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

Los almacenes de datos (data warehouses en inglés) toman cada día mayor importancia, a medida que las organizaciones pasan de esquemas de sólo recolección de datos a esquemas de análisis de los mismos. En este breve artículo se tratará de brindar una explicación general de algunas metodologías, en este caso serán la metodología Kimball y la metodología Inmon.

Abstract

Data warehouses (data warehouses in English) are becoming increasingly important, as organizations move from data-only schemes to data analysis schemes. In this short article we will try to provide a general explanation of some methodologies, in this case they will be the Kimball methodology and the Inmon methodology.

I. INTRODUCCIÓN

Actualmente las organizaciones utilizan la información y el conocimiento para apoyar la toma de sus decisiones estratégicas, y de este modo lograr sus metas y mejorar sus procesos. Uno de los desafíos que enfrentan hoy las organizaciones, es el aumento de datos, lo que ha generado dos grandes problemas; el primero, identificar los datos relevantes para dar seguimiento a su estrategia organizacional, y lograr que se cumplan los planes con las metas establecidas. Y el segundo problema, la capacidad para administrar esta gran cantidad de datos. Un almacén de datos según Inmon, es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volatil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. [1]

II. OBJETIVOS

A. General:

Dar una visión clara de BI, desde las perspectivas de los autores que sentaron las bases que son Ralph Kimball y Bill Inmon, para la mejora de las estrategias del negocio al que se desee implementar las herramientas de BI.

B. Específicos:

Describir las metodologías propuestas por los principales autores de BI desde las perspectivas de sus creadores Ralph Kimball y Bill Inmon.

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).

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. Es decir, puede haber organizaciones que tengan sólo un Data Warehouse, que sólo tengan Data Marts, o que tengan un Data Warehouse y Data Marts.

B. DATAWAREHOUSE **********

- Descomposición funcional
- Diagrama de contexto
- Diagrama de flujo de datos
- Diagrama de transición de estados
- Pseudocódigo

c. Modelo de datos. Se trabaja con 2 tipos de modelos:

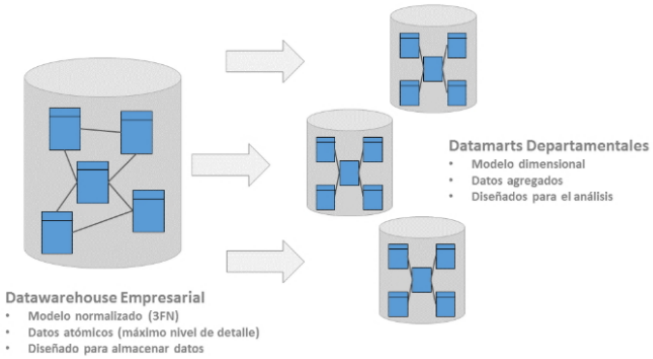
- El Modelo de datos nos muestra los datos primitivos, tomando en cuenta el elemento tiempo, se plasman los cálculos que se realicen y finalmente se muestran sus relaciones.

El Modelo de Datos del Data Warehouse. Los modelos anteriores nos deberán entregar la definición de los sujetos a los que estará orientado el Data Warehouse. Debe venir en 3 perspectivas y son explicadas en la siguiente tabla:

d. Una vez que se tiene conocimiento de este modelo se deben tomar ciertas decisiones sobre el diseño del Data Warehouse. Entre estas decisiones tenemos las siguientes:

- Normalización, debemos decidir el grado al que nuestro Data Warehouse
- Granularidad
- Particiones
- Minería de Datos

C. DATAWAREHOUSE



La arquitectura que plantea Bill Inmon consta de las siguientes partes: [2]

D. La estructura del DataWarehouse

En cuanto a la estructura interna del DataWarehouse, para Inmon la prioridad es que el modelo de datos esté construido en tercera /**********

E. DATALAKES

La metodología se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio. Este ciclo de vida del proyecto de DW, está basado en cuatro principios básicos:

- Centrarse en el negocio: Hay que concentrarse en la identificación de los requerimientos del negocio y su valor asociado, y usar estos esfuerzos para desarrollar relaciones sólidas con el negocio, agudizando el análisis del mismo y la competencia consultiva de los implementadores.
- Construir una infraestructura de información adecuada: Diseñaruna base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de requerimientos de negocio identificados en la empresa.
- Realizar entregas en incrementos significativos: crear el almacén de datos (DW) en incrementos entregables en plazos de 6 a 12 meses.
- Ofrecer la solución completa: proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios.

F. IMPLEMENTACIONES DE DATALAKES

/**********/

G. DATALAKES *********

Ralph Kimball, plantea la idea de un enfoque dimensional para el diseño de un DW, y afirma que la unión de todos los DM de una organización constituye el DW corporativo, a lo cual se le conoce como el enfoque bottom-up.

H./ 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).