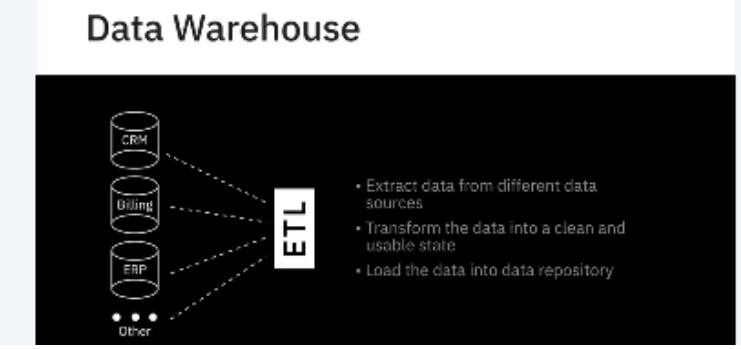
Es importante mencionar aquí dos tipos principales de bases de datos: la Base de datos relacional y las no relacionales. Las Base de datos relacional, también conocidas como RDBMS, se basan en los principios organizativos del archivo plano, con datos organizados en un formato tabular con filas y columnas siguiendo una estructura y un esquema bien definidos. Sin embargo, a diferencia del archivo plano, los RDBMS están optimizados para operaciones de datos y consultas que involucran muchas tablas y volúmenes de datos. El Lenguaje de Consulta Estructurada o SQL es el lenguaje de consulta estándar para Base de datos relacional.

Luego tenemos las bases de datos no relacionales, también conocidas como NoSQL o No Sólo SQL. Las bases de datos no relacionales surgieron en respuesta al volumen, la diversidad y la velocidad con que se generan los datos hoy en día, influenciadas principalmente por los avances en la computación en la nube, la Internet de las cosas, y la proliferación de las redes sociales. Diseñadas para la velocidad, la flexibilidad y la escala, las bases de datos no relacionales hicieron posible almacenar datos sin esquema o de forma libre. NoSQL se utiliza ampliamente para procesar grandes cantidades de datos.

Un almacén de datos funciona como un repositorio central que fusiona información proveniente de distintas fuentes y la consolida a través del proceso de extracción, transformación, y carga, también conocido como proceso ETL, en una base de datos integral para análisis e inteligencia empresarial. En un nivel muy alto, el proceso ETL le

ayuda a extraer datos de diferentes fuentes de datos, transformarlos en un estado limpio y utilizable y cargarlos en el repositorio de datos de la empresa.

Históricamente, los data marts y almacenes de datos han sido relacionales, ya que gran parte de los datos empresariales tradicionales han residido en RDBMS. Sin embargo, con la aparición de tecnologías NoSQL y nuevas fuentes de datos, los repositorios de datos no relacionales ahora también se utilizan para el almacenamiento de datos.



Sistema de administración de base de datos relacionales

Una base de datos relacional es una colección de datos organizados en una estructura de tabla, donde las tablas se pueden vincular o relacionar en función de datos comunes a cada una. Las tablas están formadas por filas y columnas, donde las filas son los registros y las columnas los atributos.

La capacidad de relacionar tablas basadas en datos comunes permite recuperar una tabla completamente nueva a partir de los datos de una o más tablas con una sola consulta. También permite comprender las relaciones entre todos los datos disponibles y obtener nuevos conocimientos para tomar mejores decisiones. Las bases de datos relacionales utilizan el lenguaje de consulta estructurado o SQL para consultar datos.

Las Base de datos relacional se basan en los principios organizativos del archivo plano, como las hojas de cálculo, con datos organizados en filas y columnas siguiendo una estructura y un esquema bien definidos.

Las bases de datos relacionales, por diseño, son ideales para el almacenamiento, la recuperación, y el procesamiento optimizado de grandes volúmenes de datos a diferencia de las hojas de cálculo que tienen un número limitado de filas y columnas.

Cada tabla de una base de datos relacional tiene un conjunto único de filas y columnas y se pueden definir relaciones entre tablas, lo que minimiza la redundancia de datos. Además, puede restringir los campos de la base de datos a tipos de datos y valores específicos, lo que

minimiza las irregularidades y conduce a una mayor consistencia e integridad de los datos. Las bases de datos relacionales utilizan SQL para consultar datos, lo que brinda la ventaja de procesar millones de registros y recuperar grandes cantidades de datos en cuestión de segundos.

La arquitectura de seguridad de las bases de datos relacionales proporciona acceso controlado a los datos y también garantiza que se puedan aplicar los estándares y políticas para gobernar los datos. Las bases de datos relacionales varían desde pequeños sistemas de escritorio hasta sistemas masivos basados en la nube. Pueden ser de código abierto y con soporte interno (AC de código abierto con soporte comercial (CID), o sistemas comerciales de código cerrado (OS).

Algunas de las Base de datos relacional en la nube más populares incluyen Amazon Relational Database Service o RDS, Google Cloud SQL, IBM DB2 en la nube, Oracle Cloud, y SQL Azure. RDBMS es una tecnología madura y bien documentada. facilitando el aprendizaje y la búsqueda de talento calificado. Una de las ventajas más significativas del enfoque de bases de datos relacionales es su capacidad de crear información significativa uniendo tablas.

Algunas de sus otras ventajas incluyen la flexibilidad. Con SQL, puede agregar nuevas columnas, agregar nuevas tablas, cambiar el nombre de las relaciones, y realizar otros cambios mientras la base de datos está en ejecución y se realizan consultas.

Las bases de datos relacionales ofrecen fácil exportación e importación, para el respaldo y la restauración. Las exportaciones pueden realizarse mientras la base de datos está en ejecución, para la restauración en caso de falla. Las bases de datos relacionales basadas en la nube realizan duplicación continua, la pérdida de datos durante la restauración se puede medir en segundos o menos.

ACID significa Atomicidad, Consistencia, Aislamiento y Durabilidad. El cumplimiento de ACID implica que los datos de la base de datos son precisos y consistentes a pesar de las fallas, y las transacciones de la base de datos se procesan. Ahora veremos algunos Casos de uso para las Bases de datos relacionales

Procesamiento de transacciones en línea Las aplicaciones OLTP se centran en tareas orientadas a transacciones que se ejecutan a altas tasas. Las bases de datos relacionales son adecuadas para aplicaciones OLTP porque pueden admitir una gran cantidad de usuarios. Admiten la capacidad de insertar, actualizar, o eliminar pequeñas cantidades de datos. Además, admiten consultas y actualizaciones frecuentes, así como tiempos de respuesta rápidos. Almacenes de datos.

Los RDBMS no funcionan bien con Datos estructurados o Datos no estructurados y, por lo tanto, no son adecuados para análisis extensivos sobre dichos datos. Para la migración entre

dos RDBMS, los esquemas y el tipo de datos deben ser idénticos entre las tablas de origen y destino.

Las bases de datos relacionales tienen un límite en la longitud de los campos de datos, lo que significa que si intenta ingresar más información en un campo de la que este puede acomodar, la información no se almacenará. A pesar de las limitaciones y la evolución de los datos en estos tiempos de big data, computación en la nube dispositivos IoT, y redes sociales, RDBMS sigue siendo la tecnología predominante para trabajar con datos estructurados.

NoSQL

Es un diseño de base de datos no relacional que proporciona esquemas flexibles para el almacenamiento y la recuperación de datos. Las bases de datos NoSQL han existido durante muchos años, pero solo recientemente se han vuelto más populares en la era de la nube, Big data y aplicaciones para dispositivos móviles: aplicación para dispositivos móviles de alto volumen.

No utilizan un diseño de base de datos tradicional de filas, columnas y tablas con esquemas fijos y, por lo general, no utilizan el lenguaje de consulta estructurado, o SQL, para consultar datos, aunque algunos pueden admitir SQL o interfaces similares a SQL.

NoSQL permite que los datos se almacenen sin esquema o en formato libre. Cualquier dato, ya sea estructurado, semiestructurado o no estructurado, se puede almacenar en cualquier registro.

Hay cuatro tipos comunes de bases de datos NoSQL

Almacén de valores clave, basado en documentos, en columnas y en gráficos. Almacén de valores clave. Los datos en una base de datos de valores clave se almacenan como una colección de pares de valores clave. La clave representa un atributo de los datos y es un identificador único. Tanto las claves como los valores pueden ser cualquier cosa, desde números enteros o cadenas simples hasta documentos JSON complejos. Los almacenes de valores clave son excelentes para almacenar datos de sesiones de usuario y preferencias de usuario, hacer recomendaciones en tiempo real y publicidad dirigida, y almacenar datos en caché en memoria.

Redis, Memcached, y DynamoDB son algunos ejemplos bien conocidos en esta categoría.

Basado en documentos. Las bases de datos de documentos almacenan cada registro y sus datos asociados dentro de un solo documento. Permiten una indexación flexible, potentes consultas, y análisis de colecciones de documentos. Las bases de datos de documentos son preferibles para Plataforma de comercio electrónico, almacenamiento de registros médicos, plataformas de CRM, y plataformas de análisis.

MongoDB, DocumentDB, CouchDB, y Cloudant son algunas de las bases de datos basadas en documentos más populares.

Basado en columnas. Los modelos basados en columnas almacenan datos en celdas agrupadas como columnas de datos en lugar de filas Una agrupación lógica de columnas, es decir, columnas a las que normalmente se accede juntas, se denomina familia de columnas.

Las bases de datos basadas en columnas pueden ser excelentes para sistemas que requieren solicitudes de escritura intensivas, almacenamiento de datos de Serie temporal, datos meteorológicos, y datos de IU. Pero si necesita utilizar consultas complejas o cambiar sus patrones de consulta con frecuencia, esta puede no ser la mejor opción para usted.

Las bases de datos de columnas más populares son Cassandra y HBase.

Basado en gráficos Las bases de datos basadas en gráficos utilizan un modelo gráfico para representar y almacenar datos. Son particularmente útiles para visualizar, analizar y encontrar conexiones entre diferentes piezas de datos. Las bases de datos gráficas son una excelente opción para trabajar con datos conectados, que son datos que contienen muchas relaciones interconectadas.

Las bases de datos de grafos son excelentes para redes sociales, recomendaciones de productos en tiempo real Diagrama de red, detección de fraudes, y gestión de accesos, pero si desea procesar grandes volúmenes de transacciones, puede que no sea la mejor opción para usted, porque las bases de datos gráficas no están optimizadas para consultas de análisis de gran volumen.

Neo4j y Cosmos DB son algunas de las bases de datos de gráficos más populares.

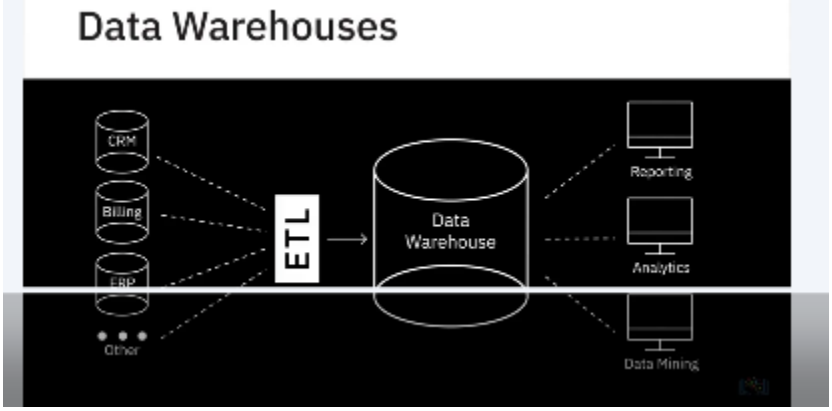
La principal ventaja de NoSQL es su capacidad para manejar grandes volúmenes de datos estructurados, semi-estructurados y Datos no estructurados. Algunas de sus otras ventajas incluyen la capacidad de ejecutarse como sistemas distribuidos escalados en múltiples centros de datos, lo que les permite aprovechar la infraestructura de computación en la nube, una arquitectura de escalamiento eficiente y rentable que proporciona capacidad y rendimiento adicionales con la incorporación de nuevos nodos. Un diseño más simple, mejor control sobre la disponibilidad, y una escalabilidad mejorada que te permite ser más ágil, más flexible, e iterar más rápidamente.

Las diferencias clave entre bases de datos relacionales y no relacionales, los esquemas RDBMS definen rígidamente cómo deben tipificarse y componerse todos los datos insertados en la base de datos, mientras que las bases de datos NoSQL pueden ser independientes del esquema, lo que permite almacenar y manipular datos no estructurados y semiestructurados. Mantener sistemas de gestión de bases de datos relacionales comerciales de alta gama es costoso, mientras que las bases de datos NoSQL están diseñadas específicamente para hardware de bajo costo.

Las bases de datos relacionales, a diferencia de la mayoría de NoSQL, admiten la conformidad con ACID, lo que garantiza la confiabilidad de las transacciones y la recuperación ante fallas. RDBMS es una tecnología madura y bien documentada, lo que significa que los riesgos son más o menos perceptibles en comparación con NoSQL, que es una tecnología relativamente más nueva.

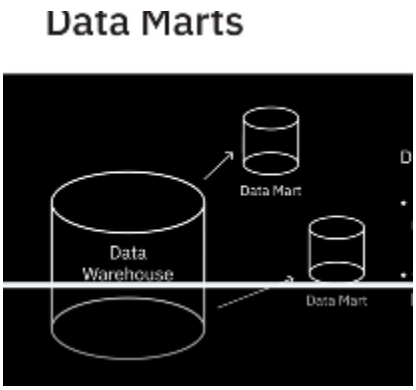
Data warehouse

Un almacén de datos funciona como un almacenamiento multipropósito para diferentes casos de uso. Cuando los datos llegan al almacén, ya han sido modelados y estructurados para un propósito específico, lo que significa que están listos para el análisis. Como organización, usted optaría por un almacén de datos cuando tenga cantidades masivas de datos de sus sistemas operativos que necesiten estar disponibles para generación de informes y análisis. Los almacenes de datos sirven como fuente única de verdad, y almacenan datos actuales e históricos que han sido limpiados, conformados, y categorizados.



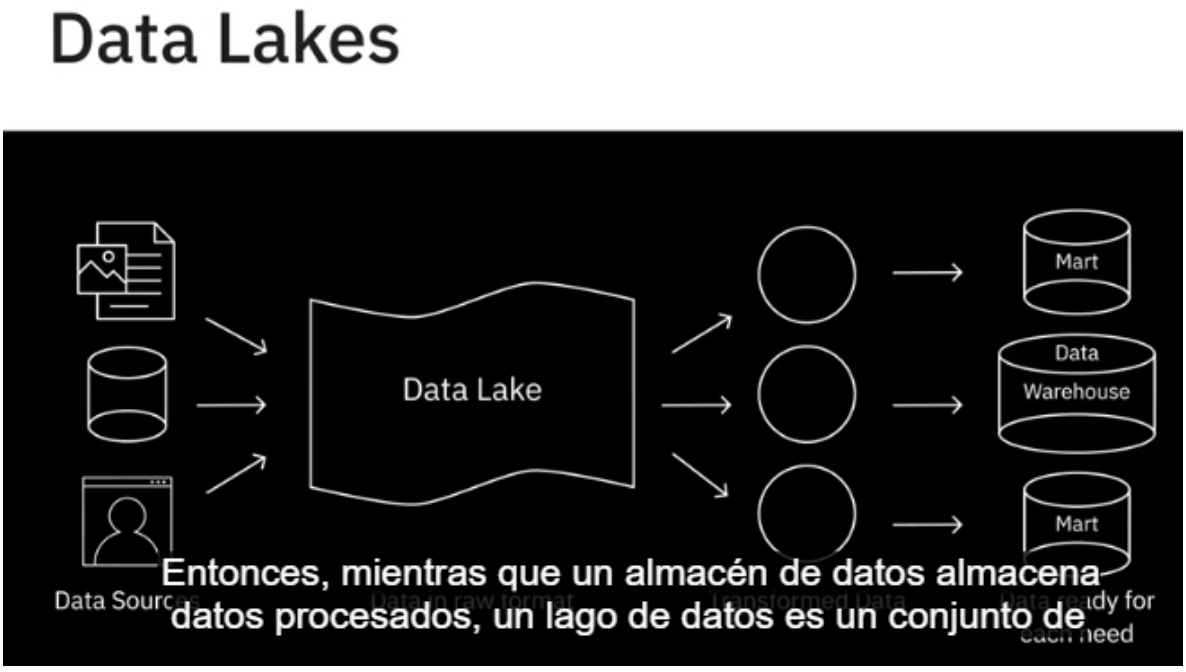
Data Marts

Un almacén de datos es un facilitador multipropósito de análisis operativos y de rendimiento. Un almacén de datos es una subsección del almacén de datos, creado específicamente para una función comercial un propósito o una comunidad de usuarios en particular. La idea es proporcionar a las partes interesadas los datos más relevantes cuando los necesitan. Por ejemplo, los equipos de ventas o finanzas que acceden a los datos para sus informes y proyecciones trimestrales. Dado que un almacén de datos ofrece capacidades analíticas para un área restringida del almacén de datos, ofrece seguridad y rendimiento aislados. La función más importante de un almacén de datos es la generación de informes y análisis específicos del negocio.



Data Lake

Un lago de datos es un repositorio de almacenamiento que puede almacenar grandes cantidades de datos estructurados, semiestructurados, y no estructurados en su formato nativo, clasificados y etiquetados con metadatos. Entonces, mientras que un almacén de datos almacena datos procesados, un lago de datos es un conjunto de datos sin procesar donde cada elemento de datos tiene un identificador único y metaetiquetas para uso futuro Optaría por un lago de datos si genera o tiene acceso a grandes volúmenes de datos de forma continua, pero no desea estar restringido a casos de uso específicos o predefinidos. A diferencia de los almacenes de datos, un lago de datos retendría todos los datos de origen sin ninguna exclusión, y los datos podrían incluir todo tipo de fuentes y tipos de datos. Los lagos de datos a veces también se utilizan como área de almacenamiento de un almacén de datos. El papel más importante de un lago de datos es el análisis predictivo y avanzado.



ETL

Proceso que está en el corazón de la obtención de valor de los datos, el proceso de extracción, transformación, y carga, o ETL. ETL es la forma en que los datos sin procesar se convierten en datos listos para el análisis. Es un proceso automatizado en el que se recopilan datos sin procesar de fuentes identificadas, y se extrae la información que se alinea con sus necesidades de informes y análisis. Limpiar, estandarizar, y transformar esos datos en un formato que pueda utilizarse en el contexto de su organización y cargarlos en un repositorio de datos. Si bien ETL es un proceso genérico, el trabajo real puede ser muy diferente en uso, utilidad, y complejidad. La extracción es el paso en el que se recopilan datos de las ubicaciones de origen para su transformación. La extracción de datos podría realizarse mediante procesamiento por lotes, lo que significa que los datos de origen se mueven en grandes fragmentos desde el sistema de origen al sistema de destino a intervalos programados.

Las herramientas para el procesamiento por lotes incluyen Stitch Blendo y SS. Procesamiento de flujo, lo que significa que los datos de origen se extraen en tiempo real de la fuente y se transforman mientras están en tránsito y antes de cargarse en el repositorio de datos. Las herramientas para el procesamiento de flujo incluyen Apache SAMSA, Apache STORM, y Apache Kafka. La transformación implica la ejecución de reglas y funciones que convierten datos sin procesar en datos que pueden usarse para el análisis. Por ejemplo, hacer que los formatos de fecha y las unidades de medida sean consistentes en todos los datos de origen, eliminar datos duplicados, filtrar datos que no necesita, enriquecer los datos, por ejemplo, dividir el nombre completo en nombre, segundo nombre y apellido. establecer relaciones clave entre tablas aplicar reglas de negocio y validaciones de datos la carga es el paso en el que los datos del proceso se transportan a un sistema de destino o repositorio de datos podría ser una carga inicial que llena todos los datos en el repositorio una carga incremental que aplica actualizaciones y modificaciones continuas según sea necesario periódicamente o actualización completa, es decir, borrar el contenido de una o más tablas y volver a cargarlas con datos nuevos.

La verificación de carga, que incluye controles de datos para detectar valores faltantes o nulos, el rendimiento del servidor, y el monitoreo de fallas de carga, es importante debido a este paso del proceso. Es vital estar atento a las fallas de carga y garantizar que existan los mecanismos de recuperación adecuados ETL se ha utilizado históricamente para cargas de trabajo por lotes a gran escala. Sin embargo, con la aparición de herramientas ETL de streaming, cada vez se utilizan más también para datos de eventos de streaming en tiempo real.

DATA PIPELINE

Es común ver los términos ETL y canalización de datos utilizados indistintamente, y aunque ambos mueven datos desde el origen al destino, canalización de datos es un término

más amplio que abarca todo el recorrido de mover datos de un sistema a otro, del cual ETL es un subconjunto. Las canalizaciones de datos se pueden diseñar para el procesamiento por lotes, para la transmisión de datos, y para una combinación de datos por lotes y de transmisión. En el caso de transmisión de datos, el procesamiento o transformación de datos ocurre en un flujo continuo. Esto es particularmente útil para datos que necesitan una actualización constante, como los datos de un sensor que monitorea el tráfico. Una tubería de datos es un sistema de alto rendimiento que soporta tanto consultas por lotes de larga duración como consultas interactivas más pequeñas. El destino de una canalización de datos suele ser un lago de datos, aunque los datos también pueden cargarse en diferentes destinos, como otra aplicación o una herramienta de visualización. Hay varias soluciones de tuberías de datos disponibles siendo las más populares Apache Beam y DataFlow.

La integración de datos incluye el acceso, la puesta en cola o la extracción de datos de sistemas operativos la transformación y fusión de datos extraídos de forma lógica o física, la calidad y la gobernanza de los datos, y la entrega de datos a través de un enfoque integrado para fines analíticos.

Mientras que la integración de datos combina datos dispares en una vista unificada de los datos, una canalización de datos cubre todo el recorrido del movimiento de datos desde los sistemas de origen hasta los de destino. En ese sentido, se puede utilizar un pipeline de datos para realizar la integración de datos, mientras que ETL es un proceso dentro de la integración de datos.

Término	Definición
Cumplimiento de ACID	Garantizar la precisión y la consistencia de los datos mediante atomicidad, consistencia, aislamiento y durabilidad (ACID) en las transacciones de bases de datos.
Plataforma de integración como servicio basado en la nube (iPaaS)	Plataformas de integración alojadas en la nube que ofrecen servicios de integración a través de nubes privadas virtuales o modelos de nube híbrida, proporcionando escalabilidad y flexibilidad.
Base de datos basada en columnas	Un tipo de base de datos NoSQL que organiza datos en celdas agrupadas como columnas, a menudo utilizada para sistemas que requieren un gran volumen de solicitudes de escritura y almacenamiento de series de tiempo o datos de IoT.

Datos en reposo	Datos almacenados y no en movimiento activo, que normalmente residen en una base de datos o un sistema de almacenamiento para diversos fines, incluida la copia de seguridad.
Integración de datos	Una disciplina que involucra prácticas, técnicas arquitectónicas y herramientas que permiten a las organizaciones ingerir, transformar, combinar y suministrar datos de diversos tipos, utilizados para fines tales como la consistencia de los datos, la gestión de datos maestros, el intercambio de datos y la migración de datos.
Lago de datos	Un repositorio de datos para almacenar grandes volúmenes de datos estructurados, semiestructurados y no estructurados en su formato nativo, lo que facilita la exploración y el análisis ágiles de datos.
Almacén de datos	Un subconjunto de un almacén de datos diseñado para funciones comerciales específicas o comunidades de usuarios, que proporciona seguridad y rendimiento aislados para análisis específicos.
Canalización de datos	Un proceso integral de movimiento de datos que cubre todo el recorrido de los datos desde los sistemas de origen hasta los sistemas de destino, que incluye la integración de datos como un componente clave.
Repositorio de datos	Término general que se refiere a los datos recopilados, organizados y aislados para operaciones comerciales o análisis de datos. Puede incluir bases de datos, almacenes de datos y almacenes de big data.
almacén de datos	Un repositorio central que consolida datos de diversas fuentes a través del proceso de extracción, transformación y carga (ETL), haciéndolos accesibles para análisis e inteligencia empresarial.
Base de datos basada en documentos	Un tipo de base de datos NoSQL que almacena cada registro y sus datos asociados dentro de un solo documento, lo que permite una indexación flexible, consultas ad hoc y análisis de colecciones de documentos.
Proceso ETL	El proceso de extracción, transformación y carga para la integración de datos implica extraer datos de diversas fuentes, transformarlos en un formato utilizable y cargarlos en un repositorio.
Base de datos basada en gráficos	Un tipo de base de datos NoSQL que utiliza un modelo gráfico para representar y almacenar datos, ideal para visualizar, analizar y descubrir conexiones entre puntos de datos interconectados.

almacén de clave-valor	Un tipo de base de datos NoSQL donde los datos se almacenan como pares clave-valor, donde la clave sirve como identificador único y el valor contiene datos, que pueden ser simples o complejos.
Portabilidad	La capacidad de las herramientas de integración de datos para usarse en diversos entornos, incluidos escenarios de nube única, nube múltiple o nube híbrida, proporciona flexibilidad en las opciones de implementación.
Conectores preconstruidos	Conectores y adaptadores catalogados que simplifican la conexión y la creación de flujos de integración con diversas fuentes de datos como bases de datos, archivos planos, redes sociales, API, CRM y aplicaciones ERP.
Bases de datos relacionales (RDBMS)	Bases de datos que organizan los datos en un formato tabular con filas y columnas, siguiendo una estructura y un esquema bien definidos.
Escalabilidad	La capacidad de un repositorio de datos para crecer y ampliar su capacidad para manejar volúmenes de datos y demandas de carga de trabajo crecientes a lo largo del tiempo.
Esquema	La estructura predefinida que describe la organización y el formato de los datos dentro de una base de datos, indicando los tipos de datos permitidos y sus relaciones.
Transmisión de datos	Los datos que se generan y transmiten continuamente en tiempo real requieren un manejo y procesamiento especializados para capturarlos y analizarlos.
Casos de uso para bases de datos relacionales	Aplicaciones como procesamiento de transacciones en línea (OLTP), almacenes de datos (OLAP) y soluciones de IoT donde sobresalen las bases de datos relacionales.
Bloqueo del proveedor	Una situación en la que un usuario se vuelve dependiente de las tecnologías y soluciones de un proveedor específico, lo que dificulta el cambio a otras plataformas.