Introducción a Big Data
Conceptos asociados

Competencias

- Conocerá las características de una solución Big Data.
- Comprender los usos del Big Data.
- Conocer casos de éxito de soluciones Big Data.
- Comprender los distintos paradigmas de soluciones Big Data.
- Conocer los conceptos de OLTP y OLAP.
- Conocer el flujo ETL y ELT.
- Conocer los conceptos de Data Lake, Data Warehouse y Data Mart.

Motivación

En la actualidad, todos los sectores buscan impulsar sus negocios en función de los datos para establecer mejores relaciones con sus clientes, optimizar sus operaciones y/o crear nuevos productos. Para lograr esto, las empresas han tenido que recurrir y aprovechar todas las fuentes de información que tienen disponibles y también descubrir otras nuevas.

Todos los días, en el mundo, se crean más de 2.5 quintillones de bytes. Estos datos provienen de múltiples fuentes: Mensajes de texto, fotos, videos, correos, transacciones de compras, señal GPS de los teléfonos, actividad en redes sociales, entre otros. Esta explosión también es impulsada por el Internet de las Cosas y el constante registro de la información de sensores para medir el clima, la humedad, la temperatura y casi cualquier otra cosa que sea posible medir.

La necesidad de almacenar todos estos datos constituye en sí un desafío, pero ha sido abordado exitosamente por los investigadores y profesionales, logrando almacenar grandes cantidades de información a bajo costo. Pero cuando hablamos de análisis de estos enormes conjuntos de datos, incluso las mejores plataformas TI se pueden ver sobrepasadas.

Afortunadamente, la comunidad Open Source ha abordado estos desafíos, realizando grandes aportes en las capas de almacenamiento y procesamiento. Principalmente, impulsados por el proyecto Apache Hadoop que surge como la principal plataforma de análisis para el Big Data. Cientos de desarrolladores que son parte de este proyecto, lo cual ha dado origen a que muchos otros proyectos se integren a él, creando un ecosistema de soluciones (Hadoop Ecosystem) a los que es posible recurrir para resolver las necesidades de análisis de datos de las empresas.

¿Pero qué es el Big Data?

Considere el siguiente ejemplo: Si usted tenía un negocio que operaba una flota de tiendas hace 20 años, la mayor parte de los datos de interacción con el cliente provendrían de sus propias cajas de facturación o registradoras en tiendas físicas. Mediante estas se podía rastrear tanto las compras (monto, artículos, descuentos) como los medios de pago.

Hoy en día es probable que esté operando de forma digital en la web y, de manera adicional a los datos precisos de compra, puede tener mucho más. Para empezar, está obteniendo registros web detallados que generan un seguimiento de cómo interactúan los clientes con su sitio web. Luego, rastreando el perfil de cada cliente o prospecto individual, se puede saber cuándo se dirigen a ciertas promociones o incluso qué artículos pueden estar interesados en comprar. Todo eso basado en el comportamiento pasado o la demografía a la que pertenecen.

Fuera de la huella digital, también puede rastrear la confianza de sus clientes en las redes sociales y en los motores de búsqueda que utilizan para acceder al sitio web (canales). Y si eso no fuera suficiente, también puede hacer un seguimiento de sus clientes en el mundo real mediante el seguimiento de direcciones IP de acceso y las identificaciones de celulares.

Sólo almacenar esta cantidad de información se convierte en un desafío importante. Además, las formas tradicionales para el análisis de datos no dan abasto con esta cantidad de información. Comienza entonces a hablarse de Big Data para referirse a grandes volúmenes de datos que cumplan con una o más de estas 3 características:

- **Volumen**: La cantidad de datos es enorme y abarca horizontes de tiempo más grandes. Normalmente estas bases de datos logran superar los varios gigabytes de información e incluso más.

- **Velocidad**: Los datos se generan con mucha frecuencia y rápidamente se acumulan para formar enormes bases de datos. Pensemos por ejemplo en la cantidad de búsquedas que se hacen en Google en un segundo o la cantidad de tweets en Twitter. Además, estos datos caducan rápidamente... volviéndose irrelevantes si no son aprovechados al poco tiempo de ser recibidos.

- **Variedad**: Tiene que ver con el hecho de que ahora existen múltiples fuentes de información que recogen elementos muy variados del negocio. A esto se agrega que mucha de esta información se almacena de distintas formas, teniendo que integrar datos estructurados, no estructurados y semiestructurados dentro de un mismo análisis.

Estas características son conocidas como las 3 V del Big Data. Si bien esto podría ser una simplificación para el entendimiento de las componentes de una solución Big Data, en el último tiempo se ha convertido en la definición más adoptada.

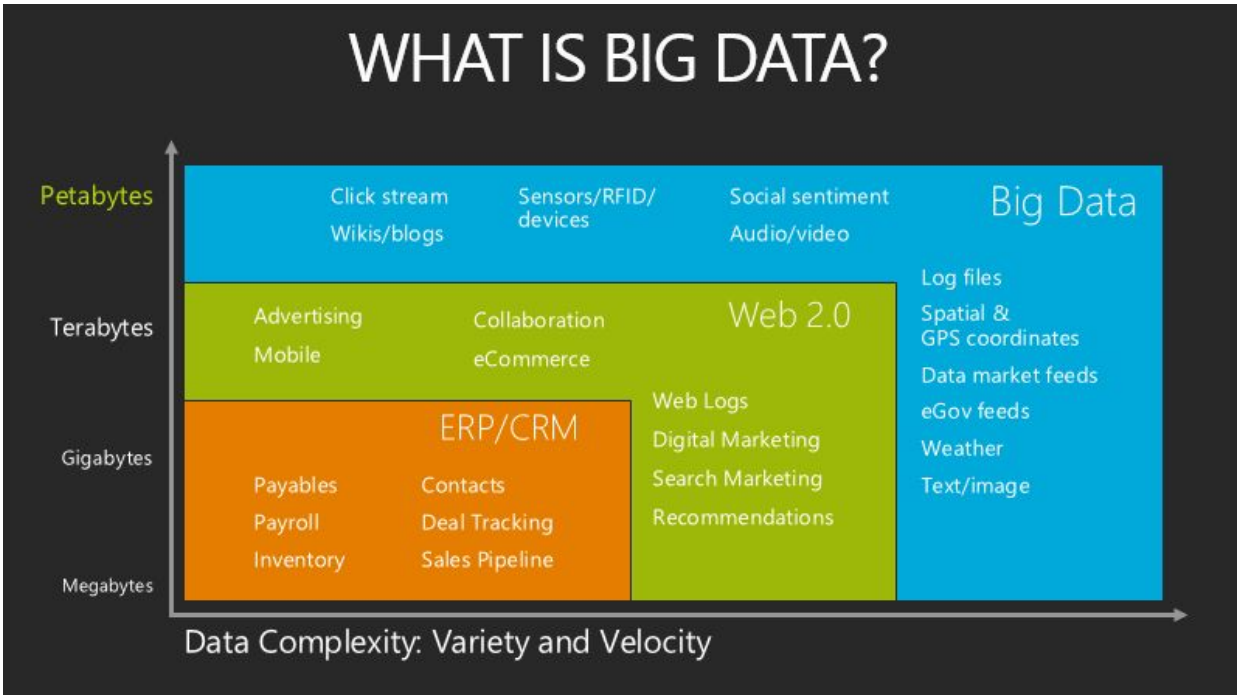

Imagen 1. What is big data.

Oportunidades e impacto del Big Data

Muchas de las oportunidades ligadas a la tecnología, se encapsulan hoy dentro del concepto de *Transformación Digital*. La transformación digital relacionada con Big Data se construyen típicamente en relación del fin que persigue respecto a la organización, y de la actividad clave que se realiza para dicho fin; Por un lado las iniciativas asociadas a Big Data pueden buscar nuevas oportunidades de negocio, o bien la reducción de costos. En la línea de nuevas oportunidades se encuentran tres categorías claves de actividades: Descubrimiento de datos, vista única, y análisis predictivo. En relación a la reducción de costos, las categorías de actividades claves son: Se debe tener en cuenta que no existe un orden para abordar esas iniciativas. No existe un paso a paso y la trayectoria hacia una arquitectura Big Data varía dependiendo de cada organización y sus necesidades.

Nuevas oportunidades de negocio

- **Descubrimiento de datos**: Muchas empresas comienzan su camino hacia Hadoop con el descubrimiento de datos; es decir, intentar resolver el problema para explorar datos que nunca han estado disponibles para ellos a esa escala (Big). Esto permite consolidar datos dispersos, explorar tipos de datos que nunca han tenido a escala, descubrir *insights* a partir de esos datos. Finalmente, se crean nuevas y mejores estrategias relacionadas al gobierno de datos. La búsqueda de este resultado es común para las empresas que tienen un largo historial de operaciones, y hay mucha diversidad en la forma en que un equipo gestiona diferentes tipos de datos, especialmente cuando se trata de gestionarlos para diferentes líneas de negocios dentro de la misma organización. Un resultado exitoso en torno al esfuerzo de descubrimiento de datos generalmente conduce a una sola vista.

- **Vista unificada**: Un descubrimiento exitoso de datos es una situación en la que las empresas se dan cuenta de que están sentados en un tesoro de datos, pero está dispersa por los distintos equipos de la organización. Incluso es posible que no se pueda analizar efectivamente los datos dentro de un silo de almacenamiento dado. Por ejemplo, para una única base de datos relacional o una solución de almacenamiento, la tarea de poder correlacionar conjuntos de datos en las ventas se convierte en un gran problema. Si volvemos al ejemplo del negocio que manejaba una cadena de tiendas, si se tienen múltiples correos electrónicos e identificadores de redes sociales, todos vinculados al mismo cliente, pero almacenados en sistemas distintos (CRM, ERP, u otro), nunca se podrá descubrir que el usuario "Carlos" puede ser atendido por distintas campañas publicitarias. En una situación como esta, la compañía comienza a usar Hadoop para unificar todos los datos y trazar una vista única de su negocio. Este tipo de unificación se llama un Data Lake basado en

Hadoop. El Data Lake crea una vista única del cliente, una vista única del paciente, una vista única de la red de telecomunicaciones, una vista única de la cadena de suministro, una vista única del producto o una vista única de casi cualquier otra cosa. Esta información completa llevada a Hadoop a través del Data Lake, lleva naturalmente al tercer propósito y oportunidad: análisis predictivo.

- **El análisis predictivo**: El análisis predictivo es el resultado empresarial más sofisticado posible con Hadoop. Las empresas ahora pueden predecir resultados futuros con mucha más certeza. Y en realidad, a veces, parece que las compañías son claramente clarividentes. Por ejemplo, las empresas pueden usar el análisis predictivo para predecir demandas en ciertos productos, o recomendar productos para comprar, o incluso modelar el riesgo financiero.

Cabe destacar, que la gran mayoría de los modelos predictivos ven aumentados su performance, a medida que se le entregan mayores cantidades de datos (Big Data). Pero, ¿Cómo aplicarlo a las necesidades de una organización? Sabemos que queremos llegar a las 3 oportunidades, ¿Pero cómo trazar un roadmap de lo que es más prioritario para nuestra realidad? Una buena metodología corresponde es pensar en cómo se relaciona nuestra empresa con los problemas de gestión de datos que ya existen.

Más específicamente y en consideración de las 3 oportunidades presentadas:

- Características de casos de uso que necesiten de descubrimiento de datos:
    - **Pequeña retención de datos**: Cuando se vuelve costoso almacenar información durante mucho tiempo y termina borrándose, utilizando simplemente lo necesario para la continuidad operacional.
    - **Alto nivel de fragmentación en los datos**: Cuando la organización está almacenando la data necesaria, pero ésta se encuentra dispersa en distintos sistemas de gestión de datos.
    - **Variedad de fuentes de información**: Cuando se quieren combinar datos de distintas fuentes pero con estructuras que son distintas y variadas.

- Características de casos de uso que necesiten de vista unificada:

    - **Puntos ciegos respecto a la visión del cliente**: Cuando los tomadores de decisiones de la organización sólo ven una parte de la relación e interacción con el cliente.

    - **Fragmentación de almacenamiento**: Cuando las plataformas de almacenamiento de datos son heterogéneas.

    - **Fallida colaboración**: Cuando diferentes equipos dentro de la misma organización preguntan lo mismo repetitivamente.

    - **Análisis de fallas y errores incompletos**: Cuando se hace difícil encontrar la causa a los errores en distintas actividades de los procesos de negocio.

- Características de casos de uso que necesiten de análisis predictivo:

    - **Falta de mantención preventiva**: Cuando el equipo tiene pocas o nulas capacidades de reparar o corregir una situación antes que ésta se vuelva un problema.

    - **Capacidad insuficiente para tomar decisiones**: Cuando la organización sólo realiza un subconjunto de las predicciones que necesita.

    - **Ocurrencia de eventos conocidos y familiares**: Cuando muchas personas dentro de una organización saben que existirá un problema, pero se ven sorprendidos cuando esto ocurre y no están preparados para afrontarlo.

    - **Falta de confianza en las decisiones basadas en datos**: Cuando una organización realiza predicciones, pero con alta dependencia del juicio experto.

Características y objetivos de una solución Big Data

1. Mejoras en el sistema de gestión de datos

Los sistemas de gestión de datos, son diseñados y optimizados en función de 2 principales casos de uso:

- **Procesamiento de transacciones**, también conocido como **OLTP** (OnLine Transactions Processing Systems).

- **Análisis de datos, procesamiento** y generación de reportes (llamado también **OLAP**, o OnLine Analytical Processing Systems).

Ambos términos cuesta diferenciarlos hoy en día. Pero de manera informal, se puede relacionar OLTP con los sistemas diseñados para manipular un gran conjunto de solicitudes finales de usuario, tan rápido como sea posible pero al mismo que mantiene la integridad del sistema. A su vez, se puede pensar en OLAP como los sistemas diseñados para manejar consultas que podrían llegar a ser mucho más complejas sobre grandes cantidades de datos con el propósito de apalancar las decisiones de la organización mediante la generación de informes. Cuando un cliente compra en una tienda online, generalmente interactúa con un sistema OLTP que administra el carrito de compra y permite el pago final. Pero el gerente de ventas del mismo comercio, y su equipo, se relacionan con los informes generados por los sistemas OLAP para tener información sobre las proyecciones del próximo trimestre.

La siguiente tabla resumen presenta las principales diferencias existentes entre los casos de uso OLTP y OLAP

Dimensión	OLTP	OLAP
Naturaleza	Transacciones constantes en cuanto a queries	Actualizaciones periódicas de largo alcance. Uso de queries multidimensionales con múltiples fuentes de datos
Ejemplos	Bases de datos de contabilidad. Bases de transacciones de retail	Reportería. Apoyo en toma de decisiones
Tipo de datos	Datos Operacionales	Datos Consolidados
Periodo de retención	Corto Plazo (2-6 meses)	Largo Plazo (2-5 años)
Almacenamiento	GB	TB/PB
Cantidad de Usuarios	Muchos	Pocos
Protección	Protección constante de datosRobust, constant data protection and fault tolerance	Protección periódica

Tabla 1. OLTP vs OLAP.

2. Reducción de costos

Incluso si la organización cree no estar lista para una transformación digital que persiga nuevas oportunidades de negocio, de igual manera puede volcar su tecnología a una arquitectura Big Data con el fin de la reducción de costos. A diferencia de las categorías anteriores, estos casos de uso no buscan crear una revolución, sino más bien optimizar los activos existentes. También proporcionan una introducción más suave de las tecnologías de Big Data en el entorno de TI existente. Las tres categorías principales de ahorro de costos corresponden a:

- **Archivo Activo**: Se centra en convertir el almacenamiento de datos de su empresa en un "archivo activo". Esto podría abarcar desde poner los registros históricos a disposición de los analistas de negocios, hasta descargar algunos de los datos de los mainframes y otros sistemas legacy. Los archivos activos almacenan datos que no requieren mucha lectura o escritura y no necesitan ser modificados, pero pueden leerse de vez en cuando. Todo eso reducirá drásticamente el costo total de propiedad de los datos y ayudará a su empresa a optimizar su balance final.

- **Explotación de ETL's**: ETL significa Extract, Transform & Load. Debido a que las soluciones de almacenamiento de datos empresariales tradicionales suelen utilizar cerca del 60% de sus ciclos en promedio solo para preparar datos para un esquema particular, moviendo esa fase, que es exactamente la fase ETL a Hadoop, les permite a estos sistemas usar sus costosos recursos para lo que realmente fueron diseñados, que es básicamente almacenamiento y análisis. Hadoop posee un esquema que permite la lectura de esquemas, lo que significa que puede almacenar datos en su formato nativo, y recuperar los datos que necesite analizar, independientemente de su origen, formato o estructura de datos. Cabe destacar que en este caso, está utilizando Hadoop y Big Data no en lugar de, sino junto a las soluciones EDW existentes.

- **Enriquecimiento de datos**: Finalmente, el caso de uso de enriquecimiento de costos más avanzado es enriquecer los datos existentes con los nuevos datos disponibles. Con estas capacidades de enriquecimiento de datos, las empresas pueden incorporar conjuntos de datos disponibles públicamente de fuentes como Twitter, Facebook, datos gubernamentales abiertos, u ofrecer nuevos productos o servicios. También pueden prevenir el fraude al encontrar nuevas correlaciones de datos que apuntan a un mal comportamiento.

Se puede apreciar que incluso si se comienza con el ahorro de costos como su principal objetivo, rápidamente obtiene un uso bastante sofisticado de Big Data, y puede comenzar a proporcionar valor más allá del objetivo inicial de ahorro de costos.

Al igual que con las otras categorías, podríamos descubrir qué oportunidad nos conviene adoptar, si nuestra organización se ve enfrentada a las siguientes problemáticas:

- Característica de casos de usos en relación a Archivo Activo:

    - **Ingesta restringida**: Cuando se niega a capturar nuevas fuentes debido a temores sobre los costos.

    - **Desafíos presupuestarios**: Cuando los costos de almacenamiento crecen más rápido que los presupuestos de almacenamiento.

- **Difícil acceso**: Cuando los archivos "fríos" son almacenados de tal manera que son difíciles de recuperar.

- **Nuevos tipos de datos**: Cuando los nuevos archivos poseen estructuras de datos difíciles de explotar.

- **Retención limitada**: Cuando la organización elimina los archivos más rápidamente de lo que quisiera.

- Característica de casos de usos en relación a Explotación de ETL's:

  - **Aplicaciones imposibilitadas de operar**: Cuando la incapacidad de integrar nuevas fuentes de datos no estructurados bloquea la creación de nuevas aplicaciones.

  - **Ignorar fuentes de datos**: Cuando la empresa simplemente ignora los datos que son demasiado difíciles de transformar.

  - **Incertidumbre analítica**: Cuando los analistas de negocios desean poder examinar los datos sin procesar para ver si se perdió algo en la transformación.

- **Asignación de capacidad ineficiente**: Cuando los recursos del Data Warehouse se consumen simplemente en preparar los datos para la ingesta del análisis.

- Característica de casos de usos en relación al Enriquecimiento de Datos:

  - **Procesos de validación**: Cuando los analistas de su negocio no detectan datos fraudulentos comparándolos con otros registros internos o externos.

  - **Llenado de datos**: Cuando sabe que se puede llenar algunos vacíos en los datos con fuentes externas.

  - **Datos obsoletos**: Cuando la organización no puede actualizar con frecuencia datos de fuentes externas.

  - **No existe conciencia del valor de los datos externos**: Cuando la ingesta de obstáculos bloquea el enriquecimiento potencial de los datos y este proceso se ve descartado.

El proceso ETL (Extract - Transform - Load)

Por ETL hacemos referencia a un proceso donde extraemos los datos de un sistema de almacenamiento, los transformamos y posteriormente los cargamos en un flujo de trabajo analítico. Este proceso es necesario dado que busca resolver dos problemas:

- Por un lado, tenemos bases de datos relacionales como MySQL y PostgreSQL que son excelentes para el desarrollo de trabajos transaccionales, así como en la lectura y actualización de registros de manera simple. Dado su énfasis en la mantención de los datos y la creación de consultas simples, no permiten conducir análisis de grandes cantidades de datos.

- Por defecto una base de datos relacional debe definir sus componentes antes de alojar los datos. El problema asociado a este punto es que en el advenimiento de grandes cantidades de datos de múltiples fuentes, un esquema predefinido no podrá soportar los grandes flujos de información.

Formas de Implementación ETL

Existen dos variantes de implementación:

- **Proceso clásico (ETL)**: Los datos se extraen de una base de datos OLTP. Posterior a su extracción son llevados a una etapa intermedia llamada staging area, donde se implementan transformaciones asociadas a la limpieza y optimización de datos para el posterior análisis. Una vez que estos datos se procesan, se cargan a una base de datos OLAP. Es en esta última donde los equipos de Business Intelligence generan queries y consultas, que eventualmente se presentarán a los usuarios finales, o serán implementadas como inputs para modelos predictivos.

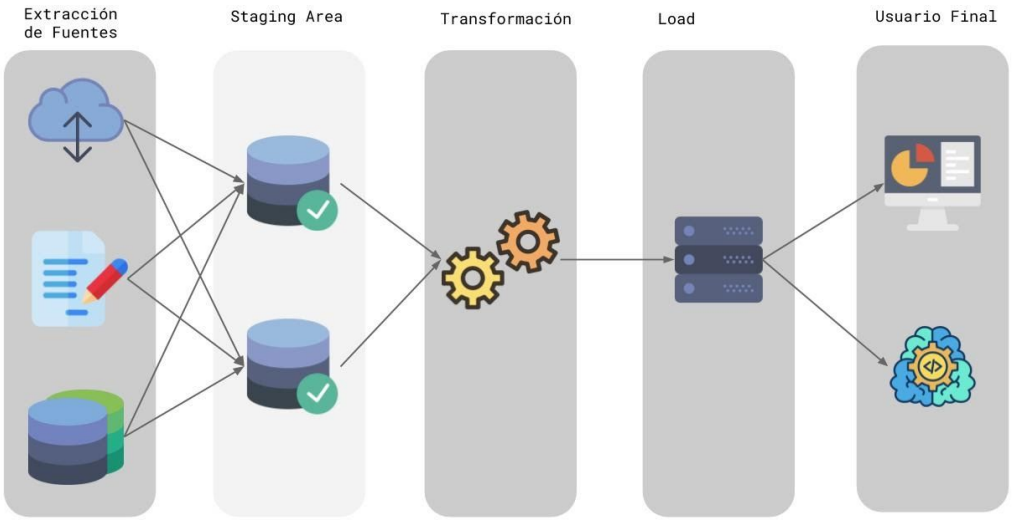

Imagen 2. ETL.

- **Proceso moderno (ELT)**: Uno de los problemas del ETL clásico es el hecho que si las consultas realizadas por el equipo de Business Intelligence no son las correctas, el proceso debe volver a generarse, siendo ineficiente en el mediano y largo plazo. En la actualidad, dado la existencia de soluciones cloud como Amazon RedShift y Google BigQuery, nos permiten desarrollar las transformaciones y carga de datos in situ, por sobre la necesidad de generar un área intermedia de staging. Así, las transformaciones por lo general se exportan en archivos de texto plano, lo que facilita la creación de soluciones más livianas con menos intermediarios y tecnologías asociadas.

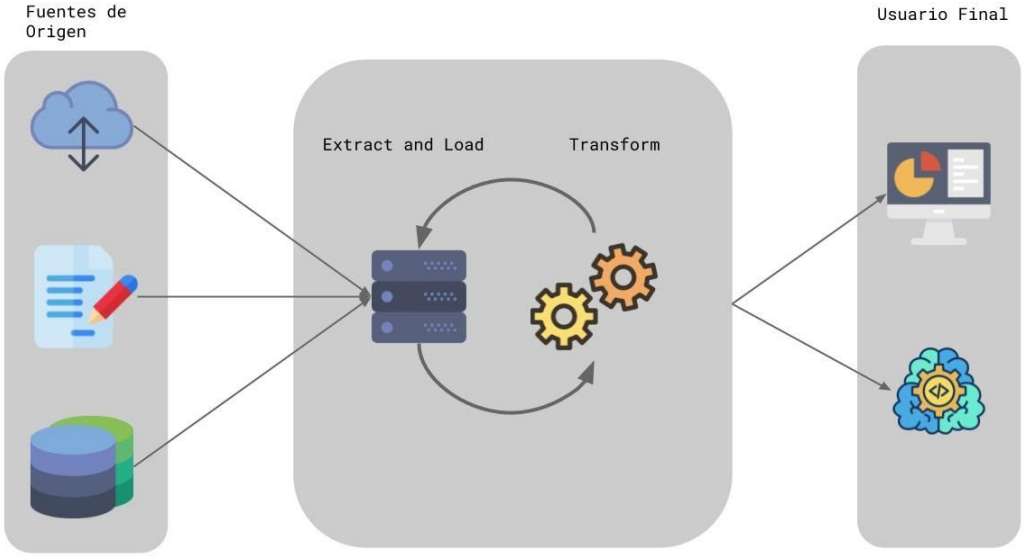

Imagen 3. ELT.

Para que un sistema ETL sea exitoso en el suministro de datos a sus usuarios finales, debe tener en consideración una serie de componentes críticos para el desempeño de la tarea:

- Un buen ETL debe tener un apoyo para capturar los cambios en los datos. La carga incremental de los datos permite actualizar los datos del warehouse sin la necesidad de generar una recarga innecesaria de éstos.

- Un ETL debe tener en consideración la existencia de un registro auditable detallado que asegure que el flujo ETL pueda ser auditado posterior a la carga de los datos.

- Un sistema ETL necesita ser capaz de recuperar los datos perdidos de manera fácil, asegurando que los datos puedan transformarse de un lugar a otro.

- Algunas decisiones de análisis deben realizarse en tiempo real. Si bien siempre existirán contratiempos en la integración de tipos de datos, el flujo ETL debe estar orientado a la minimización de la latencia.

Data Lake, Data Warehouse y Data Mart

Parte importante del gobierno de datos es la creación de una arquitectura clara que permita delegar la información a distintas partes de una empresa. Al momento de pensar en la implementación de un proceso de extracción de información, es importante saber cuál va a ser el caso de uso final asociado. Así, es importante conocer las diferencias existentes entre los conceptos Data Lake, Data Warehouse y Data Mart con los que se habla en la industria.

Data Lake

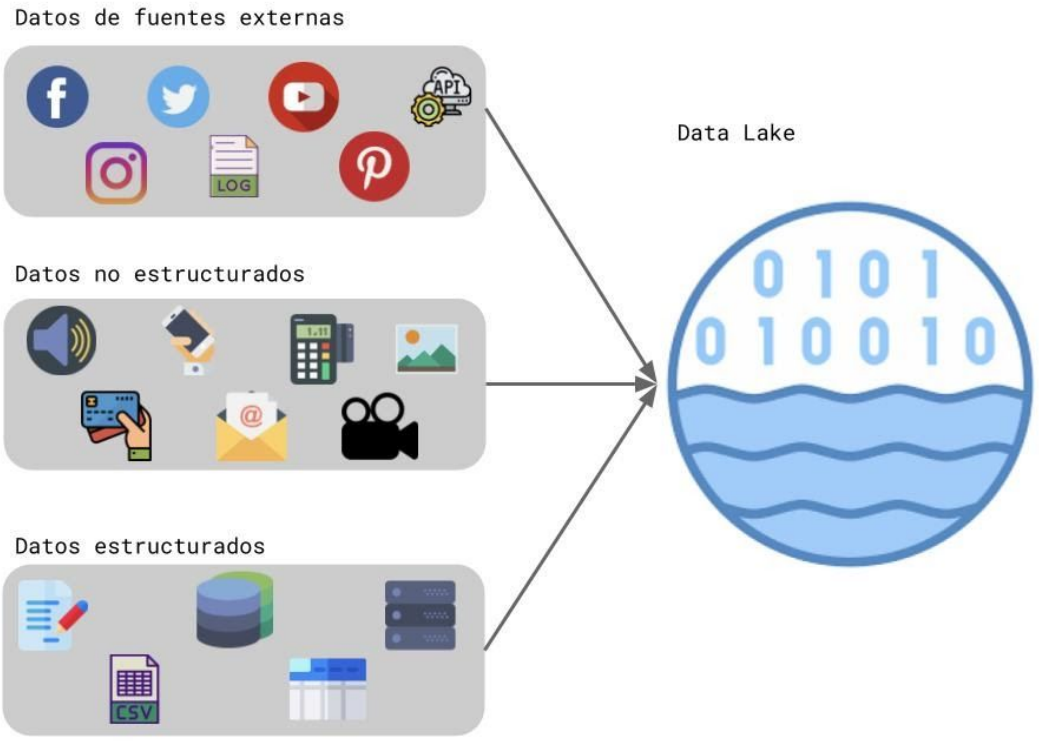

Imagen 4. Data Lake.

Un Data Lake es el lugar donde se puede alojar absolutamente todas las formas de datos generados por un negocio: Ingresos de datos estructurados, registros de sistema y de mails, imágenes, videos. Esta colección de datos no filtra información alguna ni establece criterios de selección para su ingreso. Así, datos asociados a transacciones inválidas o canceladas también serán capturadas por el Data Lake.

Un Data Lake es relevante en dos circunstancias:

- Cuando los datos de la organización no pueden ser administrados por un esquema. Así, el Data Lake presenta una alternativa lo suficientemente flexible para almacenar múltiples tipos de datos.

- Cuando la organización no tiene claridad sobre el plan de uso de datos.

Data Warehouse

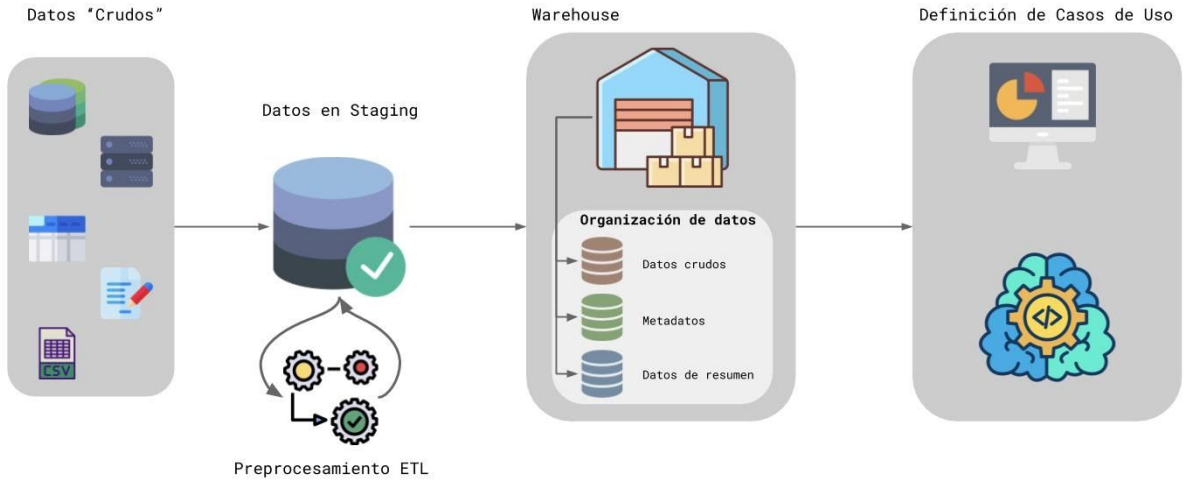

Imagen 5. DataWarehouse.

Un Data Warehouse generalmente almacena sólo aquella información que ya se encuentra modelada. Por definición, el Data Warehouse **debe** ser multipropósito y diseñado para todos los posibles casos de uso de la organización. No considera los requerimientos específicos de una unidad.

Por lo general el proceso ETL se implementa dentro de la lógica de negocios del Data Warehouse. Mediante ETL podemos extraer los datos en la fuente de origen hacia nuestro warehouse. Estos son extraídos de la fuente, transformados y validados en un formato adecuado, y cargados como datos crudos en el staging. Los datos también pueden dividirse entre **metadatos** (datos que permiten describir las relaciones entre datos) y **datos** (unidades empíricas de análisis). Esto es útil para tener en consideración todos los tipos de datos existentes en el warehouse.

Alguno de los beneficios de implementar data warehousing:

1. Al implementar un warehouse, el acceso a los datos no interfiere con bases de datos OLTP.

2. Se puede almacenar y manejar una cantidad substantiva de datos con una mayor tasa de retención que una base de datos OLTP, lo que facilita el acceso a datos históricos.

3. Dado que todos los datos están integrados en una fuente originaria, es más fácil implementar queries simples que accedan a múltiples fuentes de manera unificada.

4. Existe un modelo de datos singular, independiente de las fuentes de éstos.

Diferencias entre un data lake y un data warehouse

1. **Composición de los datos**: Un data warehouse por lo general se compone de datos extraídos de sistemas de bases de datos transaccionales, y uno de sus componentes es el uso de métricas cuantitativas para describir las relaciones existentes. Un sistema de Data Lake permite almacenar fuentes no-tradicionales de datos. Un punto a considerar es que esta no-tradicionalidad de los datos puede tener mayores costos de almacenamiento y consumo de éstos.

2. **Enfoque de la provisión de datos**: Un data warehouse es un caso de uso ideal cuando los usuarios finales desean generar reportes, analizar métricas de desempeño o manejar datos ya contenidos en archivos y hojas de cálculo. También permite brindar apoyo a usuarios que necesitan realizar análisis en los datos, incorporando formas de integración y preparación de los datos. Por el otro lado, un Data Lake permite apoyar a todos los usuarios existentes en la organización.

3. **Utilidad a priori de los datos**: En el proceso de creación del data warehouse, una porción no menor del tiempo se dedicará en analizar las fuentes de datos existentes y entender los procesos de negocio y curación de datos. Una parte importante de este procedimiento involucra realizar decisiones sobre qué datos incluir y qué datos excluir. Un Data Lake buscará mantener todos los datos existentes.

Data Mart

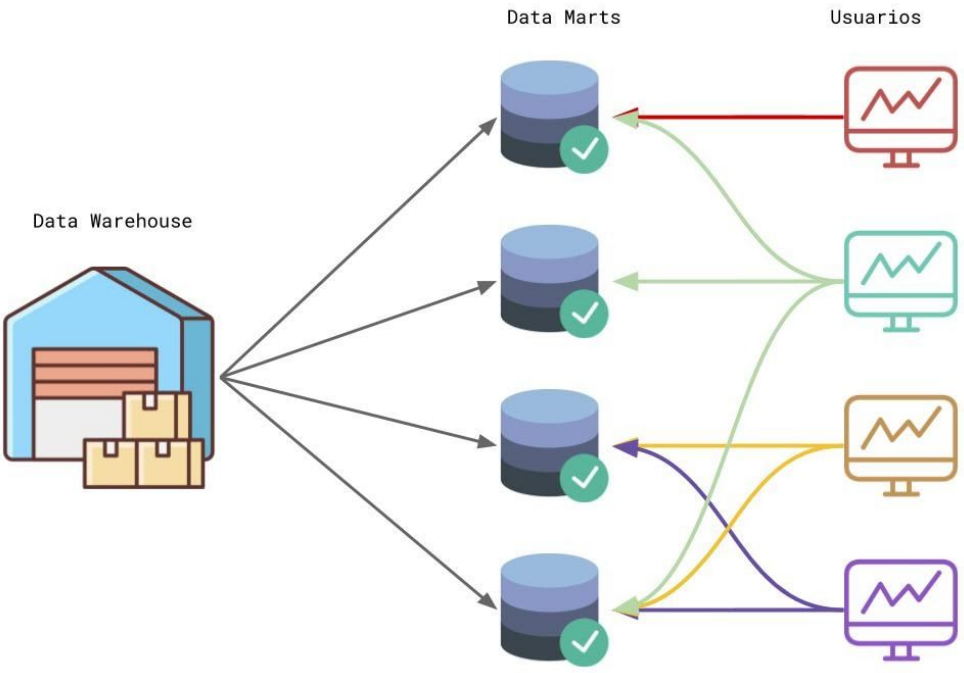

Imagen 6. Datamart.

Si bien un Data Warehouse es una unidad de almacenamiento multipropósito, un Data Mart se puede entender como una subsección del Data Warehouse, el cual está diseñado para una unidad particular dentro de la organización.

Los Data Mart se pueden entender como ambientes separados que buscan aumentar la seguridad e integridad de los datos. Dado que sólo contiene los datos pertinentes a la unidad de la organización, se asegura que el acceso y lectura de los datos responderá a una serie de políticas de implementación.

Existen tres tipos de Data Mart:

1. **Data Marts Dependientes**: Construídos a partir de un data warehouse preexistente, tienen un enfoque top-down (desde la generalidad de la organización a la especificidad de la unidad). Así, se almacenan todos los datos existentes en una ubicación centralizada y posteriormente se define una porción de los datos en función al caso de uso de la unidad.

2. **Data Marts Independientes**: Se pueden considerar como sistemas autónomos creados sin el uso de data warehouses y enfocados en una función exclusiva de la organización. Los datos se generan de fuentes internas o externas, refinados fuera de éste, y posteriormente ingresados al data mart hasta el caso de uso.

3. **Data Marts Híbridos**: Buscan integrar datos existentes de un data warehouse y sistemas operacionales adicionales. Así, combina la velocidad y foco en el usuario del enfoque top-down con la flexibilidad de incorporar nuevos cambios a nivel específico de unidad de negocios.

Diferencias entre un data warehouse y un data mart

1. Un data warehouse es una aplicación independiente, mientras que el data mart se puede entender como sistema específico de apoyo a la toma de decisiones.

2. Los datos en un warehouse se almacenan en una fuente singular centralizada. Los datos en un mart se almacenan de manera descentralizada en función a los diferentes usuarios.

3. Un data warehouse consiste de una forma general de datos, mientras que un data mart consiste de datos resumidos y fuertemente preprocesados.

Paradigmas Big Data

Si bien es cierto existen distintas soluciones Big Data, y cada una presenta sus propias características, podríamos mencionar 2 principales aspectos en los cuales se diferencian: *Batch* Processing y *Stream* Proces sing:

Batch Processing

Corresponde cuando ocurre el procesamiento por bloques de datos que ya se han almacenado durante un periodo de tiempo. Por ejemplo, el procesamiento de un conjunto de transacciones realizadas por un retail durante un fin de semana de ventas masivas. Estos datos podrían contener millones de registros que pueden ser almacenados en un archivo plano, como un csv. Este archivo se procesará al final del fin de semana para varios análisis que las distintas unidades de la empresa necesitan. Por supuesto, dependiendo del tamaño de los datos y la tarea que se desee ejecutar sobre los datos, esta tarea podría llevar bastante tiempo. Esta tarea es la que realizaría una solución Batch Processing. `Hadoop MapReduce` corresponde a una herramienta Batch Processing.

Si lo que se desea es obtener la mayor cantidad de insights de los datos, pero no en tiempo real, Batch Processing es la forma de conseguirlo.

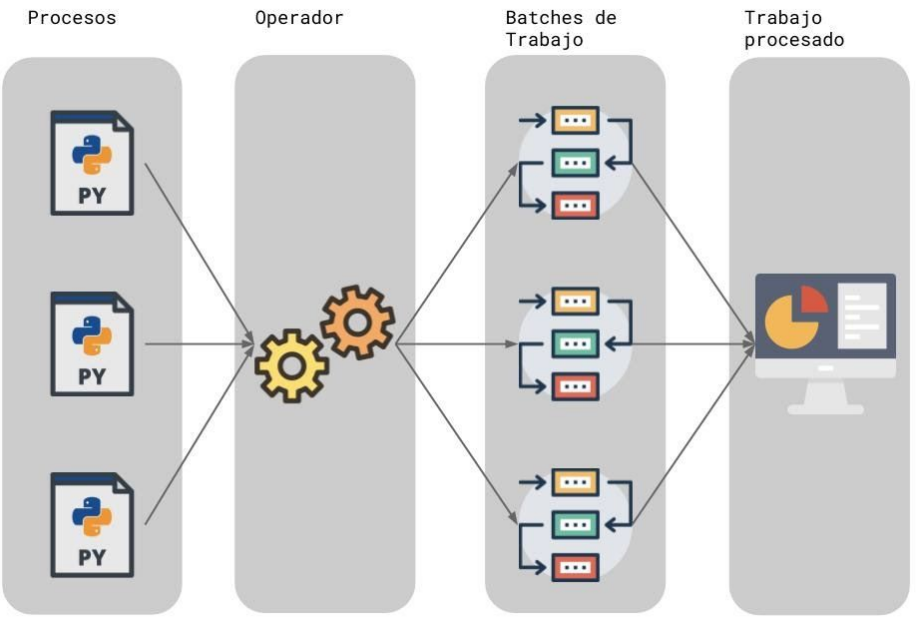

Imagen 7. Batch Processing.

Stream Processing

Stream Processing es lo que se necesita cuando los resultados son solicitados en tiempo real. Stream Processing permite procesar los datos en tiempo real, a medida que llegan y detectar rápidamente distintas situaciones en periodos de tiempo pequeños. Stream Processing también permite compartir los análisis con otros sistemas que dependen de este análisis en tiempo real. Existen distintas soluciones de Stream Processing, dentro de las cuales las que predominan son `Apache Kafka`, `Apache Flink`, `Apache Storm`, entre otras.

Stream Processing es útil para tareas de detección de fraudes o de comportamientos anómalos sobre sistemas de seguridad.

Una de las razones principales de por qué las soluciones de Stream Processing son tan rápidas es porque analizan la información antes de almacenarla en disco, lo que implica que en comparación con la arquitectura de soluciones sean muy distintas.

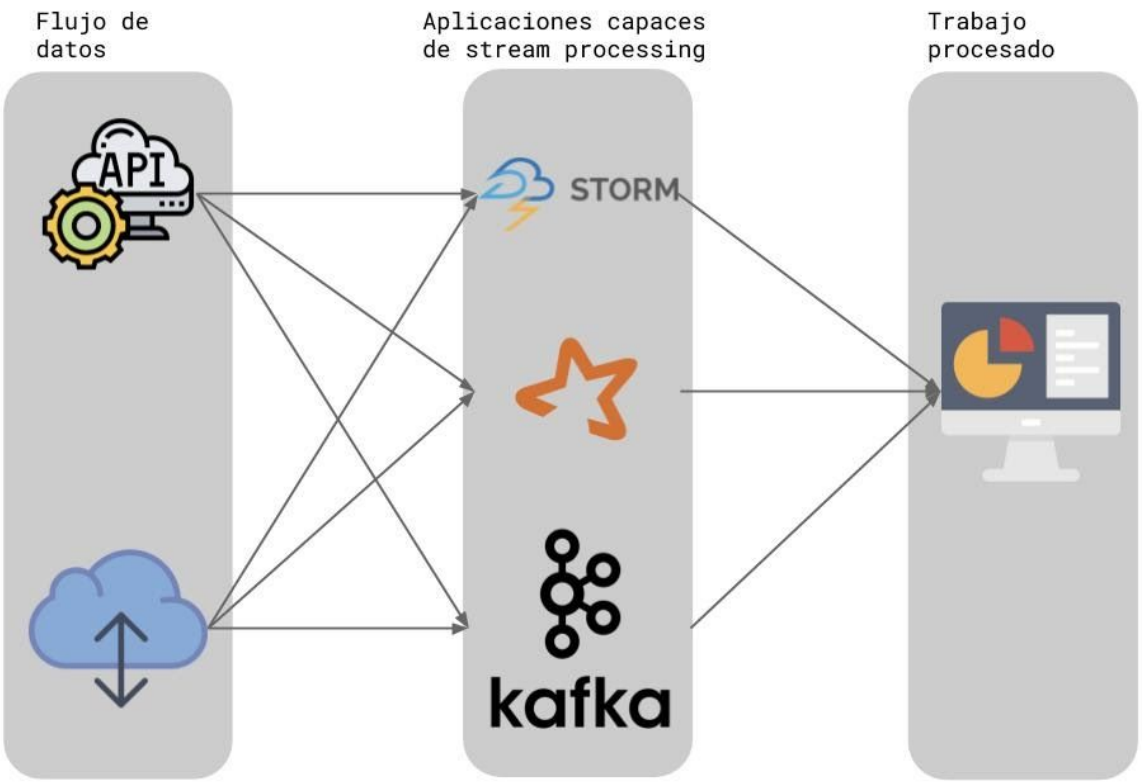

Imagen 8. Kafka.

Referencias

1. Bart Baesens, T. (2014). Big Data Analytics. Wiley.

2. Data Age 2025 (2017), IDC Analyze the Future.

3. Anantharam, D. (n.d.). Big data, Big deal. Retrieved from https://blogs.msdn.microsoft.com/data__knowledge__intelligence/2013/02/18/big-data-big-deal/