Data warehousing- Bill Inmon father of Data warehousing

Why learn Data Warehousing and Business Intelligence? All the successful companies have been investing large sums of money in business intelligence and data warehousing tools and technologies. Up-to-date, accurate and integrated information about their supply chain, products and customers are critical for their success. With the advent of Mobile, Social and Cloud platform, today's business intelligence tools have evolved and can be categorized into five areas, including databases, extraction transformation and load (ETL) tools, data quality tools, reporting tools and statistical analysis tools. This course will provide a strong foundation around Data Warehousing and Business Intelligence fundamentals and sophisticated tools like Talend, Tableau and ERwin.

Need of BI

What is ware housing

Dwh architecture- OLTP vs OLAP

-ETL extract transformation & load

-data Mart

Metadata

DWH architecture

Demo creating a DWH

BI- Planning -> data gathering->data analysis->Business action

Data warehousing is a strategies or activities, a part of BI.

BI- act of transforming raw/operational data into useful info for business analysis

It works- BI based on data warehousing tech EXTRACTS info from the company operations system

The data TRANSFORMED(cleaned and integrated) and LOADED into Data warehouse

Then the credible data used for business insights.

Without data warehouse the visualization cannot be done. The data cannot be used from data source to business analysis directly. Data need to be integrated and processed from multiple data sources or database.

DWH is a central location where consolidated data are stored from multiple databases. DWH separately maintained from organizations operational database. Because the data should not be affected.

End users access it whenever the info is needed.

DWH is not loaded every time when new data added to the database.

Operational data-> ETL(new database)-> DWH->OLAP strategy (online analytical processing) and also visualization with Tableau or qlikview ->Business users.

The data can be stored from DWH to back into databases like oracle or sql then can do the OLAP and visualization

Advantages of DWH

* Strategic quest can be answered by studying trends(past and future prediction).
* Faster and more accurate, stability in DWH

DWH not a product to be purchased, it needs to be designed according to the need of the company

Database has different types of data like real-time and legacy data. But in DWH has only legacy data. data coming from multiple sources and the tables will not be related to eachother and multiple tables for multiple teams and it cannot be integrated easily and will be a problem in visualization and analysis in database but in DWH it will be stored such that all the data will be interlinked by using schemas like star schema, snowflake schema, galaxies schema scan relate the tables and data. So can pull the data across the database.

Can put all the data and relate it with running a single query.

DWH has the data information more readable like pictorial.

**Query**

Take the data from operational system-> integrate the data from multiple sources->standardize the data& remove inconsistencies->store the data in format suitable for any access->

**Result**

DWH provides information -Processed data is called Information

Properties of DWH collection of data in support of managements decision making process

**Subject oriented- data is categorized and stored by business subject rather than an application (many questions single info)**

**Integrated- data on a given subject is collected from disparate sources and stored in a single place**

**Time-variant-Data is stored as a series of snapshots each representing a period of time**

**Nonvolatile- Typically data in the DWH is not updated or deleted**

**Key terminologies**

Information systems- OLTP(DB) vs OLAP(DWH)

**Relational database(OLTP) online transaction processing**

* **Running any kind of queries in DB called- OLTP**

Contains current data as well as past data

Useful in running a business (latest data of a customer)

Access of data is Based on entity relationship model

Provides primitive and highly detailed data (one data. But difficult to filter too many data) ,

Used for **writing data** into the database(updating the data)

Database size ranges from 100MB to 1GB (less)

Fast, provides high performance

No. of records accessed is in Tens

Ex:- all bank transactions made by customers

**Analytical data warehouse(OLAP)**

Contains only historical data

Useful in analyzing in business (details at what time which customer and which product)

Access of data Based on Star, snowflake, factsheet constellation schema

Provides summarized and consolidated data (exactly what we want)

Used for reading data from the data warehouse (only reading and do analyzation and visualization)

DWH size ranges from 100GB-1TB

Highly flexible, but not fast(gives different views of data via OLAP cube)

No. of records accessed are in millions

Ex-Bank transactions made by a customer at a particular time

ETL

Extracting a data from different sources and transforming the data to meet the requirement and load it into a target data warehouse

Xml file, flat file, cobol file , database->extract->transform->load->data warehouse

Tools for ETL- talend, informatica

Data Mart

Smaller version of the DWH dealing with a single subject

It focuses on a single area, hence they draw data from a limited number of sources

Time taken to build datamart is less than a time taken to build DWH

DWH

Enterprise wide data

Multiple subject areas

Multiple data sources

Occupies large memory

Longer time to implement

Data mart

Department wide data

Single subject area

Limited data sources

Occupies limited memory

Shorter time to implement

Datawarehouse-> sales data(data mart )

* Marketing data(datamart
* Operations data(data mart)

DWH acts as a repository for data mart

Types of data mart

Dependent – the data is first extracted from the OLTP systems and the populated in the central DWH, from DWH travel to Data mart

OLTP source->Datawarehouse->datamart

Independent – data directly received from the data source. Suitable for smaller systems and smaller organization

OLTP source->Datamart

Hybrid – The data fed from both OLTP systems and DWH

OLTP->Datamart

DWH->

Metadata- data about data (data about where the data stored, suppose the raw data, where it is stored, what size of the data, details like which source it came from and when)

In a DWH , it defines the source data that is Flat file, relational database and other objects

It is used to define which table source and target, and which concept is used to build business logic called transformation to the actual output.

Where the data coming from and where the data goes into DWH default purpose is to get the data from the source and refine and store it in DWH. Transmission(convert the data how you want to use to) epitome of DWH

DWH architecture

(Data sources)Database/ flat files->ETL->staging area (staging database)-ETL-> Datawarehouse (Metadata, raw data, aagregate data)-> Data marts(sales, purchase,stock)->user groups can access

Staging database (temporary storage) available in ETL

Demo in Talend

Populating a dwh