2

DATA LAKE Y DATA WAREHOUSE

2.1 INTRODUCCIÓN

En este capítulo describiremos dos tipos de sistemas de almacenamiento centralizado, Data Lake y Data Warehouse. Un Data Lake permite almacenar datos en su forma cruda y sin procesar. Además, admite la ingesta de datos de diferentes fuentes y formatos sin necesidad de definir un esquema rígido de antemano. En un Data Lake se sigue un esquema de ELT (Extract, Load, Transform) que consiste en recolectar los datos, cargarlos en un repositorio y luego procesarlos.

Por otro lado, un Data Warehouse es un sistema diseñado para almacenar, organizar y procesar datos estructurados con el propósito de respaldar el análisis de negocios y la toma de decisiones. A diferencia de un Data Lake, un Data Warehouse sigue un enfoque estructurado y requiere definir un esquema antes de la carga de datos. En este caso, se sigue el esquema de ETL (Extract, Transform, Load), donde los datos se extraen de las fuentes, se procesan y luego se cargan y almacenan.

2.2 DEFINICIÓN DE DATA LAKE

De acuerdo con Amazon Web Services la definición de Data Lake es: "Repositorio centralizado que permite almacenar todos los datos estructurados y no estructurados a cualquier escala. Puede almacenar los datos tal cual, sin tener que estructurarlos primero, y ejecutar diferentes tipos de análisis, desde cuadros de mando y visualizaciones hasta grandes procesamientos de datos, análisis en tiempo real y aprendizaje automático para tomar mejores decisiones."

Un Data Lake es un **repositorio centralizado** que permite **almacenar datos de cualquier tipo de estructura**, desde estructurados a no estructurados, a cualquier

escala. Almacena los datos de forma cruda sin modificarlos y sin tener que estructurarlos primero.

Los Data Lakes se implementan comúnmente utilizando servicios en la nube, como Amazon S3, Azure Data Lake Storage y Google Cloud Storage. Estos servicios ofrecen una serie de ventajas a nivel de escalabilidad prácticamente ilimitada, lo que significa que los Data Lakes pueden crecer y adaptarse fácilmente a medida que se van obteniendo más datos. Además, estos servicios generalmente ofrecen modelos de precios flexibles, lo que resulta en menores costes en comparación con la implementación y mantenimiento de infraestructuras físicas on premise.

La principal utilidad de un data lake es que **permite centralizar una gran cantidad de fuentes de información**. En la actualidad, las organizaciones necesitan hacer un manejo óptimo de la información, bien sea de sus empleados, situación del mercado y métricas del negocio. En consecuencia, un data lake podrá usarse para organizar esta información a través de carpetas específicas.

A partir de aquí, se pueden detectar oportunidades de crecimiento empresarial y mejorar la productividad. Un Data Lake es capaz de proporcionar datos a la organización para una gran variedad de procesos analíticos diferentes:

- Descubrimiento y exploración de datos.
- Análisis ad hoc.
- Análisis para la toma de decisiones.
- Realizar informes.
- Análisis en tiempo real.

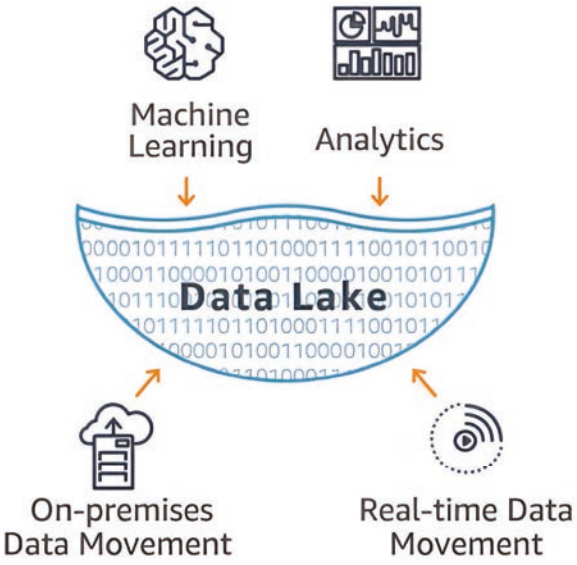

Figura 2.1. Casos de uso de un Data Lake

2.3 CARACTERÍSTICAS DE UN DATA LAKE

Un Data Lake, en su esencia, es una infraestructura de almacenamiento de datos diseñada para gestionar la creciente variedad y volumen de información generada por las organizaciones en la era digital. Su flexibilidad y capacidad para almacenar datos de diferentes formatos y estructuras lo convierten en una herramienta invaluable para la gestión y análisis de datos a gran escala.

Las características principales de un Data Lake incluyen su capacidad para almacenar datos de forma escalable y heterogénea, su esquema flexible que elimina la necesidad de definir una estructura de datos previa, y su integración con tecnologías de procesamiento distribuido y herramientas de análisis avanzado.

1. **Almacenamiento escalable**: los Data Lakes están diseñados para almacenar grandes volúmenes de datos, desde datos estructurados hasta datos no estructurados, en su forma original. Deben ser capaces de escalar horizontalmente para manejar el crecimiento de datos a largo plazo.
2. **Datos diversos y heterogéneos**: un Data Lake puede almacenar una amplia variedad de tipos de datos, incluidos datos estructurados, semiestructurados y no estructurados. Puede incluir datos de fuentes como bases de datos relacionales, registros de servidores, datos de sensores, archivos de registro, datos de redes sociales, entre otros.
3. **Esquema flexible**: a diferencia de los almacenes de datos tradicionales, los Data Lakes no requieren un esquema predefinido para los datos. Permiten la captura de datos en su forma original, lo que significa que no es necesario aplicar una estructura antes de almacenarlos.
4. **Procesamiento distribuido**: los Data Lakes suelen integrar capacidades de procesamiento distribuido para realizar análisis y consultas en grandes conjuntos de datos de manera eficiente. Esto puede incluir tecnologías como Apache Hadoop, Apache Spark, y motores de consulta distribuida.
5. **Integración con herramientas de análisis**: los Data Lakes deben ser compatibles con una variedad de herramientas de análisis y visualización para permitir a los usuarios extraer información útil de los datos almacenados. Esto puede incluir herramientas de Business Intelligence (BI), herramientas de análisis avanzado, lenguajes de programación como Python y R, entre otros.
6. **Seguridad y gobernanza**: los Data Lakes deben proporcionar mecanismos de seguridad robustos para proteger los datos almacenados, así como herramientas para gestionar el acceso y el cumplimiento normativo. También deben permitir la implementación de políticas de gobernanza de datos para garantizar la calidad y la integridad de los datos.
7. **Metadatos y descubrimiento de datos**: los Data Lakes suelen incluir capacidades de gestión de metadatos para ayudar a los usuarios a descubrir, entender y utilizar los datos almacenados. Esto puede incluir catálogos de datos, etiquetado automático de datos, y herramientas de búsqueda y navegación.

8. **Arquitectura abierta y flexible**: los Data Lakes deben ser capaces de integrarse con una variedad de tecnologías y sistemas existentes en el ecosistema de datos de una organización. Deben ser capaces de evolucionar con los requisitos cambiantes y permitir la incorporación de nuevas tecnologías y herramientas.

Las características de un Data Lake se podrían resumir en los siguientes puntos:

- Un Data lake almacena los datos tal y como se encuentran en un sistema empresarial. Esto permite guardar todos los datos independientemente de su fuente y estructura ya que se mantienen en su forma bruta y solo se transforman cuando sea necesario.
- El data lake adopta una estructura denominada "**Schema on Read**", donde la estructura no está predeterminada antes de que se almacenen los datos.
- En un data lake se almacenan todos los datos, independientemente de su estructura, tipo o usabilidad por parte de los usuarios en el momento de su recogida.
- Un data lake almacena, al menos, dos tipos de datos: los datos brutos y los datos ya procesados por los usuarios, datos que se acumulan y cambian constantemente. Esto requiere una gran capacidad de gestión de datos, que abarca las fuentes de datos, las conexiones de datos, los formatos de estos y, por último, el esquema que presentan. Este tipo de repositorio permite un almacenamiento centralizado para los datos de una empresa u organización.
- Tener una arquitectura escalable con una capacidad de crecer con el volumen de los datos, lo que les permite conservar todos los datos para cuando puedan utilizarse, añadir nuevas fuentes, etc.

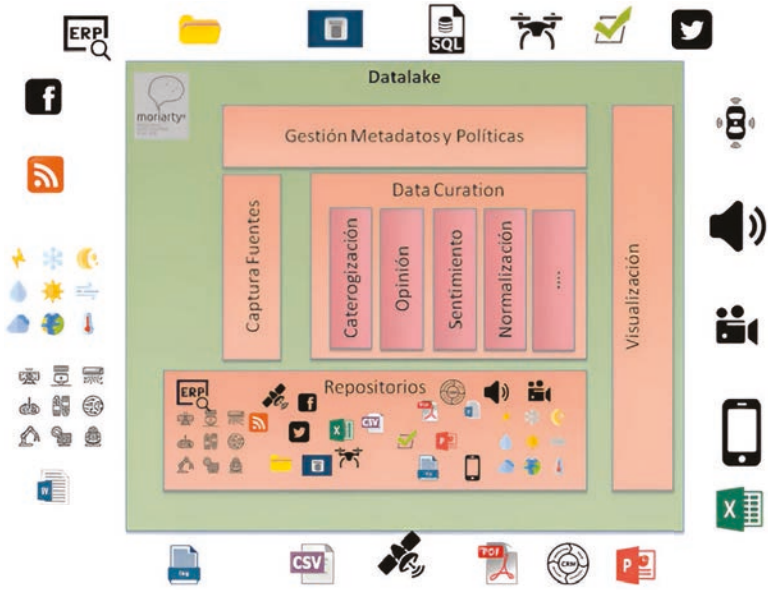

Figura 2.2. Características de un Data Lake

- Poseer herramientas para gobernar los datos, realizando una gestión de las políticas de acceso a los datos. Estas políticas deben soportar a todos los usuarios permitiendo el acceso controlado a los datos y garantizando el nivel de análisis requerido por distintos perfiles.
- Disponer de un catálogo centralizado del inventario de datos que incluya fuentes, versiones, veracidad y precisión de los datos. Sería deseable que este catálogo permitiese reflejar la cardinalidad de los datos y, además, guardar la traza de los datos.

Además, los Data Lakes deben ofrecer robustos mecanismos de seguridad y gobernanza, así como facilitar la gestión de metadatos y el descubrimiento de datos para garantizar un uso efectivo y responsable de la información almacenada. Su arquitectura abierta y flexible permite su adaptación a las necesidades cambiantes de las organizaciones y la integración con sistemas y herramientas existentes en el ecosistema de datos. En conjunto, estas características hacen que los Data Lakes sean una pieza fundamental en la infraestructura de datos moderna, capacitando a las organizaciones para obtener insights significativos y tomar decisiones informadas basadas en datos.

2.4 TIPOS DE DATA LAKES

Los Data Lakes son plataformas que permiten almacenar grandes cantidades de datos en su forma original, ya sea estructurada o no estructurada. Estos datos pueden provenir de una variedad de fuentes, como registros de servidores, datos de sensores, transacciones de negocios, redes sociales, entre otros. En términos generales, los Data Lakes están diseñados para almacenar datos de manera económica y escalable, y para proporcionar capacidades de análisis avanzadas.

En cuanto a los tipos de Data Lakes, aunque no hay una clasificación universalmente aceptada, podríamos agruparlos de la siguiente manera:

1. **Data Lakes en la nube**: se ejecutan en hardware y software en la nube de un proveedor. La mayoría sigue un modelo de suscripción de pago por uso. A medida que crecen los datos, simplemente compramos capacidad en la nube. El proveedor administra la seguridad, la confiabilidad, el respaldo de datos y el rendimiento para que podamos concentrar nuestros esfuerzos en determinar qué datos incluir y cómo analizarlos.
2. **Data Lake locales**: instala y ejecuta software para operar en servidores y almacenamiento en el centro de datos de una empresa. Se necesita una inversión de capital para comprar licencias de software y hardware, y experiencia en TI para instalarlo y administrarlo. Cada empresa es responsable de administrar la seguridad, proteger los datos y garantizar un rendimiento adecuado. Es posible que tengas que migrar el data lake a un sistema más grande a medida que crece. Sin embargo, un sistema local puede proporcionar un mejor rendimiento para los usuarios ubicados dentro de las instalaciones de la empresa.

3. **Data Lakes basados en archivos**: estos Data Lakes almacenan datos en su forma original, generalmente como archivos en sistemas de archivos distribuidos como Hadoop Distributed File System (HDFS), Amazon S3 o Azure Data Lake Storage (ADLS). Los archivos pueden ser de varios formatos, como Avro, Parquet, ORC, JSON, CSV, etc.
4. **Data Lakes basados en objetos**: similar a los Data Lakes basados en archivos, pero con un enfoque más orientado a objetos. Almacenan datos como objetos en un sistema de almacenamiento de objetos como Amazon S3, Google Cloud Storage (GCS) o Azure Blob Storage. Estos objetos pueden ser de cualquier tipo de datos y tamaño.
5. **Data Lakes basados en bases de datos**: estos Data Lakes utilizan sistemas de gestión de bases de datos (DBMS) para almacenar datos en su forma original o después de aplicar algún tipo de estructura. Ejemplos incluyen Amazon Redshift, Google BigQuery y Azure Synapse Analytics. Estos sistemas ofrecen capacidades de consulta y análisis sobre los datos almacenados.
6. **Data Lakes basados en almacenes de datos distribuidos**: estos Data Lakes combinan capacidades de almacenamiento de datos con capacidades de procesamiento distribuido. Un ejemplo popular es Apache Hadoop, que utiliza HDFS para almacenar datos y MapReduce o Apache Spark para procesarlos.
7. **Data Lakes híbridos**: estos Data Lakes combinan múltiples tecnologías y enfoques para satisfacer las necesidades específicas de una organización. Por ejemplo, podrían combinar almacenamiento basado en archivos con almacenamiento basado en bases de datos, o utilizar una combinación de servicios en la nube y en las instalaciones.

Cada tipo de Data Lake tiene sus propias ventajas y desventajas, y la elección del tipo adecuado depende de factores como los requisitos de almacenamiento, el presupuesto, la infraestructura existente y las necesidades de análisis de la organización.

2.5 ARQUITECTURA DE UN DATA LAKE

Un Data Lake consta de diversas capas, cada una desempeñando un papel en el flujo de datos y análisis:

- **Capa de ingesta**: la capa de ingesta desencadena el proceso de adquisición de datos desde una variedad de fuentes, abarcando tanto sistemas internos como externos. La recopilación de datos se hace activamente, programando extracciones periódicas y pasivamente, permitiendo la carga automática de datos al Data Lake en tiempo real.
- **Capa de almacenamiento**: una vez recopilados, los datos se almacenan en la nube utilizando servicios robustos y escalables que aseguran una gestión efectiva. En esta capa, hemos optado por aprovechar las ventajas de Azure Blob Storage

y Azure Data Lake Storage Gen2. Azure Blob Storage nos proporciona una plataforma confiable y altamente escalable para el almacenamiento de objetos, lo que garantiza que los datos se almacenen de manera segura y estén disponibles en cualquier momento. Por otro lado, Azure Data Lake Storage Gen2, con su capacidad para manejar grandes volúmenes de datos de diferentes formatos, permite una gestión ágil y eficiente de datos estructurados, semiestructurados y no estructurados.

- **Capa de procesamiento**: en la capa de procesamiento, los datos pasan por transformaciones esenciales para asegurar su calidad y utilidad. Aprovechamos la potencia de Azure HDInsight para llevar a cabo tareas de procesamiento distribuido a gran escala. Esta plataforma nos permite realizar procesamientos complejos de datos en paralelo, lo que acelera significativamente los tiempos de ejecución y mejora la eficiencia.
- **Capa de consumo**: finalmente, los resultados del análisis están disponibles para los usuarios finales en esta capa de consumo. Nos esforzamos por brindar acceso de manera eficiente y personalizada a la información valiosa que hemos extraído de los datos. Utilizamos interfaces web intuitivas y amigables que permiten a los usuarios explorar y consultar los datos según sus necesidades específicas. Además, proporcionamos APIs para aquellos que requieran una integración más profunda en sus propias aplicaciones. Para presentaciones y análisis visuales, aprovechamos herramientas líderes en la industria como Tableau y Power BI. Estas herramientas permiten la creación de visualizaciones interactivas y paneles de control que comunican de manera efectiva información compleja y resultados clave a los interesados.

En resumen, cada capa de nuestra arquitectura del Data Lake se diseña cuidadosamente para garantizar una recopilación, almacenamiento, procesamiento y consumo eficientes de datos, lo que lleva a una mejor toma de decisiones y a una ventaja competitiva. Por lo general, suele incluir los siguientes elementos:

- **Ingestión de datos**: es compatible con «conectores» y otros servicios que importan datos de múltiples fuentes estructuradas y no estructuradas.
- **Almacenamiento seguro**: debe poder almacenar y proteger un gran volumen de datos en expansión. La infraestructura que lo respalda debe escalar fácilmente y a un precio adecuado porque normalmente es imposible predecir todas las fuentes. También necesita estar protegido contra fallos del sistema y accesos no autorizados.
- **Gobernanza y conservación**: las empresas deben decidir qué datos se importan y cómo administrarlos. Los datos también deben catalogarse para que los profesionales puedan encontrarlos.
- **Procesamiento y análisis**: debe admitir una amplia gama de herramientas de análisis porque los profesionales usarán el data lake para diferentes tipos de análisis.

2.5.1 Beneficios de una arquitectura de Data Lake

El principal beneficio de un data lake es la centralización de diferentes fuentes de contenido. Una vez que estos datos están almacenados en una misma arquitectura, pueden ser combinadas y procesadas utilizando big data, búsquedas y análisis, que de otro modo no hubiera sido posible. Aun así, este tipo de arquitectura presenta algunos otros beneficios entre los que podemos destacar:

- En medidas de seguridad, un data lake permite otorgar acceso a cierta información a los usuarios que no tienen acceso a la fuente de datos original.
- Los datos son procesados según sea necesario, lo que reduce los costes de preparación sobre el procesamiento inicial.
- Una vez que los datos se encuentran almacenados en una arquitectura de data lake, pueden normalizarse y enriquecerse a través de la extracción de metadatos, conversión de formatos, aumento y extracción de entidades, agregaciones, indexaciones, entre muchas otras acciones.
- La arquitectura de los data lakes permite a las empresas generar diferentes tipos de informes, incluyendo la elaboración de informes sobre datos históricos, construcción de modelos de aprendizaje automático o aprendizaje profundo, entre otros. Estos informes permiten tomar decisiones en tiempo real y sugerir acciones o cambios para obtener mayores beneficios o un mayor rendimiento en la empresa u organización.

2.6 PROCESO DE CREACIÓN DE UN DATA LAKE

Aunque no existe una metodología estándar el proceso de creación de un data lake, se deberían considerar los siguientes pasos:

- **Adquisición de datos**: obtención de datos y metadatos, así como su preparación para una eventual inclusión en el Data Lake. Este proceso consiste en determinar qué datos, con qué granularidad (nivel de detalle), cuál es la frecuencia con la que se pueden obtener o si se pueden leer de una vez, etc. Para realizarse bien, hay que tener un conocimiento adecuado del uso que se quiere dar a los datos de cara a anticipar las necesidades de los usuarios.
- **Data Curation/Grooming data**: es el conjunto de procesos por los que los datos crudos son transformados en datos consumibles por las aplicaciones analíticas. Para ello, consideran los metadatos del paso anterior y aplican transformaciones a los datos para que puedan utilizarse.
- **Provisión de datos:** son el conjunto de procesos que permiten acceder a los datos contenidos en el Data Lake de acuerdo con las políticas que tiene establecidas.

- **Preservación de los datos:** son el conjunto de procesos y políticas que determinan qué datos deben conservarse, hasta cuándo y cuáles no. Otros objetivos de estos procesos es determinar cómo debe evolucionar la infraestructura para garantizar la disponibilidad de suficiente espacio y el rendimiento adecuado para acceder a los datos.

Un Data Lake debería permitir la ingesta de datos estructurados y no estructurados y almacenarlos con seguridad y con las debidas protecciones de acceso, incluso en tiempo real. Estas restricciones de acceso pueden ser de lectura o de modificación al nivel del dato e incluyen todas las capas de autenticación y autorización. Además, debe proporcionar un catálogo que permita a los analistas y profesionales descubrir los datos que contiene.

Es común que existan varias referencias a los mismos tipos de datos en un único Data Lake. Puede darse el caso de que los datos almacenados tengan etiquetas diferentes, pero se refieran al mismo concepto. De esta forma, se debe generar una relación para que los analistas conozcan su existencia.

Además de estos puntos, en cualquier Data Lake es necesaria su conexión con herramientas analíticas, de reporting, de procesamiento y de inteligencia artificial y de esta forma extraer valor de los datos de la organización.

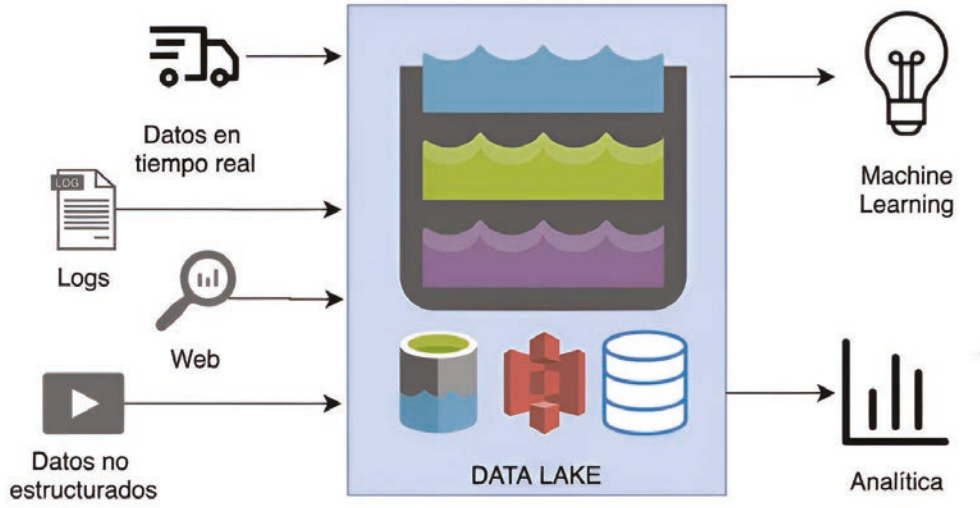

Figura 2.3. Creación de un Data Lake

Es importante anotar que, en cualquier proyecto de creación de Data Lake, hay que prestar especial atención al seguimiento y al desarrollo de un catálogo de los datos. No es suficiente con crear el Data Lake y volcar todos los datos en él, sino que debemos evaluar constantemente las oportunidades de sacar partido a estos datos.

2.7 DEFINICIÓN DE DATA WAREHOUSE

Un Data Warehouse es un sistema de almacenamiento centralizado, optimizado para almacenar datos históricos y actuales de una organización. Estos datos se estructuran y organizan de acuerdo con un esquema predefinido para facilitar el análisis y la generación de informes.

El objetivo principal de un Data Warehouse es proporcionar una fuente única y confiable de datos para la toma de decisiones empresariales y el análisis. A diferencia de otras bases de datos transaccionales, que están diseñadas para admitir operaciones de transacciones diarias, el Data Warehouse se enfoca en admitir consultas analíticas complejas y exhaustivas.

Un Data Warehouse integra datos de múltiples fuentes estructuradas, como sistemas de gestión de relaciones con clientes (CRM), sistemas de gestión de recursos empresariales (ERP) y bases de datos transaccionales. Estos datos se extraen, transforman y se cargan (ETL) en el Data Warehouse, donde se organizan y estructuran en un formato que facilite su análisis.

El esquema de datos en un Data Warehouse se establece de antemano y generalmente sigue un modelo dimensional o un modelo de estrella. Esto significa que los datos se organizan en dimensiones (características descriptivas) y hechos (medidas cuantitativas). Este enfoque facilita la realización de consultas y análisis eficientes mediante herramientas de business intelligence (BI).

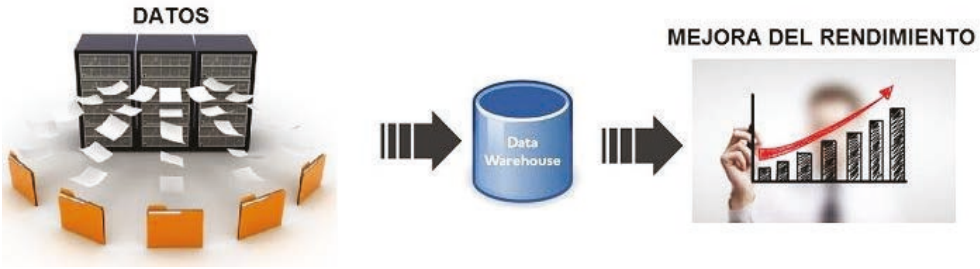

Figura 2.4. Esquema de un Data Warehouse

Los Data Warehouses también almacenan datos históricos a lo largo del tiempo, lo que permite realizar análisis retrospectivos y comparativos. Esto es esencial para identificar tendencias, patrones y comportamientos en los datos, lo que a su vez ayuda a tomar decisiones informadas y estratégicas.

Podemos definir el data warehouse como una especie de almacén digital en el que un negocio guarda gran parte de su información. Estos datos deben archivarse de forma segura, rápida y de fácil acceso. Además, en caso de pérdida debe tener la posibilidad de recuperarlo a través de algún mecanismo de recuperación. Toda esta información

puede venir de diferentes fuentes, pero tiene que almacenarse de forma organizada para acceder a ella cuando se requiera. De esta manera, los administradores pueden ejecutar los análisis que sean necesarios.

Generalmente, también se puede guardar en servidores físicos o en la nube. Este último ha sido el método más utilizado debido a su omnipresencia y mayor seguridad. De emplearse, se instalan aplicaciones específicas para extraer los datos que se necesitan en cada momento. De esta manera, los directivos pueden utilizar el data warehouse para manejar grandes cantidades de información que les permita tomar decisiones precisas. Podrán solicitar datos cuando lo requieran y modificarlos gracias al fácil manejo que tiene este tipo de sistemas.

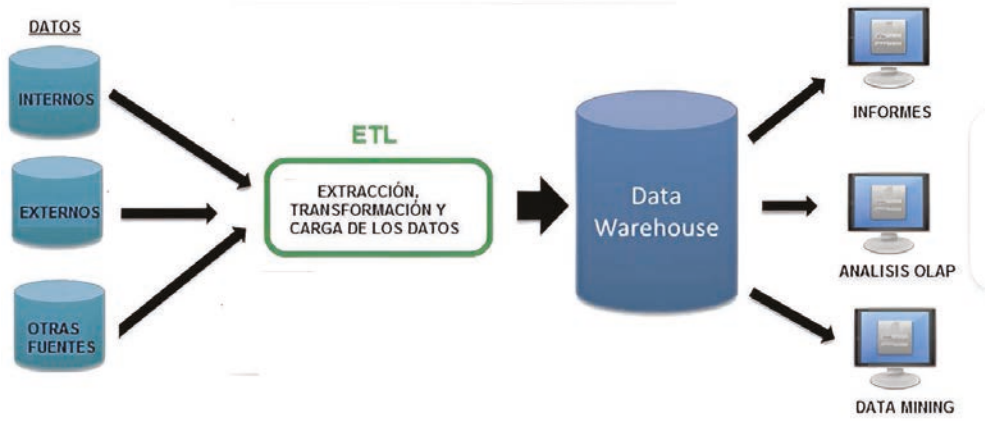

Figura 2.5. Esquema de una arquitectura Data Warehouse

2.7.1 Modelado bidimensional

El modelado dimensional es una técnica utilizada en la ingeniería de datos para diseñar la estructura de un Data Warehouse de manera eficiente y optimizada para el análisis de datos. Se basa en la creación de modelos que representan las dimensiones y los hechos de un conjunto de datos, siguiendo un enfoque intuitivo y fácil de entender. En el modelado dimensional, se distinguen dos conceptos principales: dimensiones y hechos.

- **Dimensiones**: las dimensiones representan las características descriptivas y contextuales de los datos. Pueden incluir elementos como fecha, producto, ubicación, cliente, entre otros. Cada dimensión tiene una tabla asociada en el modelo dimensional, donde se almacenan los atributos y las jerarquías correspondientes. Las tablas de dimensiones contienen columnas que describen los diferentes niveles de granularidad y los atributos relacionados con cada dimensión.

- **Hechos**: los hechos representan las medidas cuantitativas y numéricas que se analizan en el Data Warehouse. Pueden ser valores monetarios, cantidades, recuentos o cualquier otra medida relevante para el análisis. Los hechos se almacenan en una tabla de hechos en el modelo dimensional y se relacionan con las tablas de dimensiones mediante claves externas.

El enfoque central del modelado dimensional es la creación de esquemas de estrella y de copo de nieve.

- **Esquema de estrella**: en el esquema de estrella, una tabla de hechos central se conecta directamente a múltiples tablas de dimensiones. La tabla de hechos contiene las claves primarias de las dimensiones y las medidas numéricas. Este enfoque simplifica las consultas y los análisis, ya que se accede directamente a los datos en una única tabla de hechos.
- **Esquema de copo de nieve**: en el esquema de copo de nieve, las tablas de dimensiones se normalizan en múltiples niveles, lo que resulta en una estructura en forma de copo de nieve. Esto puede ayudar a reducir la redundancia de datos, pero también puede complicar las consultas al requerir más sentencias joins entre tablas para acceder a los datos.

El modelado **dimensional** tiene varias **ventajas** entre las que podemos destacar:

- **Facilidad de comprensión**: el modelado dimensional utiliza una estructura intuitiva y fácil de entender, lo que facilita la interpretación de los datos y la construcción de consultas.
- **Rendimiento optimizado**: el modelado dimensional permite un acceso rápido a los datos y un rendimiento eficiente en las consultas analíticas, ya que se minimiza la cantidad de joins necesarios y se evita la duplicación de datos.
- **Flexibilidad y escalabilidad**: el modelado dimensional es altamente flexible y se adapta bien a los cambios y adiciones futuras en las dimensiones y medidas. También es escalable, lo que significa que puede manejar grandes volúmenes de datos sin perder rendimiento.

2.7.2 Data Warehouse en la nube

Los data warehouses están atravesando actualmente dos transformaciones muy importantes que tienen el potencial de impulsar niveles significativos de innovación empresarial:

- La gran mayoría de los departamentos de TI están experimentando un rápido aumento de la demanda de datos. Los directivos quieren tener acceso a más datos históricos, mientras que, al mismo tiempo, los científicos de datos y los analistas de negocios están explorando formas de introducir nuevos flujos de datos en el almacén para enriquecer el análisis existente, así como impulsar nuevas áreas de

análisis. Esta rápida expansión de los volúmenes y fuentes de datos significa que los equipos de TI necesitan invertir más tiempo y esfuerzo asegurando que el rendimiento de las consultas permanezca constante y necesitan proporcionar cada vez más entornos para validar el valor de los nuevos conjuntos de datos.

- La segunda área de transformación gira en torno a la necesidad de mejorar el control de costes. Se necesita hacer más con cada vez menos recursos, a la vez que se garantiza que todos los datos sensibles y estratégicos estén completamente asegurados, a lo largo del ciclo de vida, de la manera más rentable.

2.8 DATA LAKES VS DATA WAREHOUSE

Es cierto que el Data Lake y el Data Warehouse tienen muchas similitudes, pero es importante que entiendas cuáles son sus diferencias. La primera de ellas es que el Data Lake conserva todos sus datos para la consulta de cualquier usuario. Esto contrasta con el data warehouse, que va excluyendo información si hay datos sin utilizar. Además, el Data Lake tiene la capacidad de soportar cualquier tipo de información sin importar la fuente. Estos se transforman solo cuando van a usarse para que el usuario final lo comprenda.

Aunque ambos paradigmas se centran en el almacenamiento de datos, hay algunas diferencias entre un Data Lake y un Data Warehouse entre las que podemos destacar:

- **Estructura de los datos**: un Data Warehouse solo recoge datos estructurados, mientras que un Data Lake recoge datos tanto estructurados como no estructurados.
- **Finalidad de los datos**: este aspecto puede estar definido o no en un Data Lake, mientras que en un Data Warehouse no hay lugar para la improvisación.
- **Flexibilidad**: en un Data Lake es más sencillo hacer cambios por no tener estructura, pero en un Data Warehouse es más complicado por estar implicados otros procesos.
- **Esquema**: los Data Lakes se basan en esquemas On Read y los Data Warehouses en los On Write.
- **Usuarios**: en un Data Lake los datos son manejados por analistas, mientras que en un Data Warehouse cualquier usuario con acceso puede manejar los datos.
- **Accesibilidad**: mientras que en un Data Lake hay una gran y fácil accesibilidad, en un Data Warehouse este apartado es más costoso y complejo.
- **Almacenamiento**: un Data Lake tiene un coste limitado con la posibilidad de ampliación en la nube, mientras que un Data Warehouse es por lo general más caro.

La diferencia entre estos dos conceptos no es estricta, ya que Data Lake se suele usar como término para herramientas de gestión y de almacenamiento de datos que no cumplen el concepto tradicional de Data Warehouse. Por tanto, el data warehouse suele ser un subconjunto de las tecnologías involucradas en el despliegue de Data Lakes. Los Data Warehouses son más lentos y complejos de implementar que los Data Lakes. A continuación, se hace un cuadro comparativo de las dos tecnologías.

DIMENSIÓN	DATA WAREHOUSE	DATA LAKE
Carga de trabajo	• Cientos de miles de usuarios concurrentes. • Realizar analíticas interactivas • Capacidades avanzadas de gestión de la carga de trabajo. • Procesamiento por lotes.	• Procesamiento por lotes de datos a escala. • Actualmente mejora sus capacidades para apoyar a usuarios más interactivos.
Esquema	• Normalmente, el esquema se define antes de almacenar los datos. • Requiere trabajo al comienzo del proceso, pero ofrece rendimiento, seguridad e integración.	• Normalmente, el esquema se define después de almacenar los datos. • Ofrece agilidad extrema y facilidad de captura de datos, pero requiere trabajo al final del proceso. • Funciona bien para tipos de datos donde no se conoce el valor de los datos.
Escala	• Grandes volúmenes de datos a un costo moderado.	• Volúmenes de datos extremos a bajo costo.
Acceso	• Datos a los que se accede mediante SQL estándar y herramientas de BI estandarizadas. • Método de búsqueda .	• Datos a los que se accede a través de programas creados por desarrolladores.
Ventajas	• Respuesta rápida. • Rendimiento consistente. • Fácil de usar. • Integración de datos. • Análisis funcional cruzado. • Cargue una vez, use muchos.	• Excelente escalabilidad. • Soporte de programación.
Costes	• Uso eficiente de CPU / IO.	• Bajo costo de almacenamiento y procesamiento.

Un Data Warehouse recopila datos de múltiples fuentes en una única ubicación independiente, contando además con una infraestructura analítica alrededor (metadatos, modelos relacionales optimizados, data lineage, etc). Se trata de sistemas costosos, con alta disponibilidad y velocidad de acceso.

Los **Data Warehouses** abrieron las puertas del **procesamiento analítico.** El pasado se convirtió en un gran predictor del futuro. El almacenamiento de datos históricos y su análisis pronto comenzó a tener un gran valor. Por otro lado, los Data Warehouses estaban diseñados para la generación de reportes y el BI, el tratamiento de datos estructurados y basados en transacciones. La evolución tecnológica siguió su curso descubriéndose limitaciones en los Data Warehouses clásicos.

Con la finalidad de almacenar todos los datos generados por una organización, surgen los Data Lakes, capaces de almacenar grandes volúmenes de datos en brutos, con la esperanza de que, llegado el momento, si esos datos son necesarios, bastaría únicamente con indagar dentro del Data Lake y, una vez encontrados, cogerlos y manipularlos. Rápidamente se descubrió que usar los datos de un Data Lake no es tan sencillo por las siguientes razones:

- Dependiendo del tipo de usuario, sus necesidades son completamente distintas. Por ejemplo, un usuario de negocio frente a un científico de datos.
- El gran volumen de datos almacenado dificulta diferenciar qué datos son útiles de los que no.
- Al guardar los datos en bruto no se incluyen metadatos de estos.
- Problemas para saber si se tratan de datos actualizados o saber cuán precisos y veraces son.
- La falta de gobernanza en los datos y de optimización en los procesos ha dado lugar a que muchos de los Data Lakes actuales no hayan sido exitosos.

2.8.1 Conversión de los datos

Durante el desarrollo de un data warehouse, se gasta una cantidad considerable de tiempo analizando las fuentes de datos, entendiendo los procesos de negocio y perfilando los datos. El resultado es un modelo de datos altamente estructurado diseñado para la generación de informes. Una gran parte de este proceso incluye tomar decisiones sobre qué datos incluir y no incluir en el almacén. Generalmente, si los datos no se utilizan para responder a preguntas específicas o en un informe definido, pueden excluirse del almacén. Esto se hace generalmente para simplificar el modelo de datos y también para conservar el costoso espacio en el almacenamiento de disco que se utiliza para hacer el data warehouse.

En contraste, el Data Lake conserva todos los datos. No solo los datos que se utilizan, sino los que podrían utilizarse algún día. Los datos también se mantienen todo el tiempo para que podamos volver en el tiempo a cualquier punto para hacer el análisis.

Este enfoque es posible porque el hardware para un data lake suele ser muy diferente del utilizado para un data warehouse. La ampliación de un data lake a terabytes y petabytes puede hacerse de manera bastante económica.

2.8.2 Soporte de los tipos de datos

Los Data Warehouse generalmente se componen de datos extraídos de sistemas transaccionales junto con métricas cuantitativas y los atributos que las describen. Las fuentes de datos no tradicionales, como los registros del servidor web, los datos de sensores, la actividad de las redes sociales, el texto y las imágenes, se ignoran en gran medida. Se siguen encontrando nuevos usos para estos tipos de datos, pero consumirlos y almacenarlos puede ser costoso y difícil.

El enfoque de un Data Lake abarca estos tipos de datos no tradicionales donde guardamos todos los datos independientemente de la fuente y la estructura. Los mantenemos en su forma bruta y sólo los transformamos cuando estamos listos para usarlos. Este enfoque se conoce como "Schema on Read" en comparación con el "Schema on Write" que es el enfoque utilizado en el data warehouse.

2.8.3 Adaptación a los cambios

Uno de los principales problemas sobre los data warehouses es cuánto tiempo se tarda en cambiarlos. Un tiempo considerable se gasta por adelantado durante el desarrollo de la estructura del almacén. Un buen diseño de almacén puede adaptarse al cambio, pero debido a la complejidad del proceso de carga de datos y al trabajo realizado para facilitar el análisis y la elaboración de informes, estos cambios podrían requerir más tiempo del inicialmente estipulado.

Muchas cuestiones relacionadas con el negocio, en ocasiones no pueden esperar a que el equipo que gestiona el data warehouse adapte su sistema para responderlas. De esta forma, la necesidad cada vez mayor de respuestas más rápidas es lo que ha dado lugar al concepto de auto-servicio de inteligencia empresarial.

En el Data Lake, por otro lado, como todos los datos se almacenan en bruto y siempre con accesibles a alguien que necesite utilizarlos, los usuarios tienen el poder de ir más allá de la estructura del almacén para explorar datos de nuevas maneras y responder a sus preguntas a su ritmo.

2.9 CAPAS DE UN DATA LAKE

Los Data Lakes se dividen típicamente en zonas o capas para organizar y gestionar los datos de manera eficiente. Estas zonas representan diferentes niveles de procesamiento y gobernanza sobre los datos almacenados. A continuación, se describen cada una de las capas de un Data Lake:

- **Landing**: la capa de "Landing", también conocida como zona de aterrizaje o de "raw data", es la primera capa en un Data Lake. Aquí los datos aterrizan desde diversas fuentes, mediante el uso de pipelines, sin transformaciones significativas. Los datos se almacenan en su forma bruta, preservando su integridad original.

En esta capa, se pueden incluir datos estructurados, semiestructurados y no estructurados, por eso los formatos a usar aquí pueden variar.

- **Trusted**: la capa Trusted es donde los datos crudos se transforman y preparan para un uso más amplio. Aquí se aplican procesos de limpieza, normalización, de duplicación, validación y otras transformaciones para mejorar la calidad y la estructura de los datos. Esta capa tiene como objetivo ofrecer datos de calidad, más estructurados y listos para los equipos de análisis y ciencia de datos. En esta capa, se suele trabajar con formatos columnares como Parquet.
- **Refined**: la capa Refined es la capa final en un Data Lake, donde se ofrecen datos enriquecidos y de valor para la organización. Aquí se aplican reglas de negocio y se realizan agregaciones, cálculos, transformaciones más avanzadas y se optimizan los datos para casos de uso específicos. Esta zona se utiliza para generar informes, paneles de control, visualizaciones y otros productos de datos que brinden información valiosa y respalden la toma de decisiones empresariales. Los datos en esta zona están altamente estructurados y están diseñados para ser utilizados por aplicaciones empresariales y usuarios finales de manera eficiente y rápida.

La división en capas proporciona una estructura lógica para gestionar y organizar los datos en el Data Lake. Cada capa tiene un propósito específico y ofrece distintos niveles de procesamiento y calidad de los datos. Las tres zonas mencionadas proporcionan una estructura común para organizar, preparar y utilizar los datos en un Data Lake. Ahora bien, es posible sumar nuevas capas de acuerdo con la necesidad de la organización. A continuación, se describen dos capas adicionales o complementarias.

- **Sensitive**. En esta zona, se almacenan y se gestionan datos sensibles que requieren un tratamiento especial debido a su naturaleza confidencial o regulaciones de privacidad. Esta zona está diseñada para garantizar la seguridad y el cumplimiento normativo de los datos sensibles. Aquí se aplican reglas y medidas específicas de seguridad, como el cifrado, el control de acceso restringido para proteger la confidencialidad y la integridad de los datos.
- **Sandbox**. Es un espacio dedicado a la experimentación y la colaboración para que los equipos de ciencia de datos puedan acceder a una copia de los datos en el data lake para realizar pruebas, prototipos y experimentos sin afectar los datos en las capas principales. La zona de sandbox proporciona un entorno seguro para probar nuevas ideas, explorar modelos de machine learning, desarrollar algoritmos y realizar investigaciones sin riesgo de impacto en las otras capas del Data Lake. Los datos en esta zona se utilizan para el desarrollo y la validación de modelos antes de implementarse.