Diccionario de Arquitecturas de Datos

Business Intelligence

Data Warehouse

Data Mart

Inmon Model

Kimball Model

Data Vault

Data Lake

Data Mesh

Lakehouse

Data Fabric

Data Governance

Data Hub

Data Centric / Data Driven

Data Pipelines

Cloud Analytics

Data Frame

Dark Data

Data Mashup

Data Sources

Data Visualization

Self Service Analytics

Data Silos

Data Health

Cada vez, están surgiendo nuevas formas de analizar los datos. Desde los sistemas tradicionales de Business Intelligence y Data Warehouse todo ha cambiado.

Os dejo un glosario que os puede ayudar, pues en muchas ocasiones, son los propios fabricantes de software, los que introducen nuevos conceptos que pueden llegar a confundir

Enlace original que publiqué en TodoBI, Enlace original

Emilio Arias

Stratebi CEO

A. Business Intelligence:

Business Intelligence es un término general que incluye las aplicaciones, la infraestructura y las herramientas, así como las mejores prácticas que permiten el acceso y el análisis de la información para mejorar y optimizar las decisiones y el rendimiento

La inteligencia de negocio (BI) tiene que ver con los datos y aplicaciones de un negocio para entenderse mejor. Semejante a la inteligencia militar, que procura entender al enemigo, la inteligencia de negocio versa sobre todo alrededor de si mismo. Específicamente, los sistemas de la inteligencia de negocio se basan en crear modelos informáticos de negocio de modo que pueda funcionar más eficientemente.

El almacenamiento de los datos está en la base de los procesos de la inteligencia de negocio. En la práctica, el Business Intelligence, se refiere generalmente al espacio entero de los sistemas de la base de datos, del software, del análisis, y de la evaluación del usuario que pretende entender y evaluar un negocio.

Los sistemas del BI se diferencian de sistemas operacionales en que están optimizados para preguntar y divulgar sobre datos. Esto significa típicamente que, en un Datawarehouse , los datos están desnormalizados para apoyar preguntas de alto rendimiento, mientras que los sistemas operacionales generalmente se normalizan completamente para apoyar integridad de referencia y para insertar datos continuamente.

Los procesos de ETL (extracción, transformación y carga) que cargan sistemas del BI tienen que traducir del sistema operacional normalizado a desnormalizado.

Y, típicamente, tienen fallos severos de funcionamiento debido a que no deben degradar el funcionamiento de los sistemas operacionales, y no deben prohibir el acceso al almacén de datos
Por eso surge el Business Intelligence, basado en nuevas estructuras de análisis, básicamente multidimensional, en contraste con el relacional.

Enlaces Recomendados:

- TodoBI (Blog Business Intelligence)
- Business Intelligence (Wikipedia)
- Qué es Business Intelligence
- Recursos en Youtube

B. Data Warehouse:

En esencia, se trata de una base de datos relacional que integra datos de múltiples fuentes dentro de una empresa. La creación de un data warehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence.

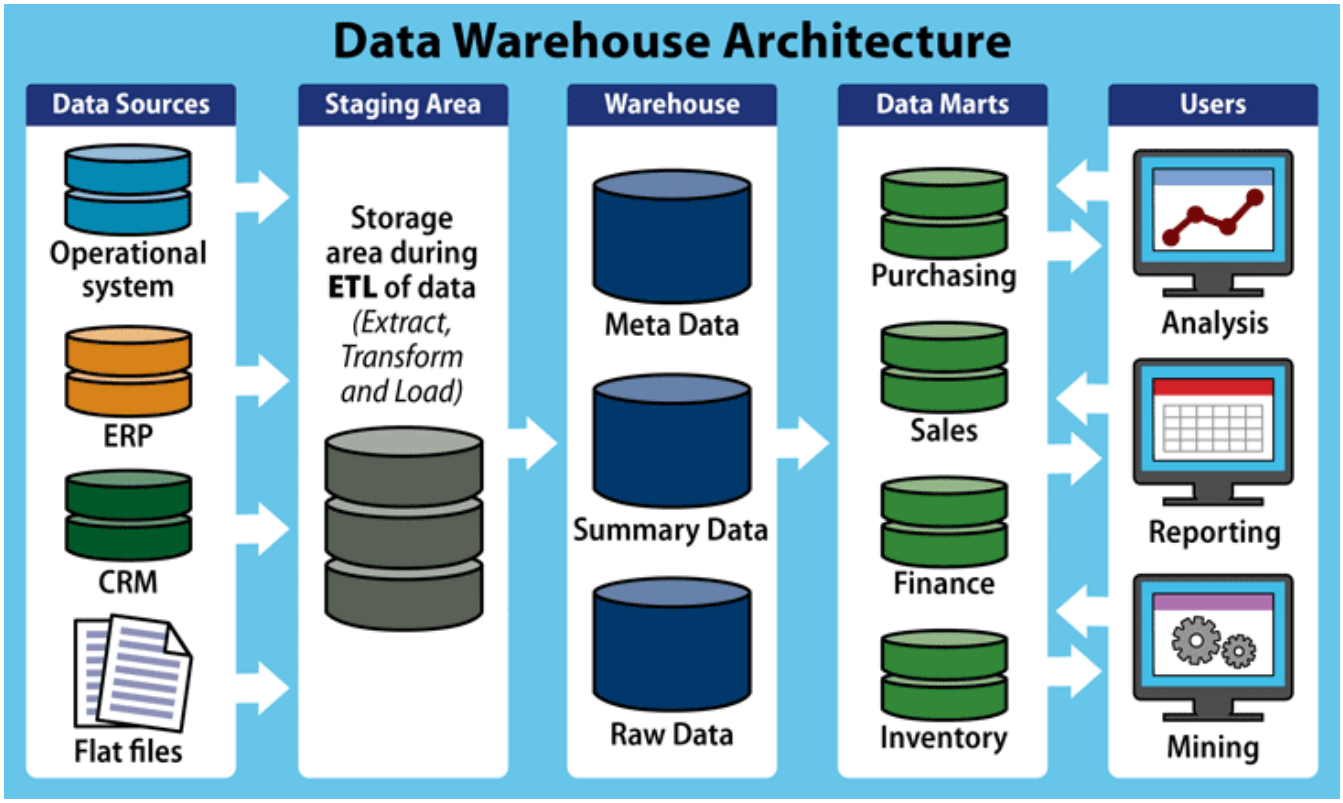
La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales... etc). Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales)

Para la creación de un Data Warehouse tradicional es necesario la ejecución de procesos ETL (Extracción, Transformación y Carga) a partir de los sistemas operaciones de una compañía:

Extracción: obtención de información de las distintas fuentes tanto internas como externas.

Transformación: filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga: organización y actualización de los datos y los metadatos en la base de datos.



Enlaces Recomendados:

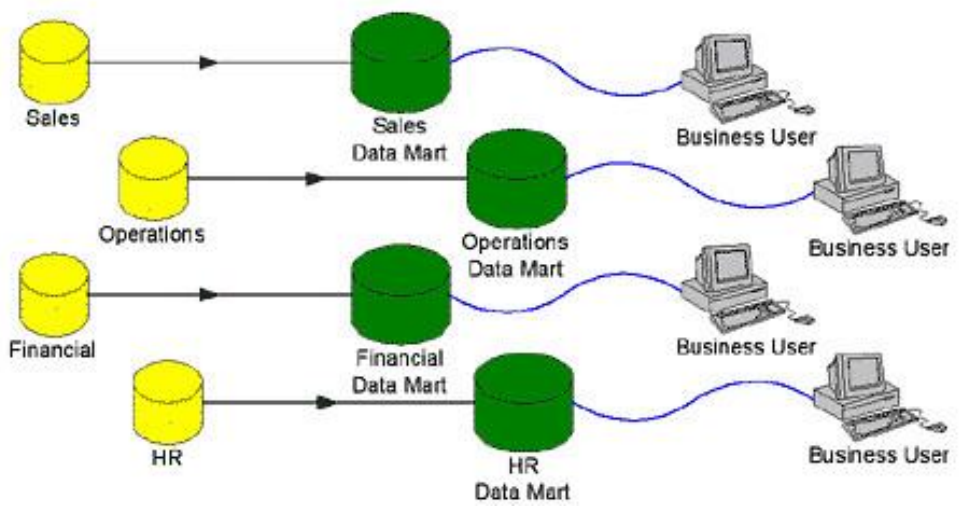
- TodoBI (enlaces sobre Data Warehouse)
- Cómo elegir la mejor solución de Data Warehouse
- Data Warehouse (Wikipedia)
- Mejores libros sobre Data Warehouse
- Videotutoriales Data Warehouse

C. Data Mart:

Un Data Mart es un almacén de datos orientado a un área específica, como por ejemplo, Ventas, Recursos Humanos u otros sectores en una organización. Por ello, también se le conoce como una base de información departamental.

Este almacén permite que una empresa pueda acceder a datos claves de un área de forma sencilla, además de realizar diversas funciones, tales como:

- La organización de información para su posterior análisis.
- La elaboración de indicadores clave de rendimiento (KPI).
- La creación de informes para un aprendizaje automático.
- La evaluación de datos sobre el cumplimiento de objetivos de un sector.



Comparación con un Data Warehouse:

Al hablar de los data marts, es inevitable la comparación con los data warehouse y al final se acaba diciendo (o entendiendo) que son como estos, pero *en pequeño*, y en cierto modo esto es así, pero esta idea suele hacer caer en los siguientes errores sobre la implementación y funcionamiento de los data marts:

- Son más simples de implementar que un Data Warehouse: *Falso*, la implementación es muy similar, ya que debe proporcionar las mismas funcionalidades.

- Son pequeños conjuntos de datos y, en consecuencia, tienen menor necesidad de recursos: *Falso*, una aplicación corriendo sobre un data mart necesita los mismos recursos que si corriera sobre un data warehouse.
- Las consultas son más rápidas, dado el menor volumen de datos: *Falso*, el menor volumen de datos se debe a que no se tienen todos los datos de toda la empresa, pero sí se tienen todos los datos de un determinado sector de la empresa, por lo que una consulta sobre dicho sector tarda lo mismo si se hace sobre el data mart que si se hace sobre el data warehouse.
- En algunos casos añade tiempo al proceso de actualización: *Falso*, actualizar el data mart desde el data warehouse cuesta menos (ya que los formatos de los datos son o suelen ser idénticos) que actualizar el data warehouse desde sus fuentes de datos primarias, donde es necesario realizar operaciones de transformación

Enlaces Recomendados:

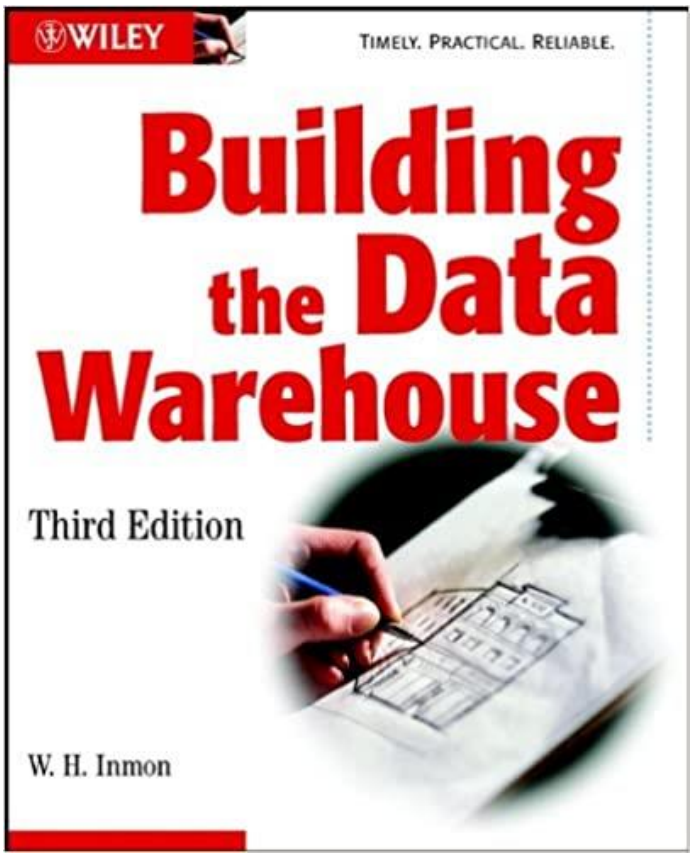
- Data Mart (Wikipedia)
- How to create a Data Mart with reports and dasboards in just 8 minutes (open source)
- Data Mart, tutoriales en Youtube
- Cómo implementar un Data Mart

D. Inmon Model:

Bill Inmon, el "padre del Data Warehousing", define un Data Warehouse (DW) como "una colección de datos orientada a un tema, integrada, variable en el tiempo y no volátil, para apoyar el proceso de toma de decisiones de la dirección".

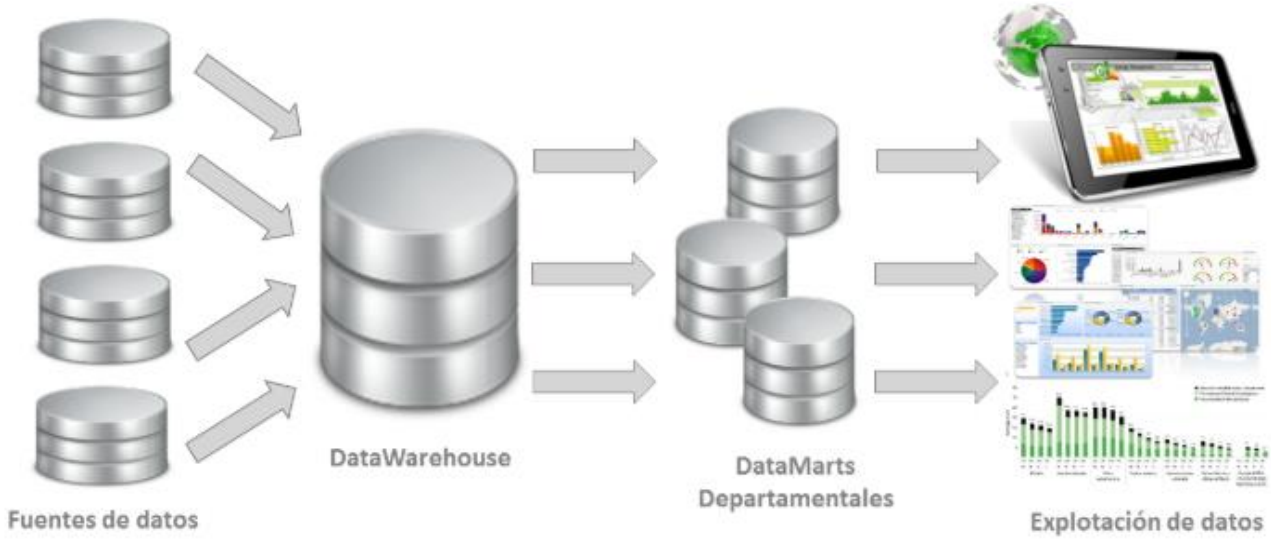
En su libro blanco, Modern Data Architecture, Inmon añade que el Data Warehouse representa la "sabiduría convencional" y es ahora una parte estándar de la infraestructura corporativa.

El Enfoque de diseño Inmon utiliza la forma normalizada para construir la estructura de la entidad, evitando la redundancia de datos tanto como sea posible. Esto da como resultado una identificación clara de los requisitos comerciales y la prevención de irregularidades en la actualización de datos. Además, la ventaja de este enfoque de arriba hacia abajo en el diseño de bases de datos es que es robusto a los cambios comerciales y contiene una perspectiva dimensional de los datos en el mercado de datos.



A continuación, se construye el modelo físico, que sigue la estructura normalizada. Este modelo de Inmon crea una única fuente de verdad para todo el negocio. La carga de datos se vuelve menos compleja debido a la estructura normalizada del modelo. Sin embargo, utilizar esta disposición para realizar consultas es un desafío, ya que incluye numerosas tablas y enlaces.

Esta metodología de data warehouse de Inmon propone la construcción de data marts por separado para cada división, como finanzas, marketing, ventas, etc. Todos los datos que ingresan al data warehouse están integrados. El almacén de datos actúa como una única fuente de datos para varios data marts a fin de garantizar la integridad y coherencia en toda la empresa.



Ventajas del método Inmon

El enfoque de diseño de Inmon ofrece los siguientes beneficios:

- El almacén de datos actúa como una fuente de verdad unificada para todo el negocio, donde todos los datos están integrados.
- Este enfoque tiene una redundancia de datos muy baja. Por lo tanto, hay menos posibilidad de irregularidades en la actualización de datos, lo que hace que el proceso de almacenamiento de datos ETL sea más sencillo y menos susceptible a fallas.
- Simplifica los procesos comerciales, ya que el modelo lógico representa objetos comerciales detallados.

- Este enfoque ofrece una mayor flexibilidad, ya que es más fácil actualizar el almacén de datos en caso de que haya algún cambio en los requisitos comerciales o en los datos de origen.
- Puede manejar diversos requisitos de informes en toda la empresa.

Desventajas del método Inmon

Los posibles inconvenientes de este enfoque son los siguientes:

- La complejidad aumenta a medida que se agregan varias tablas al modelo de datos con el tiempo.
- Se requieren recursos capacitados en el modelado de datos de almacenamiento de datos, que pueden ser costosos y difíciles de encontrar.
- La configuración preliminar y la entrega requieren mucho tiempo.
- Se requiere una operación ETL adicional, ya que los data marts se crean después de la creación del almacén de datos.
- Este enfoque requiere que los expertos administren un almacén de datos de manera efectiva.

Enlaces Recomendados:

- Let's Compare the Kimball and Inmon Data Warehouse Architectures
- Libros de Bill Inmon
- Top-Down Approach and Bottom-Up Approach for Datawarehouse Architecture|Inmon vs Kimball
- Metodologías Agiles BI/DW

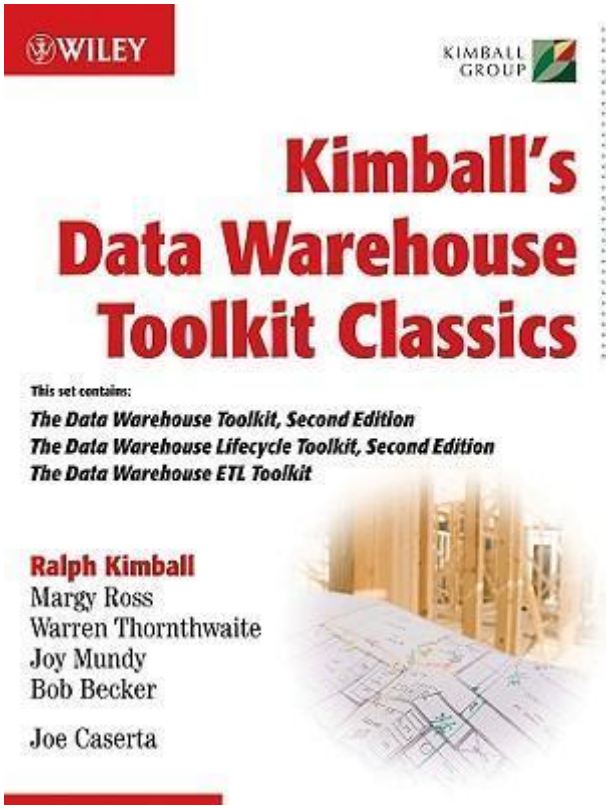
E. Kimball Model:

El modelo de datos de Kimball sigue un enfoque de abajo hacia arriba para almacenamiento de datos (DW) diseño de arquitectura en el que los mercados de datos se forman primero en función de los requisitos comerciales.

La ***Metodología Kimball***, es una metodología empleada para la construcción de un almacén de datos (data warehouse, DW) que no es mas que, una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.
La metodología se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle). Este ciclo de vida del proyecto de DW, está basado en cuatro principios básicos:

- Centrarse en el negocio
- Construir una infraestructura de información adecuada
- Realizar entregas en incrementos significativos (este principio consiste en crear el almacén de datos (DW) en incrementos entregables en plazos de 6 a 12 meses, en este punto, la metodología se parece a las metodologías ágiles de construcción de software)
- Ofrecer la solución completa (En este se punto proporcionan todos los elementos necesarios para entregar valor a los usuarios de negocios, para esto ya se debe tener un almacén de datos bien diseñado, se deberán entregar herramientas de consulta ad hoc, aplicaciones para informes y análisis avanzado, capacitación, soporte, sitio web y documentación).

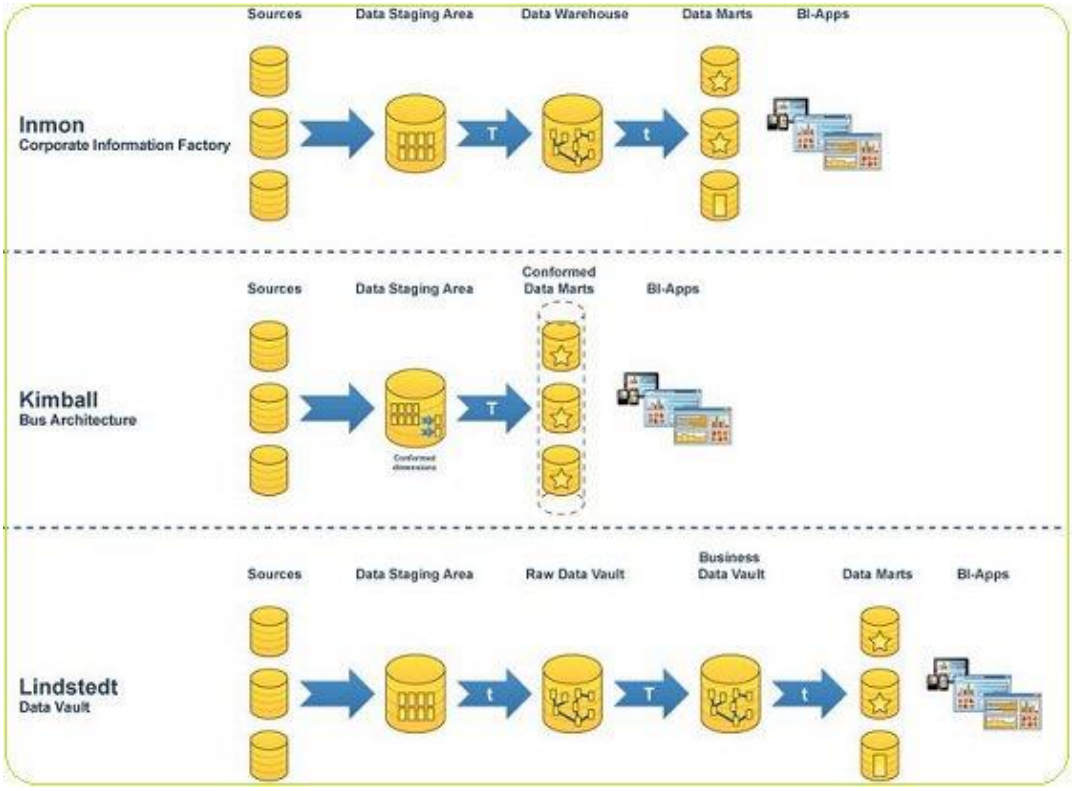
La construcción de una solución de DW/BI (Datawarehouse/Business Intelligence) es sumamente compleja, y Kimball nos propone una metodología que nos ayuda a simplificar esa complejidad



A continuación, se evalúan las fuentes de datos primarias y se utiliza una herramienta de extracción, transformación y carga (ETL) para obtener diferentes tipos de formatos de datos de varias fuentes y cargarlos en un área de ensayo del servidor de base de datos relacional.

Una vez que los datos se cargan en el área de preparación en el almacén de datos, la siguiente fase incluye la carga de datos en un modelo de almacén de datos dimensional que está desnormalizado por naturaleza.

Este modelo divide los datos en la tabla de hechos, que son datos transaccionales numéricos, o la tabla de dimensiones, que es la información de referencia que respalda los hechos.



El esquema en estrella es el elemento fundamental del modelo de almacén de datos dimensional. En este esquema en estrella, una tabla de hechos está limitada por varias dimensiones.

El modelado dimensional de Kimball permite a los usuarios construir varios esquemas en estrella para satisfacer diversas necesidades de informes. La ventaja del esquema en estrella es que las consultas de tablas dimensionales pequeñas se ejecutan instantáneamente.

Ventajas del método Kimball

Algunos de los principales beneficios de la metodología Kimball incluyen:

- El modelado dimensional de Kimball es tan rápido de construir ya que no implica normalización, lo que significa una ejecución rápida de la fase inicial del almacenamiento de datos de procesos.
- Una ventaja del esquema en estrella es que la mayoría de los operadores de datos pueden comprenderlo fácilmente debido a su estructura desnormalizada, que simplifica las consultas y el análisis.
- La huella del sistema de almacenamiento de datos es trivial porque se centra en áreas y procesos comerciales individuales en lugar de en

toda la empresa. Por lo tanto, ocupa menos espacio en la base de datos, lo que simplifica la administración del sistema.
- Permite la recuperación rápida de datos del almacén de datos, ya que los datos se segregan en tablas de hechos y dimensiones. Por ejemplo, la tabla de hechos y dimensiones para la industria de seguros incluiría transacciones de pólizas y transacciones de reclamaciones.
- Un equipo más pequeño de diseñadores y planificadores es suficiente para la gestión del almacén de datos porque los sistemas de origen de datos son estables y el almacén de datos está orientado a procesos. Además, la optimización de consultas es sencilla, predecible y controlable.
- Estructura dimensional conformada para marco de calidad de datos. El enfoque de Kimball también se conoce como el enfoque de estilo de vida dimensional empresarial porque permite que las herramientas de inteligencia empresarial profundicen en varios esquemas en estrella y genere información confiable.

Desventajas del método Kimball

Algunos de los inconvenientes del enfoque de diseño de Kimball incluyen:

- Los datos no están completamente integrados antes de la presentación de informes; la idea de una "fuente única de verdad se pierde".
- Pueden ocurrir irregularidades cuando los datos se actualizan en la arquitectura Kimball DW. Esto se debe a que, en el almacén de datos de técnicas de desnormalización, se agregan datos redundantes a las tablas de la base de datos.
- En la arquitectura Kimball DW, pueden ocurrir problemas de rendimiento debido a la adición de columnas en la tabla de hechos, ya que estas tablas son bastante detalladas. La adición de nuevas columnas puede expandir las dimensiones de la tabla de hechos, afectando su rendimiento. Además, el modelo de almacén de datos dimensional se vuelve difícil de modificar con cualquier cambio en las necesidades comerciales.

- Como el modelo de Kimball está orientado a los procesos comerciales, en lugar de centrarse en la empresa en su conjunto, no puede manejar todos los requisitos de informes de BI.
- El proceso de incorporar grandes cantidades de datos heredados en el almacén de datos es complejo.

Enlaces Recomendados:

- Kimball Group
- Libros de Ralph Kimball
- Kimball in the context of the modern data warehouse: what's worth keeping, and what's not

F. Data Vault:

El Data Vault es un método de modelado de bases de datos diseñado para proporcionar un almacenamiento histórico a largo plazo de los datos procedentes de múltiples sistemas operativos. También es un método de observación de datos históricos que aborda cuestiones como la auditoría, el rastreo de datos, la velocidad de carga y la resistencia a los cambios, además de hacer hincapié en la necesidad de rastrear la procedencia de todos los datos de la base de datos.

Esto significa que cada fila en un Data Vault debe ir acompañada de atributos de fuente de registro y fecha de carga, lo que permite a un auditor rastrear los valores hasta la fuente. Fue desarrollado por Daniel (Dan) Linstedt en 2000.

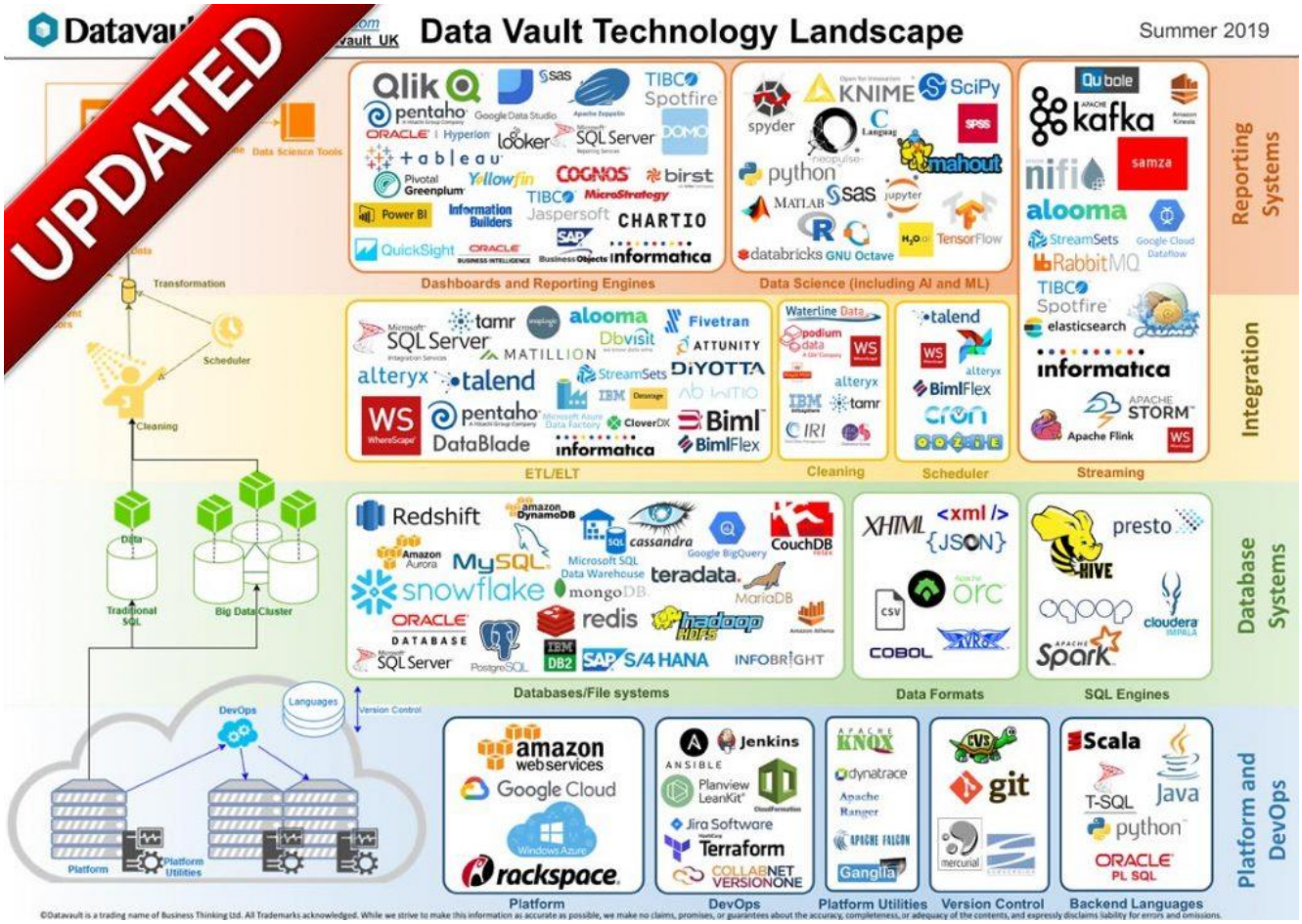
El Data Vault no distingue entre datos buenos y malos ("malos" significa que no se ajustan a las reglas del negocio).

Esto se resume en la afirmación de que Data Vault almacena "una única versión de los hechos" (también expresada por Dan Linstedt como "todos los datos, todo el tiempo"), en contraposición a la práctica de otros métodos de almacén de datos de almacenar "una única versión de la verdad en la que los datos que no se ajustan a las definiciones se eliminan o "limpian".

El método de modelado está diseñado para ser resistente a los cambios en el entorno empresarial del que proceden los datos almacenados, separando explícitamente la información estructural de los atributos descriptivos.

Data Vault está diseñada para permitir la carga paralela en la medida de lo posible, de modo que las implantaciones muy grandes puedan ampliarse sin necesidad de un rediseño importante.

Para mejorar los tiempos de implementación, se introdujo el método Data Vault para modelar el data warehouse. El principio de diseño implica separar las claves del negocio, el contexto y las relaciones en distintas tablas como hub, satélite y enlaces.



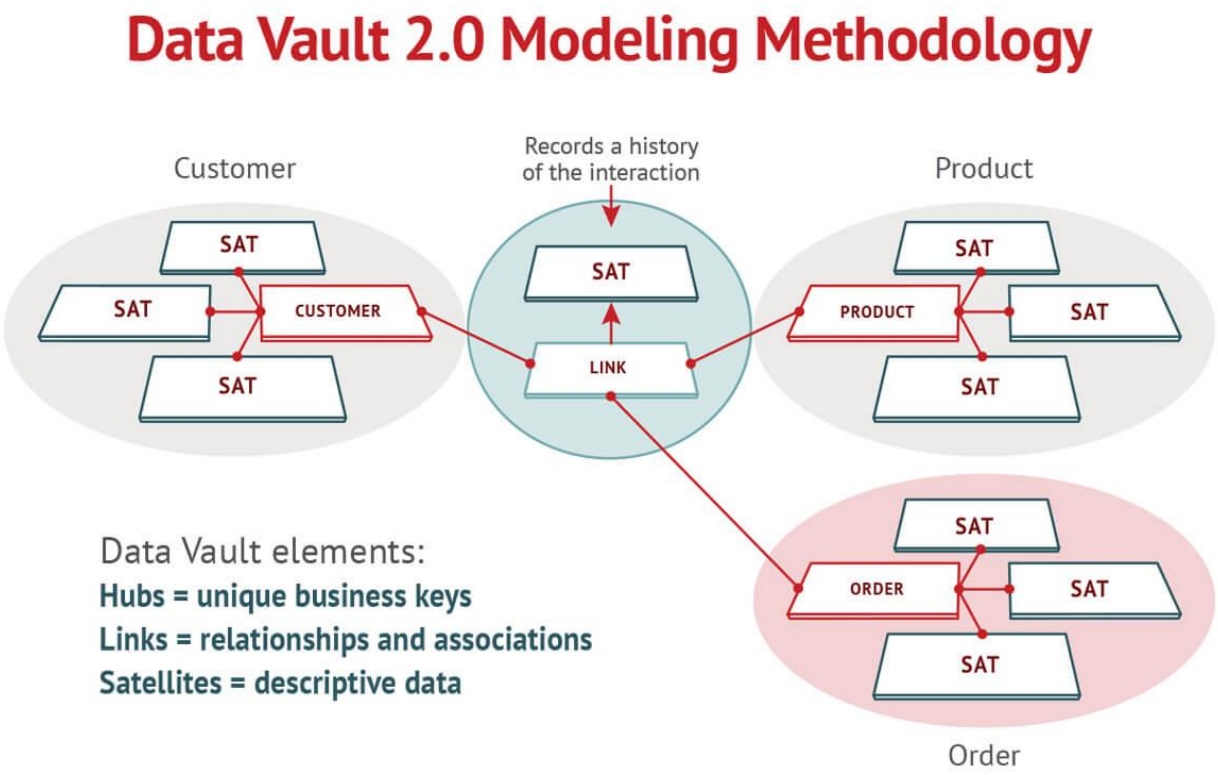
Conceptos de hub, satélites y enlaces

Un hub contiene la clave de negocio real (uno o más campos que identifican de forma exclusiva una entidad para la empresa, por ejemplo, un número de cliente) y una clave sustituta que se utiliza para conectar esta tabla con otras estructuras (equivalente a una clave primaria).

Además, también puede contener metadatos como marcas de tiempo o información sobre el origen de datos. Los enlaces a continuación conectan los hubs con una tabla simple muchos-a-muchos consistente en las claves respectivas de sustitución.

Por este medio, los hubs y los enlaces representan la parte más estable de un modelo y se enriquecen con los llamados satélites. Un satélite está conectado a un hub con su clave de sustitución y contiene uno o más atributos descriptivos que normalmente están agrupados por un sistema de origen, un contexto de negocio o una tasa de cambio.

Además, una tabla de satélites también puede comprender diferentes tipos de metadatos, como períodos de fecha válidos e información sobre su origen.



Un modelo de Data Vault básico podría estar formado por un hub de pedidos muy simple y un concentrador de clientes, así como enlaces y satélites relacionados. En este modelo, el centro de clientes podría tener dos satélites: uno con datos maestros que posiblemente proviene del sistema CRM y otro con un atributo llamado smartphone que puede provenir de un sistema de análisis web.

Beneficios de un Data Warehouse en el ámbito de Data Vault

- **La facilidad de ampliación** permite un enfoque de proyecto ágil.
- Los modelos creados son **altamente escalables**.
- **Los procesos de carga pueden ser paralelizados de forma óptima** porque hay pocos puntos de sincronización.
- **Los modelos son fáciles de auditar**

Pero junto con los muchos beneficios, los proyectos de Data Vault también presentan una serie de desafíos.

Desafíos de un Data WareHouse en el ámbito de Data Vault

- **Hay un gran aumento en el número de objetos de datos** (tablas, columnas) como resultado de separar los tipos de información y enriquecerlos con la meta información para la carga.
- **Esto da lugar a un mayor esfuerzo de modelización** que comprende numerosas tareas mecánicas no sofisticadas

Enlaces Recomendados:

- Dan Linsdted web
- Libros de Dan Lindstedt
- Data Vault en Youtube

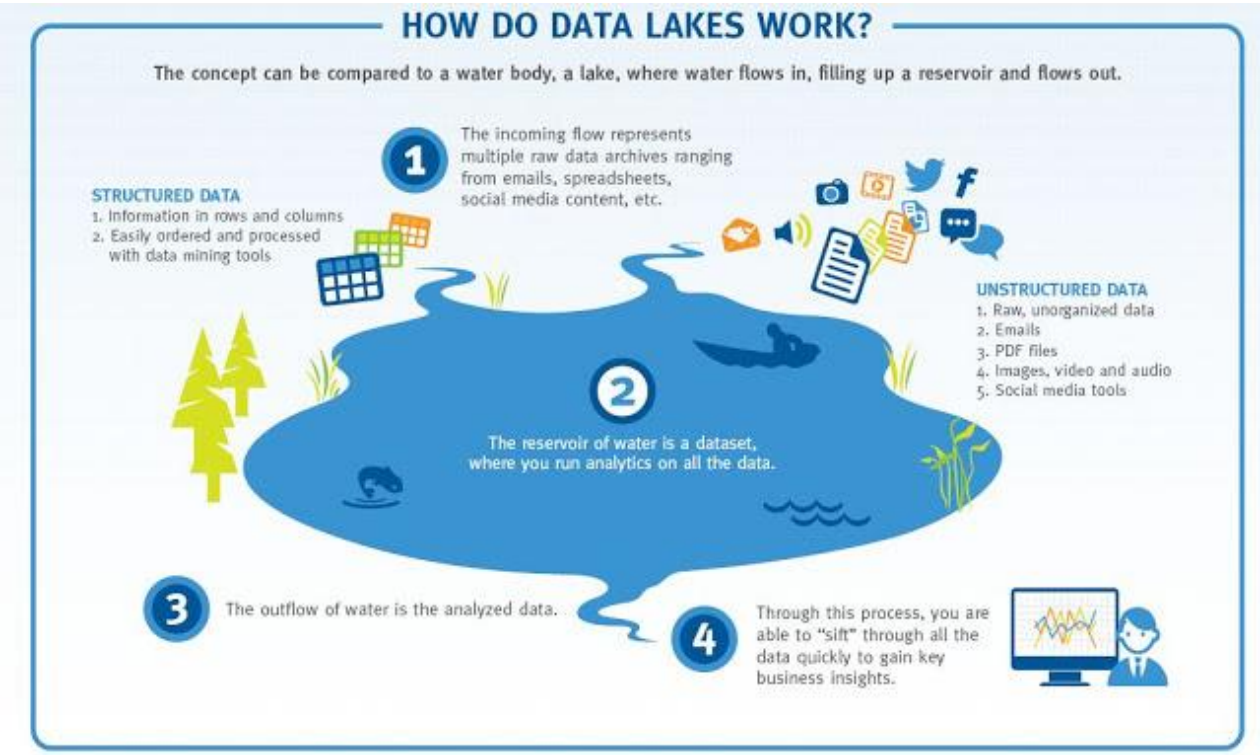
G. Data Lake:

Un Data Lake es un sistema o repositorio de datos almacenados en su formato natural/bruto, normalmente blobs de objetos o archivos. Un Data Lake suele ser un almacén único de datos que incluye copias en bruto de los datos del sistema de origen, datos de sensores, datos sociales, etc., y datos transformados que se utilizan para tareas como la elaboración de informes, la visualización, el análisis avanzado y el aprendizaje automático.

El término Data Lake fue acuñado por un antiguo amigo de este Portal, James Dixon, creador de Pentaho

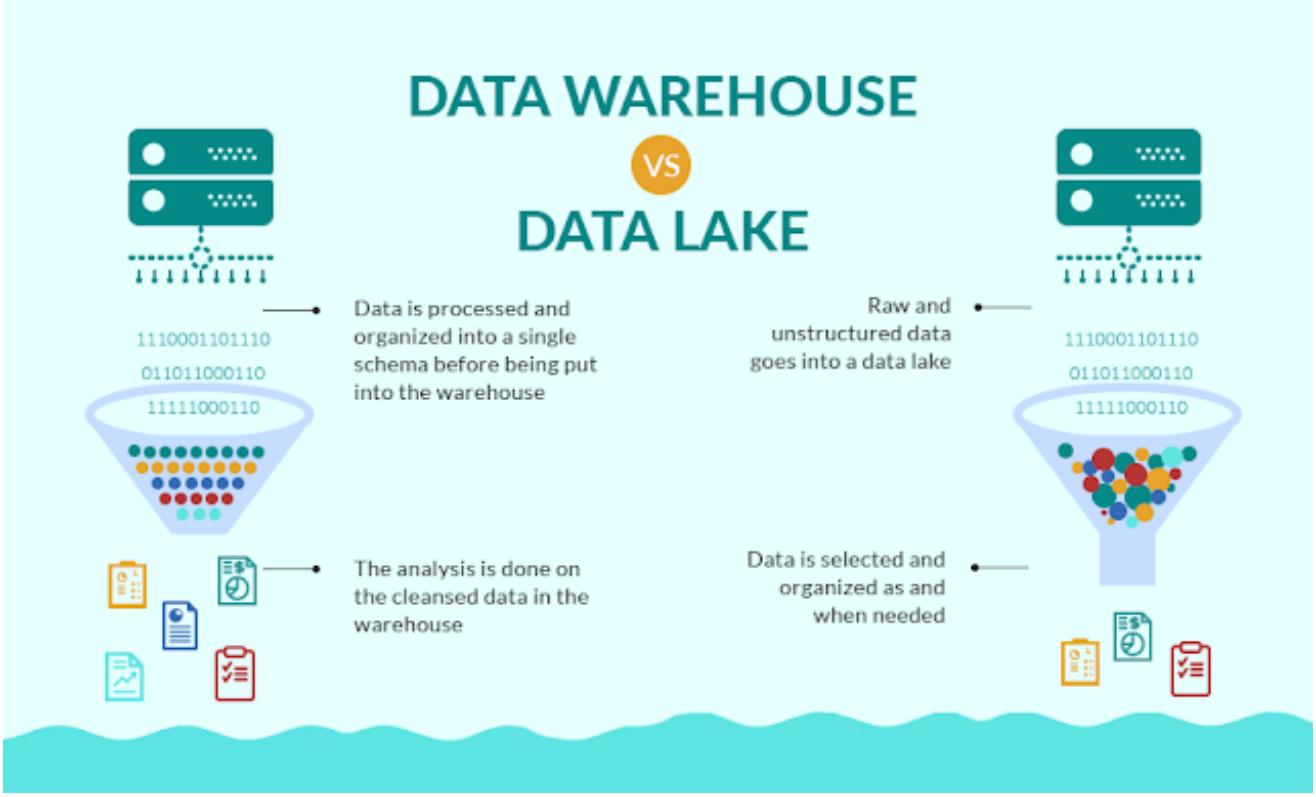
Un Data Lake puede incluir datos estructurados de bases de datos relacionales (filas y columnas), datos semiestructurados (CSV, registros, XML, JSON), datos no estructurados (correos electrónicos, documentos, PDF) y datos binarios (imágenes, audio, vídeo.

Un Data Lake puede establecerse "on premise" (dentro de los centros de datos de una organización) o "en la nube" (utilizando servicios en la nube de proveedores como Amazon, Microsoft o Google)



El término fue acuñado por James Dixon, CTO de Pentaho, y pretendía evocar un gran depósito en el que se pueden verter enormes cantidades de datos. Los usuarios empresariales de todo tipo pueden sumergirse en el Data Lake y obtener el tipo de información que necesitan para su aplicación.

El concepto ha ganado en popularidad con la explosión de los datos de las máquinas y la rápida disminución del coste del almacenamiento.
Hay diferencias clave entre los Data Lakes y los Data Warehouse que se han utilizado tradicionalmente para el análisis de datos.



En primer lugar, los Data Warehouse están diseñados para datos estructurados. En relación con esto, los Data Lakes no imponen un esquema a los datos cuando se cargan o ingestan. En cambio, el esquema se aplica cuando los datos se leen -o se extraen- del Data Lake, lo que permite múltiples casos de uso de los mismos datos.

Por último, los Data Lakes han ganado en popularidad con el aumento de los Data Scientist, que tienden a trabajar de forma más ad hoc y experimental que los analistas de negocio de antaño.

Enlaces Recomendados:

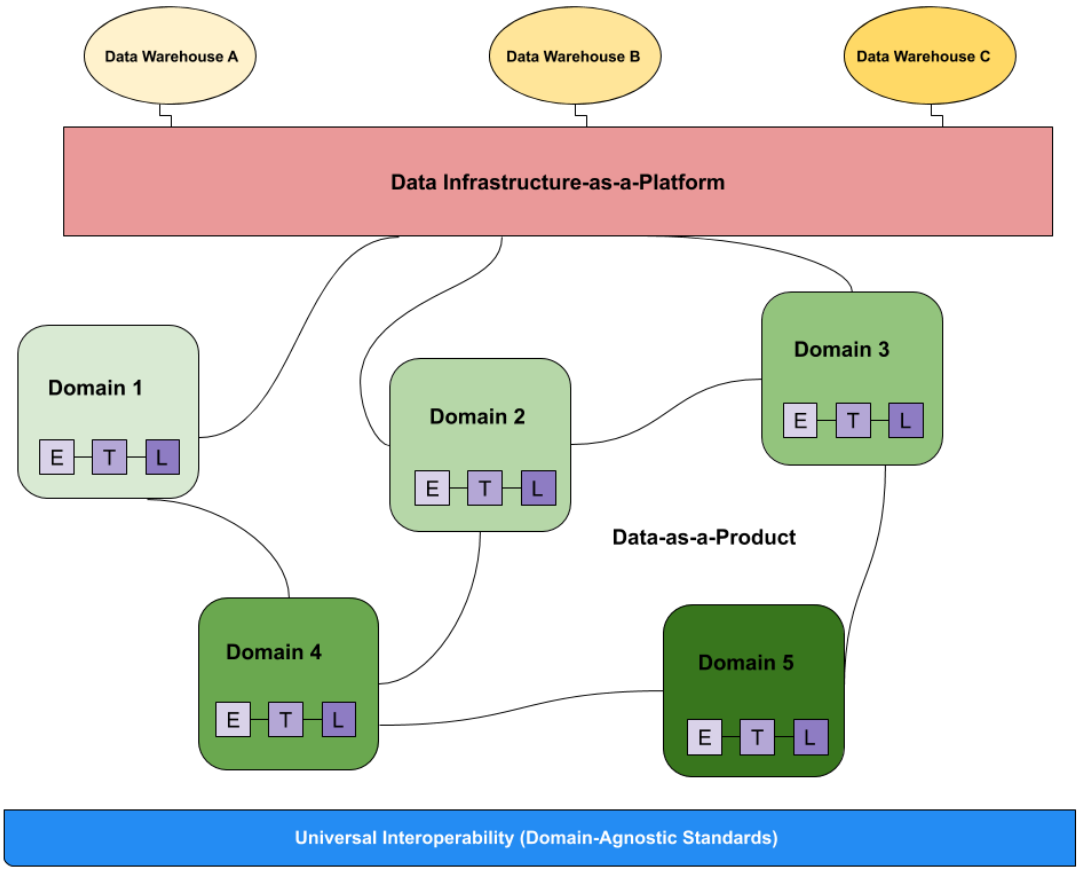
- Data Lake en TodoBI
- Data Lake en Youtube
- Libros sobre Data Lake
- Introducción al Data Lake

H. Data Mesh:

Data Mesh es un paradigma arquitectónico que desbloquea los datos analíticos a escala; desbloquea rápidamente el acceso a un número cada vez mayor de conjuntos de datos distribuidos, para una proliferación de escenarios de consumo de datos cada vez mayor, como el aprendizaje automático, la analítica o las aplicaciones intensivas en el uso de datos en toda la organización

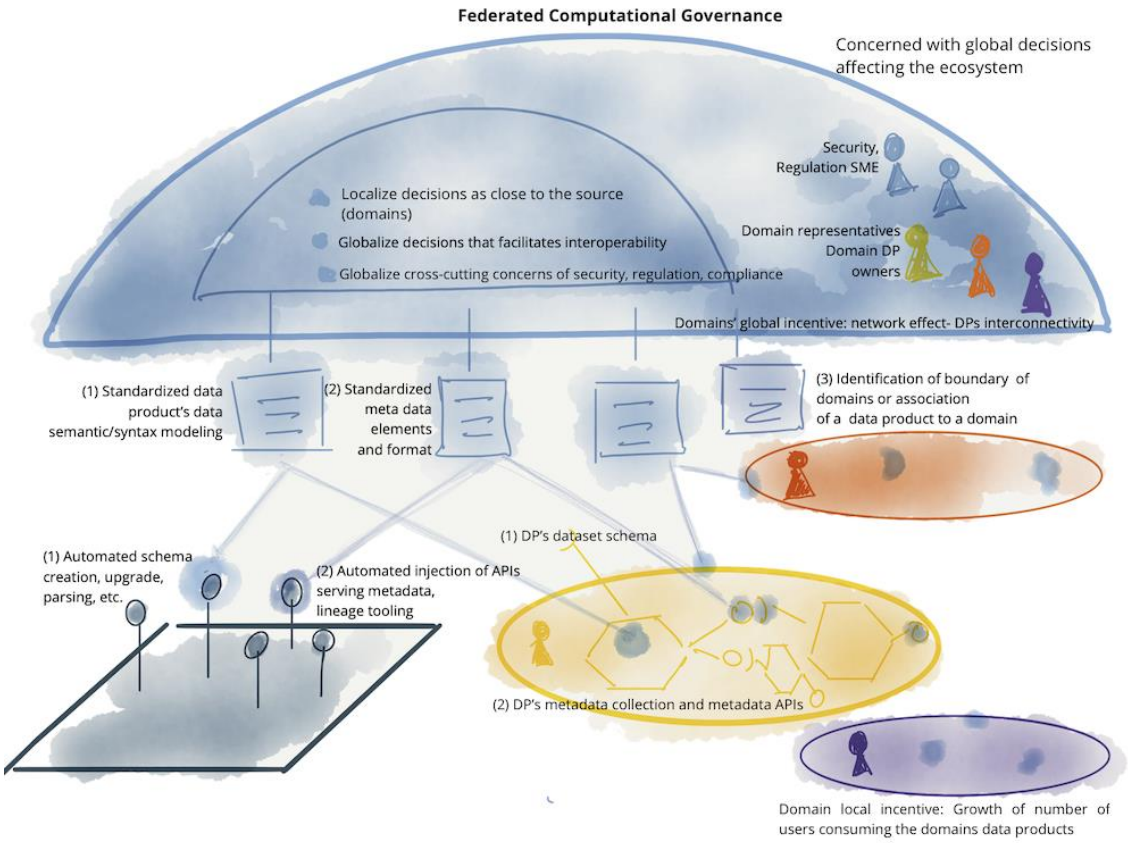
A diferencia de las infraestructuras de datos monolíticas tradicionales que gestionan el consumo, el almacenamiento, la transformación y la salida de datos en un Data Lake central, un Data Mesh admite consumidores de datos distribuidos y específicos de cada dominio y ve los "datos como un producto", y cada dominio gestiona sus propios conductos de datos.

El tejido que conecta estos dominios y sus activos de datos asociados es una capa de interoperabilidad universal que aplica la misma sintaxis y estándares de datos.



Las principales características serían:

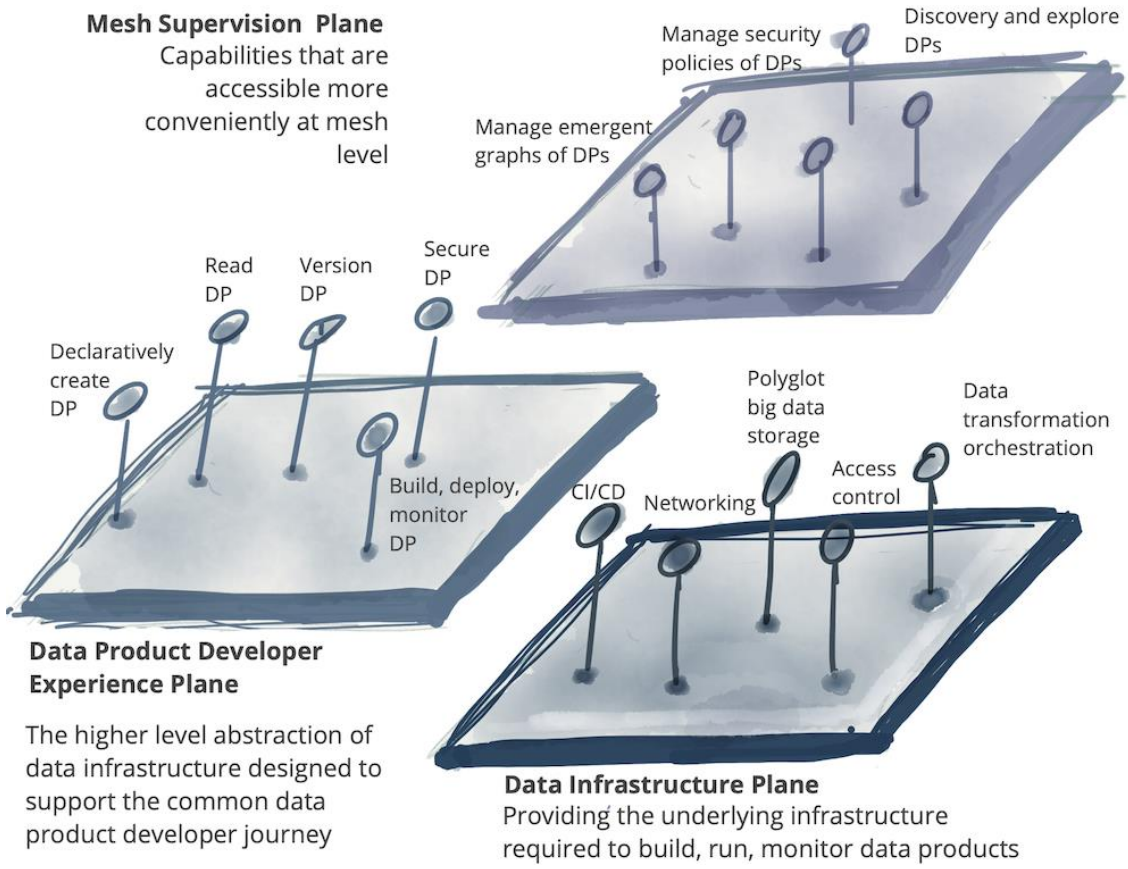
- Los datos como producto, distribuyendo la responsabilidad desde el equipo de plataforma al equipo responsable del dominio y dando la propiedad del producto y su control al dominio que tendrá que garantizar los acuerdos de servicios.
- Un gobierno federado que permita que las decisiones estén los más próximas al dominio pero conservando un control centralizado.
- El dominio como dueño de los datos, acompañado de una arquitectura que descentralice la propiedad de los datos centrada en los dominios.
- Debe ser una plataforma en autoservicio no solo en la parte de consumo de datos si no también en la creación de nuevos productos de datos.



Data Mesh es un nuevo enfoque basado en una arquitectura moderna y distribuida para la gestión de datos analíticos. Permite a los usuarios finales acceder fácilmente a los datos y consultarlos allí donde viven, sin necesidad de transportarlos primero a un Data Lake o Data Warehouse.

La estrategia descentralizada del Data Mesh distribuye la propiedad de los datos a equipos de dominios específicos que los gestionan, poseen y sirven como producto.

El objetivo principal del data Mesh es eliminar los retos de la disponibilidad y accesibilidad de los datos a escala. Data Mesh permite tanto a los usuarios de negocio como a los científicos de datos acceder, analizar y hacer operativa la información de negocio desde prácticamente cualquier fuente de datos, en cualquier lugar, sin la intervención de equipos de datos expertos.



En pocas palabras, Data Mesh hace que los datos sean accesibles, disponibles, descubribles, seguros e interoperables. El acceso más rápido a los datos de consulta se traduce directamente en un tiempo más rápido para obtener valor sin necesidad de transportar los datos.

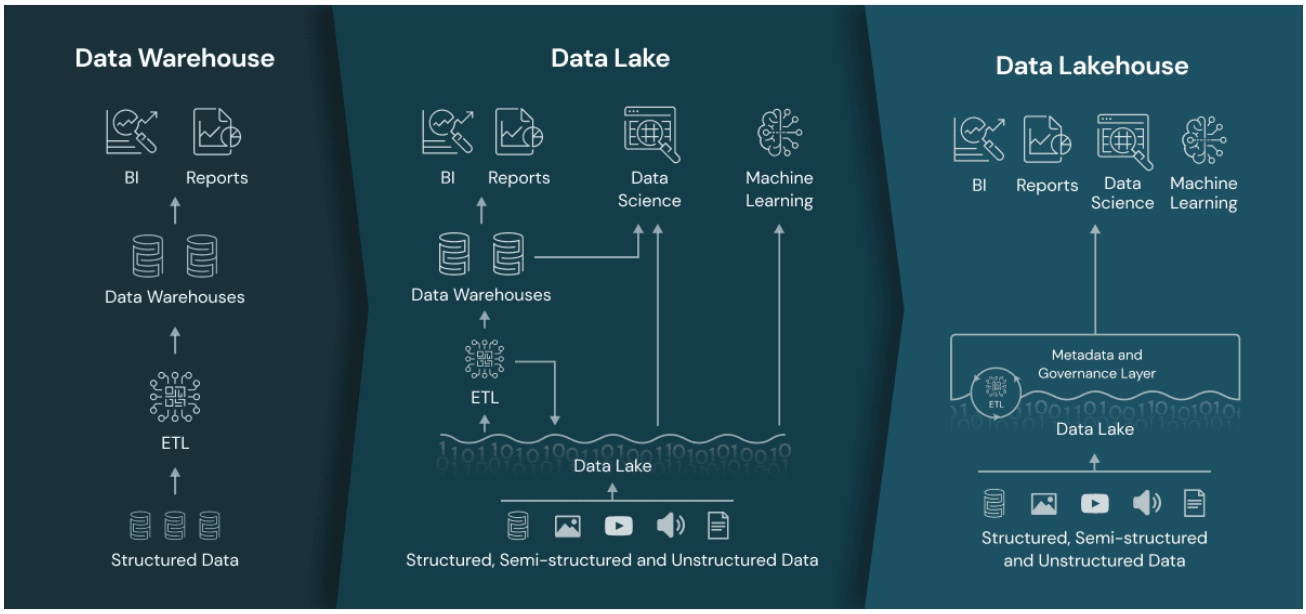
Enlaces Recomendados:

- Transición a Data Mesh
- Building a data mesh to support an ecosystem of data products
- Data Mesh Principles and Logical Architecture
- Data Mesh explained

I. Lakehouse:

Se trata de una arquitectura que está promoviendo activamente Databricks, Snowflake, etc...

Un lakehouse es una arquitectura nueva y abierta que combina los mejores elementos de los Data Lakes y Data Warehouses. Los lakehouses son posibles gracias a un nuevo diseño del sistema: la implementación de estructuras de datos y funciones de gestión de datos similares a las de un data warehouse directamente sobre el almacenamiento en la nube de bajo coste en formatos abiertos. Son lo que se obtendría si se tuvieran que rediseñar los data warehouses en el mundo moderno, ahora que se dispone de un almacenamiento barato y altamente fiable (en forma de almacenes de objetos).



Un almacén de objetos tiene las siguientes características clave

Soporte de transacciones: En un Data Lake empresarial tipo Lakehouse, muchas cadenas de datos leerán y escribirán datos simultáneamente. La compatibilidad con las transacciones ACID garantiza la coherencia cuando

varias partes leen o escriben datos simultáneamente, normalmente utilizando SQL.

Aplicación de esquemas y gobernanza: Lakehouse debe tener una forma de soportar la aplicación y la evolución de los esquemas, soportando las arquitecturas de los esquemas de DW como los esquemas en estrella o en copo de nieve. El sistema debe ser capaz de razonar sobre la integridad de los datos, y debe contar con sólidos mecanismos de gobernanza y auditoría.

Soporte de BI: Lakehouse permiten utilizar herramientas de BI directamente en los datos de origen. Esto reduce el estancamiento y mejora la actualidad, reduce la latencia y disminuye el coste de tener que poner en funcionamiento dos copias de los datos tanto en un lago de datos como en un almacén.

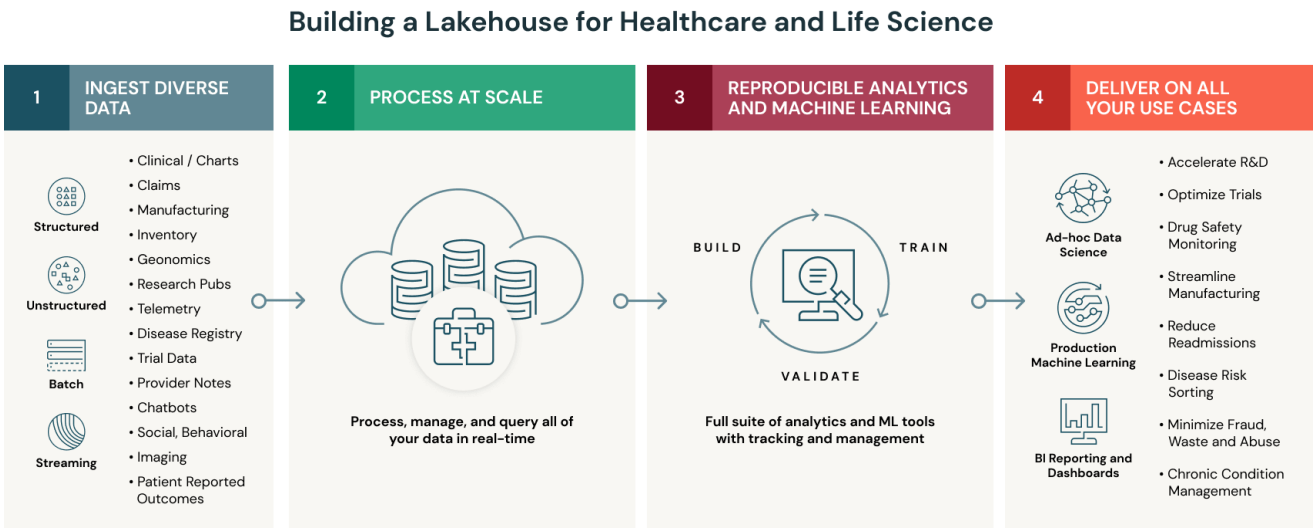
El almacenamiento está desvinculado del cálculo: En la práctica, esto significa que el almacenamiento y la computación utilizan clústeres separados, por lo que estos sistemas son capaces de escalar a muchos más usuarios concurrentes y a tamaños de datos más grandes. Algunos Data Warehouses modernos también tienen esta propiedad.

Apertura: Los formatos de almacenamiento que utilizan son abiertos y estandarizados, como Parquet, y proporcionan una API para que una variedad de herramientas y motores, incluyendo bibliotecas de aprendizaje automático y Python/R, puedan acceder directamente a los datos de manera eficiente.

Admiten diversos tipos de datos, desde los no estructurados hasta los estructurados: Lakehouse puede utilizarse para almacenar, refinar, analizar y acceder a los tipos de datos necesarios para muchas nuevas aplicaciones de datos, como imágenes, vídeo, audio, datos semiestructurados y texto.

Soporte para diversas cargas de trabajo: incluyendo la ciencia de los datos, el aprendizaje automático y el SQL y la analítica. Es posible que se necesiten varias herramientas para soportar todas estas cargas de trabajo, pero todas ellas dependen del mismo repositorio de datos.

Streaming de extremo a extremo: Los informes en tiempo real son la norma en muchas empresas. La compatibilidad con el streaming elimina la necesidad de contar con sistemas separados dedicados a servir aplicaciones de datos en tiempo real.



En pocas palabras, el sistema Lakehouse aprovecha el almacenamiento de bajo coste para mantener grandes volúmenes de datos en sus formatos brutos, al igual que los Data Lakes.

Al mismo tiempo, aporta estructura a los datos y potencia funciones de gestión de datos similares a las de los Data Warehouses mediante la implementación de la capa de metadatos en la parte superior del almacén. Esto permite que diferentes equipos utilicen un único sistema para acceder a todos los datos de la empresa para una serie de proyectos, como la ciencia de datos, el aprendizaje automático y la inteligencia empresarial.

Así que, a diferencia de los Data Warehouses, el sistema lakehouse puede almacenar y procesar muchos datos variados a un coste menor, y a diferencia de los Data Lakes, esos datos pueden ser gestionados y optimizados para el rendimiento de SQL.

La principal desventaja de un Lakehouse es que todavía es una tecnología relativamente nueva e inmadura. Por ello, no está claro si cumplirá sus promesas. Es posible que pasen años antes de que los data lakehouses puedan competir con las soluciones maduras de almacenamiento de big data.

Pero con la velocidad actual de la innovación moderna, es difícil predecir si una nueva solución de almacenamiento de datos podría llegar a usurparla.

Enlaces Recomendados:

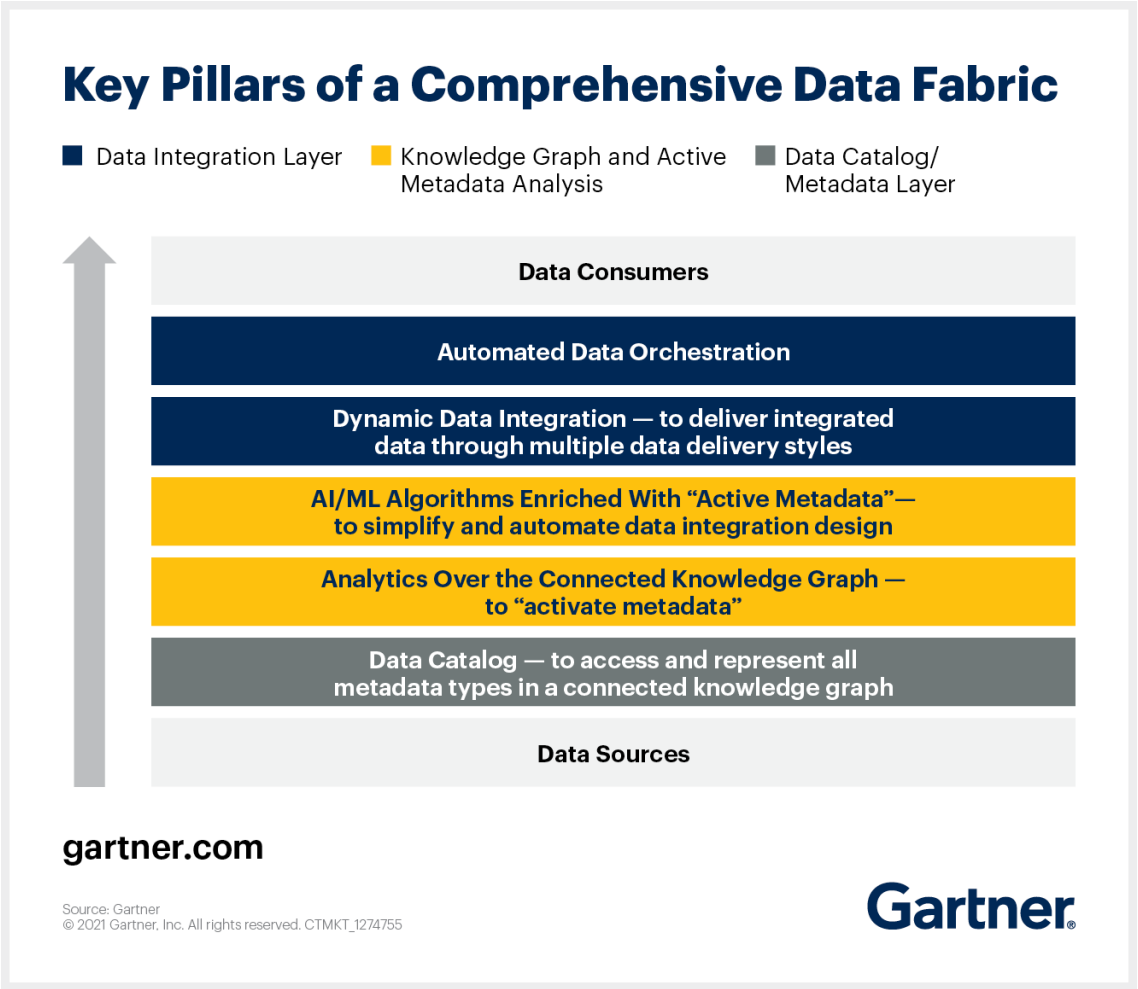
- Libro gratuito: Construyendo el Data Lakehouse
- Guía esencial del Data Lakehouse
- Data Lakehouse en Youtube

J. Data Fabric:

Data Fabric es una capa de arquitectura que conecta los datos y los procesos analíticos. Data Fabric es un marco, una estructura o un tejido, y este símil se utiliza para entender cómo se entrelazan los datos y los procesos bajo este concepto. En otras palabras, estamos hablando de una arquitectura unificada con servicios (o tecnologías) que corren por encima de ella y que ayudan a las empresas en la tarea de la gestión de datos.

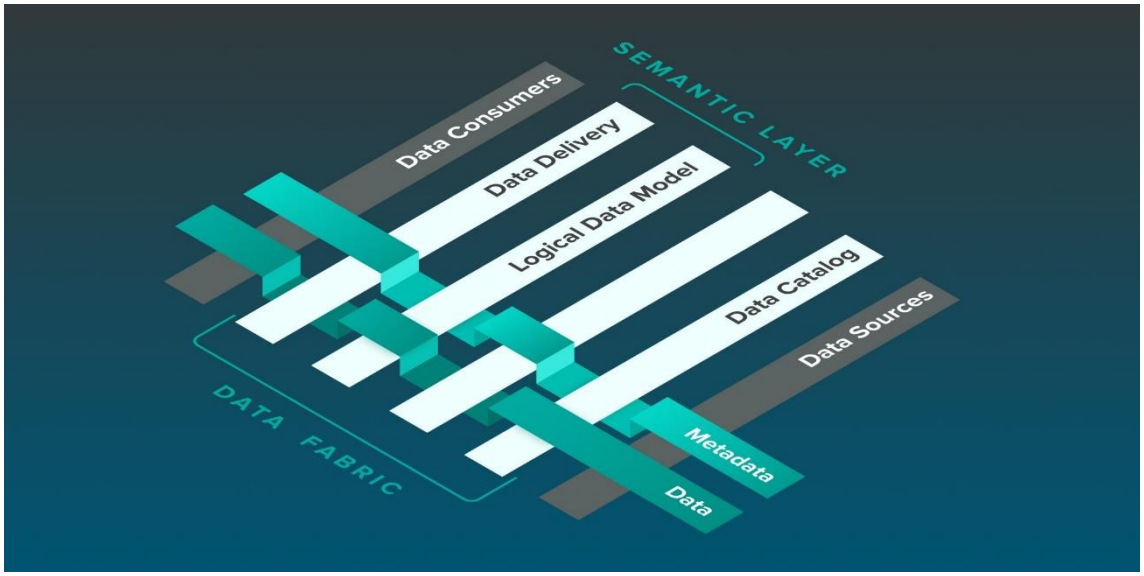
Maximizar el valor de los datos y acelerar la transformación digital son dos de sus principales beneficios, mientras que simplificar e integrar la gestión de datos en entornos Cloud (y también on-premise) son otras claras ventajas.

Data fabric simplifica e integra la administración de datos a través de la nube y en el on-premises, de esta forma acelera la transformación digital. Entrega servicios de datos para la nube híbrida de forma consistente e integrada, para otorgar visibilidad e insights, acceso y control de datos, además de seguridad y protección.



Algunas de las funciones específicas de un Data Fabric son:

- Unirse a cualquier fuente de datos mediante conectores y componentes pre-empaquetados, lo que elimina la necesidad de código.
- Proveer capacidades de integración y consumo de datos, ya sea entre fuentes o aplicaciones.
- Soporta Big Data, datos en tiempo real y por lotes.
- Administra múltiples ambientes: nube privada, híbrida, multinube, ya sean como fuente de datos o como consumidores de datos.
- Puede integrar capacidades de *data quality*, *data prepation* y *data governance*, reforzadas por machine learning.
- Soporta el intercambio de datos con stakeholders internos y externos, mediante APIs.



En última instancia, la implementación de un Data Fabric puede ayudar a una organización a superar sus retos de gestión de datos y convertirse en líderes digitales ya que permite:

- Proporcionar un único entorno para acceder y recopilar todos los datos, independientemente de dónde se encuentren y de cómo estén almacenados - eliminando los silos de datos
- Permitir una gestión de datos más sencilla y unificada, incluyendo la integración, la calidad, la gobernanza y el intercambio de datos, eliminando múltiples herramientas y proporcionando un acceso más rápido a datos más sanos y fiables
- Ofrecer una mayor escalabilidad que pueda adaptarse a los crecientes volúmenes de datos, fuentes de datos y aplicaciones
- Facilitando el aprovechamiento de la nube al soportar entornos locales, híbridos y multi-nube y una migración más rápida entre estos entornos
  Reducir la dependencia de las infraestructuras y soluciones heredadas
- Preparar la infraestructura de gestión de datos para el futuro, ya que se pueden añadir nuevas fuentes de datos y puntos finales, junto con nuevas tecnologías, a la estructura de datos sin interrumpir las conexiones o implementaciones existentes.

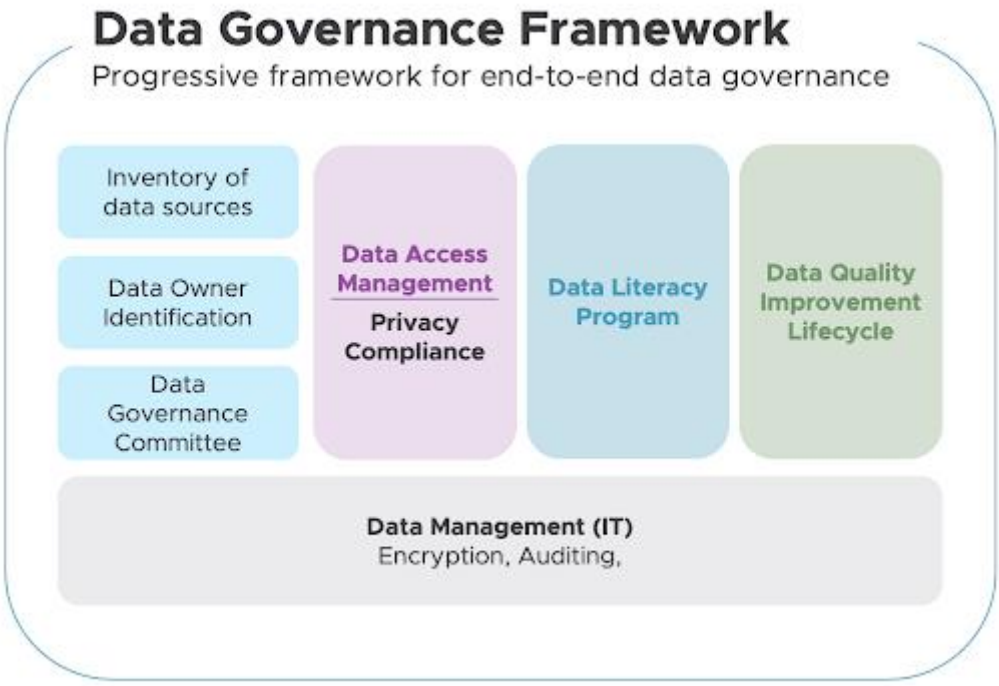
Enlaces Recomendados:

- Qué es un Data Fabric?

- Usar Data Fabric para modenizar las Arquitecturas de Datos
- Data Fabric en Youtube

K. Data Governance:

Data Governance es el proceso de organizar, asegurar, gestionar y presentar los datos utilizando métodos y tecnologías que garanticen que siguen siendo correctos, coherentes y accesibles para los usuarios verificados. Consiste en mejorar la calidad y la fiabilidad de los datos y facilitar el acceso a los mismos de acuerdo con la normativa



Data Governance es el proceso de:

Organizar: identificar todas las fuentes de datos y reunir todos los datos en un solo lugar.

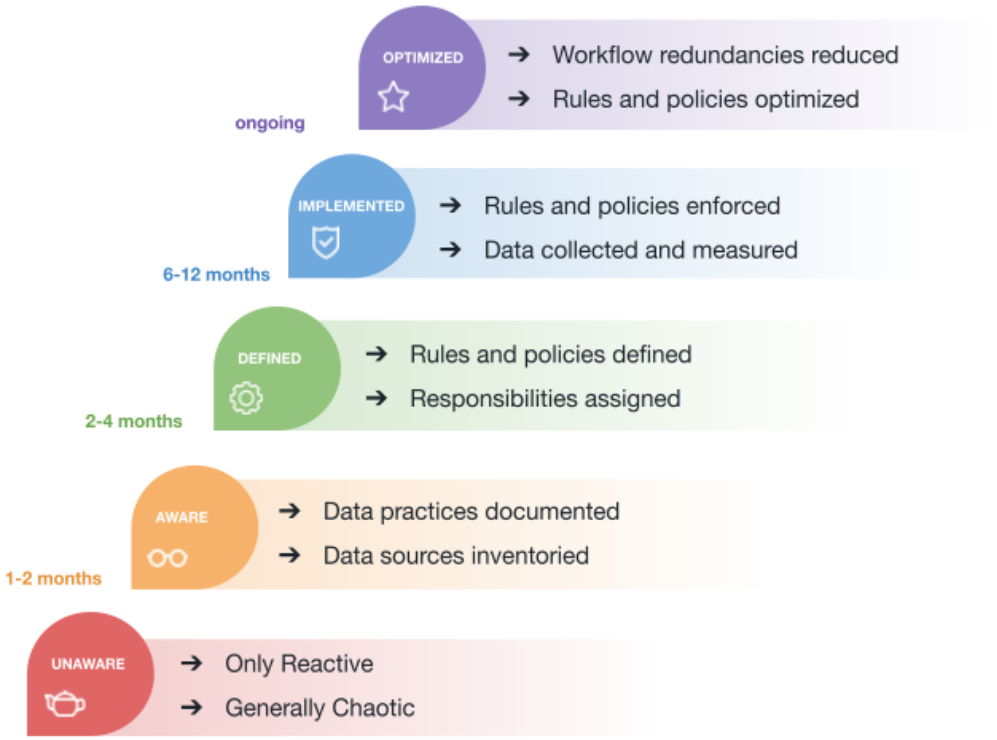
Proteger: asegurarse de que todos los datos cumplen la normativa sobre privacidad de datos y las políticas internas de la empresa.

Gestionar y presentar los datos: una vez que se han identificado los datos de la organización, hay que decidir cómo presentarlos al equipo.

Utilizar métodos y tecnologías - como las modernas plataformas de gobierno de datos, que garanticen que siguen siendo correctos, coherentes y accesibles para - las modernas plataformas de gobierno de datos. las

personas de su organización que tienen permiso para acceder a ellos, en definitiva - los usuarios verificados.

Tareas típicas de una implementación de Data Governance:



Dentro de un concepto global Data Management, el concepto de **Data Governance** se encarga de la dirección y supervisión de los datos, identificando además las siguientes áreas que están interrelacionadas:

· **Arquitectura de datos**: Define el plan de gestión de los activos de datos alineándose con la estrategia organizativa para establecer los requisitos de datos estratégicos y los diseños para cumplir con estos requisitos.

· **Modelización y diseño de datos**: Es el proceso de descubrir, analizar, representar y comunicar los requisitos de datos en una forma precisa llamada modelo de datos.

· **Almacenamiento y operaciones de datos**: Incluye el diseño, la implementación y el mantenimiento de los datos almacenados para maximizar su valor. Las operaciones se realizan a lo largo del ciclo de vida de los datos, desde la planificación hasta la eliminación de los mismos.

· **Seguridad de los datos**: La seguridad de los datos garantiza que la privacidad y la confidencialidad de los datos se mantengan, que no se violen los datos y que se acceda a ellos de manera adecuada.

· **Integración e interoperabilidad de datos**: Incluye procesos relacionados con el movimiento y consolidación de datos dentro y entre almacenes de datos, aplicaciones y organizaciones.

· **Gestión de documentos y contenidos**: Incluye las actividades de planificación, implementación y control utilizadas para gestionar el ciclo de vida de los datos y la información que se encuentran en una serie de medios, especialmente los documentos necesarios para apoyar los requisitos de cumplimiento legal y reglamentario.

· **Datos de referencia y maestros**: Incluye la conciliación y el mantenimiento continuos de datos compartidos fundamentales para permitir el uso coherente en todos los sistemas de la versión más exacta, oportuna y pertinente de la verdad sobre las entidades comerciales esenciales.

· **Data warehousing y BI**: Incluye los procesos de planificación, ejecución y control para gestionar los datos de apoyo a la toma de decisiones y permitir a los analistas de datos obtener valor de los datos mediante el análisis y la presentación de informes.

· **Metadatos**: Incluye las actividades de planificación, ejecución y control para permitir el acceso a metadatos integrados y de alta calidad que incluyen definiciones, modelos, flujos de datos y otra información crítica para comprender los datos y el sistema a través del cual se crean, se mantienen y se accede a ellos.

Calidad de los datos: Incluye la planificación y aplicación de técnicas de gestión de la calidad para medir, evaluar y mejorar la idoneidad de los datos para su uso dentro de una organización.

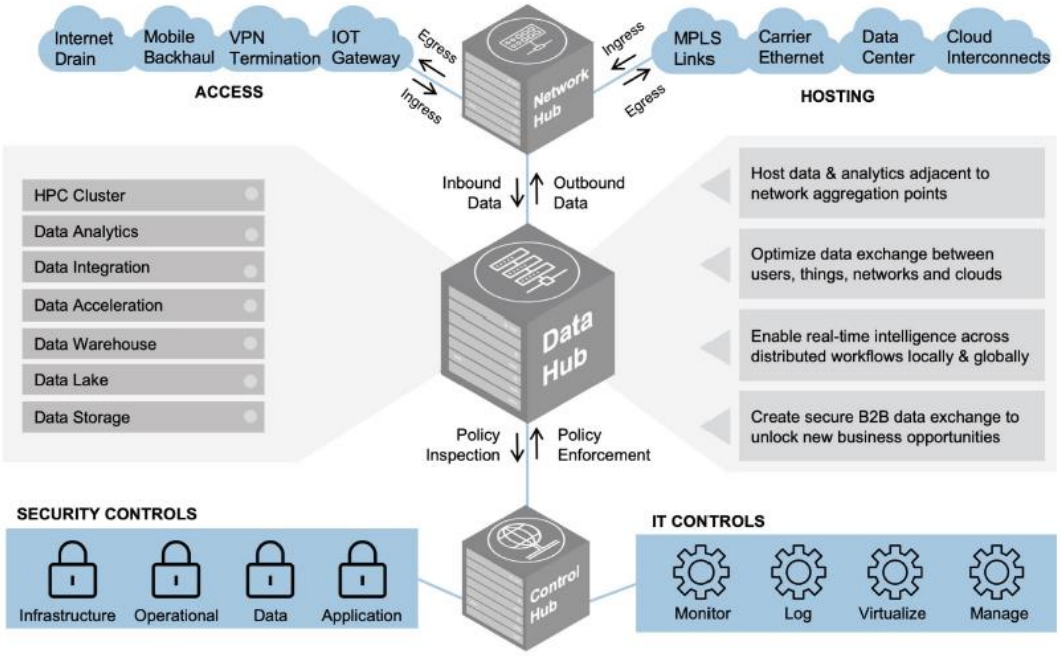
Enlaces Recomendados:

- Guía de Data Governance
- Comparación herramientas de Data Governance de Azure y Talend
- Data Governance en Youtube

L. Data Hub:

También se le conoce por su término en español, Centro de Datos. Tiene un enfoque más físico que los conceptos vistos anteriormente. Aquí el objetivo es poder integrar todos esos datos en un mismo punto, que sería el centro del que estamos haciendo referencia. En este caso los datos pueden moverse físicamente y tienen la capacidad de ordenarse de nuevo en otro sistema diferente.

A diferencia de otros sistemas, Data Hub permite que estos datos se ordenen, analicen o se descubran. El principal objetivo de Data Hub es que todos esos datos puedan contar con una fuente central teniendo en cuenta las diferentes necesidades comerciales que pueda tener una compañía.

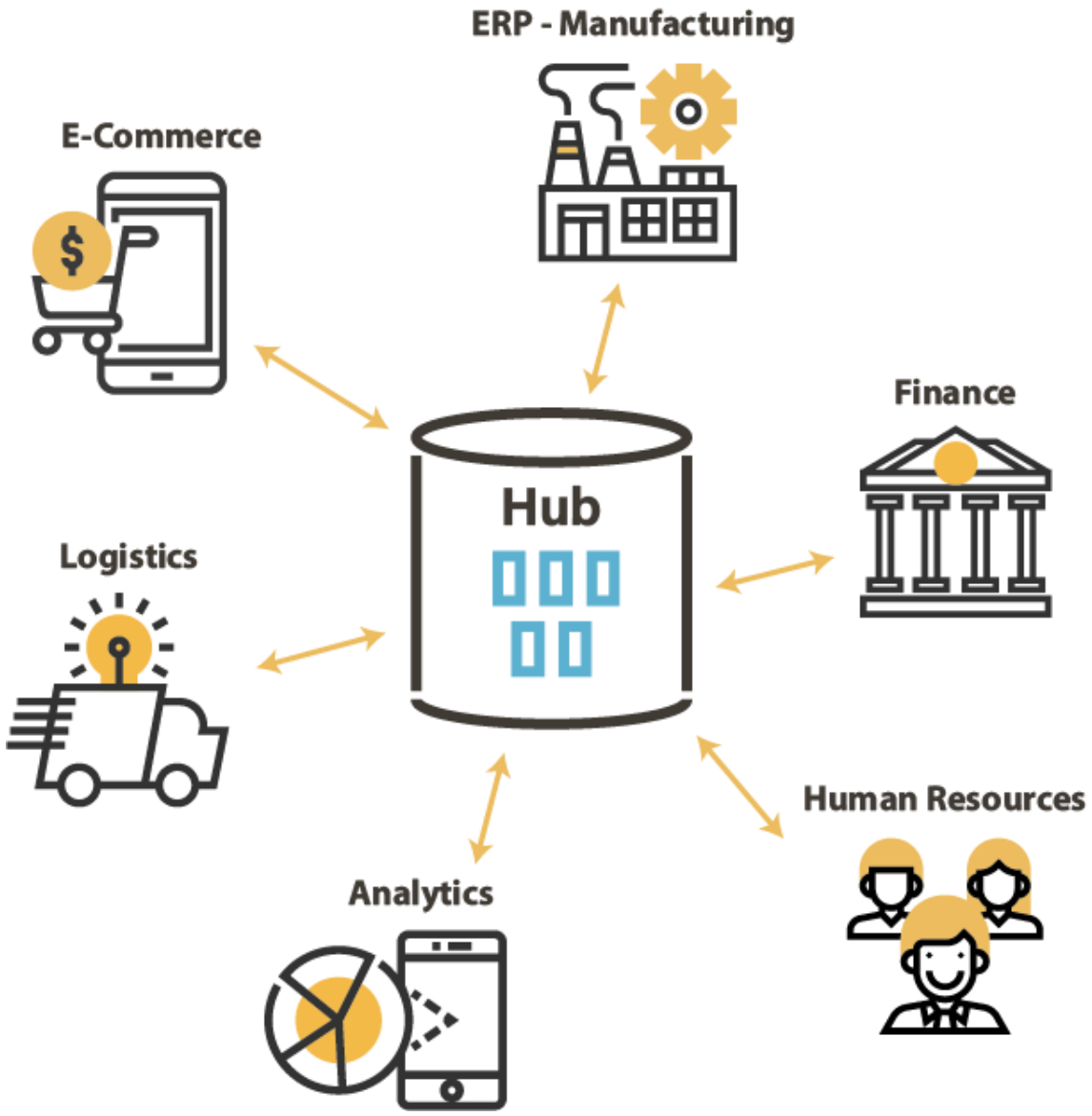


Un Data Hub es una arquitectura de almacenamiento moderna, y centrada en los datos que permite a las empresas consolidarlos y compartirlos para potenciar las técnicas de análisis y las cargas de trabajo de AI u otros. Si aún se accede a los datos con conexiones de punto a punto con silos independientes, convertir su infraestructura en un Data Hub optimizará en gran medida el flujo de datos en toda la organización.

Las ventajas frecuentes de los Data Hubs incluyen:

- Consolidación de silos en una única interfaz unificada para todos los datos
- Procesos de datos de alta velocidad, alta tasa de transferencia y alto rendimiento
- Visibilidad y accesibilidad a todos los datos
- Una interfaz unificada para la administración del almacenamiento de datos

Mientras que Data Warehouses y Data Lakes se entienden como puntos finales para la recopilación de datos que existen para respaldar el análisis de una organización, los Data Hubs sirven como puntos de intermediación e intercambio de datos.



Un centro de datos permite el intercambio de datos al conectar a los productores de datos con los consumidores de datos. Los puntos finales

interactúan con el centro de datos al proporcionar datos en él o recibir datos de él, y el centro proporciona un punto de gestión y mediación, lo que hace visible cómo fluyen los datos en la empresa.

Algunas de sus **características principales** son:

- Los Data Hubs se alimentan de una base de datos multimodelo subyacente (de la que carecen los Data Lakes y las bases de datos virtuales), lo que les confiere la capacidad de servir como sistema de verdad con toda la seguridad empresarial necesaria, incluida la confidencialidad de los datos (control de acceso), la disponibilidad de los datos (HA/DR) y la integridad de los datos (transacciones distribuidas).
- Los Data Hubs disponen de las herramientas necesarias para curar los datos (enriquecimiento, maestría, armonización) y admiten la armonización progresiva, cuyo resultado se mantiene en la base de datos.
- Los Data Hubs admiten aplicaciones operativas y transaccionales, algo para lo que los Data Lakes no están diseñados. Y, aunque las bases de datos virtuales pueden soportar transacciones, la carga se ve limitada por el rendimiento de los sistemas de bases de datos subyacentes

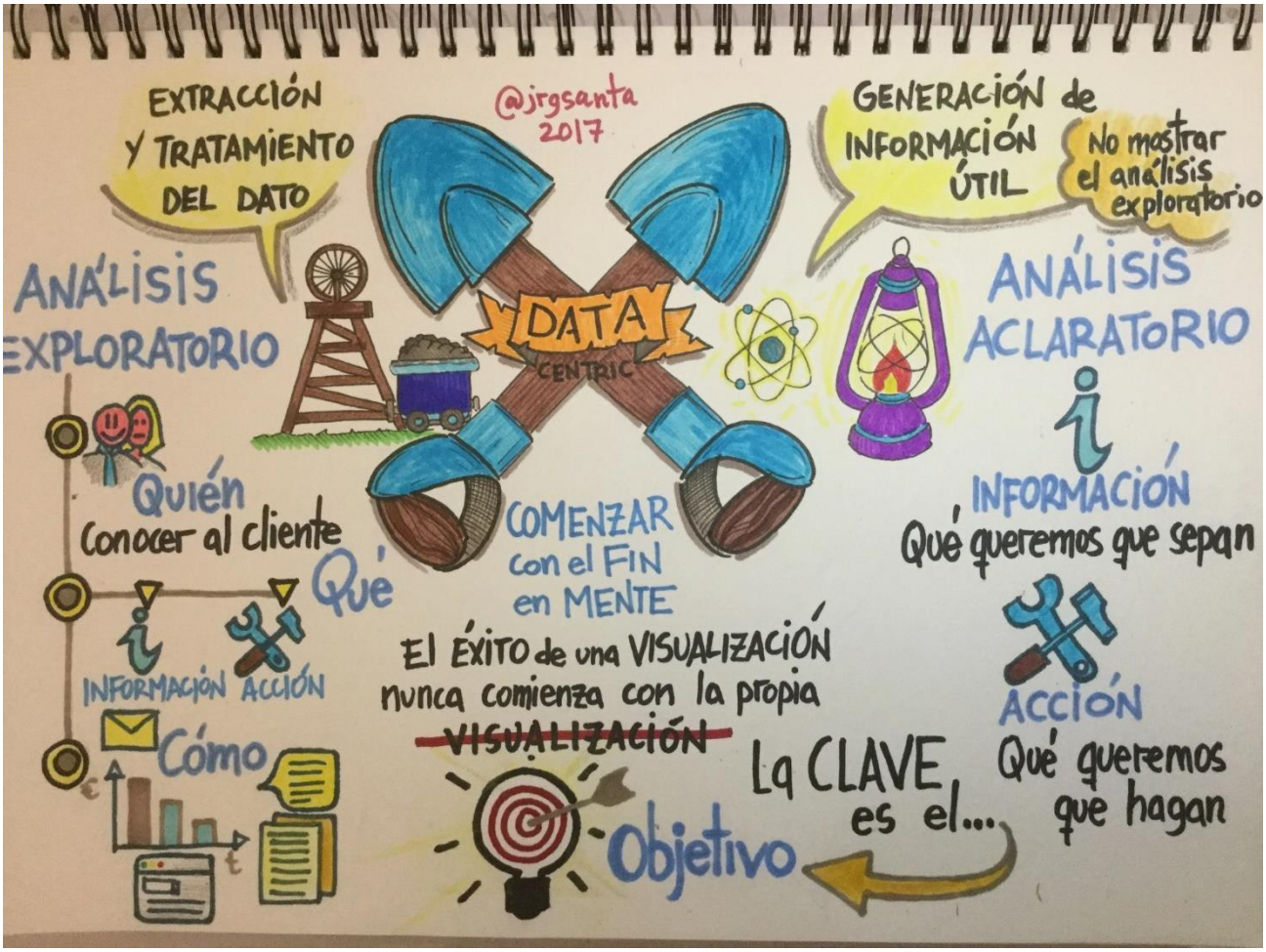
Enlaces Recomendados:

- Data Hub Wikipedia
- Tipos de Data Hubs (Gartner)
- Data Hub en Youtube

M. Data Centric/Data Driven:

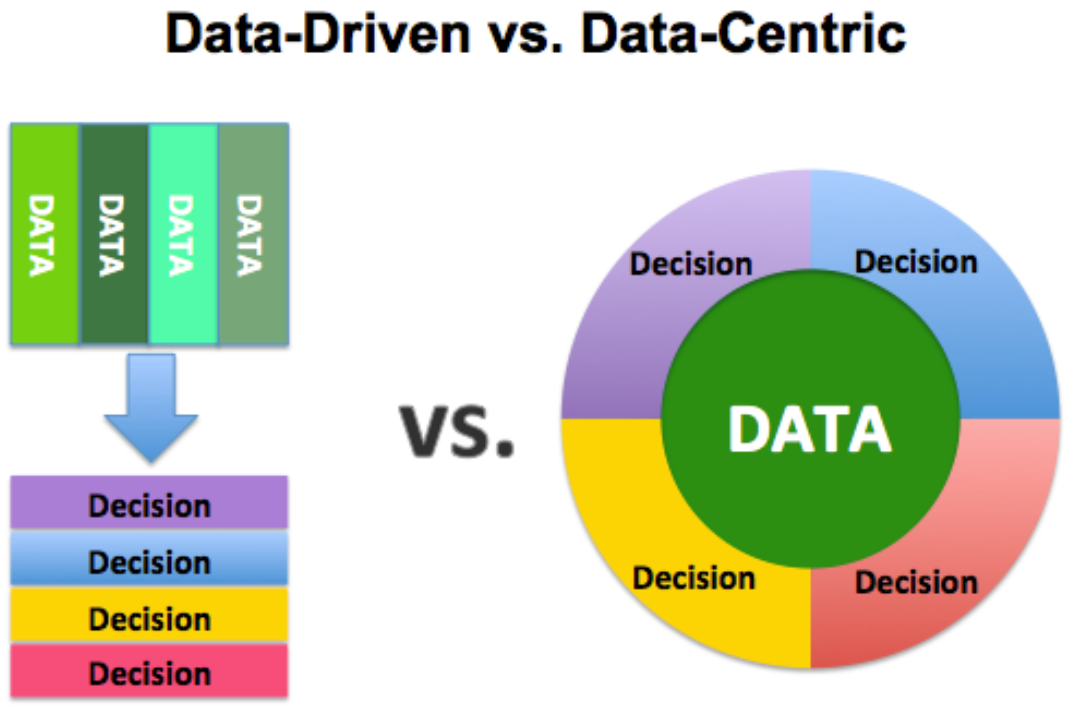
Data Centric es un concepto emergente que tiene relevancia en la arquitectura de la información y el diseño de los centros de datos. Describe un sistema de información en el que los datos se almacenan independientemente de las aplicaciones, que pueden actualizarse sin una costosa y complicada migración de datos.

La mayoría de las empresas se conforman con el desarrollo de infraestructuras y con la adquisición de herramientas que les permitan acceder y recopilar los datos. Pero solo un pequeño número de ellas va más allá y apuesta decisivamente por la creación de un departamento especializado en el dato que permita analizar esta información convenientemente, procesarla y enfocarla hacia el consumidor.



El concepto de Data Centric está mucho más ligado a la estrategia de la compañía, orientada a dar respuestas a negocio. Muy asociado al concepto de 'Data-Driven', en donde son los datos y no las áreas de negocio, usuarios, departamentos, tecnologías, los que cobran total imporancia y los que marcan las decisiones y estrategias de las compañías

Aunque se podrían encontrar ligeras diferencias, en cuanto a la secuencia e iteración del proceso, como se ve en la siguiente imagen, he preferido unir ambos conceptos en uno solo, pues tienen muchas analogías:



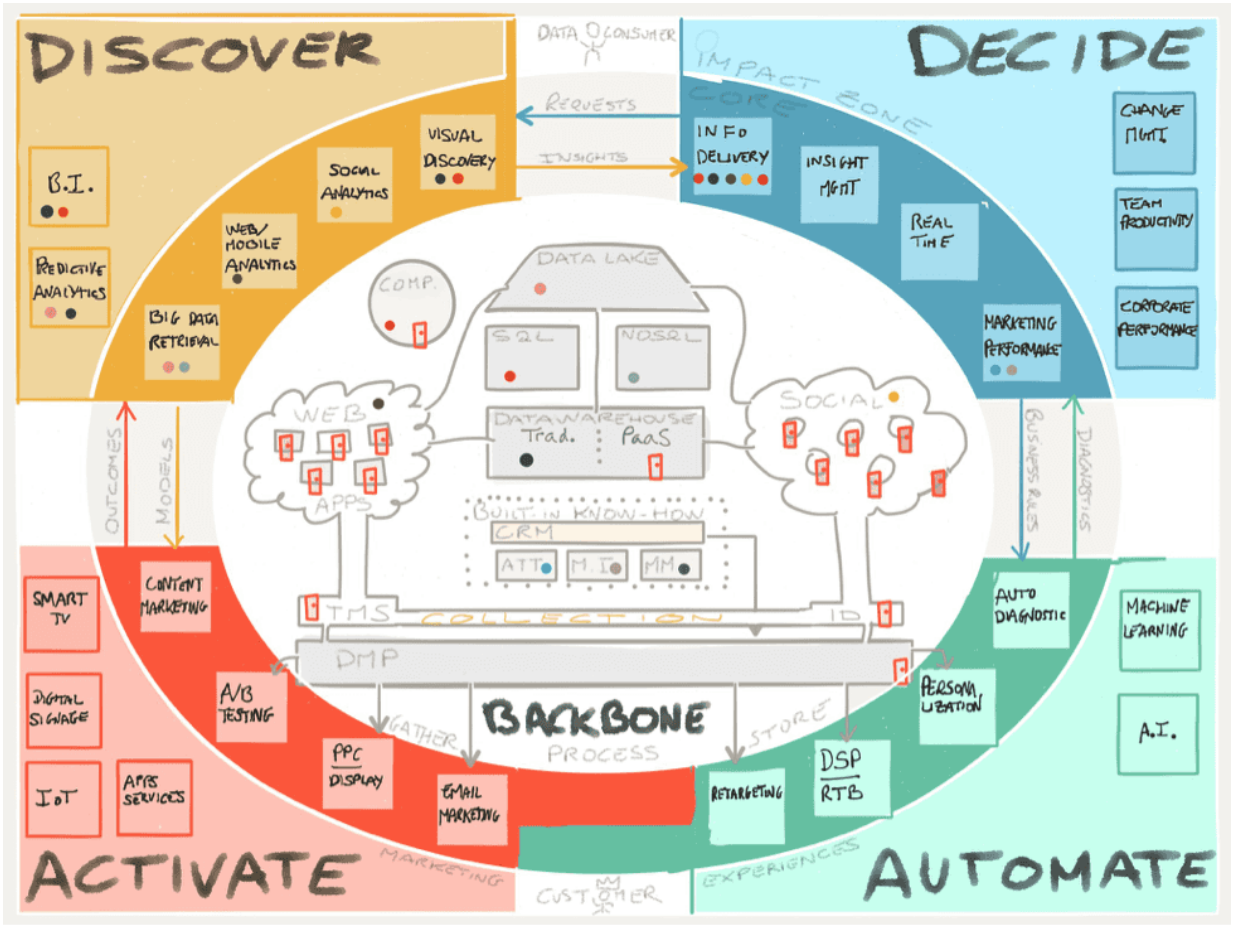
Algunas de las ventajas de adoptar esta estrategia serían las siguientes:

- Correcta organización de la información: Estructurar y distribuir adecuadamente los datos es básico para poder trabajar con ellos y tener claridad de ideas. Por lo tanto, darles su propio espacio dentro de la empresa y clasificarlos convenientemente es tremendamente útil para saber en todo momento lo que tenemos entre manos y para establecer prioridades.
- Evitar la duplicidad de datos: Contar con numerosas aplicaciones distribuidas a través de diferentes silos puede llevar a que nos encontremos con información repetida o redundante. Esto es un problema importante, porque genera costes relacionados con el almacenamiento de estos datos y puede provocar errores. Así,

centralizarlos es una estupenda forma de saber en todo momento lo que tenemos entre manos y de evitar perder el tiempo de forma innecesaria.

- Tener la capacidad de trabajar en tiempo real con los datos: Más que nunca, el tiempo es oro a la hora de tomar decisiones y la tecnología es un elemento fundamental para hacer más accesible el abordaje al Big Data. Por lo tanto, crear una infraestructura para su análisis y desarrollo también facilita la integración de herramientas basadas en la Inteligencia Artificial y el *Deep Learning*, para hacer mucho más digeribles los procesos de gran complejidad y estar en disposición de actuar en tiempo real.
- Diseñar estructuras de escalabilidad horizontal: Apostar por la modularidad ayuda a dar independencia a los microprocesos y permite potenciarlos de forma individual para que contribuyan a la estrategia global. De esta forma, se consigue una mayor flexibilidad y el desarrollo puede ser mucho más profundo.
- Dar un soporte de seguridad a los datos: Emplear la información con fines comerciales implica asumir una responsabilidad. Existen leyes para que los datos sean tratados y gestionados adecuadamente**,** y convertirnos en una empresa Data Centric es una estupenda manera que tenemos de garantizar que cumplimos con estas normas. A través de la especialización es posible entender adecuadamente la importancia de la información, así como desarrollar protocolos adecuados para asegurar un uso pertinente y regulado de los datos.

No hay que mirar muy lejos para ver casos de éxitos al utilizar datos para mejorar un servicio. Diferentes plataformas de la talla de Spotify, Netflix, Amazon y Facebook han mejorado la experiencia de los usuuarios y los contenidos que consumen gracias a la toma de decisiones *data-centric*.



Data Centric es una solución que aborda los problemas de la metodología convencional de proyectos y ofrece resultados positivos. Data Centric existe desde hace unos años y es cada vez más popular en todo tipo de sectores, donde muchos propietarios y operadores trabajan con un integrador de sistemas especializado.

La adopción de Data Centric no cambia las fases de desarrollo de un proyecto tradicional -Análisis y Diseño, backend, frontend, construcción y puesta en marcha y puesta en servicio-, sino que cambia radicalmente la forma en que se realiza el trabajo en estas fases, pues el dato, pasa a ser el protagonista

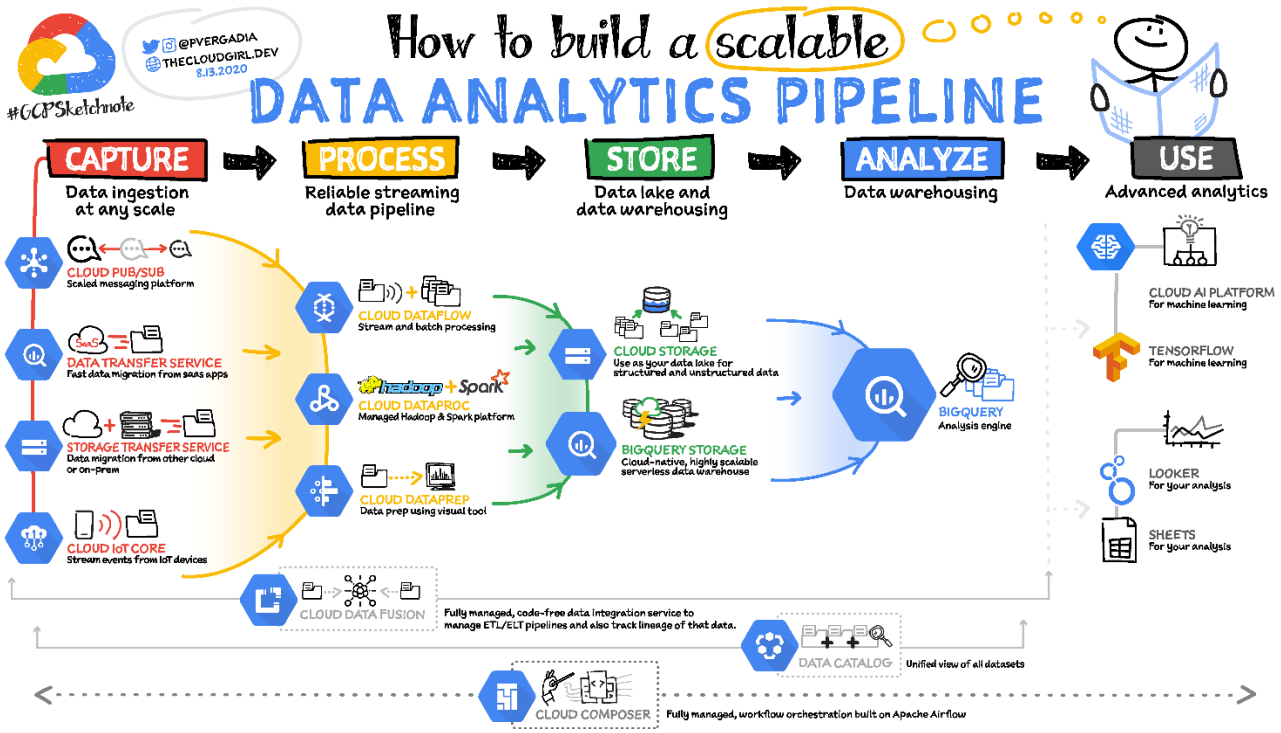
Enlaces Recomendados:

- The Data-Centric Revolution: Data-Centric vs. Data-Driven
- Libros sobre Data Driven
- Cómo convertir una empresa en Data Driven

N. Data Pipelines:

Un pipeline de datos es una construcción lógica que representa un proceso dividido en fases. Los Data Pipelines se caracterizan por definir el conjunto de pasos o fases y las tecnologías involucradas en un proceso de movimiento o procesamiento de datos.

Los Data Pipelines son necesarios ya que no debemos analizar los datos en los mismos sistemas donde se crean. El proceso de analítica es costoso computacionalmente, por lo que se separa para evitar perjudicar el rendimiento del servicio. De esta forma, tenemos sistemas OLTP, encargados de capturar y crear datos, y sistemas OLAP, encargados de analizar los datos.

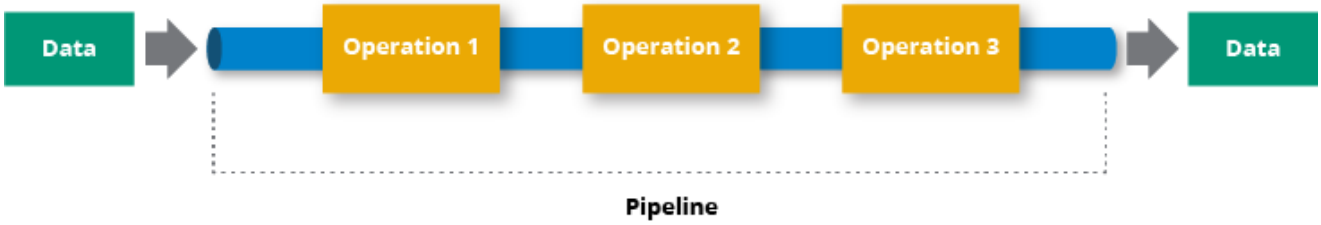


En una moderna arquitectura de datos los Data Pipelines se basan en una infraestructuras distribuidas y de alta disponibilidad diseñadas para ejecutar las actividades con tolerancia a errores.

Si se producen errores en la lógica de la actividad o en las fuentes de datos, los Data Pipelines vuelven a intentar la ejecución de la actividad automáticamente

Los Data Pipelines son, por tanto, una serie de pasos de procesamiento de datos. Si los datos no están cargados en la plataforma de datos, se introducen al principio de la cadena.

A continuación, hay una serie de pasos en los que cada paso proporciona un resultado que es la entrada del siguiente paso. Esto continúa hasta que se completa el proceso. En algunos casos, se pueden ejecutar pasos independientes en paralelo.



Los pipelines de datos constan de tres elementos clave:

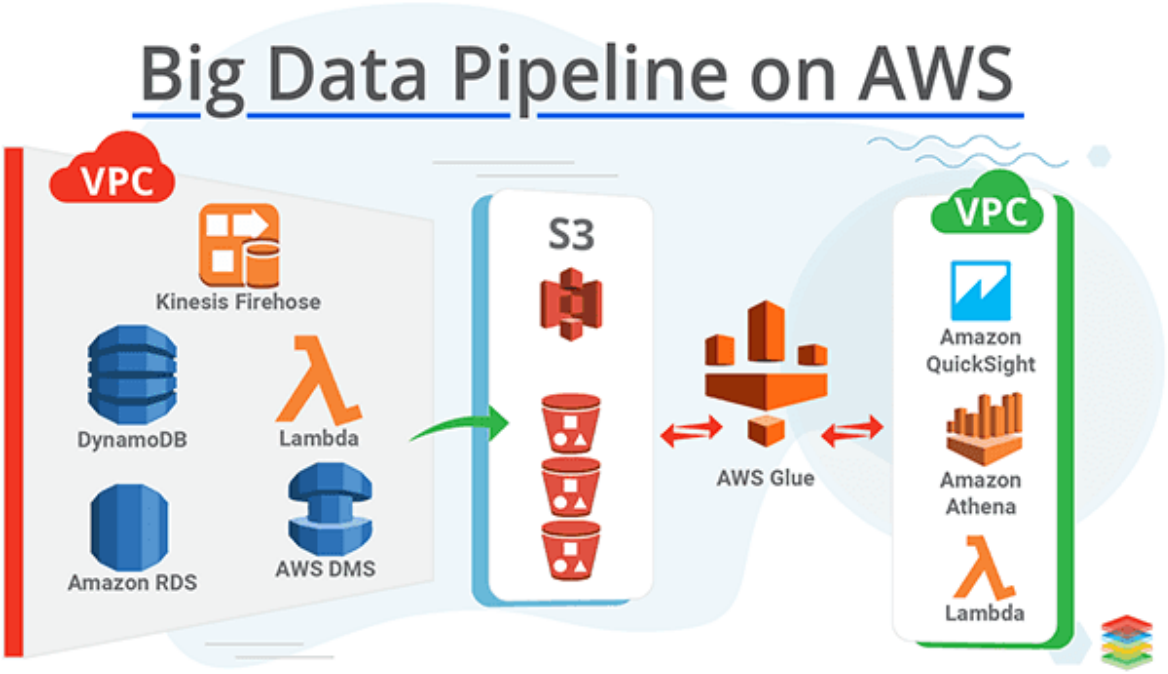
1) Una fuente

2) Un paso o pasos de procesamiento

3) Un destino.

En algunos pipelines de datos, el destino puede llamarse sumidero. Los pipelines de datos permiten el flujo de datos de una aplicación a un data warehouse, de un data lake a una base de datos de análisis, o a un sistema de procesamiento transaccional, por ejemplo.

Los pipelines de datos también pueden tener la misma fuente y el mismo sumidero, de manera que el pipeline se limita a modificar el conjunto de datos.



A medida que las organizaciones buscan construir aplicaciones con pequeñas bases de código que sirvan a un propósito muy específico (este tipo de aplicaciones se denominan "microservicios"), están moviendo datos entre más y más aplicaciones, haciendo que la eficiencia de los pipelines de datos sea una consideración crítica en su planificación y desarrollo.

Los datos generados en un sistema o aplicación de origen pueden alimentar múltiples canalizaciones de datos, y esas canalizaciones pueden tener otras múltiples canalizaciones o aplicaciones que dependen de sus salidas.

La ***diferencia entre ETL y Data Pipeline*** es interensate a considerar:

ETL se refiere a un tipo específico de canalización de datos. ETL significa "extraer, transformar, cargar". Es el proceso de mover datos desde una fuente, como una aplicación, a un destino, normalmente una Base de Datos. "Extraer" se refiere a sacar los datos de una fuente; "transformar" consiste en modificar los datos para que puedan cargarse en el destino, y "cargar" consiste en insertar los datos en el destino.

Históricamente, el ETL se ha utilizado para cargas de trabajo batch, especialmente a gran escala. Pero está surgiendo una nueva generación de herramientas ETL de flujo como parte de la canalización de datos de

eventos de flujo en tiempo real, real time, donde encaja mejor el concepto de Data Pipelines

Beneficios de usar Data Pipelines:

Las organizaciones actuales manejan cantidades ingentes de datos. Para analizar todos esos datos, se necesita una visión única de todo el conjunto de datos.

Cuando esos datos residen en múltiples sistemas y servicios, es necesario combinarlos de forma que tengan sentido para un análisis en profundidad. El flujo de datos en sí mismo puede ser poco fiable: hay muchos puntos durante el transporte de un sistema a otro en los que puede haber corrupción o cuellos de botella.

A medida que aumenta la amplitud y el alcance del papel que desempeñan los datos, los problemas no hacen más que aumentar en escala e impacto. Por ello, los Data Pipelines son fundamentales. Eliminan la mayoría de los pasos manuales del proceso y permiten un flujo de datos fluido y automatizado de una etapa a otra. Son esenciales para que los análisis en tiempo real le ayuden a tomar decisiones más rápidas y basadas en datos.

Son importantes si su organización:

- Confía en el análisis de datos en tiempo real
- Almacena datos en la nube
- Almacena datos en múltiples fuentes
- Al consolidar los datos de sus diversos silos en una única fuente de verdad, está asegurando una calidad de datos consistente y permitiendo un rápido análisis de datos para obtener información empresarial.

Enlaces Recomendados:

- What is Data Pipeline | How to design Data Pipeline ? - ETL vs Data pipeline
- Qué es un Data Pipeline?
- What is a Data Pipeline (and 7 Must-Have Features of Modern Data Pipelines)

Ñ. Cloud Analytics:

Cloud Analytics es un modelo de servicio en el que los elementos del proceso de análisis de datos se proporcionan a través de una nube pública o privada. Las aplicaciones y servicios de análisis en la nube suelen ofrecerse bajo un modelo de precios basado en la suscripción o en la utilidad (pago por uso)

Cloud Analytics es el proceso de almacenamiento y análisis de datos en la nube y su uso para extraer información empresarial procesable.

Al igual que la analítica de datos local, los algoritmos de la analítica en la nube se aplican a grandes sets de datos para identificar patrones, predecir resultados futuros y producir otra información útil para los responsables de la toma de decisiones empresariales.

Sin embargo, Cloud Analytics suele ser una alternativa más eficiente que la analítica local, que requiere que las empresas compren, alojen y mantengan costosos centros de datos.

Top Cloud Data Warehouses

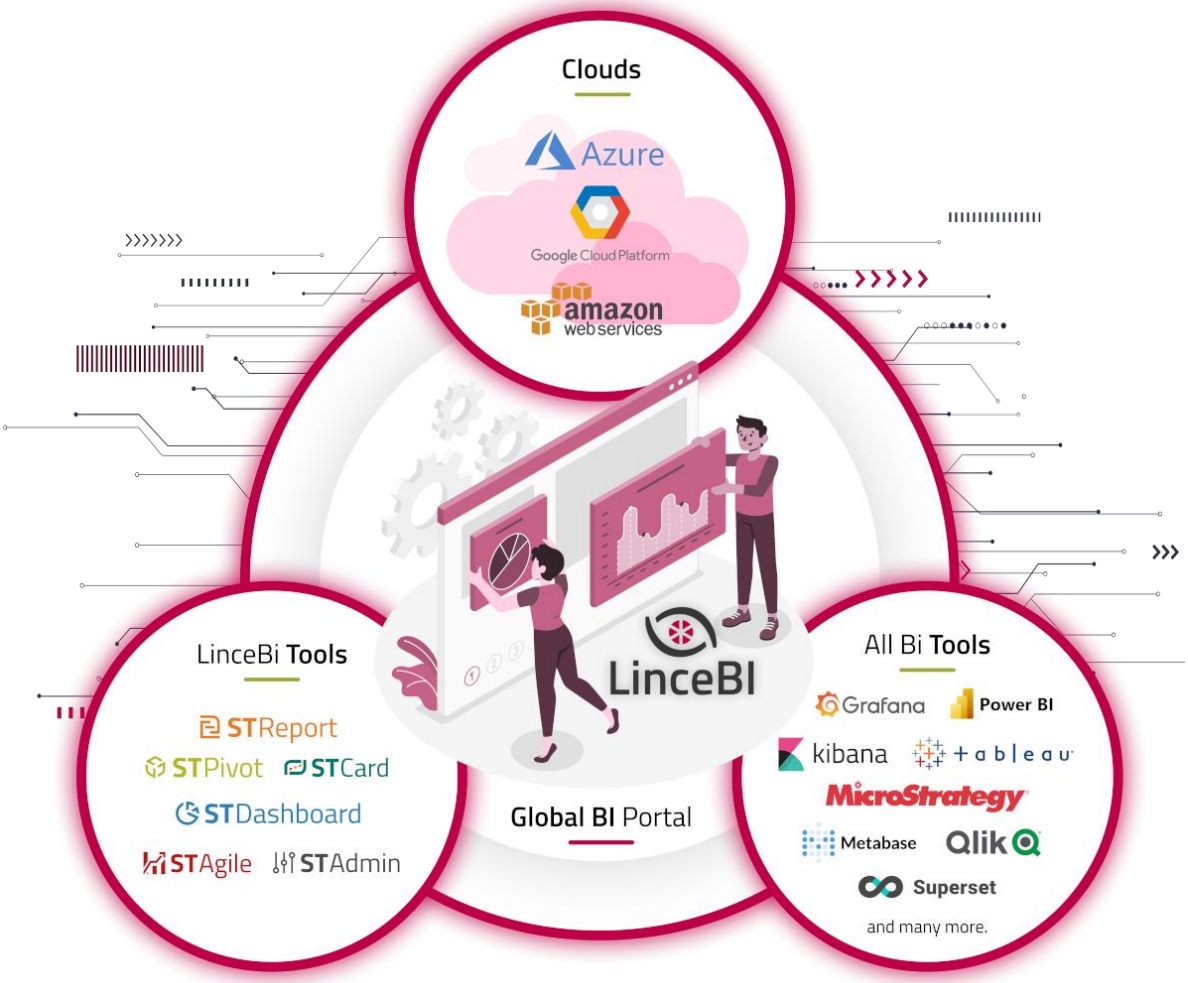
	Amazon Redshift	Microsoft Azure Synapse	Google BigQuery	Snowflake Cloud Data Platform
Initial Release	2012	2016	2010	2014
Separates Storage and Compute	No	Yes	Yes	Yes
Multi-Cloud	No	No	No	Yes
Query Language	Amazon Redshift SQL	TSQL	Standard SQL 2011 & BigQuery SQL	Snowflake SQL
Elasticity	Yes - Manual	Yes – Manual and Automatic	Yes – Automatic	Yes – Automatic
MPP	Yes	Yes	Yes	Yes
Columnar	Yes	Yes	Yes	Yes
Foreign Keys	Yes	Yes	No	Yes
Transaction	ACID	ACID	ACID	ACID
Concurrency	Yes	Yes	Yes	Yes
Durability	Yes	Yes	Yes	Yes
Automation	No	No	No	No
Website	Link	Link	Link	Link
Free Trial	Yes	Yes	Yes	Yes

Mientras que las soluciones analíticas locales ofrecen a las empresas un control interno sobre la privacidad y la seguridad de los datos, son difíciles y caras de ampliar.

La analítica en la nube, en cambio, se beneficia de la escalabilidad, los modelos de servicio y el ahorro de costes de la computación en la nube.

Las empresas generan terabytes de datos en el curso de sus operaciones diarias. Hoy en día, la mayor parte de estos datos -procedentes de sitios web, medios sociales, dispositivos informáticos y software financiero, entre otros- existen en la nube.

Las herramientas de análisis en la nube y el software de análisis son especialmente eficientes para procesar estos enormes conjuntos de datos, producir información en formatos fácilmente digeribles y crear información de los datos en la nube disponible bajo demanda, lo que resulta en una experiencia de usuario mejor y más ágil.



Las herramientas de Cloud Analytics y el software de análisis son particularmente eficientes para procesar estos enormes conjuntos de datos, produciendo perspectivas en formatos fácilmente digeribles bajo demanda que resultan en una experiencia de usuario mejor y más racionalizada.

Tipos de Cloud Analytics:

Nube pública

Las nubes públicas brindan aplicaciones como servicio a las empresas, como servidores virtuales, almacenamiento y procesamiento de datos. Son accesibles para el público en general y se basan en una arquitectura de múltiples inquilinos donde se comparten las redes de TI, pero los datos no. Las empresas pueden reducir costos y optimizar la gestión de TI como resultado de esto.

Nube privada

Las nubes privadas son nubes exclusivas que solo están disponibles para una sola empresa. Actúan como extensiones de la infraestructura de TI actual de una organización y solo están disponibles para la empresa. Se utilizan cuando la protección de datos y la confidencialidad son de suma importancia. La desventaja de este enfoque es el alto costo.

Nube híbrida

Las nubes híbridas se componen de nubes públicas y privadas. Estos marcos permiten a las empresas aprovechar la tecnología de TI bajo demanda para datos no confidenciales mientras mantienen la información confidencial en una nube privada.

Recomendaciones de plataformas Cloud Analytics según uso

- Big Data – AWS, GCP, Azure.
- Marketing digital – AWS, Azure, GCP.
- Comercio electrónico – AWS, GCP, Azure.
- Games – AWS, GCP.
- Gobiernos – AWS, Azure
- (Iot) – AWS, GCP, Azure
- Private Cloud – AWS, GCP, Azure.
- Reseller Hosting – AWS, GCP, Azure

Más allá de comparar el precio de estos tres pesos pesados de la nube, sus características también son un factor muy interesante a la hora de comparar.

En general, estas comparaciones son muy útiles cuando se considera qué partner de la nube es el más adecuado para el resultado deseado.

Por ejemplo, aunque todos pueden cubrir análisis de datos y visualización, se puede pensar que AWS es el más progresivo en esta área.

Los tres AWS, Azure y Google tienen su propia forma de categorizar los diferentes elementos, por lo que sugerimos comenzar a evaluar según las necesidades del proyecto y como cada herramienta se ajusta según sus características.

Desafortunadamente, a menudo se ven organizaciones que están tan comprometidas con Azure, por ejemplo, que no reconocen alternativas posiblemente más económicas y eficientes como AWS.

Características plataformas Cloud Analytics:

AWS: características

Al igual que los otros dos proveedores de servicios en la nube, AWS tiene diferentes algoritmos con nombres para desglosar sus productos y dividirlos en las siguientes categorías:

- Compute
- Storage
- Database
- Migration
- Networking & Content Delivery
- Developer Tools
- Management Tools
- Security, Identity & Compliance
- Analytics
- Artificial Intelligence
- Mobile Services

- Applications Services
- Messaging
- Business Productivity
- Desktop & App Streaming
- Software
- Internet of Things
- Game Development

Además de esta amplia gama de opciones, en AWS tienen productos específicos con un alto grado de categorización.

Estas soluciones cubren:

- Sitios web
- Copia de seguridad y recuperación
- Archivo
- Recuperación de desastres
- DevOps
- Big Data

Microsoft Azure: características

En la nube de Microsoft tenemos una amplia gama de opciones, muy parecidas a las de AWS con el diferencial de proporcionar ciertas capacidades basadas en usuarios.

Estos beneficios también incluyen la facturación flexible y precios competitivos, cabe destacar que Azure afirma tener un grado de certificación en estándares internacionales mayor a la de sus competidores.

Asegurar esta superioridad es una jugada bastante audaz del gigante de la computación, a pesar de tener un rango de características muy similar a AWS, sin embargo esto lo hacen con la finalidad de buscar que sus clientes depositen la confianza en su nube.

Google Cloud Platform: características

Aunque no es necesariamente el proveedor de computación en la nube más histórico, está lanzándose al mundo del Cloud Computing.

Google tiene tres puntos clave detrás de sus soluciones, destacando:

- Infraestructura a prueba de futuro
- Datos y análisis serios y potentes
- Sin servidor, solo código

Enlaces Recomendados:

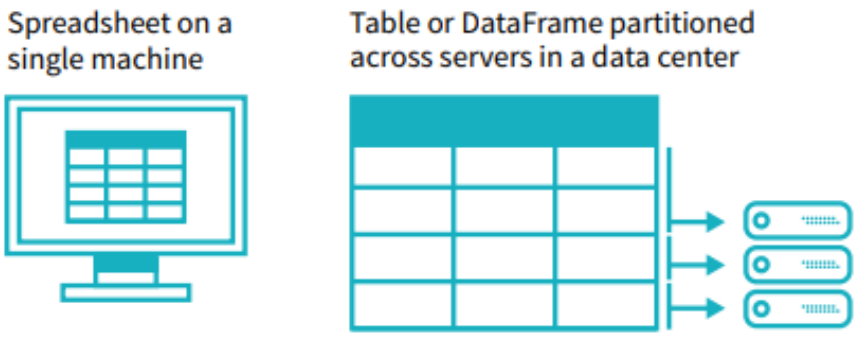
- Comparacion Amazon vs Azure vs Google vs Snowflake
- Cómo elegir la mejor solución Data Warehouse
- Comparación de arquitecturas Azure y AWS
- Videotutoriales sobre las mejores plataformas de Cloud Analytics

O. DataFrame:

Los **dataframes** son estructuras de datos de dos dimensiones (rectangulares) que pueden contener datos de diferentes tipos, por lo tanto, son heterogéneas. Esta estructura de datos es la más usada para realizar análisis de datos y es habitual si has trabajado con paquetes estadísticos.

El dataframe es la estructura fundamental para manipular conjuntos de datos en R. El dataframe se utiliza para guardar tablas de datos. Se puede considerar una lista de vectores de igual longitud que no tienen por qué ser del mismo tipo.

Un DataFrame es una estructura de datos que organiza los datos en una tabla bidimensional de filas y columnas, muy parecida a una hoja de cálculo. Los DataFrames son una de las estructuras de datos más utilizadas en la analítica de datos moderna porque son una forma flexible e intuitiva de almacenar y trabajar con los datos.



Cada DataFrame contiene un plano, conocido como esquema, que define el nombre y el tipo de datos de cada columna. Los DataFrames de Spark pueden contener tipos de datos universales como StringType y IntegerType, así como tipos de datos específicos de Spark, como StructType.

Los valores ausentes o incompletos se almacenan como valores nulos en el DataFrame.

Una analogía sencilla es que un DataFrame es como una hoja de cálculo con columnas con nombre. Sin embargo, la diferencia entre ellos es que mientras una hoja de cálculo se encuentra en un ordenador en una ubicación específica, un DataFrame puede abarcar miles de ordenadores. De este modo, los DataFrames permiten realizar análisis de big data, utilizando clusters de computación distribuidos.

In [10]:
```
ventas = pd.DataFrame({
    "Entradas": [41, 32, 56, 18],
    "Salidas": [17, 54, 6, 78],
    "Valoración": [66, 54, 49, 66],
    "Límite": ["No", "Sí", "No", "No"],
    "Cambio": [1.43, 1.16, -0.67, 0.77]
},
    index = ["Ene", "Feb", "Mar", "Abr"]
)
ventas
```

Out[10]:

	Entradas	Salidas	Valoración	Límite	Cambio
Ene	41	17	66	No	1.43
Feb	32	54	54	Sí	1.16
Mar	56	6	49	No	-0.67
Abr	18	78	66	No	0.77

La razón para poner los datos en más de un ordenador debería ser intuitiva: o bien los datos son demasiado grandes para caber en una sola máquina o simplemente se tardaría demasiado en realizar ese cálculo en una sola máquina.

El concepto de DataFrame es común en muchos lenguajes y marcos de trabajo diferentes. Los DataFrames son el principal tipo de datos utilizado en pandas, la popular biblioteca de análisis de datos de Python, y los DataFrames también se utilizan en R, Scala y otros lenguajes.

Ventajas del uso de Dataframes:

1.DataFrame es una colección distribuida de datos. En los DataFrames, los datos se organizan en columnas con nombre.

2. Son conceptualmente similares a una tabla en una base de datos relacional. Además, tienen optimizaciones más ricas.

3. Los DataFrames potencian las consultas SQL y la API de DataFrame.

4. Podemos procesar formatos de datos estructurados y no estructurados a través de él. Como por ejemplo: Avro, CSV, búsqueda elástica y Cassandra. Además, trata con sistemas de almacenamiento HDFS, tablas HIVE, MySQL, etc.

5. Hay bibliotecas generales disponibles para representar árboles. En cuatro fases, DataFrame utiliza la transformación de árboles:

- Análisis del plan lógico para resolver las referencias
- Optimización del plan lógico
- Planificación física
- Generación de código para compilar parte de una consulta a bytecode Java.

6. Las API de DataFrame están disponibles en varios lenguajes de programación. Por ejemplo, Java, Scala, Python y R.

7. Proporciona compatibilidad con Hive. Podemos ejecutar consultas Hive no modificadas en el almacén Hive existente.

8. Puede escalar desde kilobytes de datos en un solo portátil hasta petabytes de datos en un gran cluster.

9. DataFrame proporciona una fácil integración con las herramientas y el marco de Big data a través del núcleo de Spark.

Enlaces Recomendados:

- Cómo trabajar con DataFrames en R
- Cómo usar pandas para crear Dataframes
- Tutoriales de Dataframes en R

P. Dark Data:

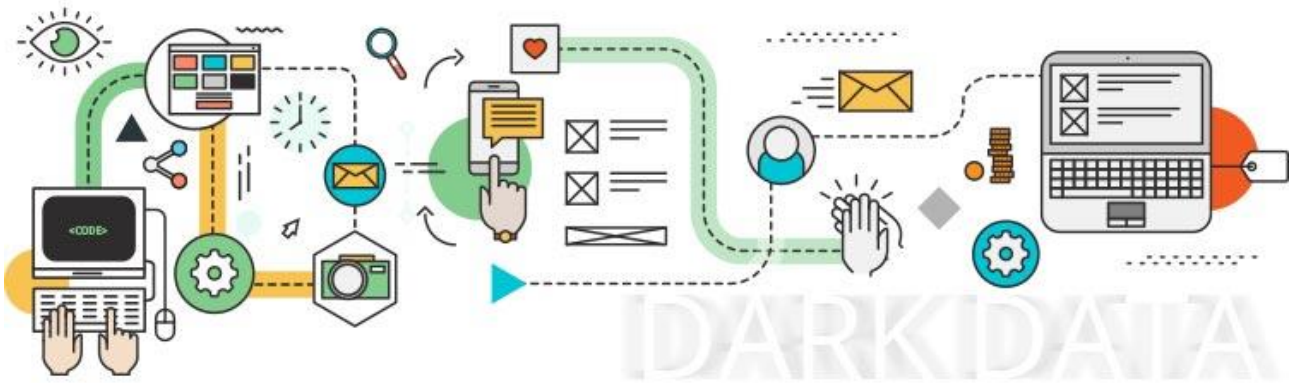
Dark data son datos adquiridos a través de operaciones en la red informáticas pero no usados para lograr conocimientos o para tomar decisiones. La capacidad que tenga una organización para recoger datos puede superar el throughput con el cual se puede analizar los datos. Hay casos en los que una organización puede no estar consciente de la acumulación y colección de datos

Las organizaciones suelen conservar el Dark Data a meros efectos de cumplimiento. Almacenar y proteger datos suele significar incurrir en más gastos (y en ocasiones mayor riesgo) que en valor.

El Dark Data comprende un tipo de datos no estructurados, no etiquetados y desaprovechados, que se encuentran en los almacenes de datos y no se han analizado ni tratado.

Son parecidos a los macrodatos pero se diferencia en que en su mayoría, los administradores de negocio y de TI ignoran el valor que tienen.

Con la explosión tecnológica, se crean datos constantemente. Todas las acciones y procesos que se realizan digitalmente –a través de Internet o de los sistemas y dispositivos que se utilizan en la organización– dejan una huella en forma de datos.



Además, en el mismo término se incluirían informes y documentos que las empresas guardan en los archivos físicos. La tarea de digitalizarlos impide

muchas veces tenerlos en cuenta, cuando constituyen un buen *background* de la actividad.

Una vez tengamos claro que los datos –incluido el dark data– pueden ser fuente de conocimiento para una organización, llega el momento de definir un objetivo, buscar los que necesitamos y tratarlos adecuadamente. Y es que de su identificación al verdadero uso hay un proceso y tecnología de por medio.Precisamente otro de los obstáculos a superar, además del desconocimiento del valor de los datos, son las barreras tecnológicas.

Muchas empresas no aprovechan la información porque no tienen la tecnología adecuada para tratar datos masivos.

Desde el IIC, trabajamos con diferentes clientes en el desarrollo de una infraestructura para almacenarlos y procesarlos, contando con la actitud proactiva de la organización en cuestión. Después, con la interpretación y análisis de los datos, implementamos distintas **soluciones tecnológicas** en diferentes ámbitos y sectores: energía, salud, legal, RR. HH., redes sociales y organizacionales, etc.

Con el objetivo de cada organización, nuestros expertos saben qué datos buscar, dónde buscarlos y cómo tratarlos adecuadamente para **extraer valor** de toda la información disponible. Todo esto sin olvidar la cantidad y calidad necesarias para que los resultados sean óptimos y ayuden a resolver problemas o mejorar diferentes aspectos del negocio.

Ejemplos de uso del Dark Data:

Considerad una solicitud bancaria para la aprobación de una tarjeta de crédito. La atención se centra en la información del cliente y en su elegibilidad. Sin embargo, la forma en que el usuario llegó al formulario en línea podría ser una información útil. Estos datos están disponibles pero no se utilizan.

En el ámbito médico, los historiales en papel están digitalizados. Los médicos los utilizan. Al estar almacenados como imágenes escaneadas, los sistemas de recuperación o análisis de información los ignoran.

Otro ejemplo es cuando una empresa adquiere usuarios o recoge opiniones a través de diferentes canales. Los registros del centro de llamadas se capturan como archivos de audio, mientras que las visitas al sitio web se

capturan como registros del servidor web. Esta es otra fuente de datos oscuros en la que no se puede acceder a nuevas perspectivas sólo porque las fuentes de datos existen de forma independiente y no se pueden combinar fácilmente.

Pensemos en un departamento de cuentas por pagar. Los datos oscuros podrían incluir resúmenes en el sistema ERP, transacciones anteriores, historial de cuentas, información de CRM, etc. Estos datos no sólo no se utilizan, sino que suponen un riesgo debido a su carácter sensible y confidencial.

¿Cuáles son los diferentes tipos o fuentes dark data?

Los datos oscuros (dark data) son diferentes en cada industria. Algunas categorías de datos oscuros incluyen información de clientes, información de exempleados, archivos de registro, datos de encuestas, estados financieros, notas, presentaciones, correos electrónicos, archivos adjuntos de correos electrónicos, bases de datos inactivas, versiones antiguas de documentos, transcripciones de centros de llamadas, comentarios de clientes, etc.

El Dark Data han surgido porque las organizaciones se dan cuenta de que los datos son valiosos y el almacenamiento es barato. Así que acaban almacenando muchos de ellos sin utilizarlos adecuadamente. Los data lakes y sus tecnologías asociadas ayudan a almacenar grandes volúmenes de datos multiformato.

Una razón importante para el Dark Data es el Big Data. Las características de los big data (volumen, variedad y velocidad) dificultan su procesamiento por parte de las organizaciones. Por eso lo almacenan y aplazan el procesamiento hasta más adelante. Desde esta perspectiva, se ha utilizado el término "datos polvorientos".

Cuando la credibilidad de los datos es esencial, los datos que no pueden rastrearse hasta su fuente no se utilizarán para el análisis.

Enlaces Recomendados:

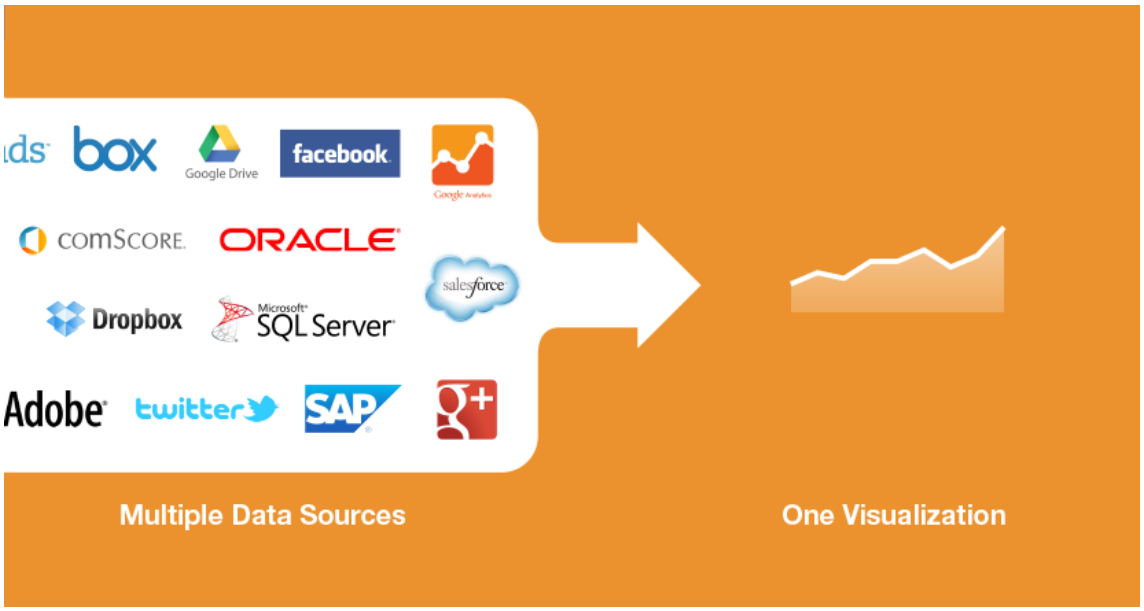
- Entender Dark Data
- Cómo extraer valor del Dark Data
- Es el Dark Data un riesgo para tu empresa?
- Cómo aplicando Dark Data se pueden conseguir miles de clientes ocultos

Q. Data Mashup:

El término "mashup" surgió en la esfera de los medios de comunicación en la última década, cuando la gente aprovechó las oportunidades que ofrecían las nuevas tecnologías de software y hardware de medios. Se hizo mucho más fácil combinar fragmentos de canciones, vídeos o gráficos de diferentes fuentes para crear contenidos nuevos y diversos.

Más recientemente, el concepto y el término se ampliaron para incluir aplicaciones de contenidos web, como portales definidos por el usuario que pueden combinar fuentes RSS, otros contenidos sindicados o incluso contenidos no sindicados para crear un producto personalizado de alto valor para su distribución y consumo.

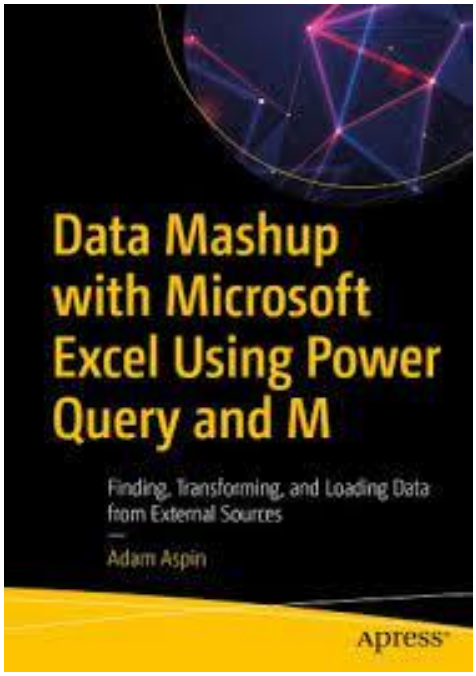
Un Data Mashup empresarial es la integración de datos y aplicaciones digitales heterogéneos procedentes de múltiples fuentes con fines empresariales. Un mashup empresarial también se conoce a veces como mashup de negocios o, con menos precisión, como mashup de datos.



El término mashup empresarial se utiliza a menudo para diferenciar un mashup relacionado con la empresa de un mashup web, que suele ser para el público en general. En general, un mashup se crea a partir de

componentes modulares que el usuario final puede ensamblar y reensamblar como desee para satisfacer sus necesidades actuales.

En un mashup empresarial, el producto suele ser una combinación de datos y aplicaciones corporativas internas con datos de origen externo, SaaS (software como servicio) y contenido web.



Características:

Otras características que diferencian al mashup empresarial son la integración con el entorno informático de la empresa, la gobernanza de los datos, la inteligencia empresarial (BI) / el análisis empresarial (BA), herramientas de programación más sofisticadas y medidas de seguridad más estrictas.

Mashup es una palabra que proviene de un término musical en inglés, que significa la creación de una nueva canción a partir de la mezcla o pedazos de otras canciones. Desde este concepto se basa el mashup de software.

La Wikipedia lo define como “una aplicación o sitio web que combina contenido de una o más fuentes dentro de una nueva experiencia de usuario o manejo de información”.

Los mashups son un producto de la Web 2.0 donde el usuario es el centro de todo. Va de la mano con conceptos como colaboración y distribución de la información.

Las características que podemos encontrar actualmente en el Web 2.0 son:

> Hecho por el usuario para él mismo.
> Capacidad dinámica de integración de información con otras fuentes.
> Integración concurrente limitada.
> Utilización de servicios Web públicos.
> Orientado al consumidor.

Sin embargo, conforme el Web 2.0 comienza a ser adoptado en las empresas, empezamos a ver su evolución, que llevará a las siguientes características en el futuro:

> Hecho por el usuario para él y compartirlo con más usuarios.
> Capacidad dinámica de compartir e integrar de la misma manera con otras fuentes.
> Utilización tanto de servicios Web públicos, así como servicios internos.
> Orientado hacia la empresa, sus clientes y aliados de negocio.

Enlaces Recomendados:

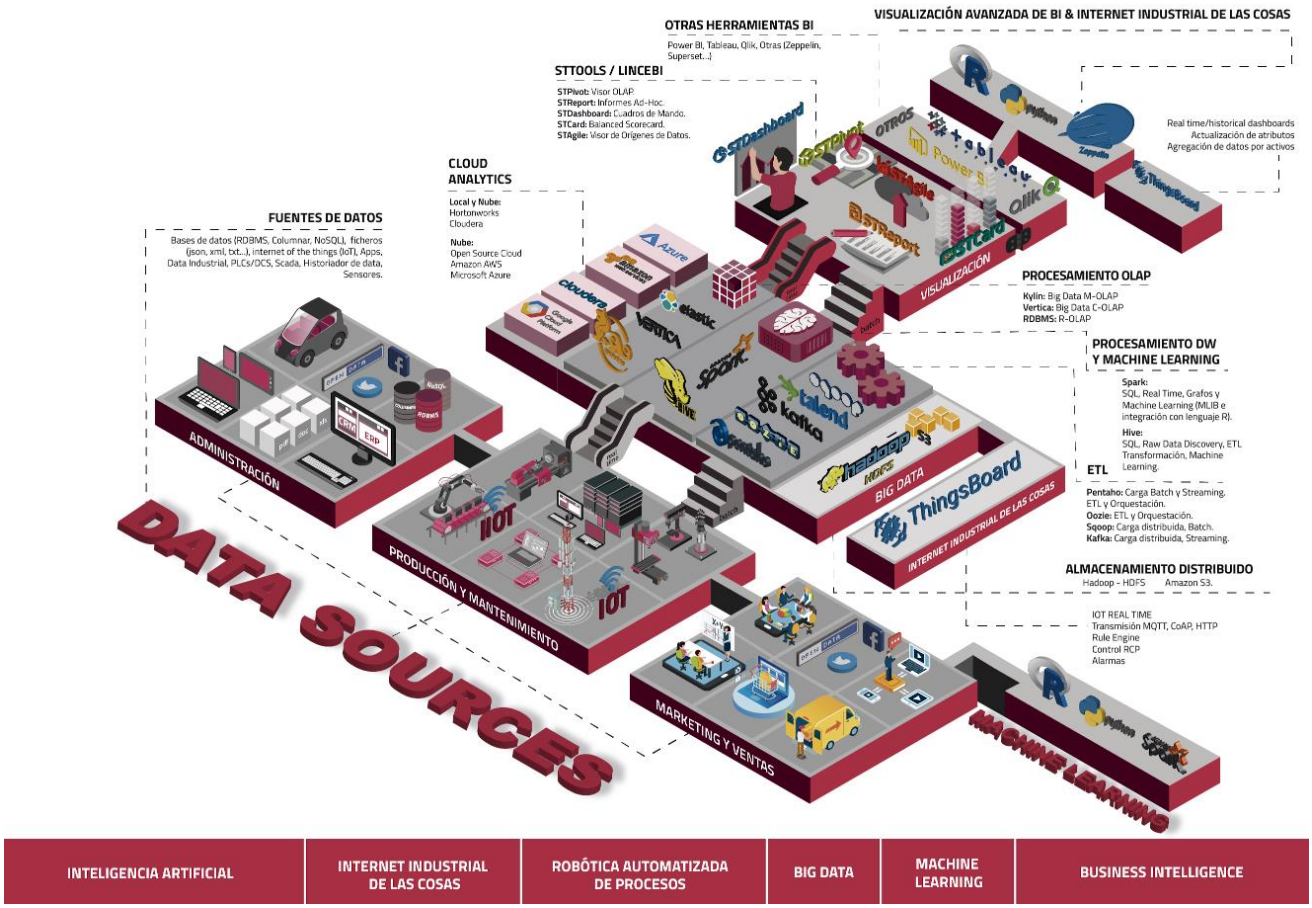
- What Is Data Mashup and What Are Its Benefits?
- What is Mashup - Definition, meaning and examples
- Data Mashup con Microsoft
- Business Intelligence Data Mashup

R. Data Sources:

Una fuente de datos es un lugar donde se obtiene la información. La fuente puede ser una base de datos, un archivo plano, un archivo XML o cualquier otro formato que un sistema pueda leer. La entrada se registra como una colección de registros que contienen información utilizada en el proceso empresarial. Esa información puede incluir detalles de los clientes, cifras de contabilidad, ventas, logística, etc.

Una data source, en sentido amplio, puede ser mcuhas cosas:

(1) El lugar físico o digital donde se almacenan los datos en cuestión en forma de tabla de datos (u otro formato),
(2) El grado de originalidad de una tabla de datos
(3) Una marca de proveedor de datos
(4) Los datos utilizados a través de una herramienta de datos de autoservicio como Excel, Tableau o Power BI
(5) El tipo de almacenamiento informático, es decir, fuente de datos de archivo o fuente de datos de máquina
(6) Una base de datos técnica como Amazon AWS o Microsoft Azure
(7) Una fuente de datos heredada con un nombre propio dentro de una organización
(8) Un tipo de datos como acciones, contabilidad o indicadores económicos.



Las fuentes de datos pueden diferir según la aplicación o el campo en cuestión. Las aplicaciones informáticas pueden tener definidas múltiples fuentes de datos, dependiendo de su propósito o función.

Aplicaciones como los sistemas de gestión de bases de datos relacionales e incluso los sitios web utilizan bases de datos como fuentes de datos primarias.

El hardware, como los dispositivos de entrada y los sensores, utilizan el entorno como fuente de datos primaria.

Un buen ejemplo es un sistema de control de la temperatura y la presión de un sistema de circulación de fluidos, como los que se utilizan en las fábricas y las refinerías de petróleo, que toman todos los datos relacionados del entorno o de lo que estén controlando; así que la fuente de datos aquí es el entorno.

Los datos, como la temperatura y la presión del fluido, son tomados por sensores regularmente y luego almacenados en una base de datos, que luego se convierte en la fuente de datos principal para otra aplicación informática que manipula y presenta estos datos.

Una fuente de datos se utiliza más comúnmente en el contexto de las bases de datos y los sistemas de gestión de bases de datos o cualquier sistema que se ocupa principalmente de los datos, y se conoce como un nombre de fuente de datos (DSN), que se define en la aplicación para que pueda encontrar la ubicación de los datos.

Simplemente significa lo que las palabras significan: de dónde vienen los datos.

Tipos de fuentes de datos:

- **Biométricos.** Son los referidos a la identificación automática de una persona basada en sus características anatómicas o trazos personales, como la **firma biométrica.** Hablamos tanto de reconocimiento facial pero también genético (ADN).

- **Máquina a máquina.** Se refiere a Internet de las Cosas, son aquellas tecnologías que permiten la conexión de diferentes dispositivos entre sí. Un ejemplo son los GPS, pero también los denominados chips NFC (aquella tecnología que se sustenta en la comunicación inalámbrica y que permite la transmisión de datos de forma segura: integrada fundamentalmente en smarphone y tablets) . Todo un mundo de posibilidades que puede hallarse también en parquímetros, cajeros, máquinas expendedoras…

- **Datos de transacciones.** Los datos que se registran en los departamentos de facturación forman parte de las operaciones normales que se producen en las transacciones habituales. También están los centros de llamada, mensajería, reclamaciones, presentación y registro de documentos y los que se generan con los pagos por tarjeta, pago online.

- **Generados por los humanos.** Todas aquellas grabaciones a operadores de atención al cliente: Call Center, también los e-mail o los registros médicos electrónicos.

- **Web y medios sociales.** Son los que se originan en la red y configuran, según los expertos, el trozo más grande del pastel llamado Big Data y es una de las fuentes de datos más utilizadas en la actualidad. Hablamos de la información que se genera sobre clicks en vínculos y elementos. Pero también de toda aquella contenida en las búsquedas que realizamos por ejemplo en Google, las publicaciones en las Redes sociales (Twitter, Facebook, Linkedin…) y el contenido web como páginas, enlaces o imágenes.

Los datos que las empresas analizan a través de la inteligencia empresarial provienen de un tipo de fuentes de datos muy diversas. Los ejemplos más comunes son:

- Databases
- Flat files
- Web services
- Other sources, como RSS feeds

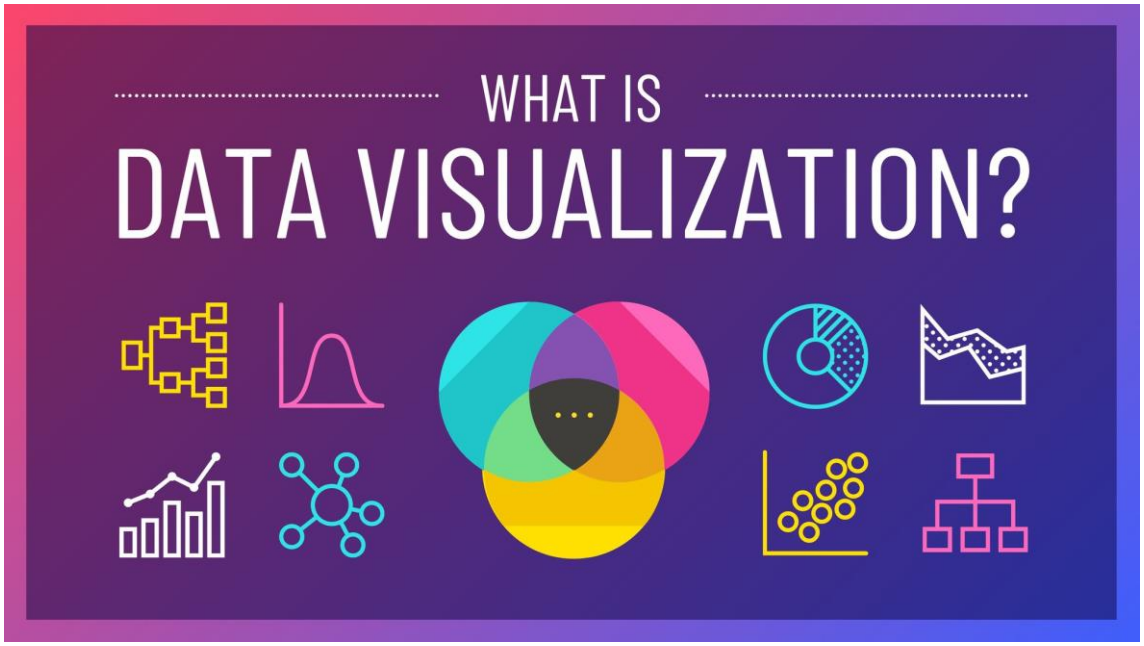
Enlaces Recomendados:

- 70 Amazing Free Data Sources You Should Know
- Public Health Research Guide: Primary Data Sources
- Cómo encontrar las mejores fuentes de conjuntos de datos públicos y gratuitos
- Fuentes de Datos usadas en Business Intelligence

S. Data Visualization:

La visualización de datos es la práctica de trasladar la información a un contexto visual, como un mapa o un gráfico, para que el cerebro humano pueda entender los datos y extraer conclusiones. El objetivo principal de la visualización de datos es facilitar la identificación de patrones, tendencias y valores atípicos en grandes conjuntos de datos

La visualización de datos e información (data viz o info viz) es un campo interdisciplinar que se ocupa de la representación gráfica de datos e información. Es una forma especialmente eficaz de comunicar cuando los datos o la información son numerosos, como por ejemplo una serie temporal



También es el estudio de las representaciones visuales de datos abstractos para reforzar la cognición humana.

Los datos abstractos incluyen tanto datos numéricos como no numéricos, como texto e información geográfica. Está relacionada con la infografía y la visualización científica. Una distinción es que es visualización de la información cuando se elige la representación espacial (por ejemplo, el

diseño de la página de un gráfico), mientras que es visualización científica cuando se da la representación espacial

Desde un punto de vista académico, esta representación puede considerarse como un mapeo entre los datos originales (normalmente numéricos) y los elementos gráficos(por ejemplo, líneas o puntos en un gráfico).

El mapeo determina cómo varían los atributos de estos elementos en función de los datos. En este sentido, un gráfico de barras es un mapeo de la longitud de una barra a una magnitud de una variable. Dado que el diseño gráfico del mapeo puede afectar negativamente a la legibilidad de un gráfico, el mapeo es una competencia fundamental de la visualización de datos.

La visualización de datos e información tiene sus raíces en el campo de la estadística y, por tanto, suele considerarse una rama de la Estadística descriptiva. Sin embargo, dado que para visualizar eficazmente se requieren tanto habilidades de diseño como de estadística y computación, muchos sostienen que es tanto un arte como una ciencia.

La investigación sobre cómo las personas leen y malinterpretan diversos tipos de visualizaciones está ayudando a determinar qué tipos y características de las visualizaciones son más comprensibles y eficaces para transmitir información

Características de las visualizaciones eficaces:

Características de las presentaciones gráficas eficaces

El mayor valor de una imagen es cuando nos obliga a fijarnos en lo que no esperábamos ver.

Edward Tufte ha explicado que los usuarios de las visualizaciones de información ejecutan determinadas tareas analíticas, como hacer comparaciones. El principio de diseño del gráfico informativo debe apoyar la tarea analítica. Por ejemplo, los gráficos de puntos y de barras superan a los gráficos circulares

En su libro de 1983 The Visual Display of Quantitative Information, Edward Tufte define las "visualizaciones gráficas" y los principios para una visualización gráfica eficaz en el siguiente pasaje "La excelencia en los gráficos estadísticos consiste en comunicar ideas complejas con claridad, precisión y eficacia.

Las visualizaciones gráficas deben:

- Mostrar los datos
- Inducir al espectador a pensar en la sustancia y no en la metodología, el diseño gráfico, la tecnología de producción gráfica o cualquier otra cosa
- Evitar distorsionar lo que dicen los datos
- Presentar muchas cifras en poco espacio
- Hacer que los grandes conjuntos de datos sean coherentes
- Animar al ojo a comparar diferentes datos
- Revelar los datos en varios niveles de detalle, desde una visión general hasta la estructura fina
- Servir a un propósito razonablemente claro: descripción, exploración, tabulación o decoración
- Estar estrechamente integrados con las descripciones estadísticas y verbales de un conjunto de datos.

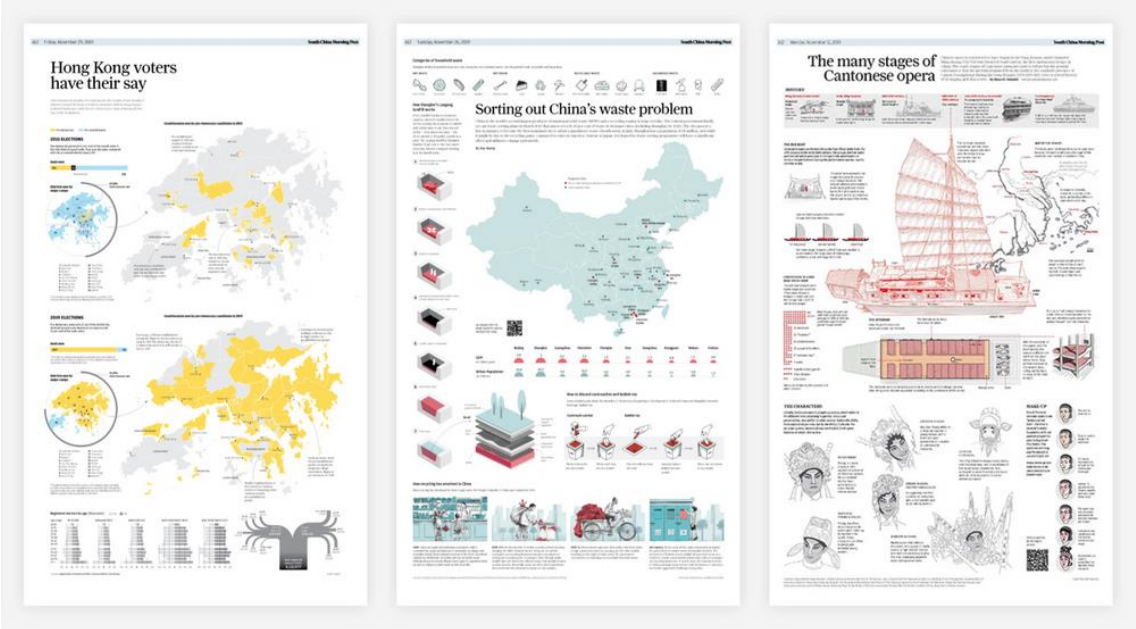
Los gráficos revelan los datos. De hecho, los gráficos pueden ser más precisos y reveladores que los cálculos estadísticos convencionales

La visualización de datos se utiliza habitualmente para estimular la generación de ideas en los equipos. A menudo se aprovechan durante las sesiones de brainstorming o Design Thinking al comienzo de un proyecto, ya que ayudan a recopilar diferentes perspectivas y a destacar las preocupaciones comunes del colectivo.

Aunque estas visualizaciones suelen estar sin pulir y sin refinar, ayudan a sentar las bases dentro del proyecto para garantizar que el equipo esté alineado con el problema que buscan resolver para las partes interesadas clave.

La visualización de datos para la ilustración de ideas ayuda a transmitir una idea, como una táctica o un proceso.

Se suele utilizar en entornos de aprendizaje, como tutoriales, cursos de certificación o centros de excelencia, pero también puede emplearse para representar estructuras o procesos organizativos, facilitando la comunicación entre las personas adecuadas para tareas específicas. Los gestores de proyectos suelen utilizar diagramas de Gantt y diagramas de cascada para ilustrar los flujos de trabajo.



El descubrimiento visual y la visualización diaria de datos están más alineados con los equipos de datos.

Mientras que el descubrimiento visual ayuda a los analistas de datos, a los científicos de datos y a otros profesionales de los datos a identificar

patrones y tendencias dentro de un conjunto de datos, la visualización de datos diaria apoya la narración posterior después de que se haya encontrado una nueva perspectiva.

La visualización de datos es un paso fundamental en el proceso de la ciencia de datos, ya que ayuda a los equipos y a los individuos a transmitir los datos de forma más eficaz a sus colegas y a los responsables de la toma de decisiones.

Sin embargo, es importante recordar que es un conjunto de habilidades que puede y debe extenderse más allá de su equipo de análisis principal.

Enlaces Recomendados:

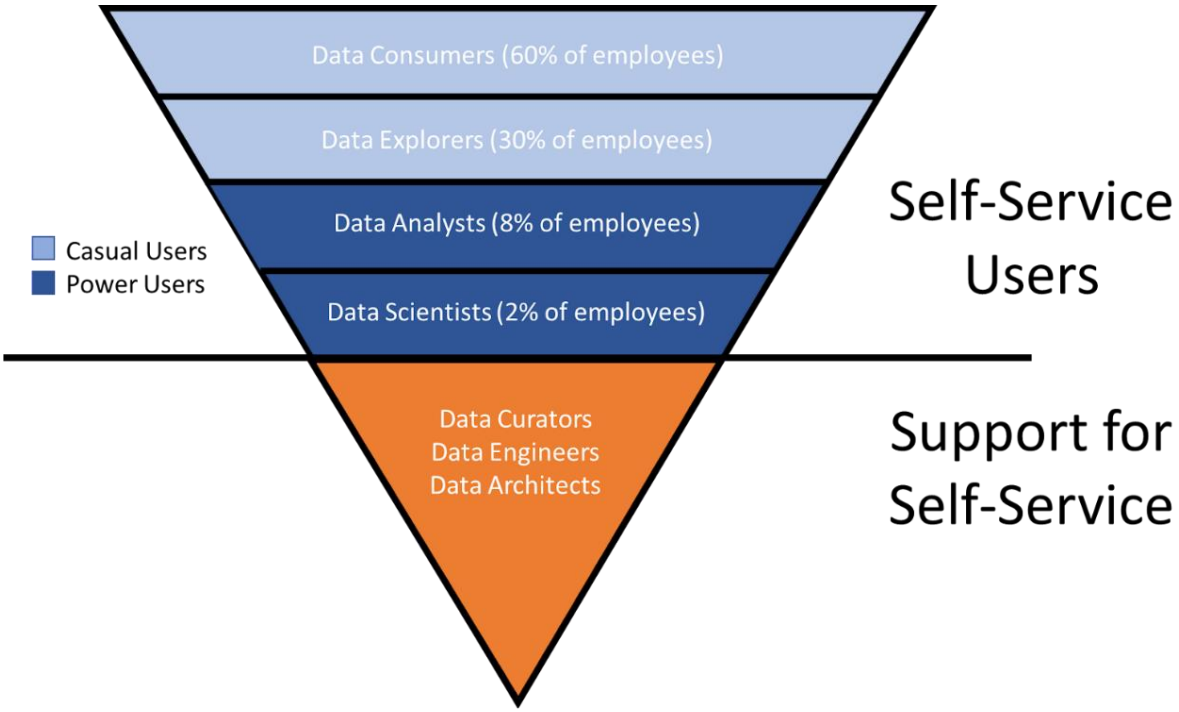
- Guía para elegir los mejores colores en tus visualizaciones
- (Libro y VideoTutorial) Storytelling with Data: A Data Visualization Guide for Business Professionals
- Libros sobre Data Visualization
- Los mejores video tutoriales sobre Data Visualization

T. Self Service Analytics:

Self-Service Analytics es una forma de inteligencia empresarial (BI) en la que se permite y anima a los profesionales y usuarios de negocio a realizar consultas y generar informes por sí mismos, con un apoyo puntual de TI

Self-service analytics es la habilitación de los usuarios de una organización para que accedan a los datos y generen ideas sin la intervención de un experto técnico profundo o sin depender de TI.

La diferencia clave entre la inteligencia empresarial tradicional y Self-service analytics es que, en lugar de que el departamento de TI proporcione BI a los usuarios finales, el papel del departamento de TI es facilitar el BI a esos mismos usuarios.



En un entorno de BI tradicional, un usuario final puede presentar una solicitud a TI para que consulte los datos y genere informes u hojas de cálculo de Excel con los datos solicitados.

Este método permite al departamento de TI mantener un mayor control sobre la calidad de los datos, pero limita la creatividad y la información que la empresa puede generar rápidamente.

Por otro lado, el Self-service analytics elimina al departamento de TI como intermediario y permite a los usuarios finales realizar sus propios análisis sin renunciar al control de la ingesta y la gestión de los datos. Eliminar el cuello de botella de TI de las actividades de análisis de datos es un primer paso clave para permitir el autoservicio, pero no significa adoptar un enfoque de no intervención.

Por el contrario, el papel de TI debe ser educar continuamente a los usuarios de la empresa para que entiendan e interactúen con los datos utilizando herramientas o procedimientos ya establecidos, ayudar a mantener una única fuente de verdad y permitir la colaboración entre ambas partes.

Claves para que funcione el self-service analytics:

Adoptar una cultura de autoservicio es un maratón, no un sprint, y la realización del valor sigue aumentando con la madurez y la capacidad de los usuarios. Las organizaciones que logran habilitar el autoservicio lo hacen a través de algunos principios clave:

- IT da prioridad a los datos limpios y precisos para el análisis
- Los usuarios de la empresa están capacitados y formados para llevar a cabo el análisis de los datos.
- Se permite y fomenta la colaboración entre los usuarios y el departamento de TI.
- Las empresas adoptan un modelo formal de análisis federado

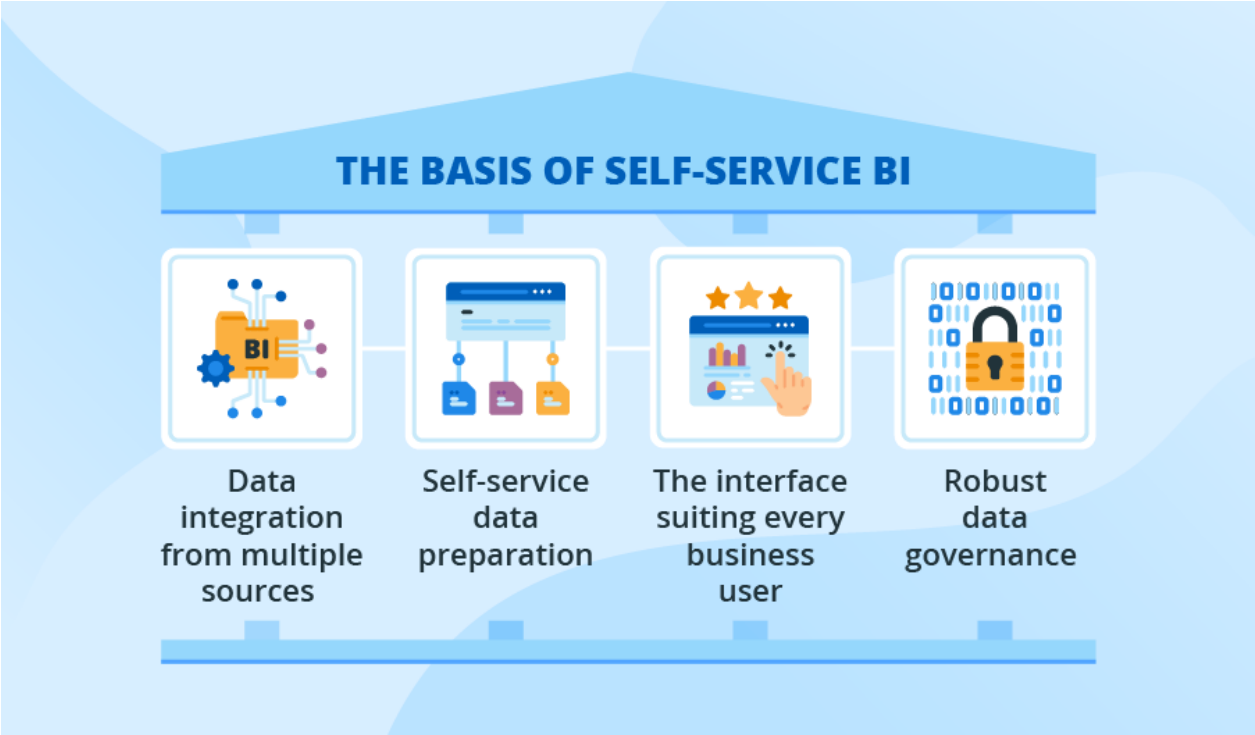
En la analítica de autoservicio es responsabilidad del departamento de TI garantizar que los ingredientes ya estén listos antes de que los usuarios de la empresa se dispongan a consumir los datos.

Esto se consigue a menudo mediante la creación de una capa semántica o un catálogo de datos, que simplemente mapea datos complejos de múltiples fuentes en cubos o categorías familiares que pueden ser fácilmente identificados por un usuario de negocios que nunca ha visto los datos de origen en bruto.

Un catálogo de datos permite al departamento de TI desmitificar la complejidad de los datos y, al mismo tiempo, mantener la gobernanza y el control necesarios sobre lo que se informa.

Un catálogo de datos exitoso debe ofrecer la posibilidad de buscar y filtrar múltiples conjuntos de datos para encontrar rápidamente lo que se busca.

Además de crear un catálogo de datos, los usuarios de la empresa deben saber que existe y tener acceso a formación sobre cómo consumir, analizar e informar sobre los datos. Aunque esta clave parece más fácil que la primera, es importante fomentar una cultura de conocimiento de los datos para que el despliegue del autoservicio tenga éxito.

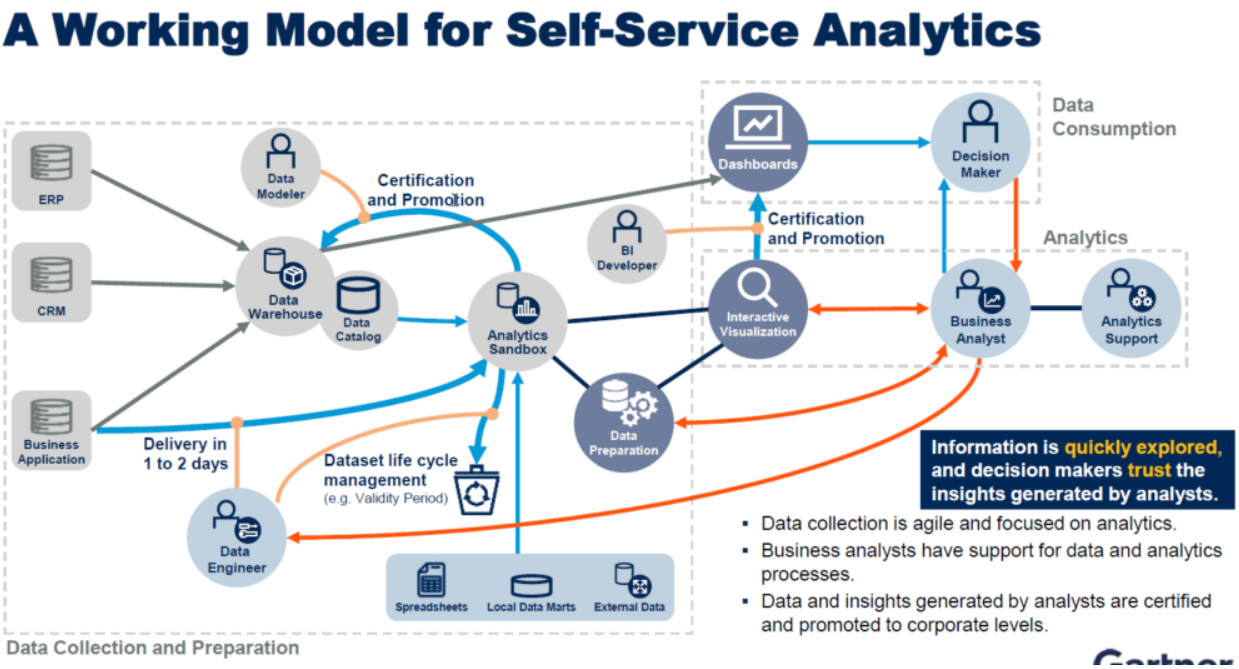


Por último, la creación y gestión de un canal de comunicación abierto entre los usuarios de TI y los de la empresa es crucial para el éxito del modelo de análisis de autoservicio.

Esto puede ser tan sencillo como adoptar el enfoque del Genius Bar de Apple para responder a las preguntas relacionadas con los datos, organizar llamadas semanales o mensuales de la "comunidad" con sus usuarios de análisis, crear un sencillo sistema de tickets para responder a las solicitudes de disponibilidad de datos o una combinación de los tres.

Otra ventaja de la comunicación simplificada entre el departamento de TI y la empresa es que proporciona información clave sobre los objetivos prácticos de la empresa. Si TI es responsable de habilitar el BI en todas las unidades de negocio, una mayor comunicación y transparencia entre las partes ayudará a informar a TI de en qué debe centrarse.

Independientemente del enfoque que elija, el cumplimiento de estas claves de éxito le ayudará a fomentar una cultura de análisis de autoservicio.



Beneficios para las empresas que aplican self-service analytics:

1) ***Menos dependencia de TI***: Cuando su equipo de TI pasa horas atendiendo solicitudes de informes, tiene menos tiempo para trabajar en proyectos de alto valor y a largo plazo, como el gobierno de datos y el desarrollo de aplicaciones. Permitir que los usuarios generen su propia información se traduce en menos solicitudes de TI y más innovación técnica.

2) Información sin esperas: superar a la competencia requiere procesos ágiles y una toma de decisiones informada. Con la analítica de datos de autoservicio, los usuarios de la empresa pueden dedicar más tiempo a explorar los datos, sopesar las perspectivas y tomar medidas decisivas, y pasar menos tiempo yendo y viniendo con el departamento de TI sobre los requisitos y debatiendo sobre la ciencia de datos frente a la analítica de datos.

3) Descubrimiento ilimitado de datos: Cada usuario de la empresa tiene necesidades únicas, por lo que debe ser capaz de cortar sus datos de la manera que considere oportuna. Pero si se limitan a un análisis lineal basado en consultas, es probable que sus conocimientos se queden cortos. Con el motor asociativo de Qlik, los usuarios pueden explorar los datos libremente y descubrir perspectivas que nunca esperaron.

4) Mayor conocimiento de los datos: Ayudar a los empleados a aumentar sus habilidades analíticas significa darles la formación que necesitan para leer, manipular, analizar y argumentar con los datos. Esa es la definición de alfabetización en datos, y la base de una cultura basada en datos. Con una plataforma de análisis de autoservicio, puede aumentar la alfabetización en datos en toda su empresa.

Enlaces Recomendados:

- 5 claves de self-service analytics
- Glosario de TDWI sobre self-service analytics
- Comparando 'guided' analytics con 'self-service analytics'
- Videotutoriales sobre 'self-service analytics'

U. Data Silos:

Un silo de datos es un repositorio de datos controlado por un departamento o unidad de negocio y aislado del resto de la organización, al igual que la hierba y el grano en un silo agrícola están cerrados a los elementos externos. Los datos en silo suelen almacenarse en un sistema independiente y a menudo son incompatibles con otros conjuntos de datos. Esto dificulta el acceso y la utilización de los datos por parte de los usuarios de otras partes de la organización.

A) Por qué los Data Silos son un problema?

Los silos de datos obstaculizan las operaciones empresariales y las iniciativas de análisis de datos que las respaldan.

Los silos limitan la capacidad de los ejecutivos de utilizar los datos para gestionar los procesos empresariales y tomar decisiones de negocio informadas. También impiden que los agentes de los centros de llamadas, los representantes de ventas y otros trabajadores operativos accedan a los datos pertinentes sobre los clientes, los productos, las cadenas de suministro, etc.

Estos son los principales ejemplos de los problemas que causan los Data Silos a las organizaciones:

1. Conjuntos de datos incompletos. Los silos de datos bloquean los datos de los usuarios que no pueden acceder a ellos. Como resultado, las estrategias y decisiones empresariales no se basan en todos los datos disponibles, lo que puede conducir a una toma de decisiones errónea.

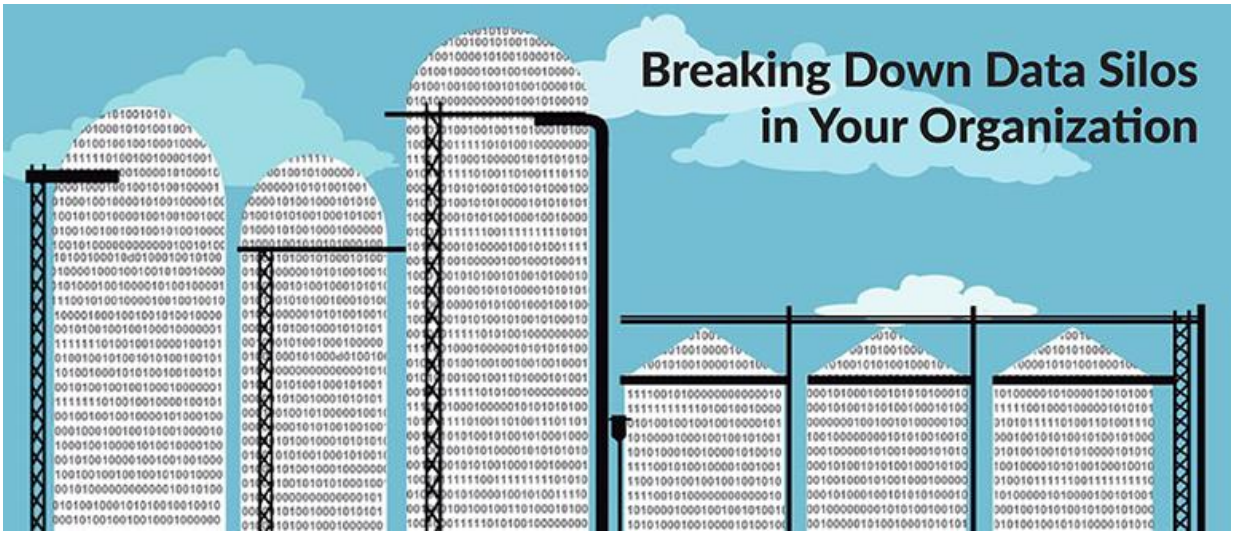
Los silos también pueden hacer fracasar los esfuerzos por crear almacenes y lagos de datos que integren diferentes conjuntos de datos para aplicaciones de inteligencia empresarial (BI) y análisis.

2. Datos incoherentes. Muchos silos de datos no son coherentes con otros conjuntos de datos. Por ejemplo, un equipo de marketing puede formatear los datos de los clientes de forma diferente a los de otros departamentos. Los errores de datos de un equipo de ventas pueden no ser identificados y corregidos. Las actualizaciones de datos en otros sistemas no se realizan en un silo de servicio al cliente.

Estas incoherencias crean problemas de calidad, precisión e integridad de los datos que afectan a los usuarios finales tanto en las aplicaciones operativas como en las analíticas.

3. Duplicación de plataformas y procesos de datos. Los silos de datos aumentan los costes de TI al incrementar el número de servidores y dispositivos de almacenamiento que una organización necesita comprar.

En muchos casos, esos sistemas también son desplegados y gestionados por separado por los departamentos en lugar del equipo de gestión de datos de una organización. Esto aumenta aún más el gasto y el uso ineficiente de los recursos de TI.



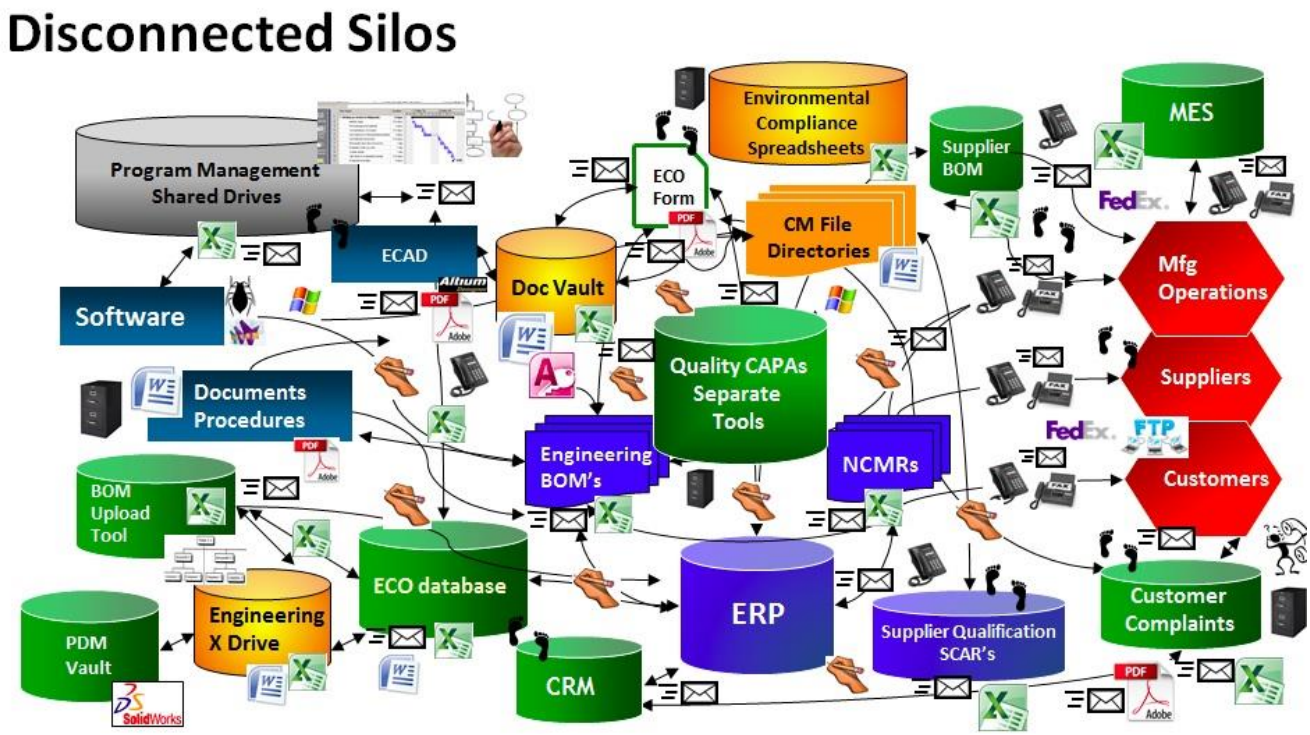
4. Menos colaboración entre los usuarios finales. Los conjuntos de datos aislados en silos reducen las oportunidades de intercambio de datos y colaboración entre usuarios de diferentes departamentos. Es más difícil colaborar eficazmente cuando las personas no tienen visibilidad de los datos aislados.

5. Una mentalidad de silo en los departamentos. Los silos de datos contribuyen a la formación de silos organizativos: departamentos y unidades de negocio que guardan sus datos de cerca y son reacios a compartirlos con otros.

También pueden resistirse a los programas de gobernanza de datos que pretenden acabar con los silos de datos y garantizar que los datos sean coherentes y correctos en todos los sistemas de una organización.

6. Problemas de seguridad de los datos y de cumplimiento de la normativa. Algunos silos de datos son almacenados por usuarios individuales en hojas de cálculo de Excel o en herramientas empresariales en línea como Google Drive, a menudo en dispositivos móviles.

Esto aumenta los riesgos de seguridad y privacidad de los datos para las organizaciones si no tienen los controles adecuados. Los silos también complican los esfuerzos para cumplir con las leyes de privacidad y protección de datos.



B) Razones por las que aparecen Data Silos en las organizaciones:

1. Estrategia de TI y despliegue de tecnología. Algunas organizaciones han descentralizado las decisiones de compra de TI y permiten que los departamentos y las unidades de negocio compren tecnologías por su cuenta.

Esto suele llevar a la implantación de bases de datos y aplicaciones empresariales que no son compatibles o no están conectadas con otros sistemas. Lo mismo puede ocurrir cuando los equipos de TI corporativos participan en las decisiones de compra si un departamento necesita una tecnología concreta.

La variedad de plataformas de datos disponibles en la actualidad también contribuye a eliminar los silos de datos: además de las principales bases de datos relacionales, las organizaciones pueden desplegar plataformas de big data, bases de datos NoSQL, servicios de almacenamiento de objetos en la nube y bases de datos de propósito especial para satisfacer diferentes necesidades empresariales.

2. Estructura organizativa y de gestión. Los silos de datos suelen producirse cuando las unidades de negocio están totalmente descentralizadas y se gestionan como entidades separadas.

Esto es más común en grandes organizaciones con diferentes filiales y empresas operativas, pero puede ocurrir en otras más pequeñas con una estructura y un enfoque de gestión similares.

3. Cultura y principios corporativos. Incluso cuando las operaciones de TI y de negocio se gestionan de forma más unificada, la cultura de la empresa puede estimular la creación de silos de datos.

Hay menos incentivos para evitarlos si el intercambio de datos no es una norma cultural y una organización no tiene objetivos y principios comunes para la gestión de datos.

Los departamentos también pueden considerar sus datos como un activo que les pertenece y controlan, lo que fomenta aún más el desarrollo de silos de datos.

4. Crecimiento empresarial y adquisiciones. Las organizaciones en crecimiento son propensas a los silos de datos. A medida que una empresa se expande, es posible que haya que atender rápidamente nuevas necesidades de negocio y que se creen unidades de negocio adicionales.

Ambas situaciones son incubadoras naturales de silos de datos. Las fusiones y adquisiciones también introducen silos en una organización, algunos conocidos y otros que pueden estar ocultos.

C) Cómo conseguimos acabar con los Data Silos

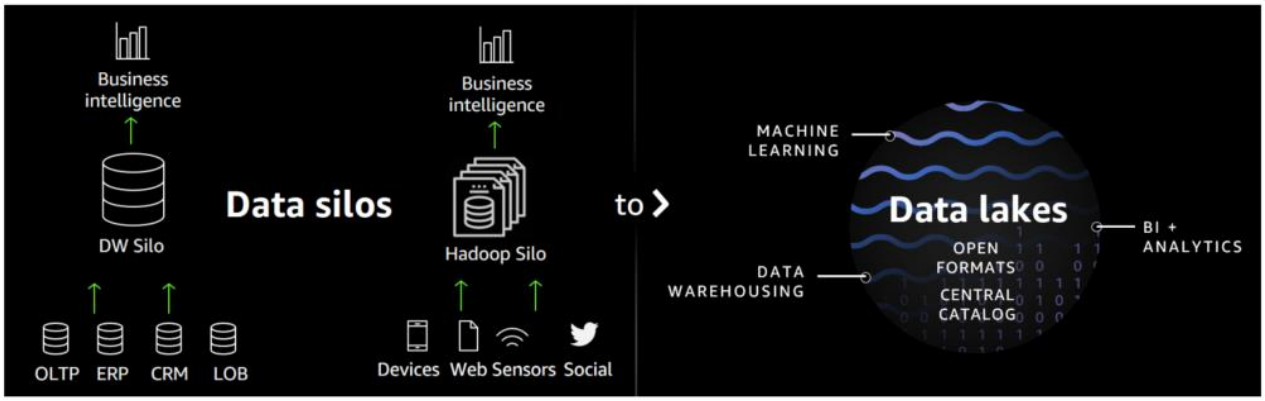
1. Integración de datos. Integrar los silos de datos con otros sistemas es la forma más directa de romperlos. La forma más popular de integración de datos es la extracción, transformación y carga (ETL), que extrae los datos de los sistemas de origen, los consolida y los carga en un sistema o aplicación de destino.

Otras técnicas de integración de datos que pueden utilizarse contra los silos son la integración en tiempo real, la virtualización de datos y la extracción, carga y transformación, una variación de ETL.

2. Data Warehouses y Data Lakes. El sistema de destino más común en los trabajos de integración de datos es un almacén de datos, que almacena datos de transacciones estructurados para aplicaciones de BI, análisis e informes.

Cada vez más, las organizaciones también construyen lagos de datos para albergar conjuntos de big data, que pueden incluir grandes volúmenes de datos estructurados, no estructurados y semiestructurados utilizados en aplicaciones de ciencia de datos. Estos dos tipos de plataformas

proporcionan repositorios centralizados para los datos de diferentes sistemas, lo que los convierte en una forma natural de abordar los silos.



3. Data Governance. En última instancia, lo mejor es no sólo eliminar los silos de datos existentes, sino también evitar que se creen otros nuevos. Una estrategia de gestión de datos más completa ayuda a conseguir ambos objetivos.

Por ejemplo, el diseño de la arquitectura de datos documenta los activos de datos, traza los flujos de datos y crea un plan para la implantación de plataformas de datos.

Una estrategia de datos empresarial alinea mejor el proceso de gestión de datos con las operaciones del negocio. Y un sólido programa de gobierno de datos puede reducir directamente el número de silos de datos en una organización y promover normas y políticas de datos comunes.

4. Cambio de cultura. Para acabar realmente con los silos de datos, puede ser necesario cambiar la cultura de una organización. Los esfuerzos para hacerlo pueden formar parte del proceso de desarrollo de la estrategia de datos o de una iniciativa de gobierno de datos.

En algunos casos, puede ser necesario un programa de gestión del cambio para aplicar los cambios culturales y garantizar que los departamentos y las unidades de negocio los adopten.

Enlaces Recomendados:

- What are Data Silos?
- 5 riesgos de los silos de datos en las empresas
- Qué es un Data Silo y por qué es malo para tu organización
- Videotutoriales sobre Data Silos

V. Data Health:

Data health es el estado de los datos de una empresa y la forma en que apoyan las decisiones eficaces y oportunas y los objetivos empresariales. Para saber que los datos de su organización son saludables, debe poder demostrar que son válidos, completos y de suficiente calidad para producir análisis en los que los responsables de la toma de decisiones se sientan cómodos para tomar decisiones empresariales.

Como cualquier sistema de atención sanitaria, la salud de los datos implica el seguimiento y la intervención a lo largo de todo el ciclo vital. Pensamos en la salud de los datos en un marco de prevención, tratamiento y apoyo a la comunidad:

- Atención preventiva: identificar los problemas de datos de forma preventiva.
- Tratamientos eficaces: curar sistemáticamente los problemas y riesgos de fiabilidad de los datos
- Cultura de apoyo: establecer una disciplina de cuidado de los datos en colaboración

Con las métricas de salud de datos para demostrar el valor empresarial de

los datos, una organización puede mejorar casi cualquier aspecto de sus operaciones:

- Mejorar los análisis de ventas y marketing
- Abordar la gobernanza y el cumplimiento de los datos
- Mejorar los procesos empresariales
- Transformar la experiencia del cliente
- Impulsar el compromiso de 360 grados
- Habilitar el aprendizaje automático y la IA

Sin datos saludables, todos esos procesos van mal. No es posible dirigirse a los clientes adecuados, acortar los ciclos de ventas o mejorar los procesos si los datos disponibles en los que se basa el trabajo son inexactos, no están controlados o no están actualizados.

Los datos poco fiables cuestan a las empresas tiempo y calidad en su toma de decisiones, lo que añade costes y puede afectar negativamente a los ingresos. A medida que se amplía el uso de big data, la salud de los datos es cada vez más importante.

Es fundamental que las empresas que trabajan con big data instituyan métricas de salud

Cómo tener una buena salud en los Datos:

Factores de riesgo en la auditoría de salud: lo primero que hay que hacer es identificar qué riesgos internos (aplicaciones, herramientas, procesos, metodologías y personal de su empresa) y qué riesgos externos (socios, proveedores e incluso clientes) pueden intervenir en la calidad de los datos.

Podríamos calificar como factor de riesgo el hecho de que diferentes departamentos intervengan en el proceso de información sin que se registren los cambios.

Establecer medidas de prevención: una vez que sepamos dónde están los puntos débiles de nuestra empresa en cuanto a la gestión de los datos, debemos crear protocolos para evitar cualquier riesgo, por ejemplo, un correcto etiquetado (que se sepa de dónde viene cada dato y por qué está en las bases de datos) y unas normas específicas para cada departamento que intervenga en su tratamiento.

Automatización de la gestión: en la era del big data, una media del 60% del tiempo productivo de un gestor de datos se dedica a tareas de limpieza y depuración de bases de datos. El aprendizaje automático puede entrenar a sus sistemas para que reconozcan los datos incorrectos y las fuentes

sospechosas antes de que puedan contaminarlos, aumentando la productividad del personal.

Evaluación constante: Cuanto antes se detecte un problema de salud de los datos, mayor será la posibilidad de intervenir eficazmente. Por lo tanto, es una buena práctica crear perfiles de datos continuos, además de las evaluaciones de todos los datos entrantes y las comprobaciones periódicas de los lotes.

Previsión continua: Una evaluación debe ir seguida de una receta que determine cómo actuar tanto de forma preventiva como reactiva. Debemos conocer en todo momento no sólo el estado de nuestras fuentes, sino también cómo y desde dónde se está interviniendo.

Equilibrio permanente: a estas alturas deberíamos ser conscientes de la cantidad de factores que intervienen en la gestión de datos. Tenerlos todos bajo control implica tiempo y modulación de riesgos como no cumplir con la seguridad y su normativa (por ejemplo, permitir el libre acceso a todas las bases de datos). Una buena gestión de datos debe sopesarlos y equilibrarlos (por ejemplo, con restricciones de acceso por perfil).

Espero que os haya sido de utilidad y recordad!!

Si te ha parecido interesante, compártelo!!