University of South-Eastern Norway

Lecture 2: Data Warehouse Concepts

Veralia Gabriela Sánchez
Associate Professor

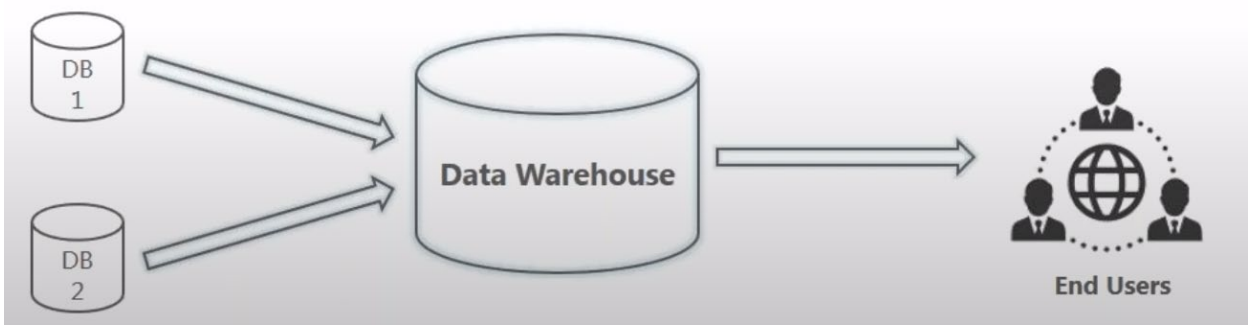
Outline

- Basic Data Warehouse Concepts
- Motivations and Characteristics
- Data Warehouse Architectures
- Employment Opportunities

Basic Data Warehouse Concepts

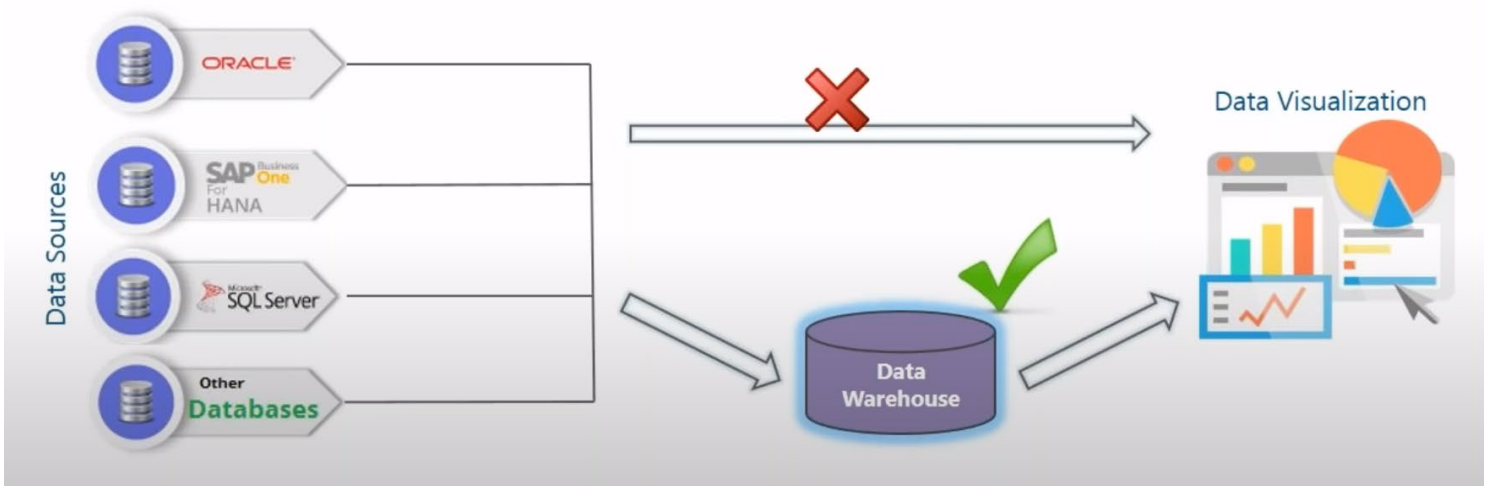
Business Intelligence and Data Warehouse

- Business Intelligence (BI) is the act of transforming raw/operational data into insightful and actionable business information.
- **How Does BI Works?**
  1. BI systems **collect** information from several data sources (the company operational DBs).
  2. This data is **transformed** (cleaned and integrated) and **loaded** into a **Data Warehouse**.
  3. Since the data in DW is credible, it is used for business insights.



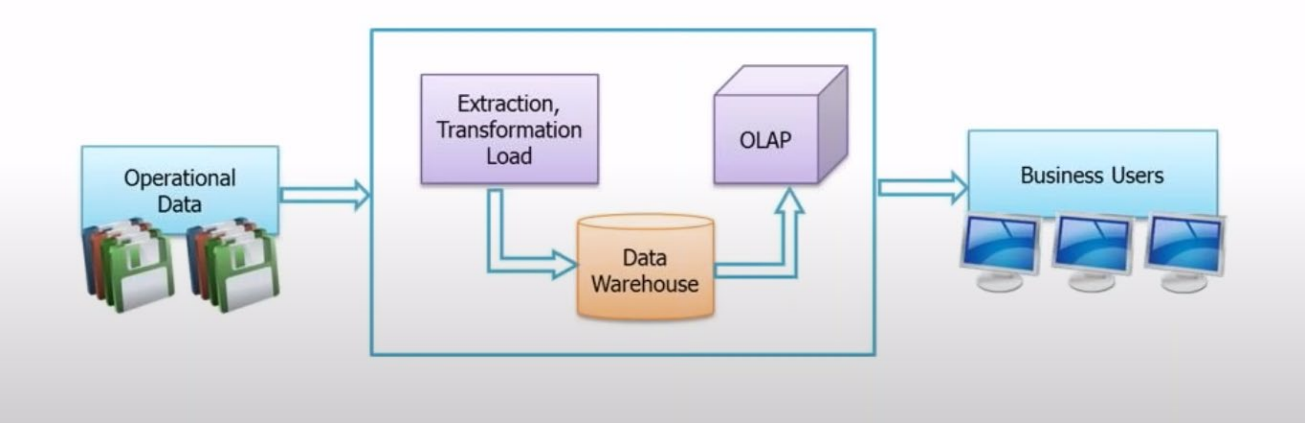
Why a Data Warehouse?

- Data collected from heterogenous sources cannot be directly visualized.
- The data first needs to be integrated and then processed before visualization takes place.



What is a Data Warehouse?

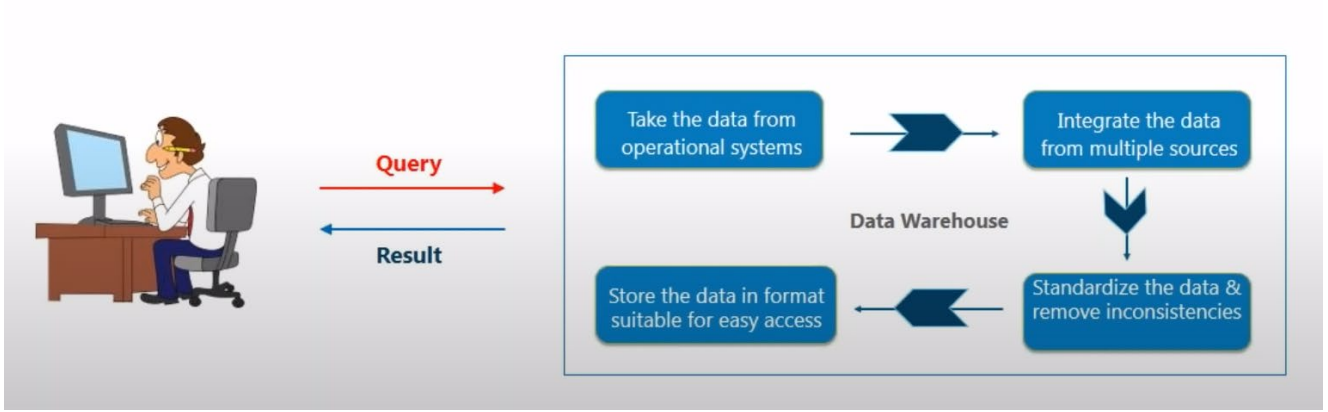
- A central location where transformed data from multiple sources (DBs) are stored.
- Data Warehouse is maintained separately from an organization's operational database.
- End users access DW whenever any information is needed.
- **Note that** a Data Warehouse is not loaded every time new data is added to database. Generally, the loading process occurs automatically once per day during the night.



Motivations and Characteristics

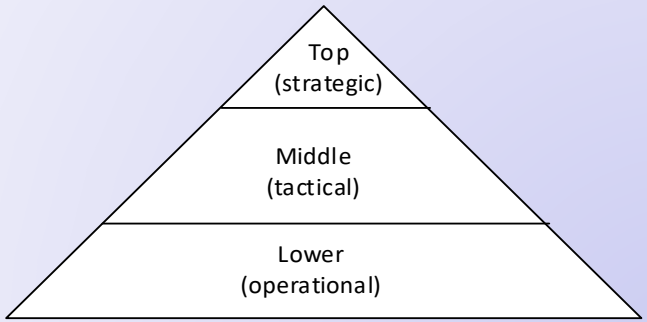
Advantages of a Data Warehouse

- A Data Warehouse can answer strategic and tactical Business questions.
- Data Warehouse is fast and accurate in responding to queries.
- **Note that** a Data Warehouse is not a product that a company can purchase. It needs to be designed and implemented according to the company business requirements.



Decision Making Hierarchy

Decision making hierarchy



Typical decisions

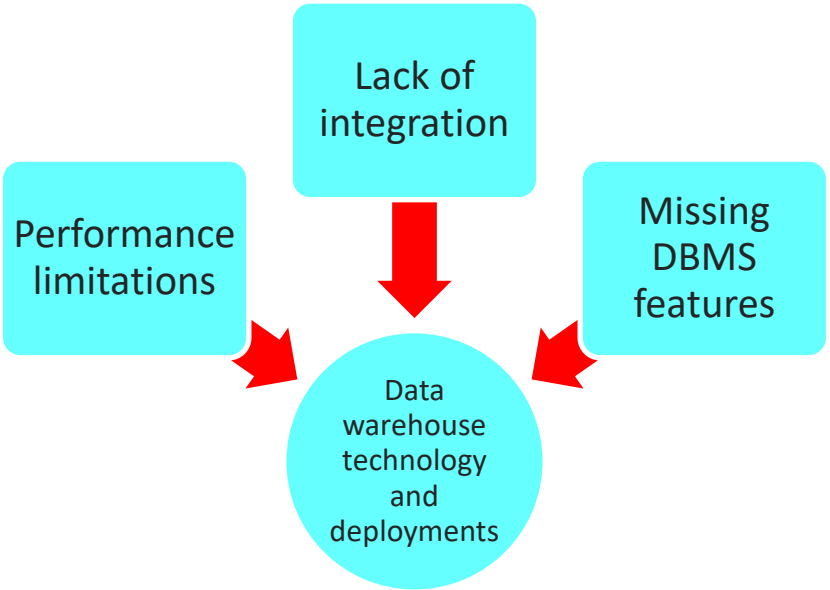
Identify new markets, choose store locations

Choose suppliers, forecast sales

Resolve order delays, schedule employees

Technology and Deployment Limitations

- Why Databases are not suitable for carrying BI tasks?
  - DB cannot be integrated with other transaction DB and other external sources.
  - Using the same DB for transaction processing and BI tasks results in high workload for DB and lead to performance drop.
  - DB does not offer the features requested from Datawarehouse, e.g., summary data.



Data Warehouse Definition

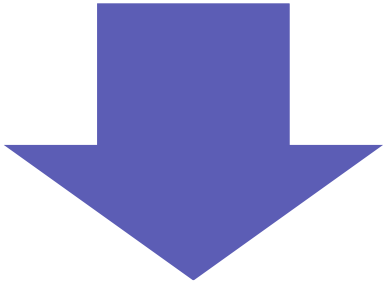
- DW is an essential part of infrastructure for business intelligence
- DW is logically centralized repository for decision making
  - Populated from operational databases and external data sources
  - Integrated and transformed data
  - Optimized for reporting and periodic integration

Data Warehouse Characteristics

- "A Data Warehouse is a ***subject-oriented***, ***integrated***, ***time-variant*** and ***nonvolatile*** collection of data in support of management's decision-making process." *Bill Inmon, Father of Data Warehousing*

  - *Subject-oriented:* Organized around business entities (e.g., customers, products, and employees) rather than business processes
  - *Integrated:* many transformations to unify source data from independent data sources (units of measure, data formats, naming conventions)
  - *Time-variant:* historical data (time stamped); snapshots of business processes captured at different points in time
  - *Nonvolatile:* existing data is not changed; new data are appended periodically; warehouse data may be archived after its usefulness declines

Comparison of Processing Environments

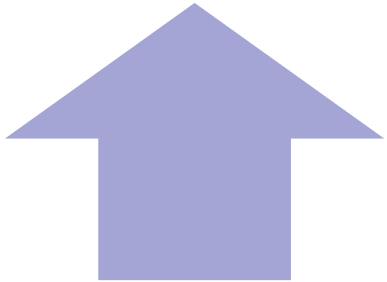


Transaction processing

- Primary data from transactions
- Daily operations and short term decisions

Business intelligence processing

- Transformed secondary data
- Medium and long-term decisions

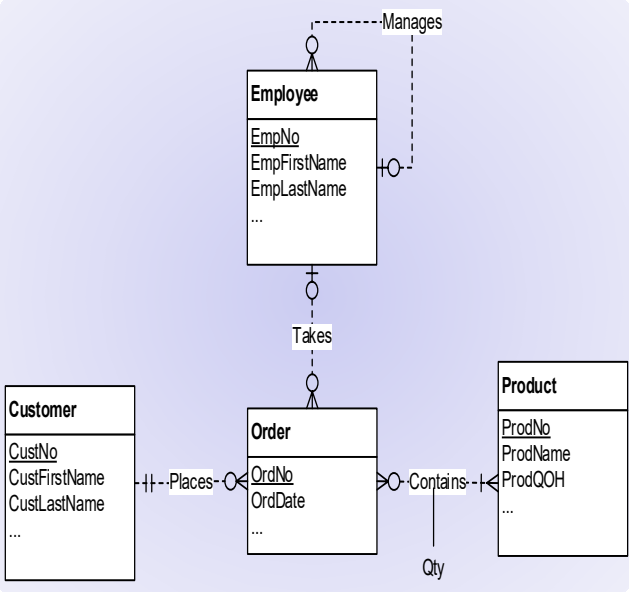


Data Comparison

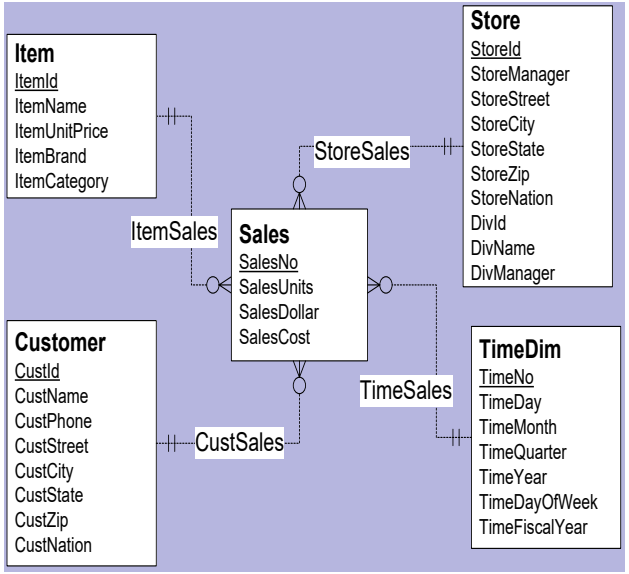
Characteristic	Operational Database	Data Warehouse
Currency	Contains **current** data	Contains **historical** data
Details level	Individual	Individual and summary
Orientation	Process-oriented	Subject-oriented
Records per request	Few	Thousands
Normalization level	Mostly normalized	Normalization relaxed
Update level	Highly volatile	Mostly refreshed (nonvolatile)
Data model	Relational	Relational (star schemas) and multidimensional (data cubes)

Schema Comparison

- Operational database



- Data warehouse



Data Warehouse Architectures

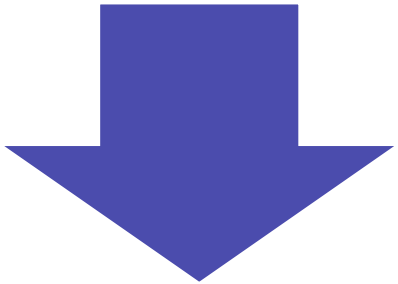
Architecture Issues

- The choice of the architecture for a DW is decided by the organization (the company), this choice is not limited by the technology
- The choice of DW architecture depends on Data warehouse scope
  - Number of data sources
  - Number of organizational units

Architecture Issues

- The choice of DW architecture depends on the integration level
  - Coordination and cooperation among business units
  - Find common entities
  - Enforce standards: units of measure, naming conventions
  - Reconcile differences such as revenue and cost recognition
  - Sometimes modify source systems

Architecture Choices

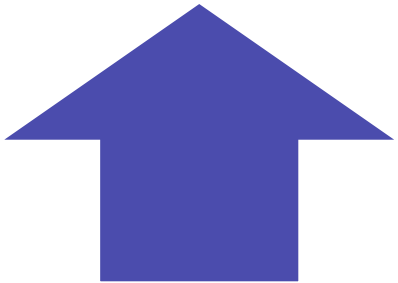


Top Down

- Enterprise data warehouse
- Higher integration levels
- Logically centralized
- Larger project scope

Bottom Up

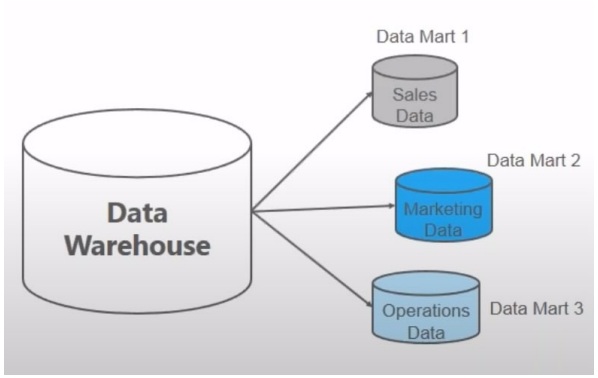
- Independent data marts
- Lower integration levels
- Logically decentralized
- Smaller project scope



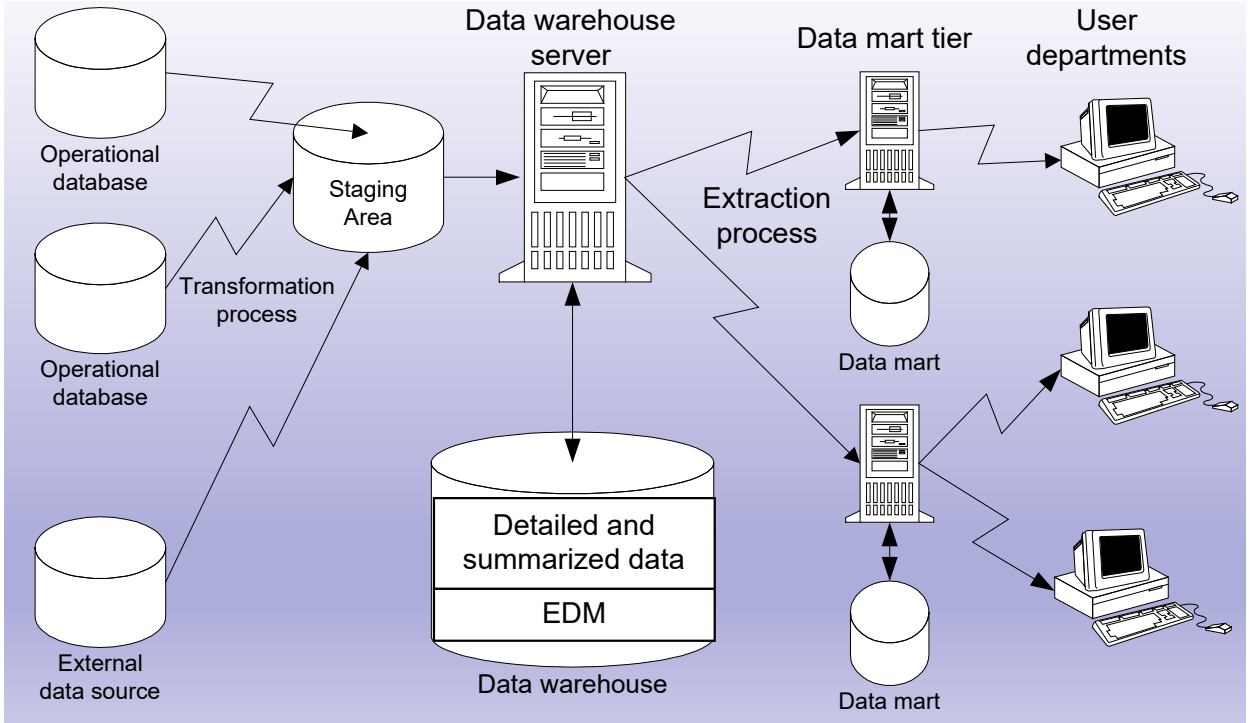
Data Mart

- Data mart is a smaller version of the Data Warehouse which deals with a single subject area.
- Since Data marts focus on one area, they collect data from a limited number of sources
- The time taken to build a Data Mart is very less compared to the time taken to build a Data Warehouse

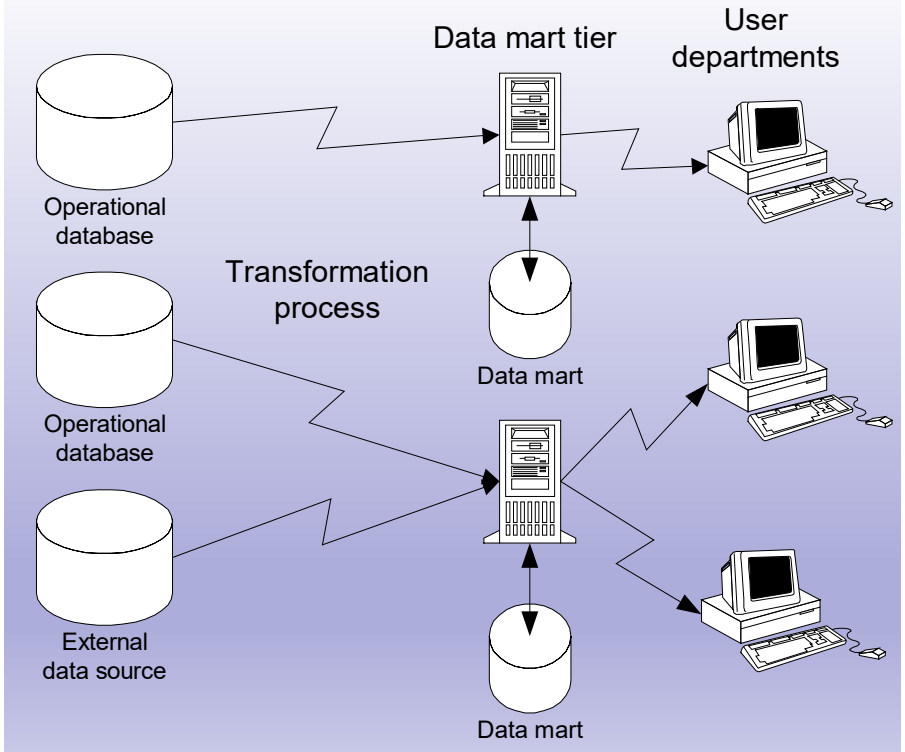
Data Warehouses	Data Marts
Enterprise-wide data	Department-wide data
Multiple subject areas	Single subject area
Multiple data sources	Limited data sources
Occupy large memory	Occupy limited memory
Longer time to implement	Shorter time to implement



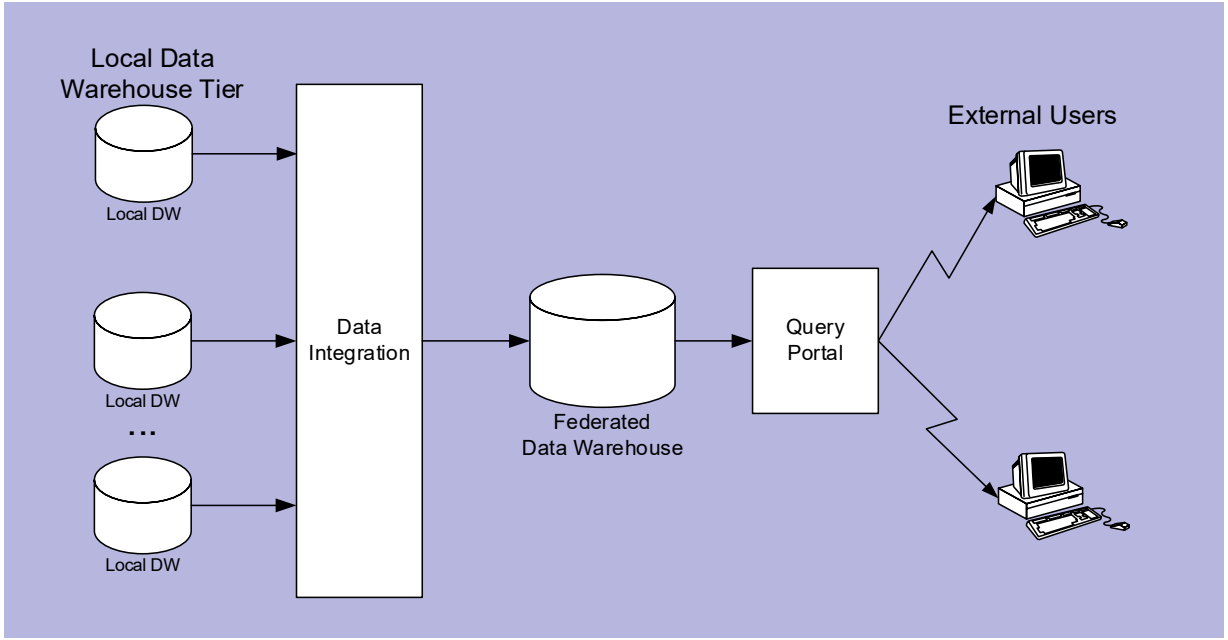
Top-Down Architecture



Bottom-up Architecture



Federated Architecture



Federated Architecture

- For highly decentralized or independent organizations, the federated data warehouse architecture provides another compromise approach.
- As depicted in this diagram, the federated data warehouse approach supports two levels of data warehouses.
- Each organization independently maintains one or more data warehouses using any of the architectures.
- To provide inter-organizational sharing, each organization contributes to the federated data warehouse.

Federated Architecture

- Typically, another layer of data integration and a query portal support data sharing in the federated data warehouse.
- Depending on the environment, participation can be voluntary or compulsory (typically required by government agencies).
- Some users of a federated data warehouse may be external stakeholders, not members of participating organizations.
- A possible use case is a consortium formed by several companies can adopt the Federated Architecture

Architecture Selection Factors

- Learning effects
  - Project risk
  - Intangible business value
- Strategic view of information technology
  - Level of sponsorship
  - Information independence
  - Task routineness

Employment Opportunities

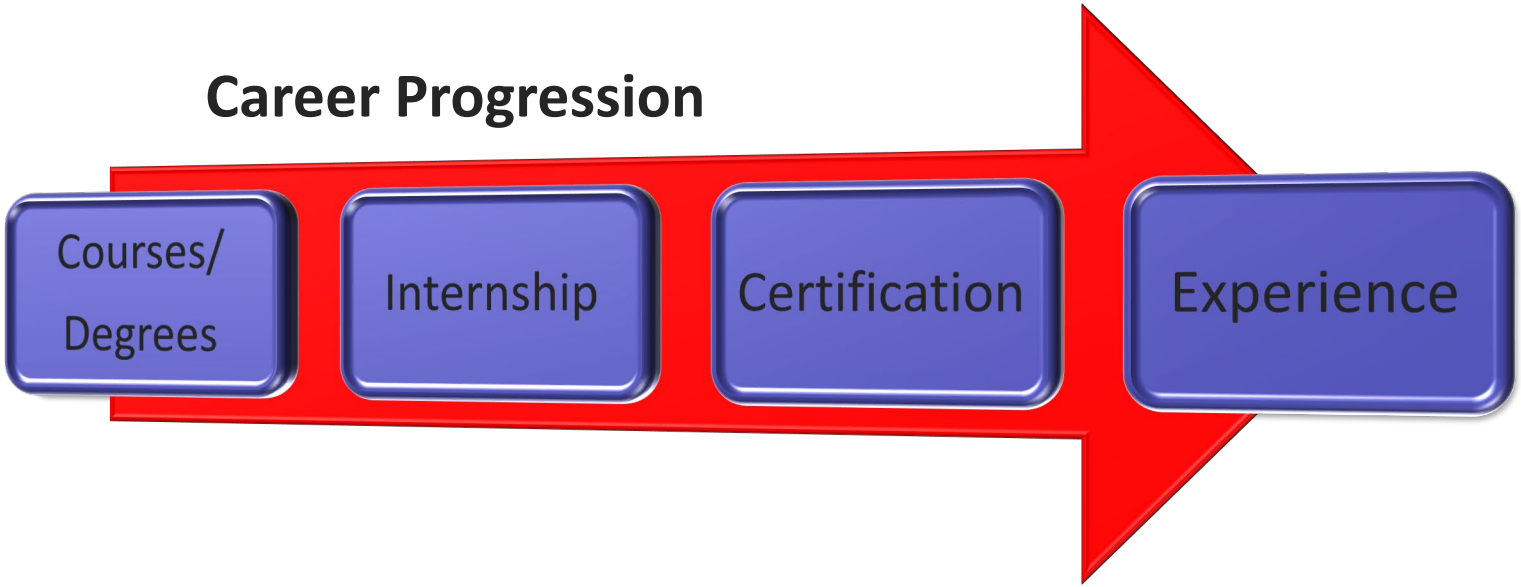
Employment Opportunities

DW Analyst	• Recommend technology solutions • Define user interfaces • Collaborate with business analysts and DW managers
DW Manager	• Design, develop, and maintain data warehouses • Ensure conformance to enterprise standards • Develop and implement data integration procedures
BI Analyst	• Develop data analysis and reporting solutions • Mine and analyze data from multiple sources • Communicate results to management • Prepare data (reduction and missing values)
Data Analyst	• Document data elements • Use reporting tools • Collaborate with business analysts and data architects • Develop data extraction procedures

Skill-Position Mapping

Competency	Position		
	DW Manager	*DW Analyst*	*BI Analyst*
Communication	■	■	■
Data cube tools	■	■	■
Dashboards		■	■
Data mining	■		■
Data integration tools	■	■	
DW schema design	■	■	
Performance analysis	■		
Quantitative modeling			■
SQL extensions	■	■	■

Competency Acquisition



Salary Percentiles (USA) in 2018

Job Title	25th Percentile	Midpoint	75th Percentile
DB manager	$107,000	$127,000	$152,250
DB developer	$97,950	$116,000	$139,000
Data analyst	$81,000	$96,000	$115,000
DW manager/architect	$110,00	$130,000	$156,000
DW analyst	$75,250	$93,500	$117,000
BI analyst	$83,750	$104,000	$130,250