Stockholms universitet

Presentation: Introduction to Data Warehousing

Erik Perjons

DSV, Stockholm University

Stockholms universitet

Problems in Classic Information Systems

Problems to address

- ”We collect tons of data, but we can't access it"
  - "**Business people** need to **get the data easily**”
  - **Lack of technical tools for extracting required information in a proper format** (that is, a format easy to understand) **for business users**

- We need ***Accessibility*** to the data for users not very familiar with IT and data structures

(Kimball & Ross, 2013)

Problems to address

- "We spend entire meetings arguing about who has the right numbers rather than making decisions"
  - Different operational systems (databases) **devoted to specific business areas** within one enterprise easily leads to **inconsistencies between the systems** (databases)

- We need ***Data integration*** on the basis of a standard enterprise model

(Kimball & Ross, 2013)

Problems to address

- "We want to slice and dice the data every which way!"
  - **Decision making processes are often ad hoc** and we do not know in advance what is needed

- We need ***Query flexibility*** to maximise advantages obtained on the run

(Kimball & Ross, 2013)

Problems to address

- "Just show me just what is important!"
  - **Examining data on a maximum level of detail is self-defeating** because it takes focus from meaningful information (although details are sometimes needed)

- We need ***Information conciseness*** allowing target oriented and effective analysis

(Kimball & Ross, 2013)

Problems to address

- "We want to people to use information to support more fact based decision making"
  - **Data is often incorrect or unavailable**

- We need ***Correctness*** and ***Completeness*** of integrated data

(Kimball & Ross, 2013)

Problems to address

- We want to see trends
  - In business analytics, **trend and time-series analysis are essential to predict** and meet future demands.

- We need a solution that captures ***Historical data***

(Navathe, 2010)

Problems to address

- We do not want to wait for data
  - **Fast read access even over a large volume of data** is necessary in business analytics

- We need to optimize ***Read performance***

(Navathe, 2010)

Stockholms universitet

Solutions to the problems

Solutions to the problems: DW/BI

- Data warehouse (DW)
  - "A decision support database that is maintained separately from the organisation's operational databases."

    (Navathe)
- Data warehousing (DW)
  - " a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions."

    (S. Chaudhiri & U. Dayal)
- Business Intelligence (BI)
  - "Business intelligence (BI) is an umbrella term that is commonly used to describe the technologies, applications, and processes for gathering, storing, accessing, and analyzing data to help users make better decisions"

    (Wixom and Watson)
- DW/BI
  - Usually used together: BI applications utilizes data from DW: not all BI requires DW

The bedrock requirements for the DW/BI systems

- The DW/BI system must make information **easily accessable**
  - that is, data/info simple to understand and possuble to retrieve fast for the end user
- The DW/BI must **present information consistently**
  - that is, not different definitions for the same term in different departments, or different facts
- The DW/BI system must **present information in a timely way**
  - that is, present data within an hour, a day, a week, etc, after received in the transactional/operational system, depending on the requirement from business)
- ….

(Kimball & Ross, 2013)

The bedrock requirements for the DW/BI systems

- ….
- The DW/BI system must **adapt to change**
  - that is, manage organizational change due to changes in environment etc
- The DW/BI system must serve as the **authoritative and trustwothy foundation for improved decision making**
- The **business community must accept the DW/BI system** to deem it successfully

(Kimball & Ross, 2013)

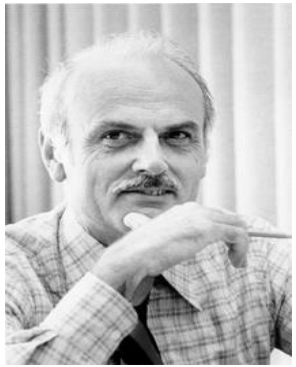
Two categories of systems

- Operational system (database)
  - Manage transactions
  - **Online Transactional Processing (OLTP)**
- Data warehouse
  - Manage data analysis
  - **Online Analytical Processing (OLAP)**

Decisions on different levels

Some influental people

- **Codd**
  - The relational model
  - OLTP/OLAP



- **Imnon**
  - Data Warehousing
  - DW 2.0

- **Kimball**
  - The Kimball Approach

The Data Warehouse - definition

Stockholms universitet

”A data warehouse is a **subject oriented**, **integrated**, **non-volatile**, and **time-variant** collection of data in support of management’s decisions”.

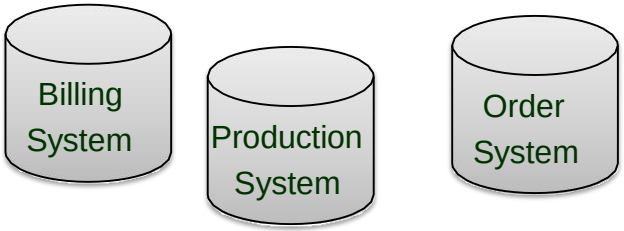
(Inmon)

These four concepts can be seen as major requirements on the DW/BI systems

However, it is not clear what the meaning of these four terms are

Data Warehouse - Subject-Oriented

- **Classical operation systems** is **organized around applications and functions:**
  - order handling
  - billing
  - sales
  - delivery



- **Data Warehouse** systems is **organized around subjects** (”major nouns”):
  - customer
  - supplier
  - product
  - office
  - date

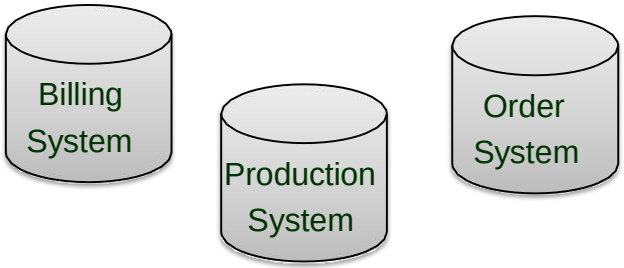
Answer questions such as **Who, What, Where, When regarding business events/transaction** (such as a sale or payments events)

Data Warehouse - Subject-Oriented

Stockholms universitet

- Focusing on the **modeling and analysis of data for decision makers, not on daily operations or transaction processing** … (**Query flexibility**)
- … to provide a **simple and concise view on particular subject issues** … (**Accessibility**)
- … by **excluding data that are not useful** in the decision support process (**Information Conciseness**)

Operational systems
- focussing on daily operations and transaction processing



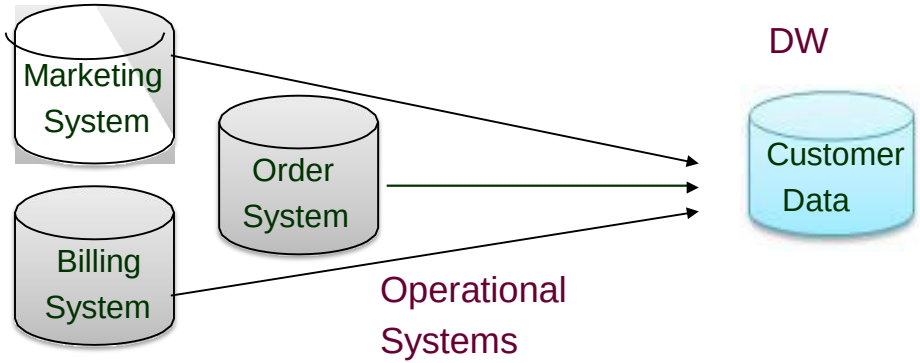
Data Warehousing systems
- data organized around on subject, that is major business objects and events

Data Warehouse - Integrated

- Constructed by **integrating multiple, heterogeneous data sources -** relational or other databases, flat files, external data, etc **- consolidating the organization's info** (**Data integration**)
- Integrated are often described as "consolidated", ¨"reconciled", "centralized" or "in a consistent format"

There may be different definition of the term "customer" in the different systems

There may be different attribute/data about a customer in the different systems



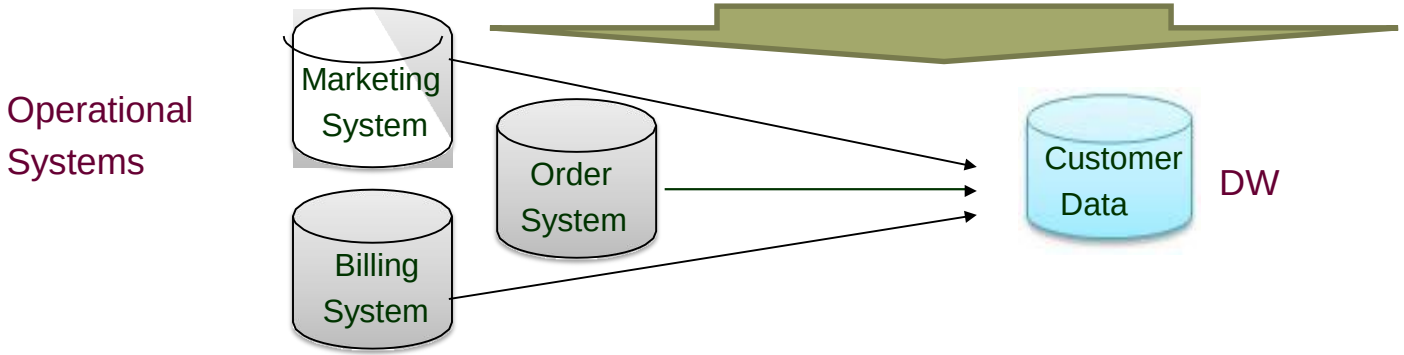
One definition of the term "customer"

All attributes/data about a customer is gathered

Data Warehouse - Integrated

- Data cleaning and data integration techniques are applied (**Correctness** and **Completeness**)
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
  - When data is moved to the warehouse, it is converted to the same format

- Cleansing
- Integration techniques
- Conversion to the same format
- Restructure



Data Warehouse – Time Variant

Stockholms universitet

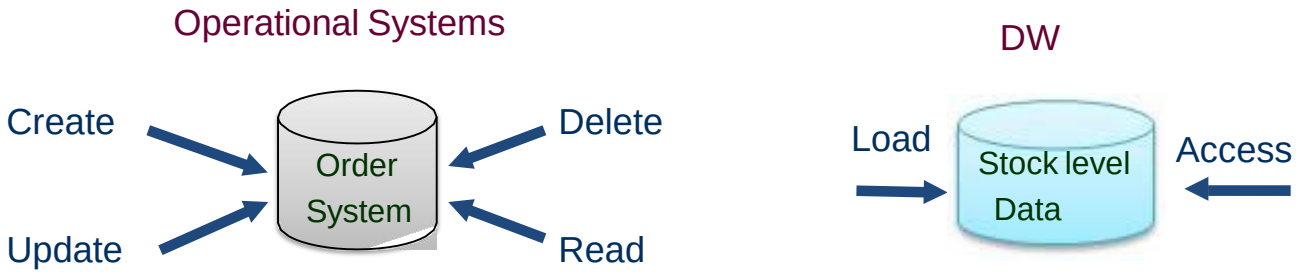
- Time variant means that **every unit of data in the data warehouse is accurate as of some moment in time**
- That is, we should have the **ability to move the time and access and analyse the data at a certain point in time**
- **This require that every transaction and snapshot stored in the DW must have a time-stamp**

Data Warehouse - Time Variant

- Note also, the **time horizon for the data warehouse is significantly longer than that of operational systems** (Historical data)
  - Operational database: **current data**
  - Data warehouse data: **provide information from a historical perspective** (e.g., past 5-10 years)
- Every core structure of data in the data warehouse contains an element of time

Data Warehouse - Non-Volatile

- Non-volatile implies that **when new data is added to the system** the **old data are never removed or changed**, instead they just remain there
- Once data is in the data warehouse, it will not change.



Data Warehouse - Non-Volatile

- **Operational update of data does not occur in the data warehouse environment**
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only: loading and access (read) of data

Summary: OLTP vs OLAP

Stockholms universitet

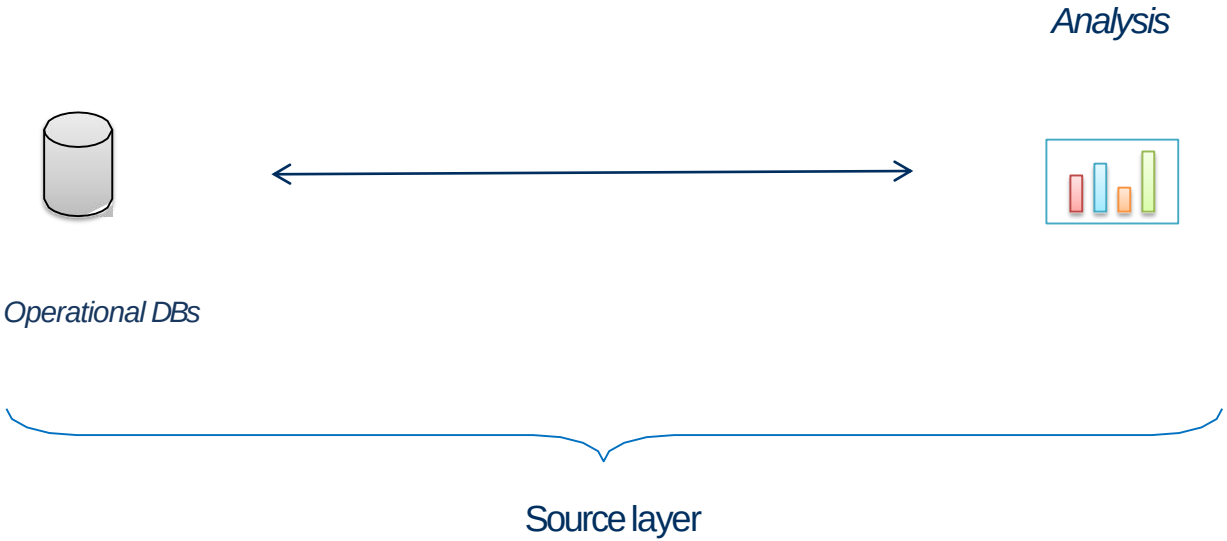
OLTP	vs.	OLAP
holds current data	←→	holds historic and integrated data
stores detailed data	←→	stores detailed and summarised data
data is dynamic	←→	data is largely static
repetitive processing	←→	ad-hoc and unstructured processing
predictable pattern of usage	←→	unpredictable pattern of usage
transaction driven	←→	analysis driven
application based	←→	subject based
support day-to-day decisions	←→	supports strategic decisions
operational users	←→	managerial users

Stockholms universitet

Layers for Analysis – DW architecture

(based on Golfarelli & Rizzi Data Warehouse Design, 2009)

Single-Layer Architecture for Analysis



Single-Layer Architecture for Analysis

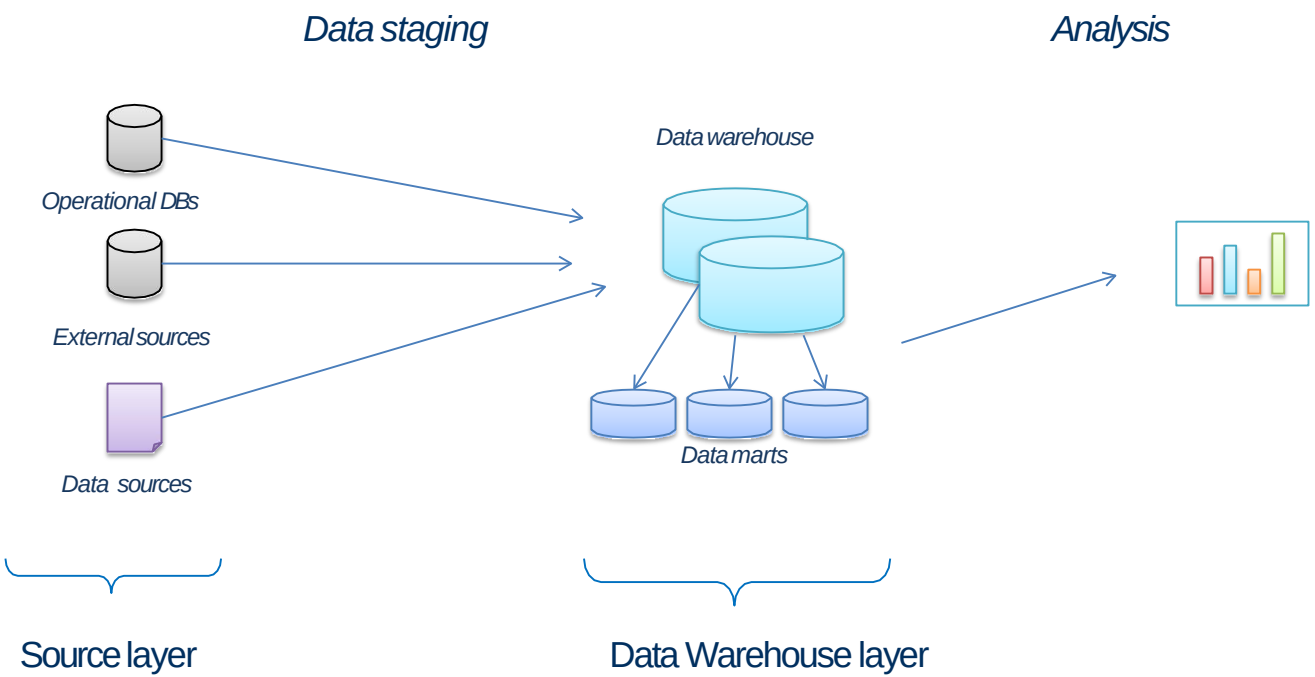
Stockholms universitet

- Rarely used architecture
  - **Does not separate analytical processing from transaction processing**
  - **Queries will affect transactional workloads (and disturb the operational system with complex queries)**
  - Not effiecient in **managing integration and correctness issues**

Why separate Data Processing from Analysis?

- The **operational systems/transactional systems** are tuned to **support known daily operations**
- **DW requires consolidating data from many heterogeneous sources**
- **DW requires** special **data structures, access methods and implementation methods**
- **DW may requires data that may be missing** in the operational systems. Therefore, we need to acquire additional external data

DW Architectures: Two-Layer for Analysis



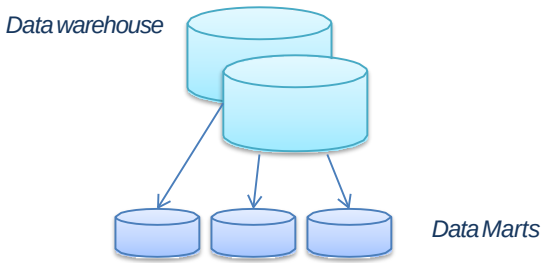
DW Architectures: Two-Layer for Analysis

- **Data is stored separately in a data warehouse and/or data marts**
  - Information is available even when sources are not
  - Queries for analysis do not affect transaction management in source systems
  - Data is structured to facilitate analysis rather than transactions
  - Data can be summarized
  - Historical data can be managed
  - Data warehouses can have specific design solutions for performance optimatition of analysis and report applications

Data Mart

- "**A data mart is a subset or an aggregation of the data stored to a primary data warehouse**. It includes a set of information pieces relevant to a **specific business area, corporate department**, or **category of users**."

Golfarelli & Rizzi



- **Data marts are useful:**
  - They can be used as building blocks when incrementally building the data warehouse
  - They mark out information required by specific groups of users
  - Better performance as they are smaller than primary data warehouses

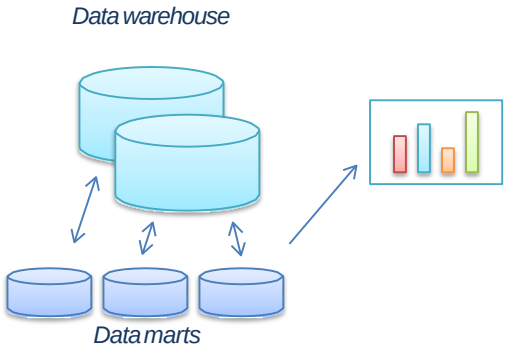
Data Warehouse vs. Data Mart

Stockholms universitet

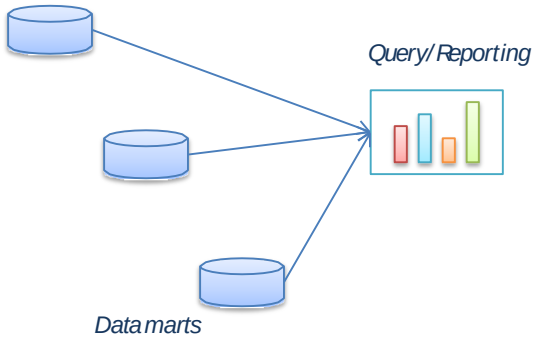
- **Enterprise warehouse:**
  - Collects all information about subjects that span the entire organization
    - (*customer, product, assets, personnel, supplier*)
  - Requires extensive business modelling
  - May take years to design and build
- **Data Mart:**
  - Departmental subsets that focus on selected subjects:
    - **Marketing data mart:** customer, product, sales
    - **Ordering data mart:** Supplier, product, stock level
  - Faster roll-out
  - Complex integration in the long term

Data Mart architecture

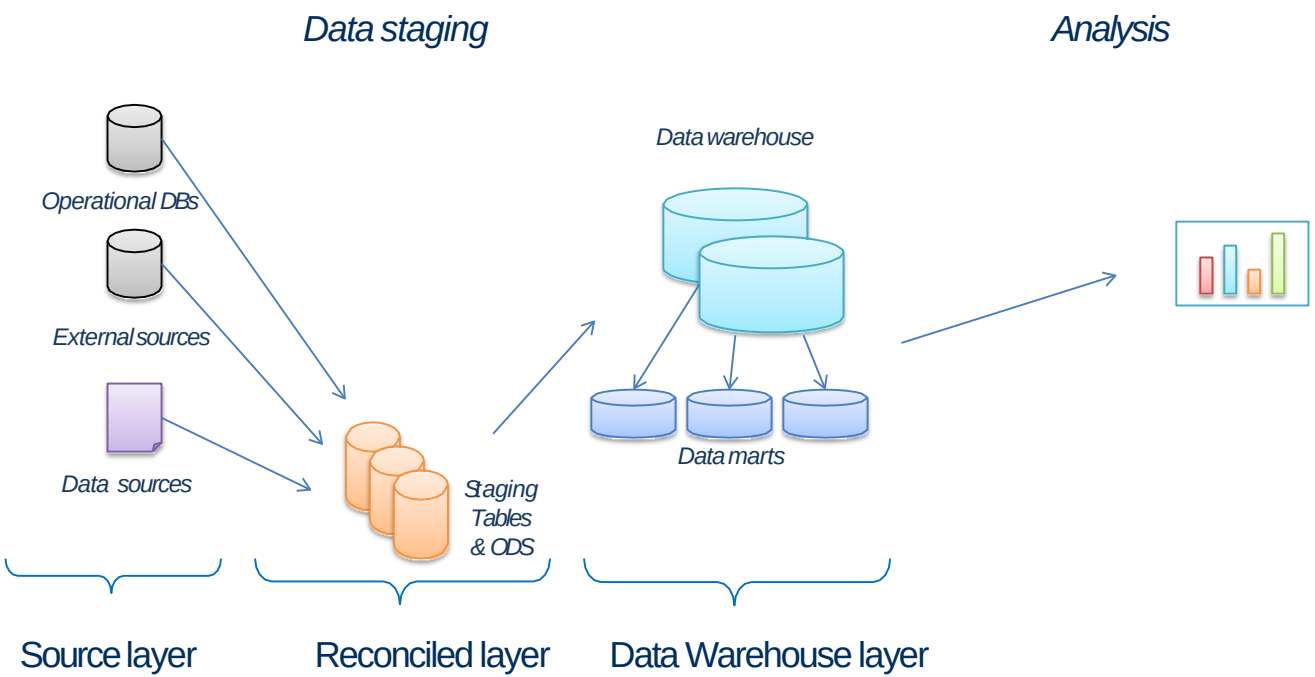
Dependent architecture



Independent architecture



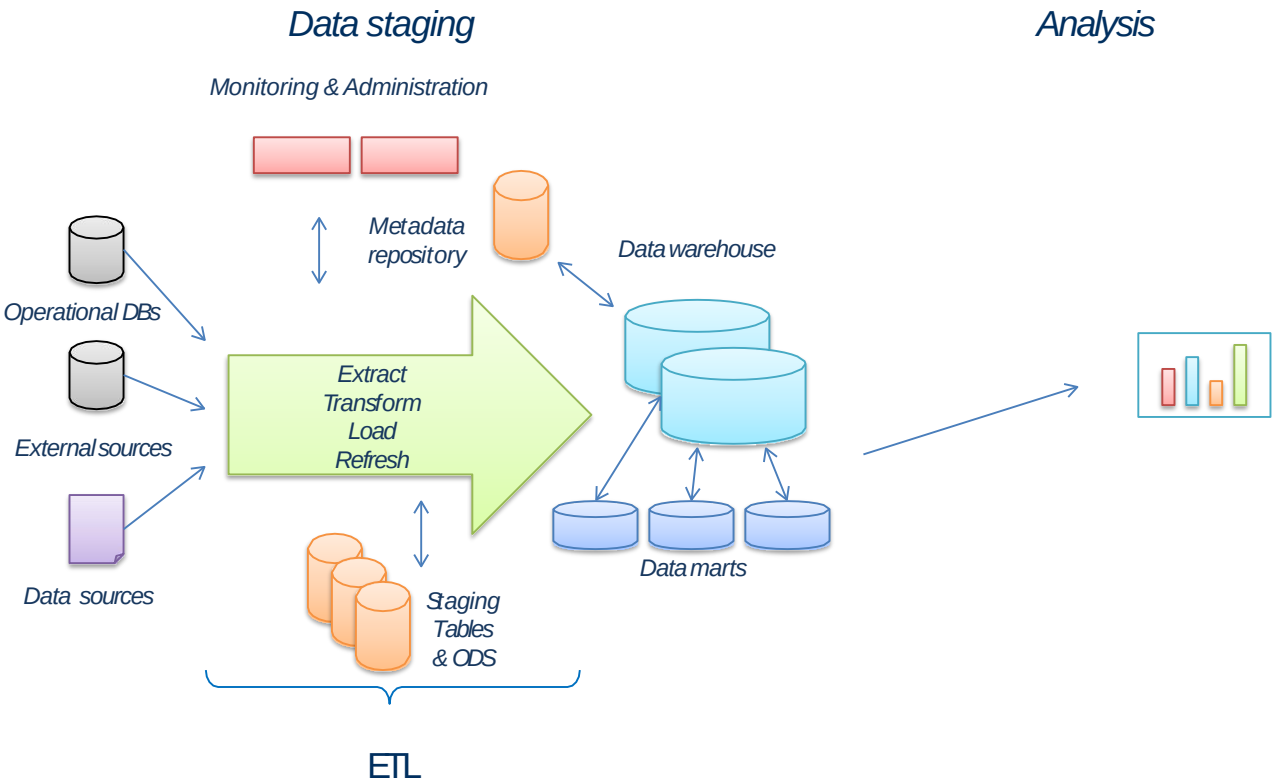
DW Architectures: Three-Layer for Analysis



DW Architectures: Three-Layer for Analysis

- **Reconciled data layer**
  - Can be a so called Operational Data Store (ODS)
  - ODS can be used in organisation if the organization need to access this data before it reach the DW – for example a bank needs to access data in the end of the data that is hard to get from the operational system/transactiional system
  - Holds operational data before it is manipulated for the DW
  - Can be an relational database in 3NF or flat files or both
  - Are used by ETL tools for transforming data from source systems. Such extracted data can be stored as flat files, staging tables, or in a database i 3NF

DW Architecture: The ETL Process



DW Architecture: The ETL Process

Stockholms universitet

- **Extract (E)**
  - Copying the data needed for the data warehouse into the staging area for further manipulation,
  - Usually from many (types of) sources
- **Transform (T)**
  - Data conversion
  - Data cleaning
  - Data enrichment

DW Architecture: The ETL Process

Stockholms universitet

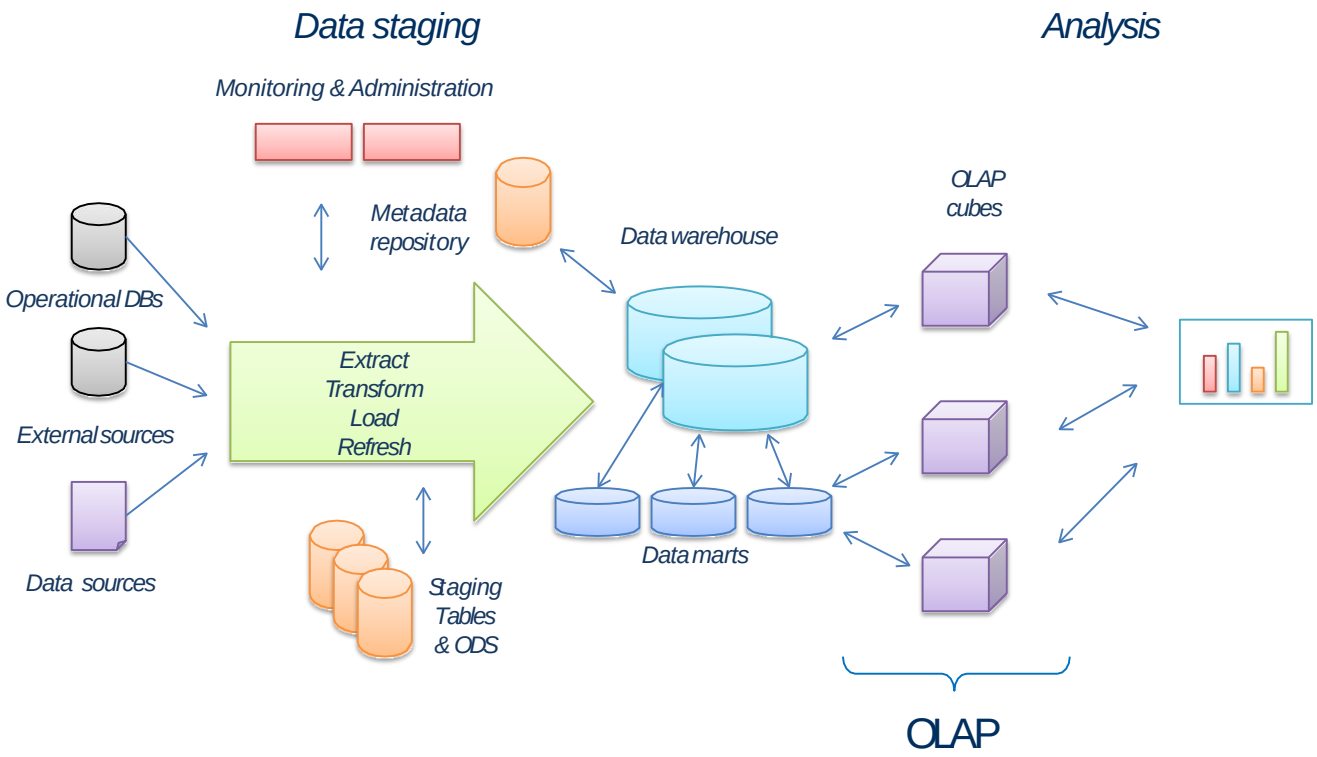
- **Load (L)**
  - data sorting
  - assigning warehouse keys
  - aggregation/summarisation
  - building indexes

DW Architecture: The ETL Process

Stockholms universitet

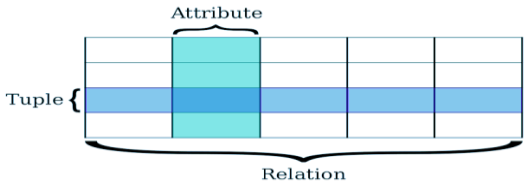
- **Refresh**
  - Incremental uploads
  - Replication server techniques
    - Data shipping (Push)
      - relying on triggers
    - Transaction shipping (Pull)
      - relying on transaction logs

DW Architecture: OLAP cubes

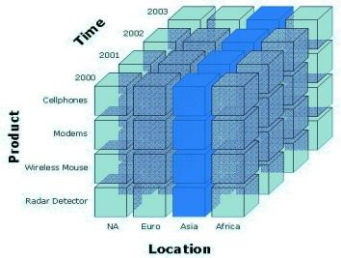


How to store data in a DW?

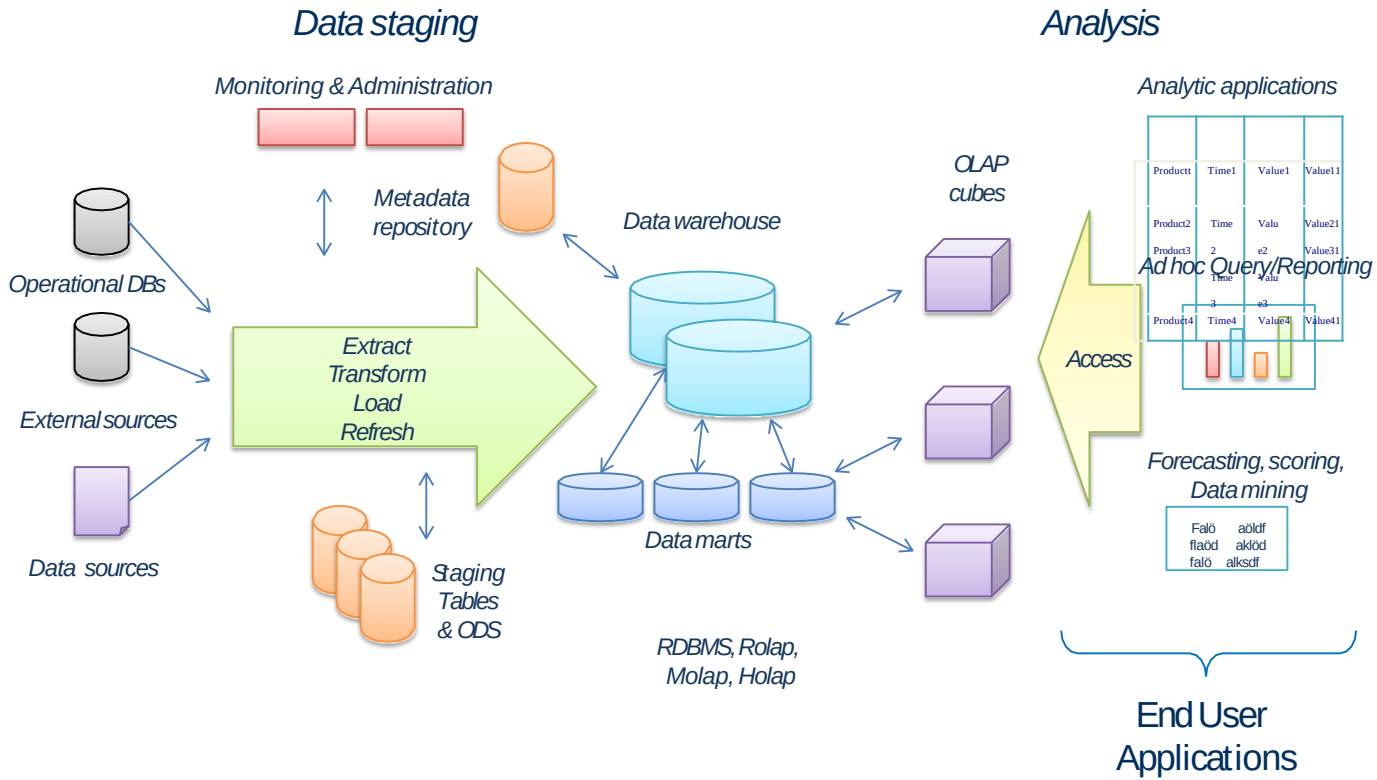
- Two dimensional: Relational Star schema



- Multidimensional OLAP Cubes



DW Architecture: End User Applications



DW Architecture: End User Applications

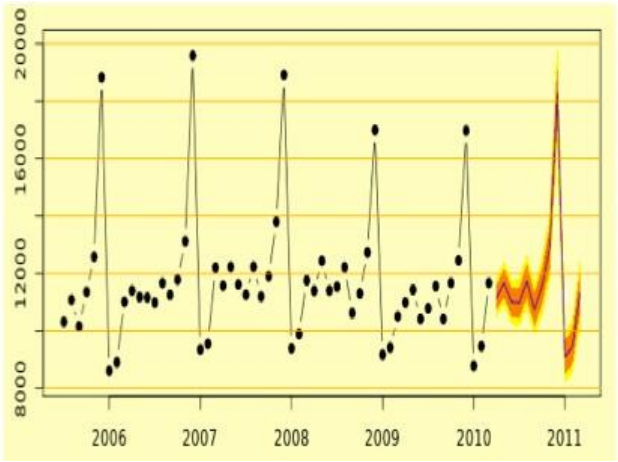
- Query / Reporting
  - E.g sql output as reports and spread sheets
- BI apps and OLAP tools
  - Microsoft Power BI, IBM Cognos, QlikView, Microsoft SSAS, Jaspersoft, SAP NetWeaver BI, Pentaho BI, Tableau BI
- Data mining
  - e.g. Text mining

DW Architecture: End User Applications

Dashboard examples

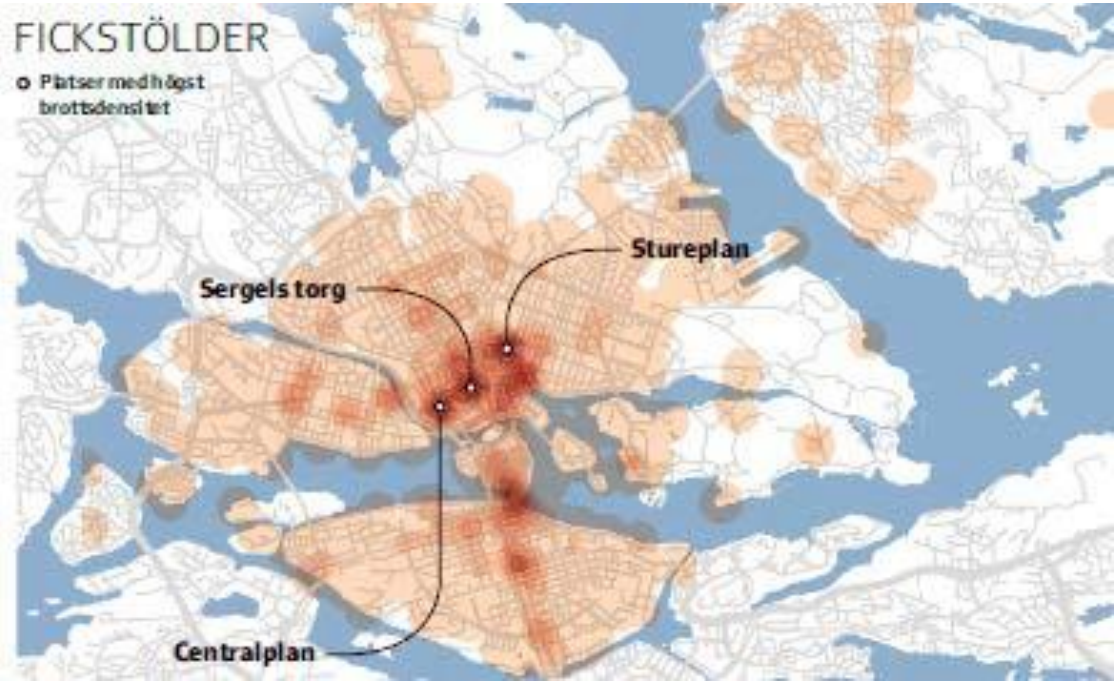
On the fly analysis

Prediction sales



Stockholms universitet

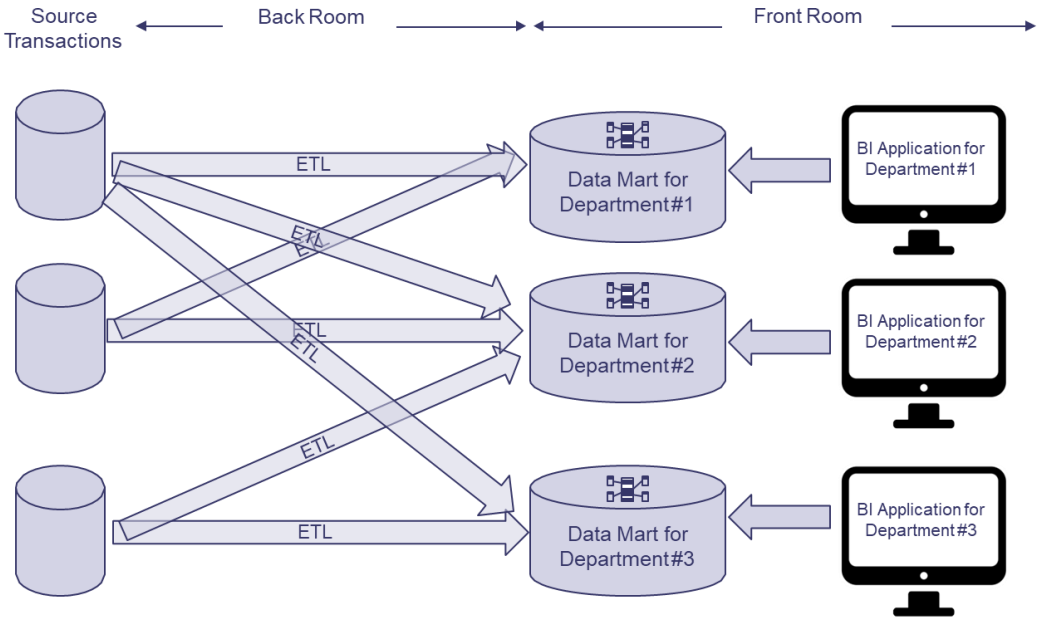
GIS analysis



ESRI & SvD©

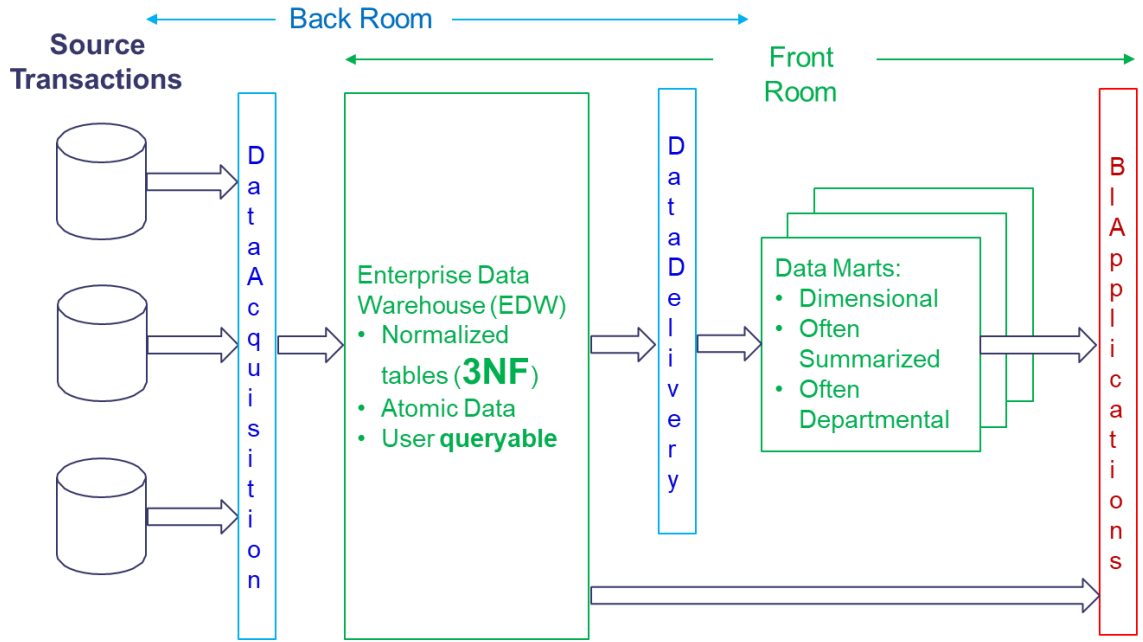
Four types of Data Warehouse architectures

Independent Data Mart Architecture



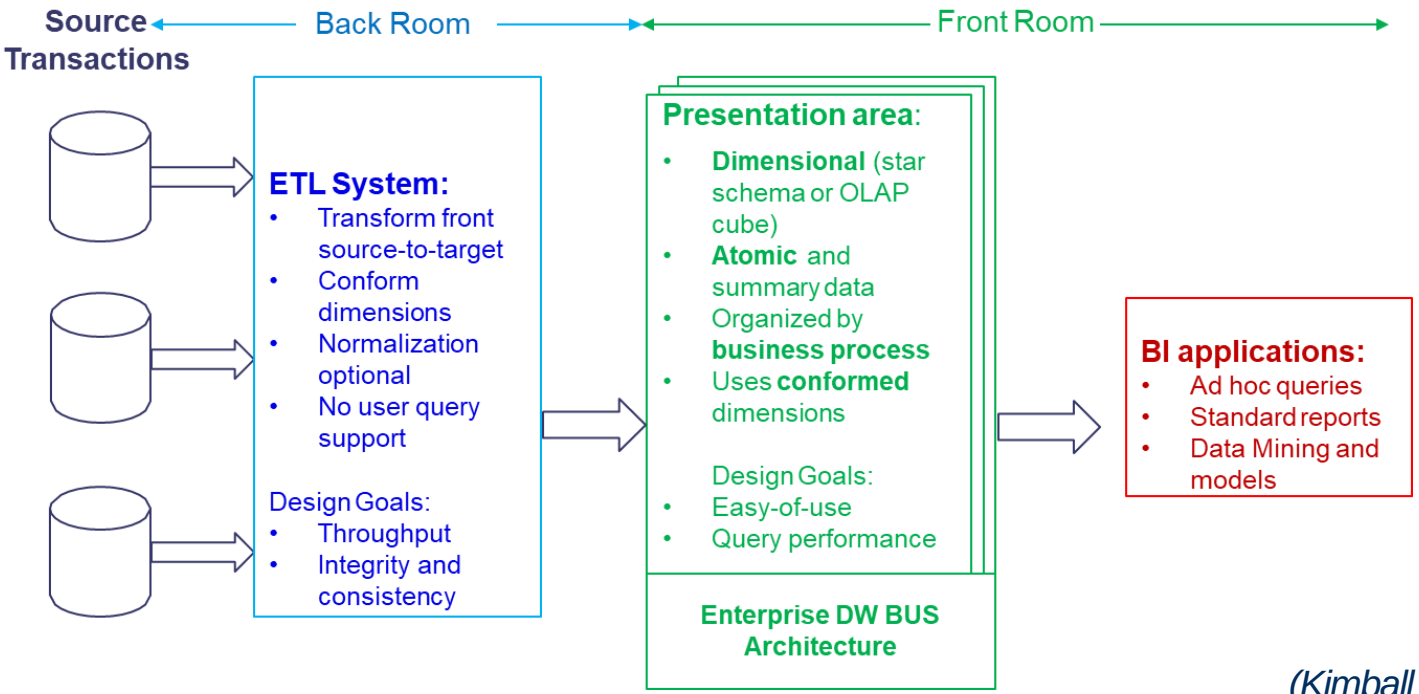
(Kimball & Ross, 2013)

Hub-and-Spoke Corporate Information Factury Archiecture (Inmon)



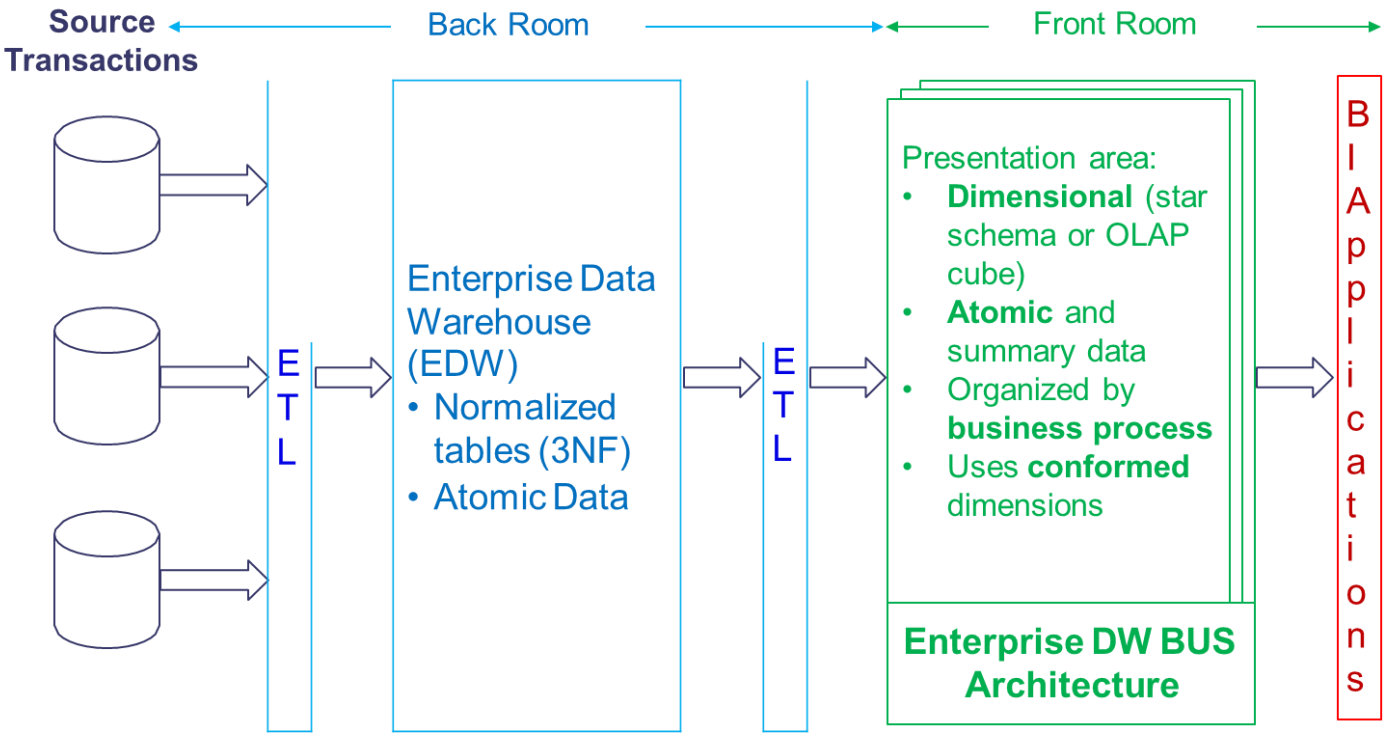
(Kimball & Ross, 2013)

Kimball/Ross DW/BI Architecture



(Kimball & Ross, 2013)

Hybrid Architecture



(Kimball & Ross, 2013)