Comparativa Datawarehouse vs DataLakes

Robles Flores, Anthony Richard (2016056192), Estrella Palacios, Katherine Lizbeth (2015050948), Sosa Bedoya, Sharon Fiorela (2016054460), and Torres Beltran , Johanna Andrea (2020067849)
Universidad Privada de Tacna \Facultad de Ingenieria \Escuela Profesional de Ingenieria de Sistemas

Resumen

Realizar la comparación entre Data Lake y Data Warehouse es algo muy común entre aquellas empresas que se disponen a implantar soluciones de big data. Una de las diferencias que marcan la comparación entre Data Lake vs Data Warehouse es que, este último, por lo general, sigue una metodología como las definidas por Inmon y Kimball, a diferencia de Data Lake.

Abstract

Making the comparison between Data Lake and Data Warehouse is very common among those companies that are preparing to implement big data solutions. One of the differences that make the comparison between Data Lake vs Data Warehouse is that the latter, generally, follows a methodology such as those defined by Inmon and Kimball, unlike Data Lake.

I. INTRODUCCIÓN

Por mucho tiempo, los datawarehouses funcionaron como depósitos fundamentales de datos integrados que facilitaron las decisiones informadas. En el pasado reciente, el campo de la gestión de datos ha sido testigo de un cambio catastrófico con la llegada de Big Data que ha creado la necesidad de procesar un gran volumen y variedad de datos junto con el deseo de un acceso más rápido a los datos. Datalakes se emplea ampliamente para cumplir con estos requisitos y se promociona como una ventana única para cumplir con todos los requisitos de datos de una organización. Esta suposición ha cuestionado el papel del Datawarehouse tradicional. Este documento describe las características esenciales de los datawarehouses y datalakes. También compara y contrasta los dos sistemas y concluye que los dos no son tecnologías competidoras sino complementarias. Juntos pueden atender de manera efectiva y eficiente las necesidades de gestión de datos de una organización.

II. OBJETIVOS

A. General:

Brindar una perspectiva de comparación entre Datawarehouse y DataLakes, basada en la la teoría recopilada sobre ambos

B. Específicos:

- Describir el marco teórico del Datawarehouse
- Describir el marco teórico del DataLakes

III. MARCO TEÓRICO

A. DATAWAREHOUSE

Un Data Warehouse es una base de datos corporativa en la que se integra información depurada de las diversas fuentes que hay en la organización. Dicha información debe ser homogénea y fiable, se almacena de forma que permita su análisis desde muy diversas perspectivas, y que a su vez dé unos tiempos de respuesta óptimos. Para ello la información se encuentra altamente desnormalizada y modelada de una forma bastante diferente a los sistemas transaccionales, principalmente se utilizan los modelos en estrella (star schema) y en copo de nieve (snowflake schema).[**?**]

1. Características DataWarehouse

Un Data Warehouse es mucho más que lo que hemos comentado hasta el momento. Según Bill Inmon se caracteriza por ser:

- **Orientado a temas:** los datos están organizados por temas para facilitar el entendimiento por parte de los usuarios, de forma que todos los datos relativos a un mismo elemento de la vida real queden unidos entre sí. Por ejemplo, todos los datos de un cliente pueden estar consolidados en una misma tabla, todos los datos de los productos en otra, y así sucesivamente.
- **Integrado:** los datos se deben integrar en una estructura consistente, debiendo eliminarse las inconsistencias existentes entre los diversos sistemas operacionales. La información se estructura en diversos niveles de detalle para adecuarse a las necesidades de consulta de los usuarios.

Algunas de las inconsistencias más comunes que nos solemos encontrar son: en nomenclatura, en unidades de medida, en formatos de fechas, múltiples tablas con información similar.

- **Histórico (variante en el tiempo):** los datos, que pueden ir variando a lo largo del tiempo, deben quedar reflejados de forma que al ser consultados reflejen estos cambios y no se altere la realidad que había en el momento en que se almacenaron, evitando así la problemática que ocurre en los sistemas operacionales, que reflejan solamente el estado de la actividad de negocio presente. Un Data Warehouse debe almacenar los diferentes valores que toma una variable a lo largo del tiempo.

- **No volátil:** la información de un Data Warehouse, una vez introducida, debe ser de sólo lectura, nunca se modifica ni se elimina, y ha de ser permanente y mantenerse para futuras consultas.

Adicionalmente estos almacenes contienen metadatos (datos sobre los datos), que aportan un valor adicional, permitiendo tener información sobre su procedencia (sobre todo cuando tenemos múltiples fuentes), la periodicidad con la que han sido introducidos, la fiabilidad que nos ofrecen, etc. Todo ello nos aporta una ayuda adicional, tanto al usuario final como a los técnicos responsables de su mantenimiento, ayudando a estos últimos en aspectos como su auditoría y administración.

Kimball determinó que para él un Data Warehouse no era más que un conjunto de los Data Marts de una organización. Un Data Mart es una copia de las transacciones específicamente estructurada para la consulta y el análisis. Defiende por tanto una metodología Bottom-up a la hora de diseñar un almacén de datos.

La diferencia de un Data Mart con respecto a un Data Warehouse es solamente en cuanto al alcance. Mientras que un Data Warehouse es un sistema centralizado con datos globales de la empresa y de todos sus procesos operacionales, un Data Mart es un subconjunto temático de datos, orientado a un proceso o un área de negocio específica. Debe tener una estructura óptima desde todas las perspectivas que afecten a los procesos de dicha área. Es más, según Ralph Kimball, cada Data Mart debe estar orientado a un proceso determinado dentro de la organización, por ejemplo, a pedidos de clientes, a compras, a inventario de almacén, a envío de materiales, etc.

La existencia de un Data Warehouse no descarta la existencia de Data Marts, ni viceversa. [1]

B. Data Warehouse en la nube

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

- La primera área de transformación es el impulso para aumentar la agilidad general.

- La segunda área de transformación gira en torno a la necesidad de mejorar el control de costes.

La gran mayoría de los departamentos de TI están experimentando un rápido aumento de la demanda de datos. Los directivos quieren tener acceso a más y más datos históricos, mientras que al mismo tiempo, los científicos de datos y los analistas de negocios están explorando formas de introducir nuevos flujos de datos en el almacén para enriquecer el análisis existente, así como impulsar nuevas áreas de análisis. Esta rápida expansión de los volúmenes y fuentes de datos significa que los equipos de TI necesitan invertir más tiempo y esfuerzo asegurando que el rendimiento de las consultas permanezca constante y necesitan proporcionar cada vez más entornos para equipos individuales para validar el valor comercial de los nuevos conjuntos de datos.

Existe una creciente necesidad de hacer más con cada vez menos recursos, al mismo tiempo que se garantiza que todos los datos sensibles y estrategicos estén completamente asegurados, a lo largo de todo el ciclo de vida, de la manera más rentable.[2]

C. Migracion en la nube

La importancia de la nube para el data warehouse está directamente vinculado a tres factores clave:

- Mayor agilidad

  Es más rápido, más fácil y más eficiente iniciar estos nuevos proyectos centrados en los datos utilizando los servicios en la nube.

- Mejor control de costes

  Los servicios en la nube deben facilitar todo esto, siendo más rentables ya que todas las características de seguridad se pueden habilitar de forma predeterminada y mejorada, y actualizada de forma transparente.

- Co-localización

  Ubicar conjuntamente el data warehouse junto con los sistemas fuente que ya se estén ejecutándose en la nube

entre las ventajas de mover el data warehouse a la nube

- Más fácil consolidación y racionalización
- Monetización más rápida de los datos en la nube
- La nube ofrece mejor protección

D. DATALAKES

Un DataLakes es un repositorio centralizado que le permite almacenar todos sus datos estructurados y no estructurados a cualquier escala.
Puede almacenar sus datos tal cual, sin tener que estructurar primero los datos, y ejecutar diferentes tipos de análisis, desde paneles y visualizaciones hasta procesamiento de grandes datos, análisis en tiempo real y aprendizaje automático para guiar mejores decisiones.

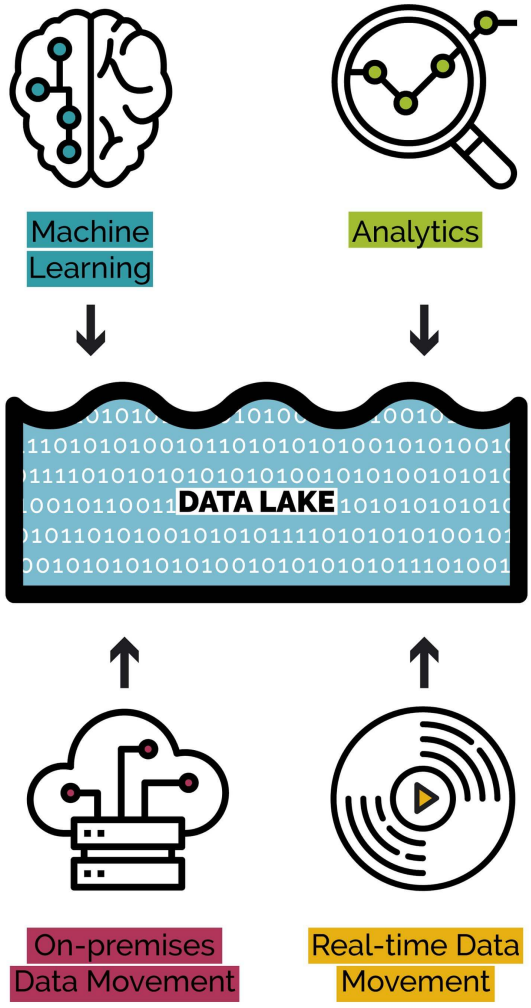
Los expertos de la industria estiman que los datos de casi todas las industrias básicamente se duplican cada dos años. Capturar, almacenar y utilizar todos estos datos es un verdadero desafío.

Beneficios de DataLakes

- **Centralización de fuentes de contenido dispares:** Las fuentes de contenido dispares a menudo contienen información confidencial que requerirá la implementación de las medidas de seguridad apropiadas.
- **Medidas de seguridad:** se otorga acceso a cierta información a los usuarios del data lake que no tienen acceso a la fuente de contenido original.
- Una vez que el contenido está en el data lake, puede normalizarse y enriquecerse.
- Los usuarios, de diferentes departamentos, potencialmente dispersos por todo el mundo, pueden tener acceso flexible a un data lake y a su contenido desde cualquier lugar.
- Un data lake pone la información de toda la empresa en manos de muchos más empleados para hacer a la organización un todo más inteligente, más ágil y más innovadora.

E. COMPONENTES DE DATALAKES

- **Ingestión y almacenamiento de datos:**Es la capacidad de adquirir datos en tiempo real o en lotes, y también es la capacidad de almacenar datos y hacerlos accesibles.
- **Procesamiento de datos:** Es la capacidad de trabajar datos sin procesar para que puedan analizarse a través de procesos estándar. También tiene la



función de soluciones de ingeniería, que pueden extraer valor de los datos utilizando procesos regulares y automáticos generados por las operaciones de análisis.

- **Análisis de datos:** Es la creación de módulos para extraer información de los datos de manera sistemática; esto puede suceder en tiempo real o mediante un proceso que se ejecuta periódicamente.

- **Integración de datos:** es la capacidad de conectar aplicaciones a la plataforma; En primer lugar, las aplicaciones deben permitir consultar el Data Lake para extraer los datos en el formato correcto, en función del uso que desee hacer de ellos.

F. ARQUITECTURA DE UN DATA LAKE

- **Data lakes en Hadoop:** Hadoop es la opción que se asocia mayormente con los data lakes, por defecto.

  • Mayor familiaridad entre los tecnólogos
  • Más económicos, porque son de código abierto
  • Muchas herramientas ETL disponibles para la integración con Hadoop
  • Fáciles de escalar
  • La localidad de los datos permite una computación más rápida

- **Data lakes en AWS:** AWS tiene un paquete exhaustivo de ofertas de producto para su solución de data lake.

  • Paquete de productos exhaustivo y repleto de prestaciones
  • Flexibilidad a la hora de elegir productos a partir de requisitos particulares
  • Precios bajos
  • Nivel de seguridad y conformidad alto
  • Separación de computación y almacenamiento para poder ajustar su escala según convenga
  • Colaboración con empresas de APN (AWS Partner Network), como Talend, que garantiza una etapa inicial de funcionamiento con AWS exenta de problemas

- **Data lakes en Azure:** Azure es un data lake de Microsoft. Tiene una capa de almacenamiento y de analítica.Tanto el almacenamiento como la computación en cloud facilitan su manejo.

  • Servicios analíticos robustos con funcionalidades potentes
  • Fácil de migrar desde un clúster Hadoop existente
  • Muchos expertos en big data conocen Hadoop y sus herramientas, de modo que resulta fácil encontrar trabajadores cualificados.
  • La integración con Active Directory garantiza que no son necesarios esfuerzos adicionales para gestionar la seguridad

G. COMPARACION DE DATAWAREHOUSE Y DATALAKES

IV. CONCLUSIONES

- Es necesario revisar la categorías de un lago de datos y un almacén de datos, para analizar cual se adapta mejor al caso de uso en el que se desea trabajar.

	Data Warehouses	Data Lakes
Tipos de datos	Datos históricos que se han estructurado para ajustarse a un esquema de base de datos relacional.	Datos no estructurados y estructurados de varias fuentes de datos de la compañía.
Propósito	Análisis para decisiones comerciales.	Almacenamiento rentable de big data.
Los usuarios	Analistas de datos y analistas de negocios.	Científicos e ingenieros de datos.
Tareas	Por lo general, consultas de solo lectura para agregar y resumir datos.	Almacenamiento de datos y análisis de big data, como aprendizaje profundo y análisis en tiempo real.
Talla	Solo almacena datos relevantes para el análisis.	Almacena todos los datos que podrían usarse, pueden ocupar petabytes.
Almacenamiento	Un almacén de datos consistirá en datos que se extraen de sistemas transaccionales o datos que consisten en métricas cuantitativas con sus atributos. Los datos se limpian y transforman.	En el lago de datos, todos los datos se mantienen independientemente de la fuente y su estructura. Los datos se mantienen en su forma cruda. Solo se transforma cuando está listo para usarse.

- Cuando se construyen tuberías de datos, es necesario implementar una combinación de ambas soluciones de almacenamiento.

- El lago de datos integran diferentes tipos de datos para generar preguntas completamente nuevas, ya que es probable que estos usuarios no usen almacenes de datos porque pueden necesitar ir más allá de sus capacidades, mientras que en el almacén de datos la mayoría de los usuarios de una organización están operativos. Este tipo de usuarios solo se preocupa por los informes y las métricas clave de rendimiento.

[1] Imhoff C, y. G. (2003). Mastering data warehouse desing, relational and dimensional techiniques. *Wiley Publishing Inc.*, pages 28–29.

[2] Leonard Brizuela, E. I. and Castro Blanco, Y. (2013). Metodologías para desarrollar almacén de datos. *Revista de arquitectura e Ingeniería*, 7(3).