BSc in IT: Specialising in Data Science

IT3021: Data Warehousing and Business Intelligence

Lecture 01

Introduction to DW & BI

Content

- Module overview
  - Course content
  - Learning outcomes
  - Assessment criteria
  - Prerequisites
- Introduction to DW & BI
- OLTP vs DW (OLAP)
- Implementation steps

Module Overview

Course Content

- Introduction to DW/BI
- OLTP vs DW (OLAP)
- Data Warehousing Architectures
- Data Warehouse Designs & Concepts
  - Dimensional Data Modelling
- ETL/ELT Process (Data Ingestion Flows)
- OLAP Cubes and Related Concepts
- Business Intelligence
- Testing & Tuning
- Emerging Trends in Data Engineering

Learning Outcomes

- **LO 01:** Describe the role of BW & BI in today’s marketplace.
- **LO 02:** Develop familiarity with the data warehouse modelling concepts, and various technologies and tools required to implement a data warehouse.
- **LO 03:** Apply existing methods and tools for extracting, transforming, and loading data.
- **LO 04:** Design and implement BI solutions for real world problems.
- **LO 05:** Design and implement testing and tuning processes for DW & BI solutions.

Assessment Criteria

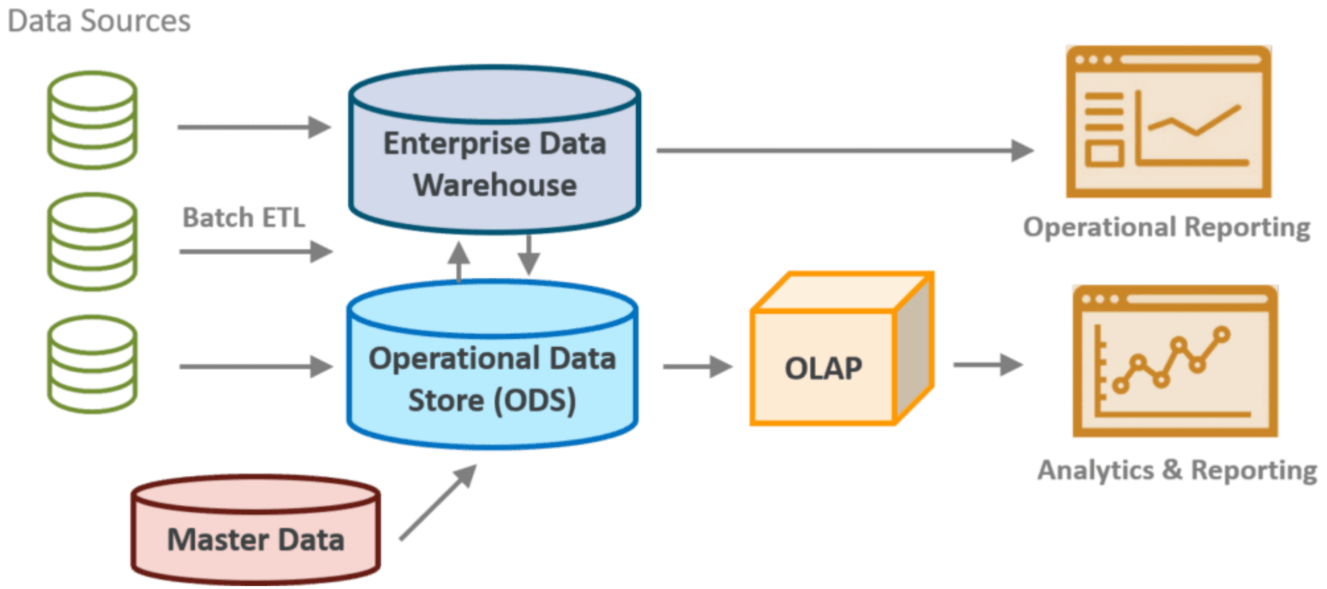
- Continues assessment
  - Practical Assignment 1: 20% **(L01 – L03)**
  - Practical Assignment 2: 20% **(L04 – L05)**

- End semester assessment
  - Final Examination - 60% **(L01 – L05)**

Prerequisites

- Basic concepts of database systems
- Database design process and E-R model (ERD)
- Normalization/De-normalization
- Relational database design
- Structured Query Language (SQL)

Introduction to DW & BI



Data

- One of the most important assets of any organization
- Purposes of data:
  - Operational record keeping (OLTP)
  - Analytical decision making (OLAP)
- Why different solutions for multiple purposes?
  - Different types of users
  - Different requirements
  - Performance reasons

What is DW & BI?

- Data Warehousing (DW)
  - It is a set of processes, architectures and technologies for collecting and managing data from various sources to support deriving meaningful business insights from raw data
  - Data collection involves data gathering, transforming and storing
  - It also includes database creation and data integration process development along with ‘data profiling’ and business validation rules
  - High level tasks include data acquisition, metadata management, data cleansing, data transformation, data distribution and data recovery/backup planning

What is DW & BI?

- Business Intelligence (BI)
  - It is a set of processes, architectures, and technologies for converting raw data into meaningful information and knowledge that supports profitable business actions
  - BI helps finding insights which portray business’s current picture (as-is) and historical picture (as-was)
  - Enable interested parties (end-users and down stream systems) to consume organization’s data by providing access to a consolidated data store (DW)
  - Deals with OLAP, data visualization, and data mining and query/reporting tools

History of Data Warhousing

- **1960 -** Dartmouth and General Mills in a joint research project, develop the terms dimensions and facts
- **1970 -** Nielsen and IRI introduces dimensional data marts for retail sales
- **1983 -** Tera Data Corporation introduces a database management system which is specifically designed for decision support
- **Late 1980s -** IBM worker Paul Murphy and Barry Devlin developed a Business Data Warehouse which is considered as the start of Data Warehousing
- **1990 -** the real concept was given by Inmon Bill (father of data warehouse). He had written about a variety of topics for building, usage, and maintenance of the warehouse & the Corporate Information Factory

Definition of Data Warehouse

- The term "Data Warehouse" was first coined by Bill Inmon in 1990
- According to Inmon:

  ‘a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization’

Characteristics of Data Warehouse

- Subject-Oriented
  - Offers information regarding a theme instead of organization’s ongoing operations. These subjects can be sales, marketing, distributions, etc. Subjects contain their unique and also common set of entities.
  - Emphasis on modelling and analysis of data for decision making.
- Integrated
  - A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. Moreover, it must keep consistent naming conventions, format, and coding.
  - Establishment of a common unit of measure for all similar data from dissimilar databases. The data also needs to be stored in the Data Warehouse in common and universally acceptable manner.
  - This integration helps in effective analysis of data. Consistency in naming conventions, attribute measures, encoding structure etc. have to be ensured.

Characteristics of Data Warehouse

- Time-Variant
  - The time horizon for data warehouse is quite extensive compared with operational systems. The data collected in a data warehouse is recognized with a particular period and offers information from the historical point of view. It contains an element of time.
  - In general, once data is inserted in the warehouse, it can't be updated or changed.

- Non-volatile
  - Previous data is not erased when new data is entered in it.
  - Data is periodically refreshed. This also helps to analyse historical data and understand what & when happened. It does not require transaction process, recovery and concurrency control mechanisms.
  - Activities like delete, update, and insert which are performed in an operational application environment are omitted in Data Warehouse environment.

Goals of DW & BI

- Make information accessible easily
  - Understandability
  - Obviousness
  - Users' vocabulary
  - Easy to use tools
- Present information consistently and timely
  - Credible data
  - Quality assured
  - Based on business requirements
  - Weekly, daily, hourly, near-real-time, etc.

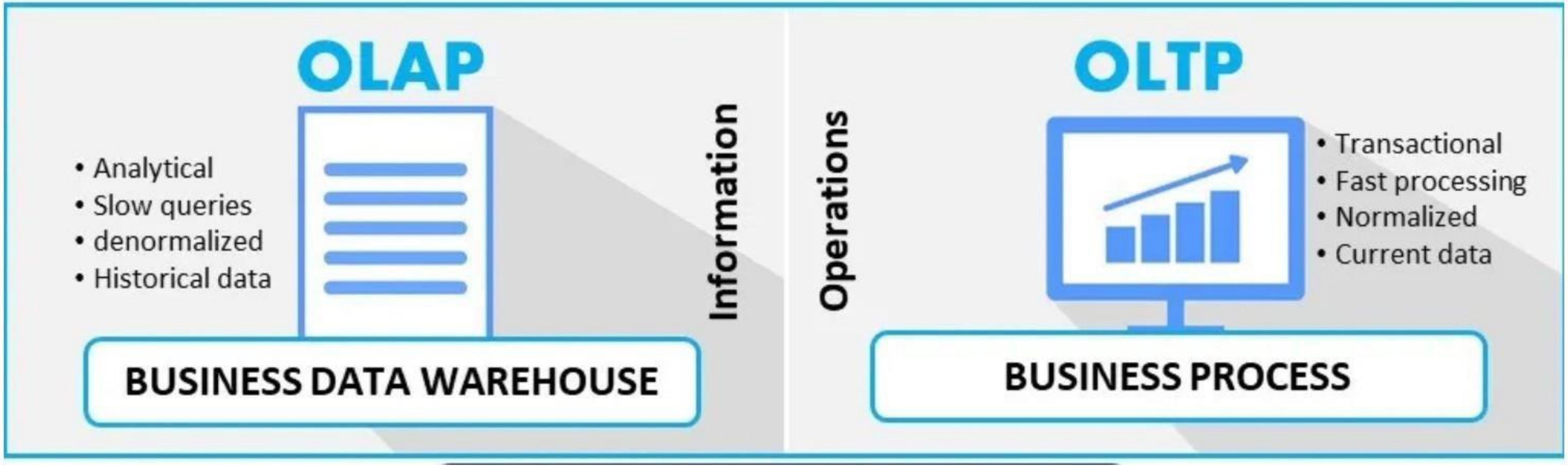
Goals of DW & BI

- Provides security
  - Access control
  - Data masking
- Acceptability of the solution
  - Active users
- Adapt to change
  - Requirements
  - Business logics
  - Technology

Who Needs Data Warehouses?

- People who wants a systematic approach for decisions making and relies on mass amount of data
- People who uses customized, complex processes to obtain information from multiple data sources
- People who relies on simple technologies to access the data
- People who wants fast performance on a huge amount of data, which is a necessity for visualization
- Who wants to discover ‘hidden patterns’ of data-flows and groupings

OLTP vs. DW (OLAP)



OLAP

- Analytical
- Slow queries
- denormalized
- Historical data

BUSINESS DATA WAREHOUSE

Information

OLTP

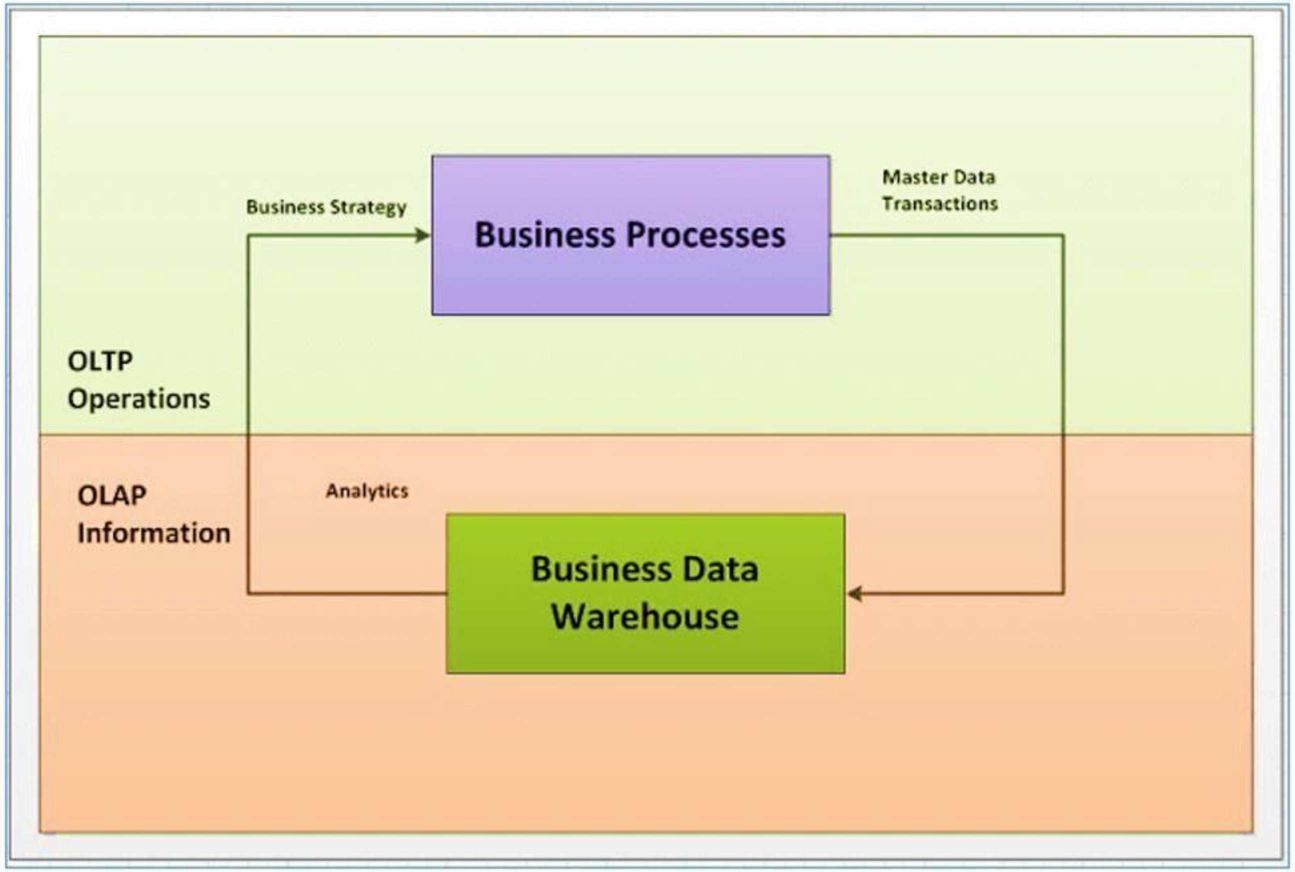
- Transactional
- Fast processing
- Normalized
- Current data

BUSINESS PROCESS

Operations

OLTP vs. DW (OLAP) Comparison

- **OLTP**: primary objective is data processing and not data analysis
- **OLAP**: primary objective is data analysis and not data processing

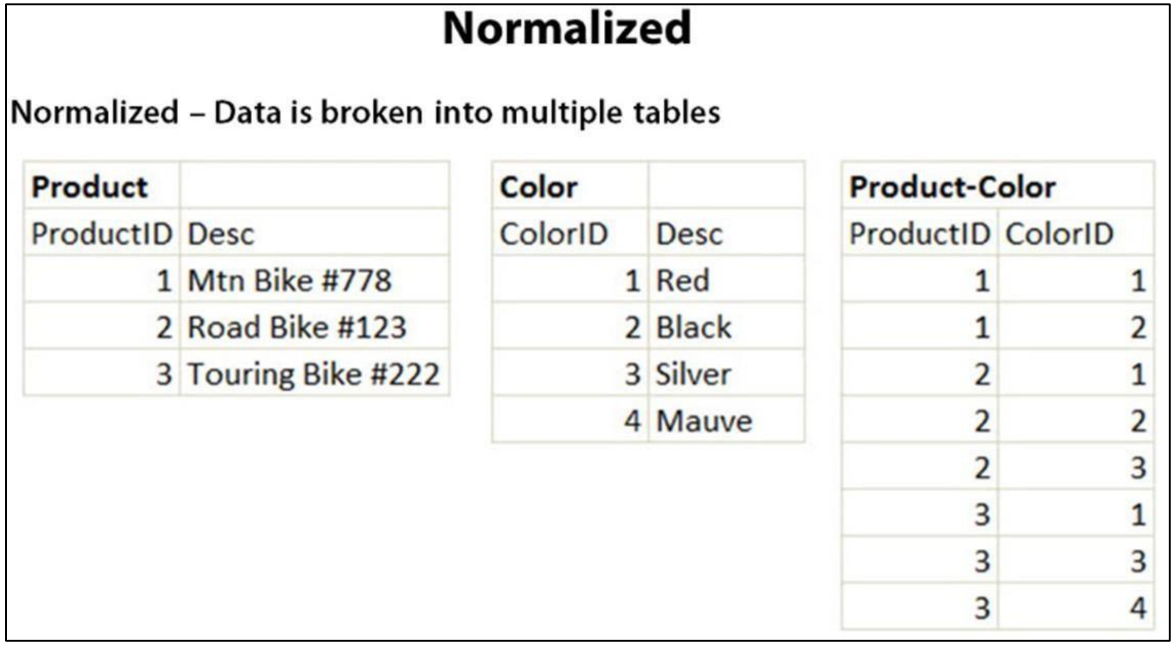


OLTP vs. DW (OLAP) Comparison

Parameter	OLTP	DW (OLAP)
Design	Application oriented	Subject oriented
Purpose	Operational and real-time/Transactional	Analytical
Data processing	Optimised for mostly updates and for required reads to support the operation	Optimised for reads. Rarely writes
Volume of data	Data required to support current operations of business	Vast amount of data to support historical analysis. Includes historical snapshots
Level of data	Elemental data required for day-today transactions	Raw data preserving history and summarised data
Data Intigerity	Maintains PKs	Business PKs are not must. Integrity managed using different mechanisms (SK)

OLTP vs. DW (OLAP) Comparison

Parameter	OLTP	DW (OLAP)
Data model	Normalized relational model	De-normalized dimensional model and multidimensional views



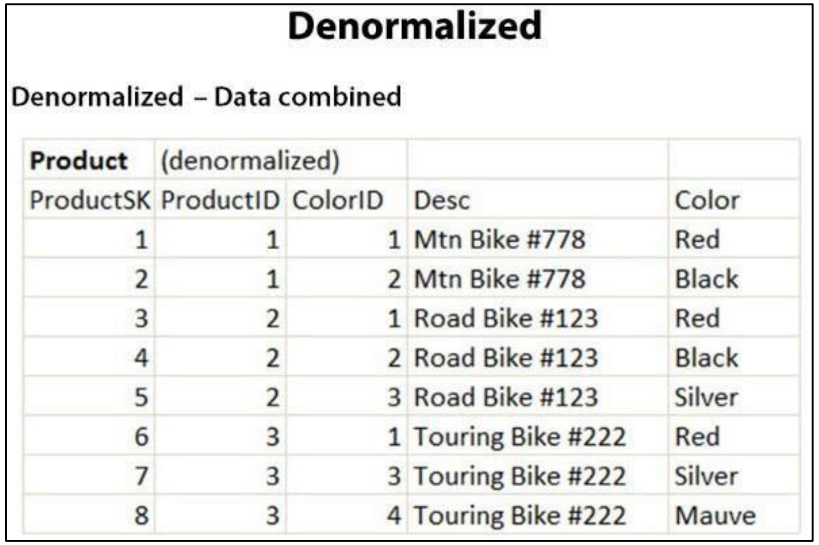
Normalized

Normalized – Data is broken into multiple tables

Product	
ProductID	Desc
1	Mtn Bike #778
2	Road Bike #123
3	Touring Bike #222

Color	
ColorID	Desc
1	Red
2	Black
3	Silver
4	Mauve

Product-Color	
ProductID	ColorID
1	1
1	2
2	1
2	2
2	3
3	1
3	3
3	4



Denormalized

Denormalized – Data combined

Product	(denormalized)			
ProductSK	ProductID	ColorID	Desc	Color
1	1	1	Mtn Bike #778	Red
2	1	2	Mtn Bike #778	Black
3	2	1	Road Bike #123	Red
4	2	2	Road Bike #123	Black
5	2	3	Road Bike #123	Silver
6	3	1	Touring Bike #222	Red
7	3	3	Touring Bike #222	Silver
8	3	4	Touring Bike #222	Mauve

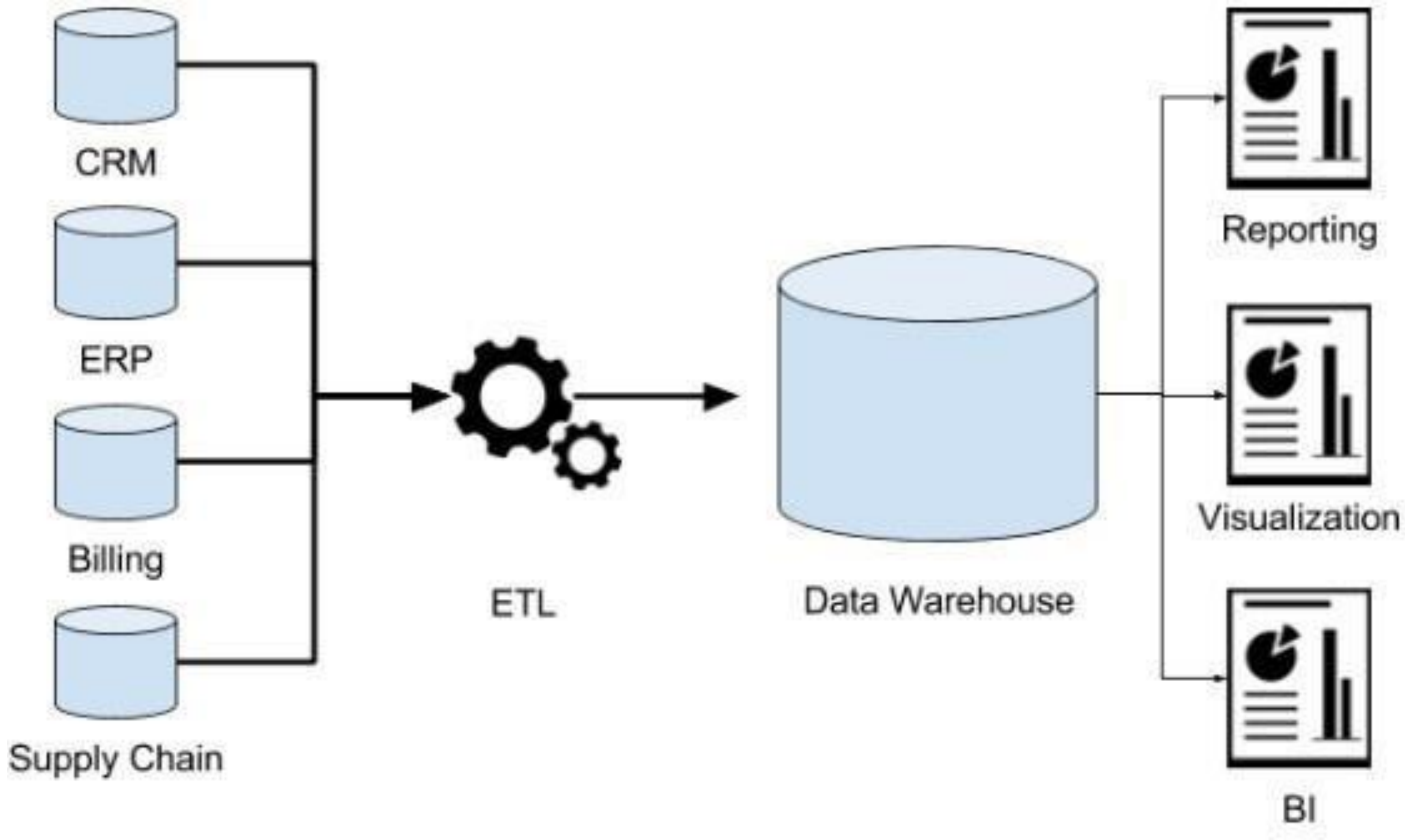
OLTP vs. DW (OLAP) Comparison

Parameter	OLTP	DW (OLAP)
Response times	Miliseconds	Seconds to minutes
Usefulness	To support and control intended business operations	To support business deceisions
User audience	People who runs the business operations	Decision makers and analysts
Number of users	High. Anyone who who is involved in business operations	Low. Executive level decision makers, top management, analysts
Back-up	Regular backups.	Not a must compared to OLTP. But time-to-time backups are taken
Tools	Traditional DBMS	DBMS for data warehouses (sometimes specilized appliances), OLAP engines, ETL tools, BI tools
Performance metric	Transactional throughput	Query throughput

Why Not OLTP for Analytics?

- Frequent updates
    - Great deal of locking
- Highly normalized data
    - Many table joins
- Too complex to support ad-hoc queries
    - Many tables to work with
- Slowness & impact on the transaction system

High-Level Architecture



Implementation Steps

Step	Tasks	Deliverable
1	Determine business objective and define scope	Scope definition
2	Collect and analyze rerequirements (business & technical), and identification of required architectural components	Architectural documents
3	Analyse source systems to understand data (this will help us understanding data quality requirements too!)	Data profile and analysis report
4	Data model building for data layer components (data warehouse, staging databases, operational data store, semantic layer inclusing OLAP cubes	Conceptual models Logical models *discussed in lecture 02
5	Identify required tools and technologies to implement the solution	Implementation plan
6	Detailed data model for the data warheouse and other data layer components	Physical model *discussed in lecture 02
7	Install/configure necessary tools and softwear (this could be cloud based tools or services!)	Documented implementation details and readied environment

Implementation Steps

Step	Tasks	Deliverable
8	Implement the data warehouse and other data layer components	Physically developed data models
9	Design and develop ETL process flow in each layer as applicable	Developed, ready to deploy set of process flows (ETLs)
10	Design and develop OLAP layer	Developed, ready to deploy set of OLAP layer
11	Design and develop BI layer components (required visualizations: reports/dashboards, self-service BI platform, front end BI application, etc.)	Developed, ready to deploy set of BI components
12	Initial data loading from sources to data layer	Components with data ingested
13	Process scheduling automation	Automated processes
14	Monitor, tuning, and enhancements	

Please note, testing is not mentioned in above tasks list, and should take place in most of the steps as required!