Data Warehousing & Business Intelligence - I

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

School Of Computer Engineering

Datamining and Data warehousing (CS 2004)

Dr. Amiya Ranjan Panda
Assistant Professor [II]
School of Computer Engineering,
Kalinga Institute of Industrial Technology (KIIT),
Deemed to be University,Odisha

3 Credit

Lecture Note 06

Acknoledgement

A Special

Thanks to

J. Han and M. Kamber.

&

Tan, Steinbach, Kumar

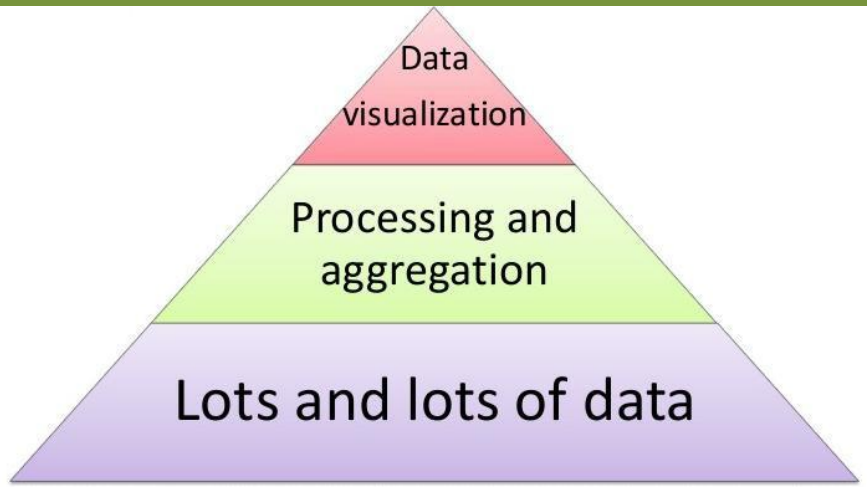
for their slides and books, which I have used for preparation of these slides.

Chapter Contents

- What is the need for BI (Business Intelligence)?
- What is Data Warehousing?
- Key Terminology to DWH Architecture
  - OLTP Vs OLAP
  - ETL
  - Data Mart
  - Metadata
- DWH Architecture

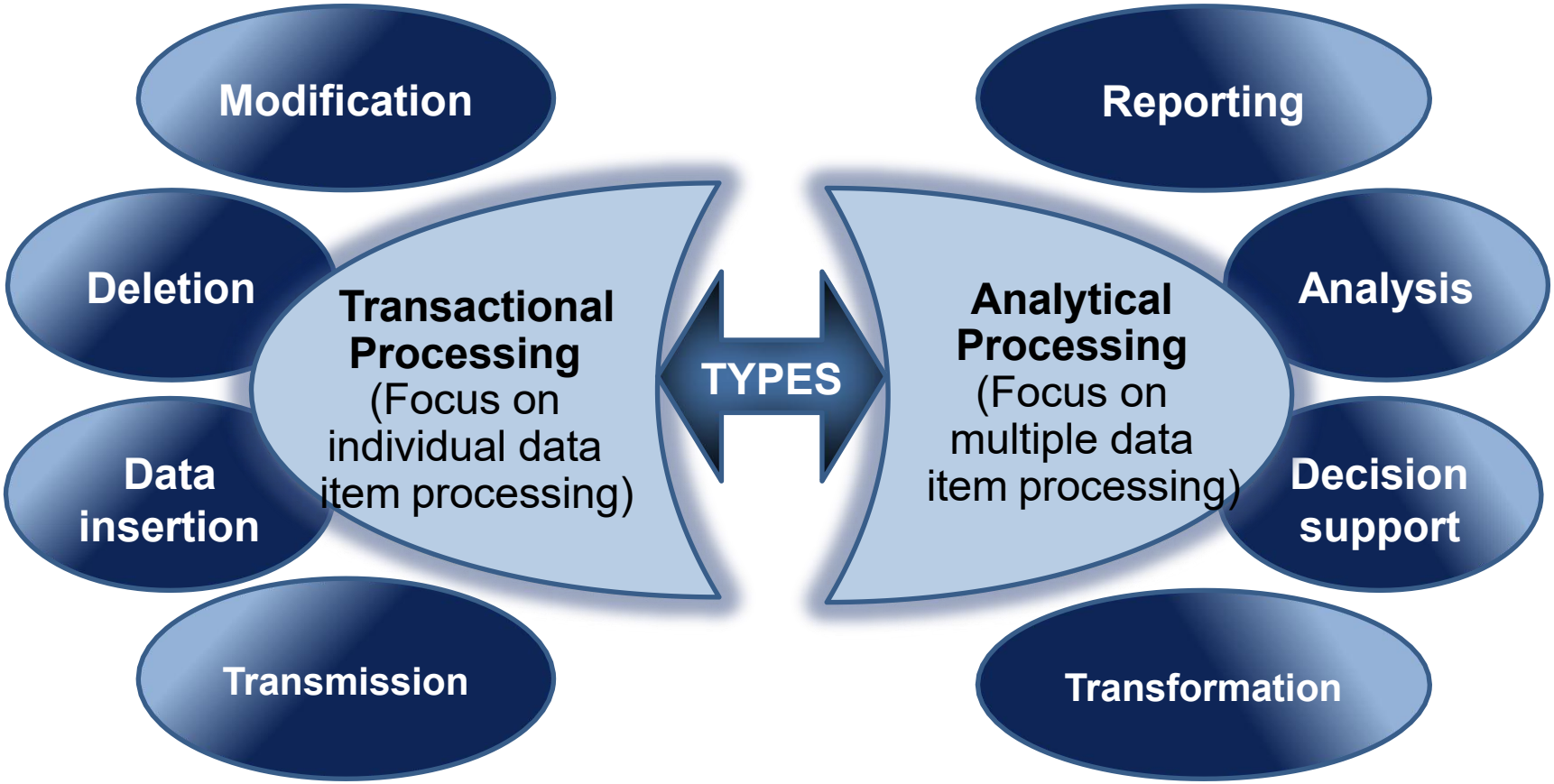
What is Business Intelligence (BI) ?

- Data, Information, Decision
- BI as a decision process
- BI as an information system



- Business Inteligence is a set of theories, methodologies, architectures, and technologies that ***transforms raw data into meaningful and useful information*** for business purposes.
- It provides reports but it can't predict future.
- Helps to answer the questions of current problem.
- Input- past data, Output- present solution.

Information Processing



In this sense, BI focuses on analytical data processing

What is Data Warehouse

- A data warehouse is a **central repository** of data management system that collects, manages data from various sources designed to enable and support business intelligence (BI) activities, especially analytics. It is used to help the organization in taking decisions.
  - A data warehouse centralizes and consolidates large amounts of data from multiple sources.
  - They store current and historical data in one single place.
  - Data in the data warehouse must have strong analytical characteristics.
  - Creating data to be analytical requires that it be **subject- oriented, integrated, time-referenced, and non-volatile**.
  - Support information processing by providing a solid platform
  - of consolidated, historical data for analysis.

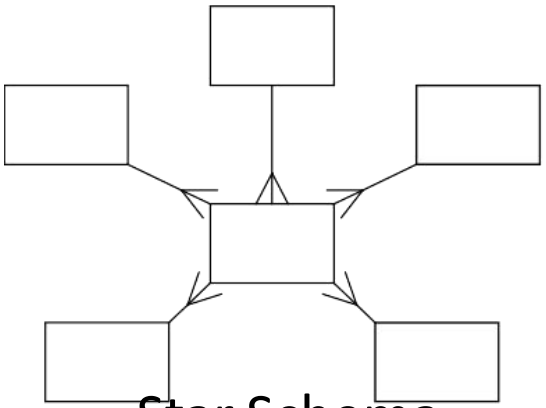
Data Warehouses Vs Operational Database Systems

- A special database system called data warehouse or data mart is often used to store enterprise data.
  - The purpose of a data warehouse is to organize lots of stable data for ease of analysis and retrieval.
  - It deals with analyzing data and making decisions, often major, about how the enterprise will operate now, and in the future.
- **OLAP: On Line Analytical Processing:** Describes processing at warehouse
- Traditional (operational) relational databases facilitate data management and transaction processing. They have two limitations for data analysis and decision support
  - Performance
    - ✓ They are transaction oriented (data insert, update, move, etc.)
    - ✓ Not optimized for complex data analysis
    - ✓ Usually do not hold historical data
  - Heterogeneity
    - ✓ Individual databases usually manage data in very different ways, even in the same organization (not to mention external data sources which may be dramatically different).
- **OLTP: On Line Transaction Processing:** Describes processing at operational

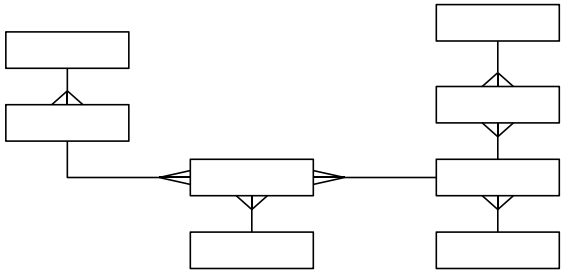
Data Warehouses Vs Operational Database Systems - Functional point of view

Key	Data warehouse	Operational Database
Basic	A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose	Operational Database are those databases where data changes frequently
Data Structure	Data warehouse has denormalized schema	It has normalized schema
Transaction Optimization	Optimized for bulk loads and large complex, unpredictable queries.	Optimized for a common and known set of transactions.
Performance	It is fast for analysis queries	It is slow for analytics queries
Type of Data	It focuses on historical data	It focuses on current transactional data
Uses Case	It is used for OLAP	It is used for OLTP
Data Updates	Batch updates	Continuous updates
Query Handling	Usually very complex queries	Simple to complex queries

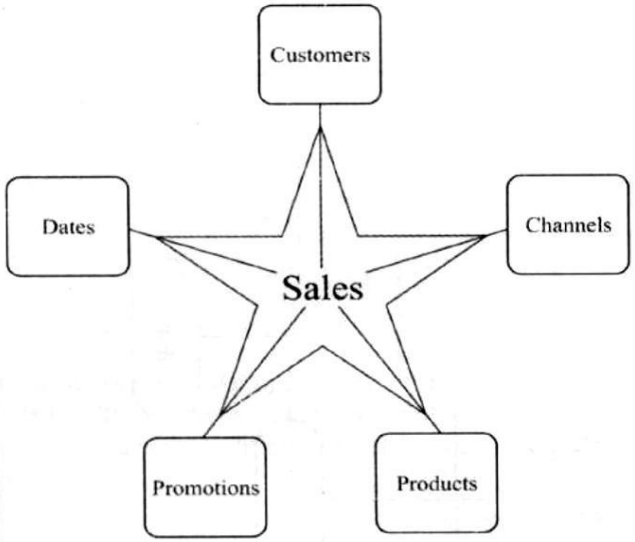
Design point of view

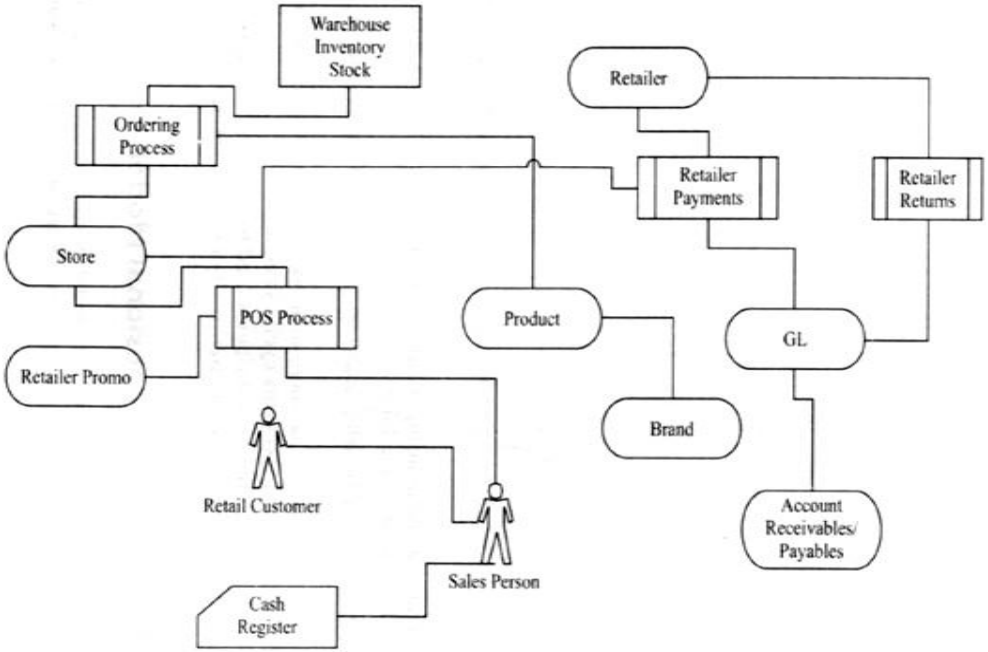


Star Schema



ER Diagram





Why Do We Need Data Warehouses?

- Consolidation of information resources
- Improved query performance
- Separate research and decision support functions from the operational systems
- Foundation for data mining, data visualization, advanced reporting and OLAP tools
- **Example of Data Warehousing:**
  - E-commerce website. They have a warehouse. Where the organization project the demand for the products, that were to procure from the supplier and store it in the warehouse. The good is dispatched by the supplier.
  - The customers directly interact with the e-commerce website for the product. The managers of these e-commerce business houses directly orders the supplier and get the product.
  - The products come to the business houses from different sources by maintaining different databases.
    - ✓ Grocery items (database1)
    - ✓ Fashion items like clothes, shoes, cosmetics (database2)
    - ✓ Computer, laptops, mobile and electronic goods (database3)

Data Warehouses Example...

- Now , lets consider 100 suppliers will contact the manufacturer for a specific product which is demanded by the consumers, imagine the strain that is in manufacturer to supply those products.
- From the customers prospective, there is a delay in getting the product which you will be never acceptable.
- If you place an order in an e-commerce website, you must except the product should be deliver in 24-48 hrs. If it is delayed by 2-3 days it impact the customers satisfaction levels. So the business house may lose the customers.

Applications of Data Warehousing

Sector	Usage
Airline	Helps in airline system management operations like crew assignment, analyzes of route, frequent flyer program discount schemes for passenger, etc.
Banking	It is used in the banking sector to manage the resources available on the desk effectively.
Healthcare sector	Used to strategize and predict outcomes, create patient's treatment reports, etc. Advanced machine learning, big data enable datawarehouse systems can predict ailments.
Insurance sector	Used to analyze data patterns, customer trends, and to track market movements quickly.
Retail chain	Helps you to track items, identify the buying pattern of the customer, promotions and also used for determining pricing policy.
Telecommunication	Used for product promotions, sales decisions and to make distribution decisions.

Database vs Data Warehousing

	Database	Data Warehouse
Purpose	Is designed to record	Is designed to analyze
Processing Method	The database uses the Online Transactional Processing (OLTP)	Data warehouse uses Online Analytical Processing (OLAP).
Usage	The database helps to perform fundamental operations for your business	Data warehouse allows you to analyze your business.
Tables and Joins	Tables and joins of a database are complex as they are normalized.	Table and joins are simple in a data warehouse because they are denormalized.
Orientation	Is an application-oriented collection of data	It is a subject-oriented collection of data
Storage limit	Generally limited to a single application	Stores data from any number of applications
Availability	Data is available real-time	Data is refreshed from source systems as and when needed
Usage	ER modeling techniques are used for designing.	Data modeling techniques are used for designing.
Technique	Capture data	Analyze data
Data Type	Data stored in the Database is up to date.	Current and Historical Data is stored in Data Warehouse. May not be up to date.
Storage of data	Flat Relational Approach method is used for data storage.	Data Ware House uses dimensional and normalized approach for the data structure. Example: Star and snowflake schema.
Query Type	Simple transaction queries are used.	Complex queries are used for analysis purpose.
Data Summary	Detailed Data is stored in a database.	It stores highly summarized data.

Comparison of OLTP and OLAP Systems

Feature	OLTP	OLAP
Characteristics	Operational Processing	Informational processing
Orientation	Transaction	Analysis
User	Clerk,DBA, Database professional	Knowledge worker (manager, Executive, Analyst)
Function	day-to-day operation	long-term informational requirements decision support
DB design	ER-based, application oriented	star/snowflake, subject oriented
Data	current, guaranteed up to date	historic, accuracy maintained over time
Summarization	Primitive, highly detailed	summarized, consolidated
view	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read and write	mostly read
Focus	data in	information out
Operations	index / hash on primary key	lots of scans
# of records access	tens	millions
Number of users	thousands	hundreds
DB size	GB to High order GB	>= TB
Piority	High performance and high-availability	High flexibility, end-user autonoy
Metric	Transaction throughput	query throughput, response time

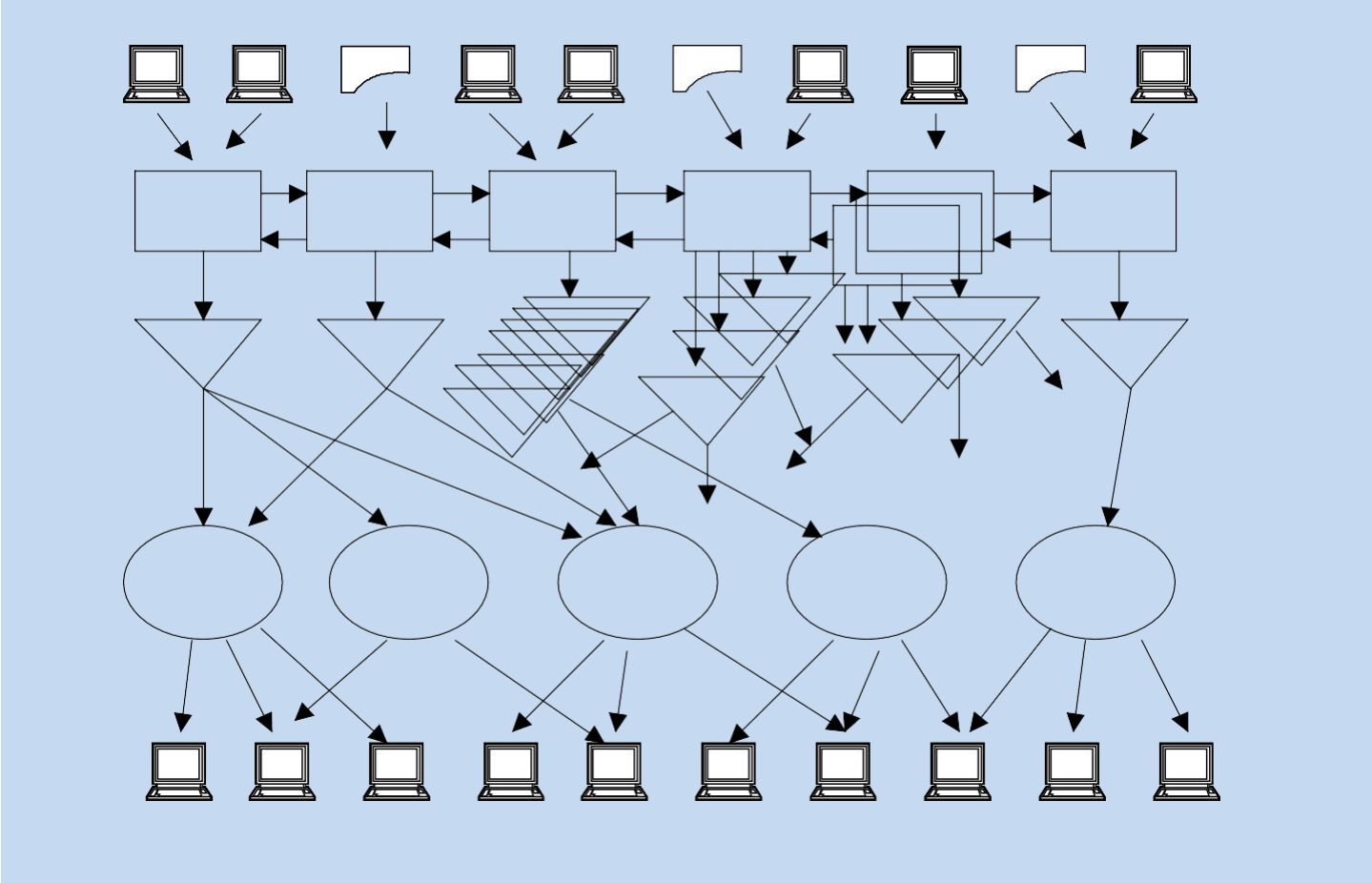
Examples of OLTP and OLAP

- OLTP
  - A railway reservation server which records the transaction of the passengers.
  - A supermarket server which records the every product
  - purchased from the market.
  - A bank server which records every time the transactions made by any account
- OLAP
  - An insurance company wants to know the number of policies each agent has sold.
  - A bank Manager wants know how many customers are utilizing the ATM of that branch.

Examples of OLTP and OLAP...

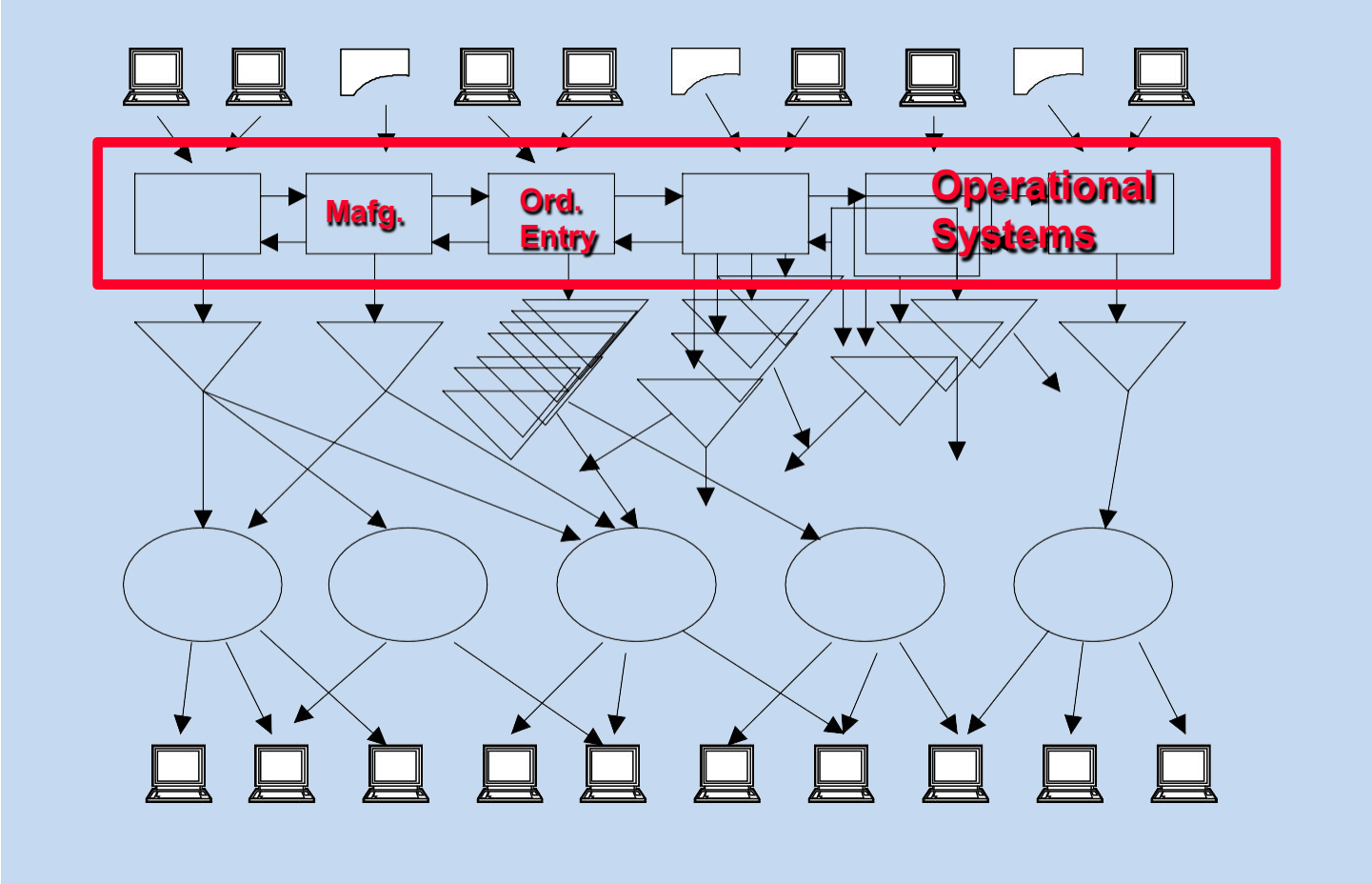
- OLTP- To record each and every transactions. Take a real life example of ATM. Every transaction is recorded in a OLTP system. ATM is not the only feeder to this OLTP system. When you will go to the bank and perform some transaction that is also recorded. So there are multiple sources to feed to this OLTP. The disadvantages are:
  - To have a query, one has to combine this multiples sources which
  - are different formatting of their own.
  - The number of transactions. (100s of customers, 1000s of queries)
  - So this only to record the transactions.
  - Railway Reservation System:
    - ✓ For reservation, one can opt internet, mobile, railway station.
    - ✓ Can go to 'n' number of agents spread across the town. These are multiple desperate sources. When u will use internet to book the ticket the format is totally different from u use a mobile.
    - ✓ The data types are changing. So this multiple desperate sources is difficult for queering. It is an hindrance to Analytical Processing. So for reporting at the end of the day we need a OLAP system. It is an alternate system.

Operational vs. Informational Systems



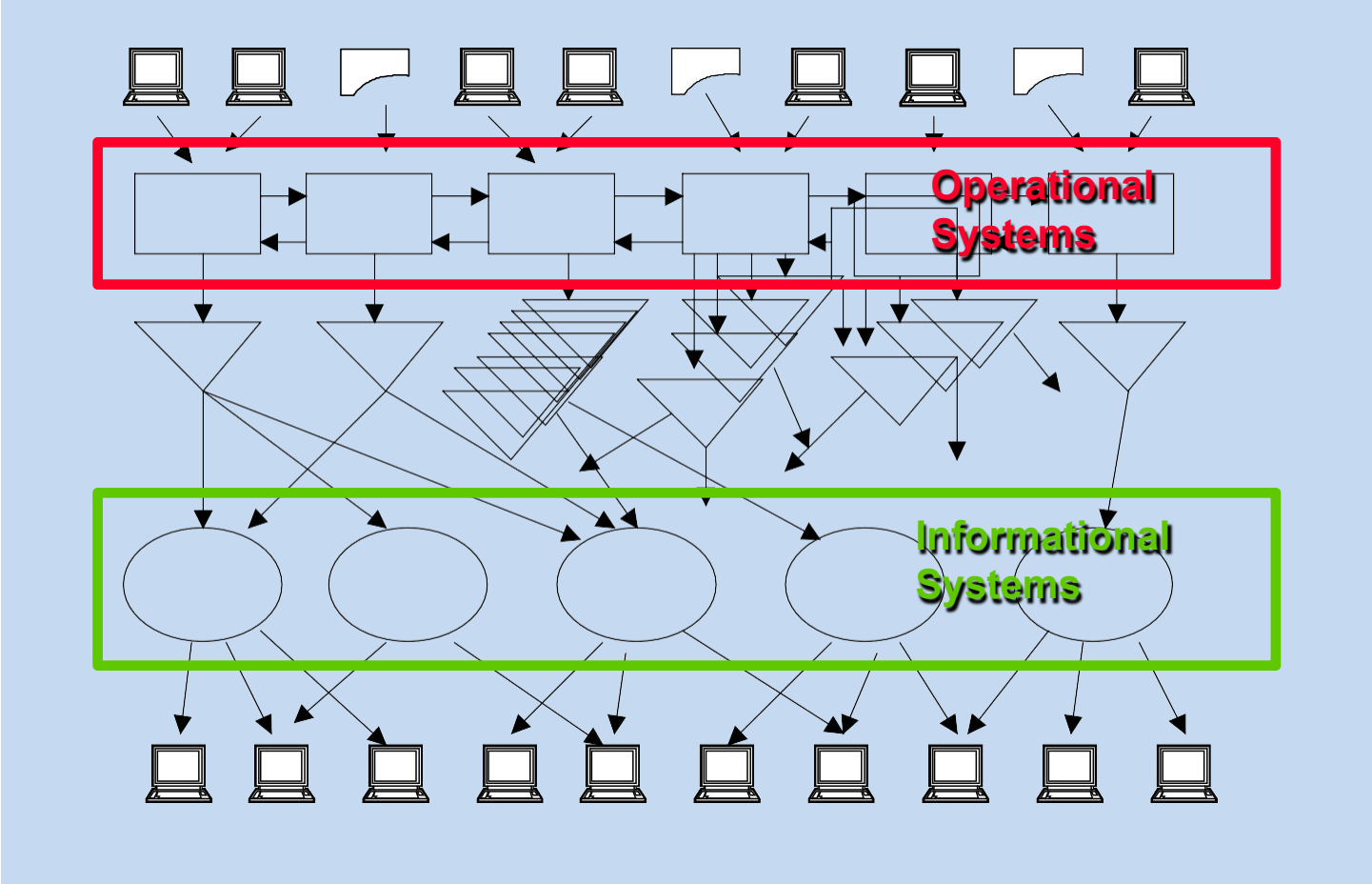
Information Access Today

Operational vs. Informational Systems



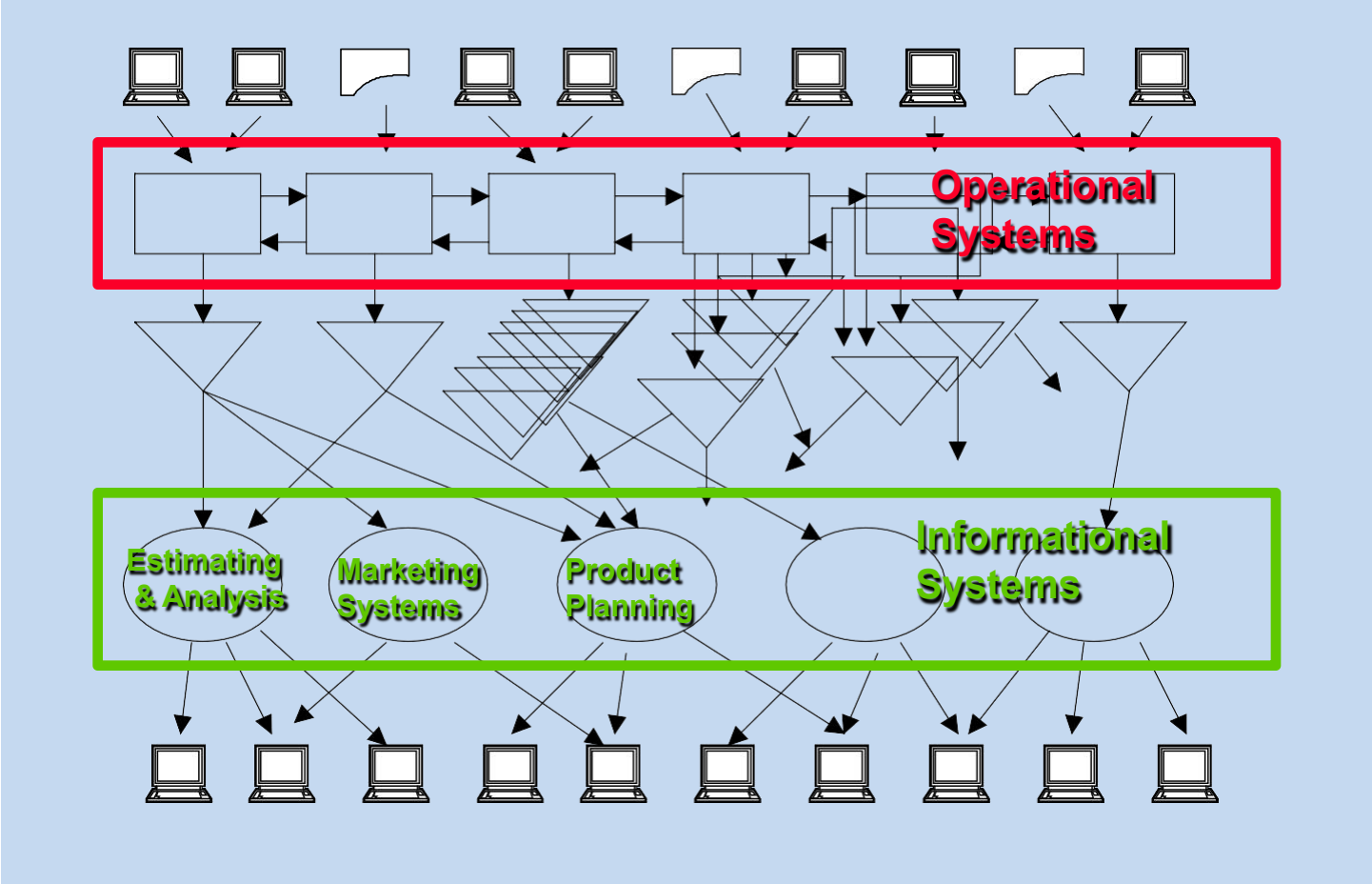
Information Access Today

Operational vs. Informational Systems



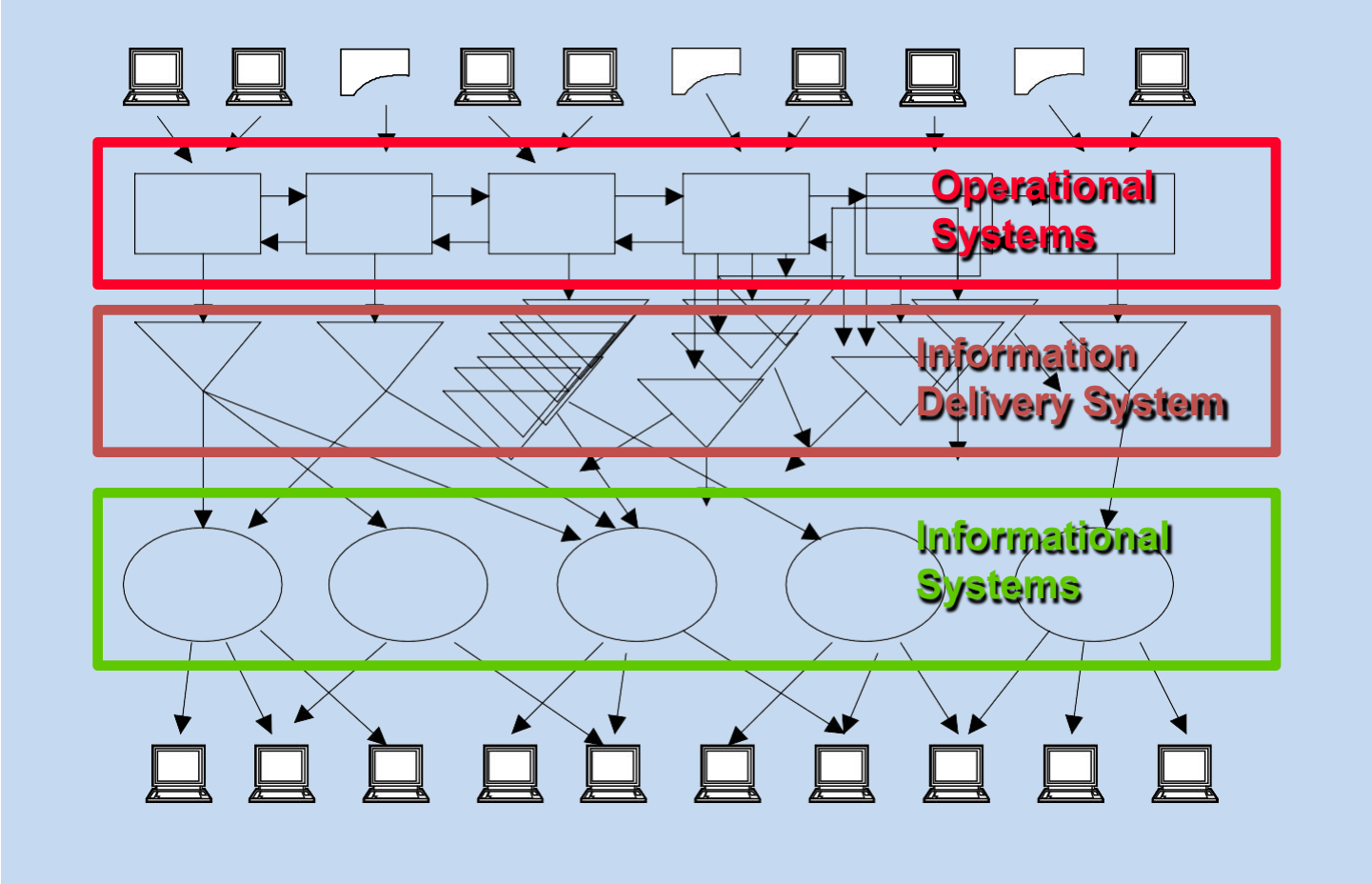
Information Access Today

Operational vs. Informational Systems



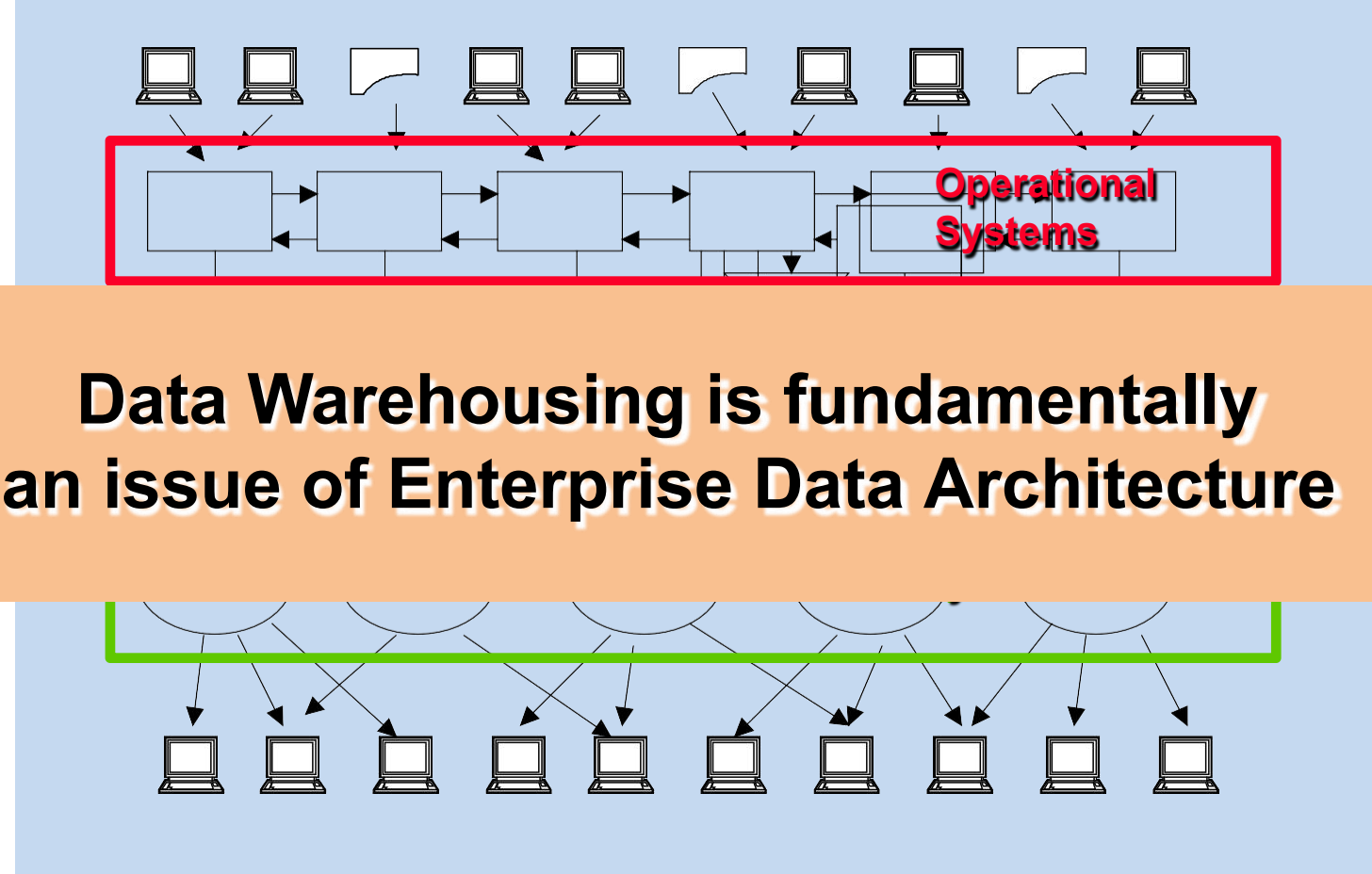
Information Access Today

Operational vs. Informational Systems



Information Access Today

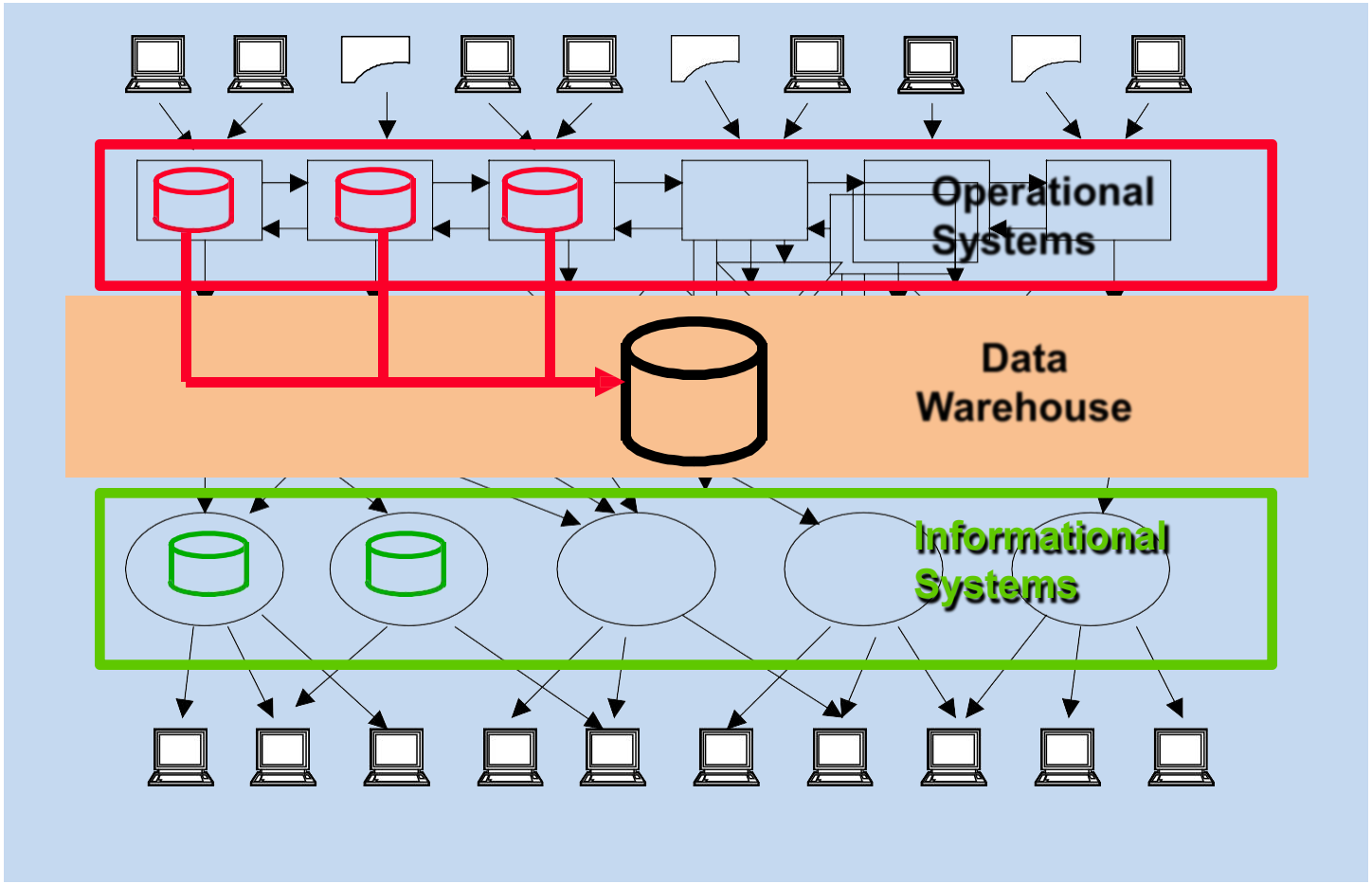
Operational vs. Informational Systems



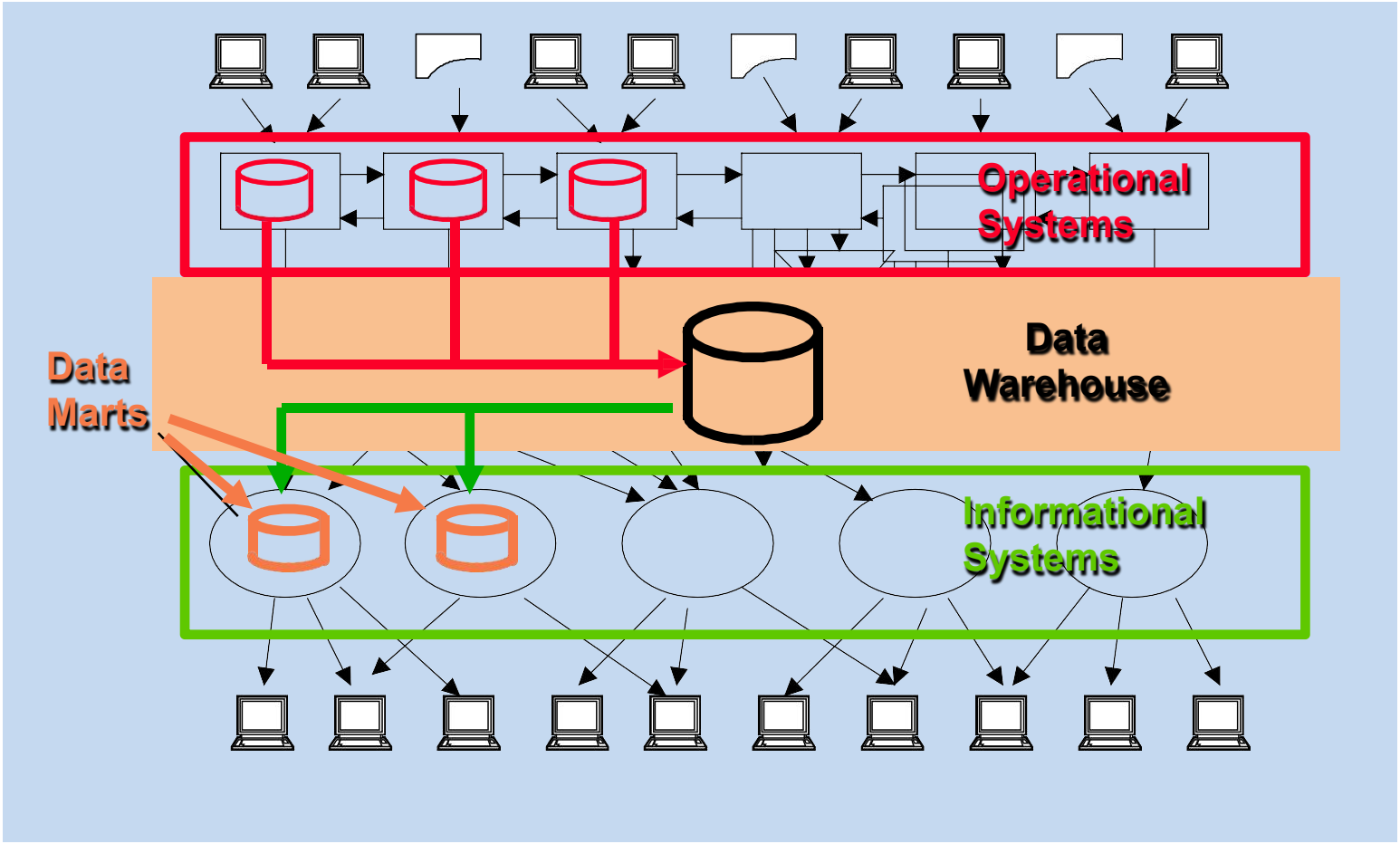
Data Warehousing is fundamentally an issue of Enterprise Data Architecture

Information Access Today

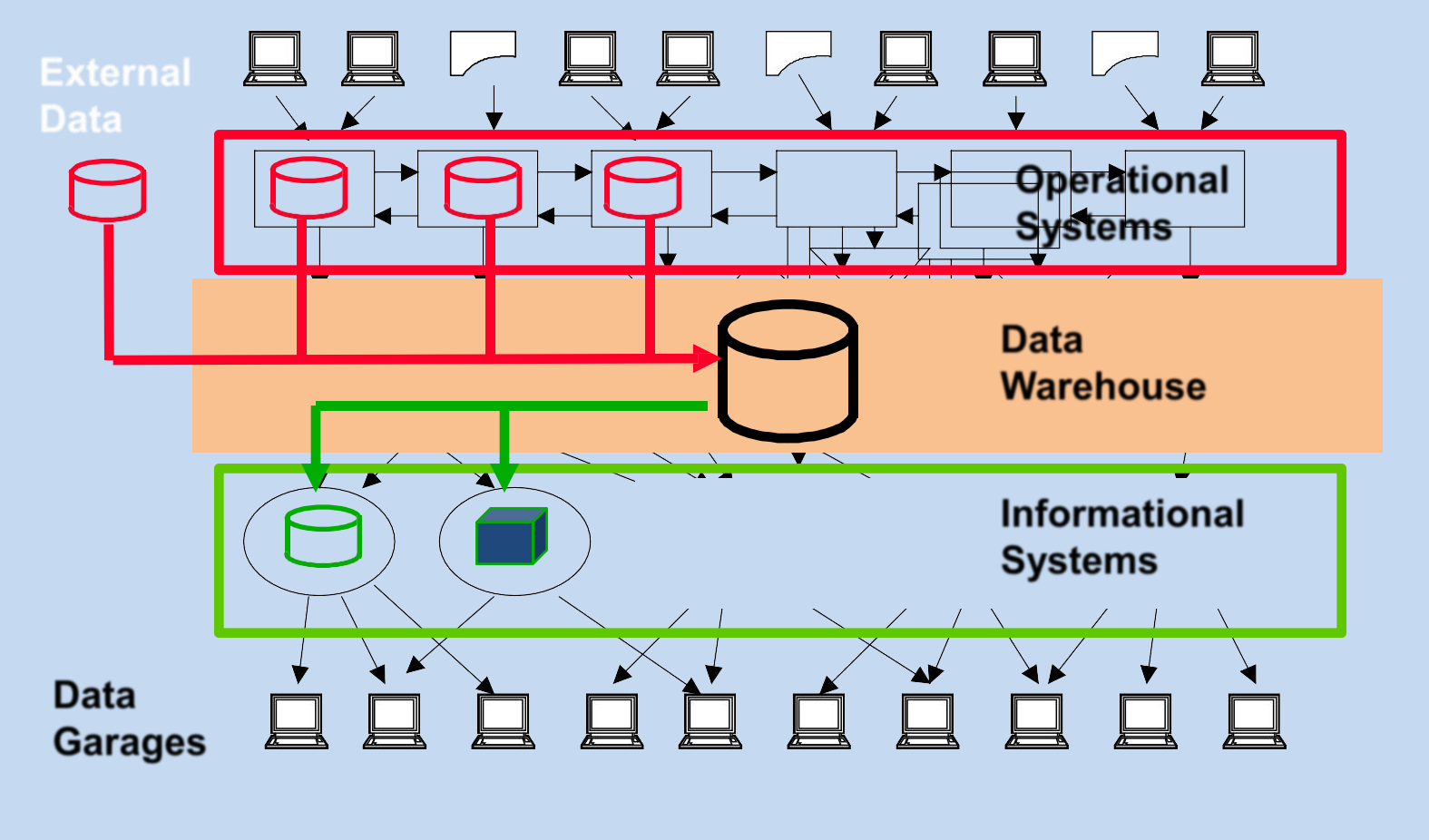
Operational vs. Informational Systems



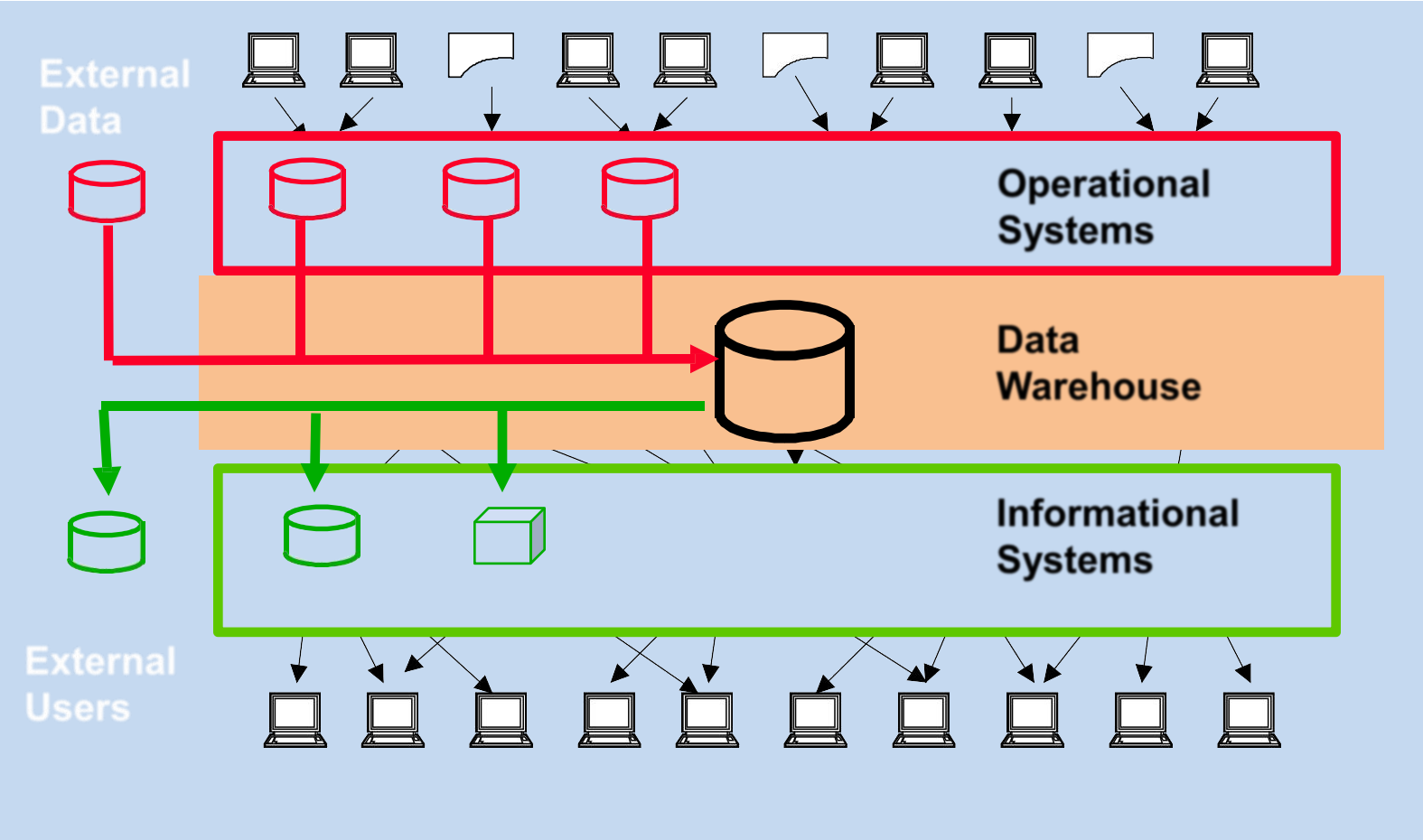
Operational vs. Informational Systems



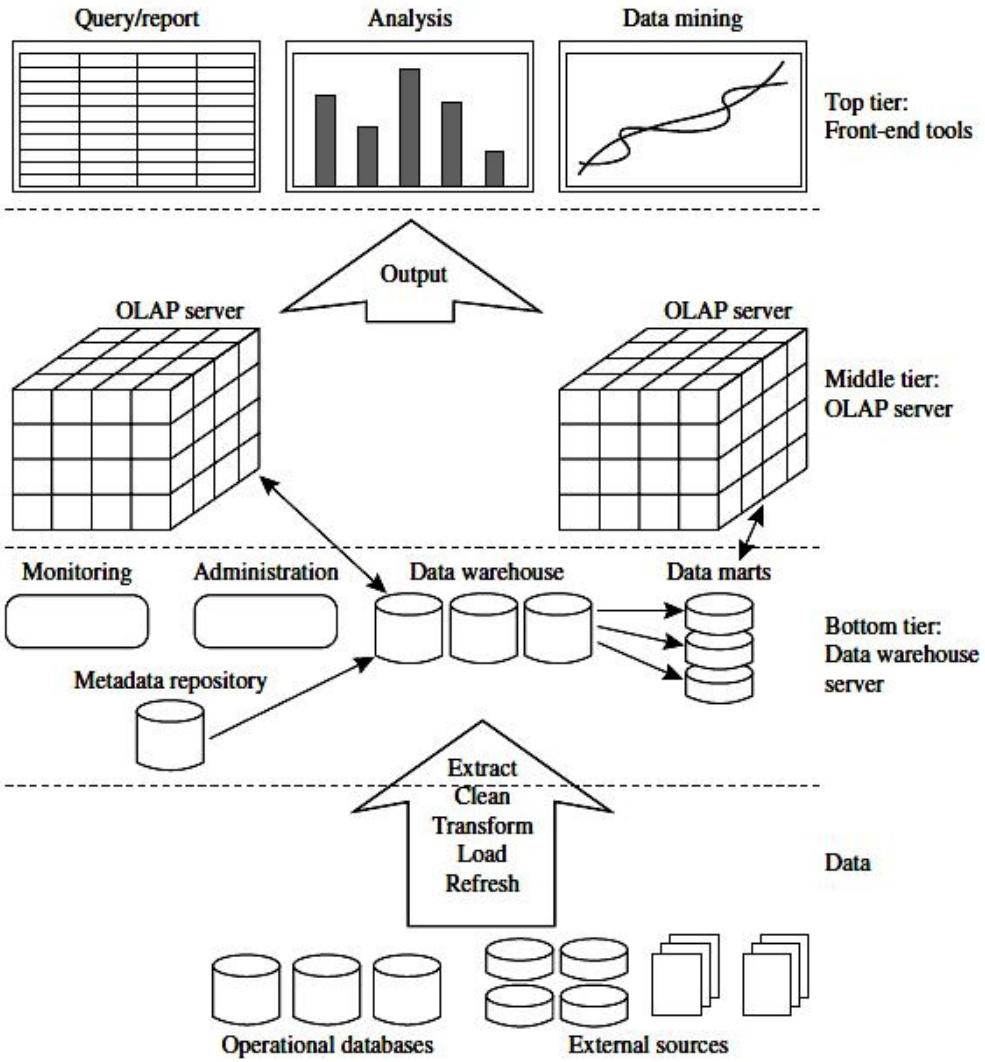
Operational vs. Informational Systems



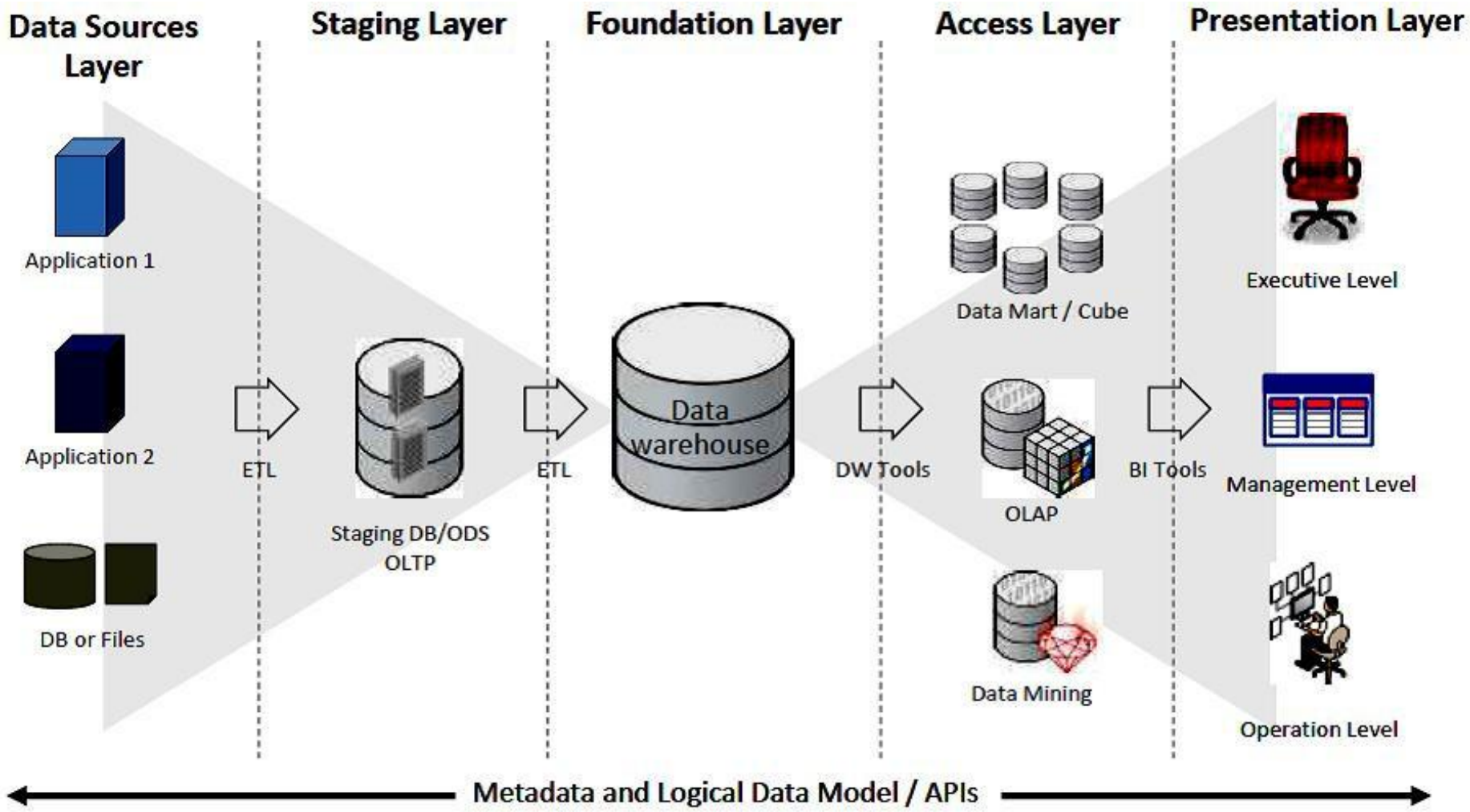
Operational vs. Informational Systems



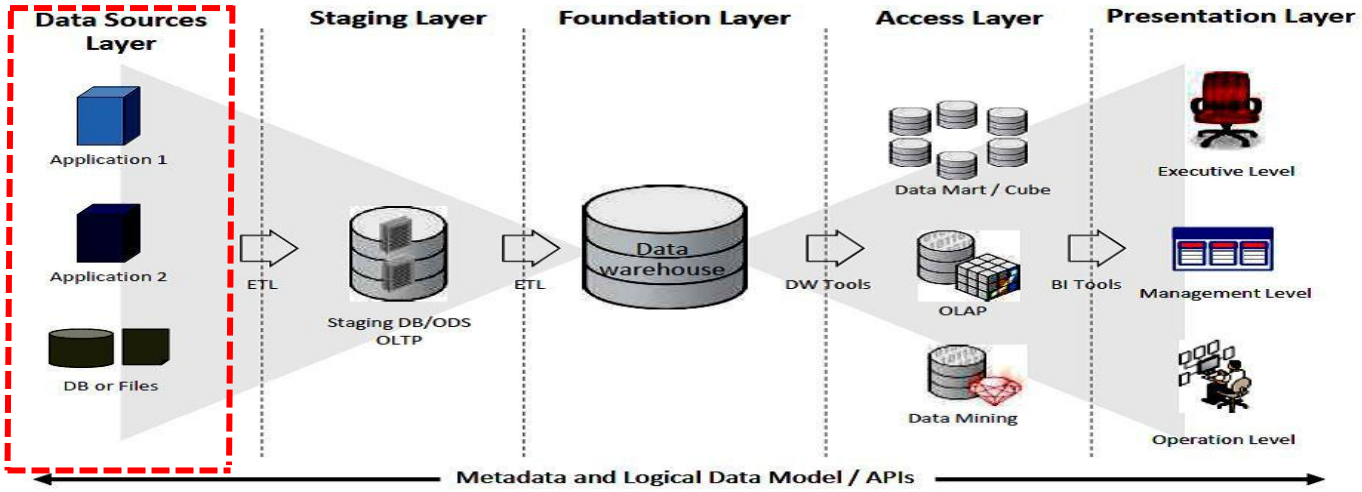
Three-tier Data Warehousing Architecture



Traditional Data Warehousing and Business Intelligence

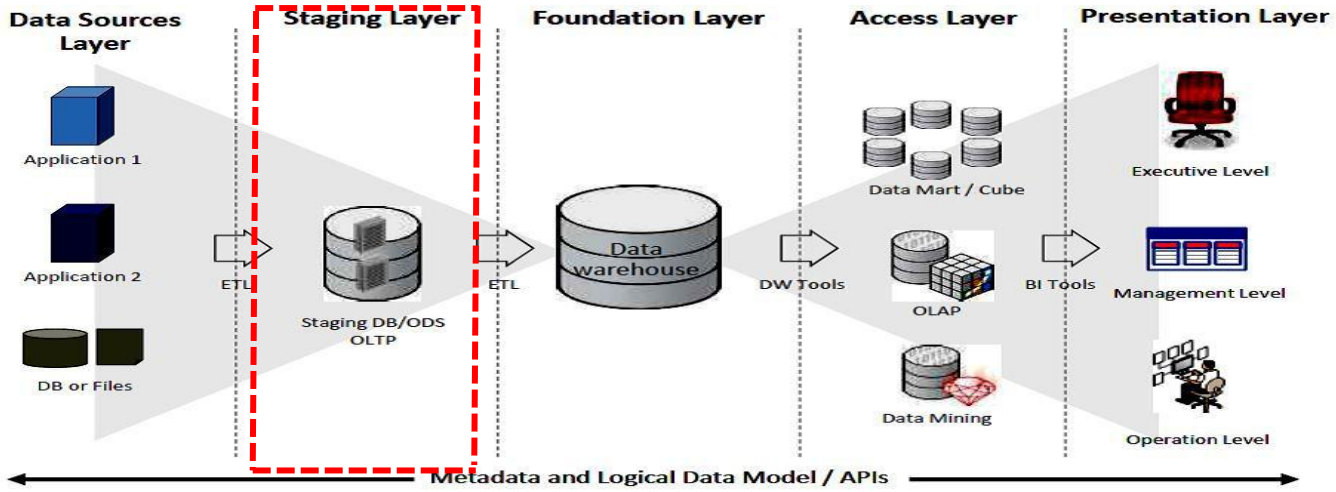


Traditional Data Warehousing and Business Intelligence



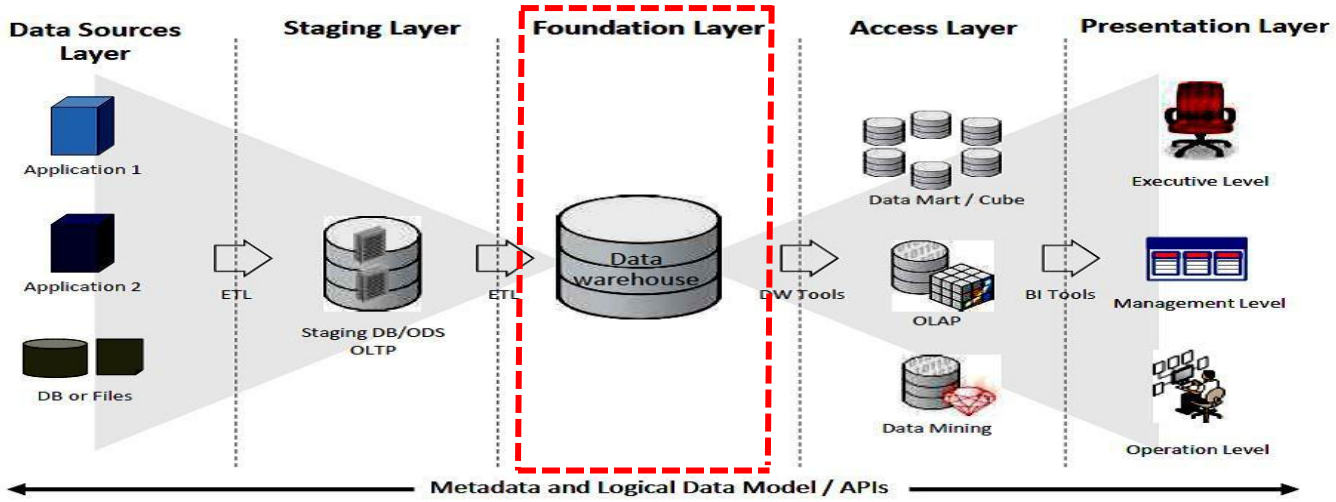
- **Data Source Layer:** defining which data will be loaded into the system and analyzed.
  - Text Files
  - OLTP, Databases
  - XML
  - JSON
  - Spreadsheet Files
- Source Data Examples
  - Retail POS system
  - Web Site
  - DBMS

Traditional Data Warehousing and Business Intelligence



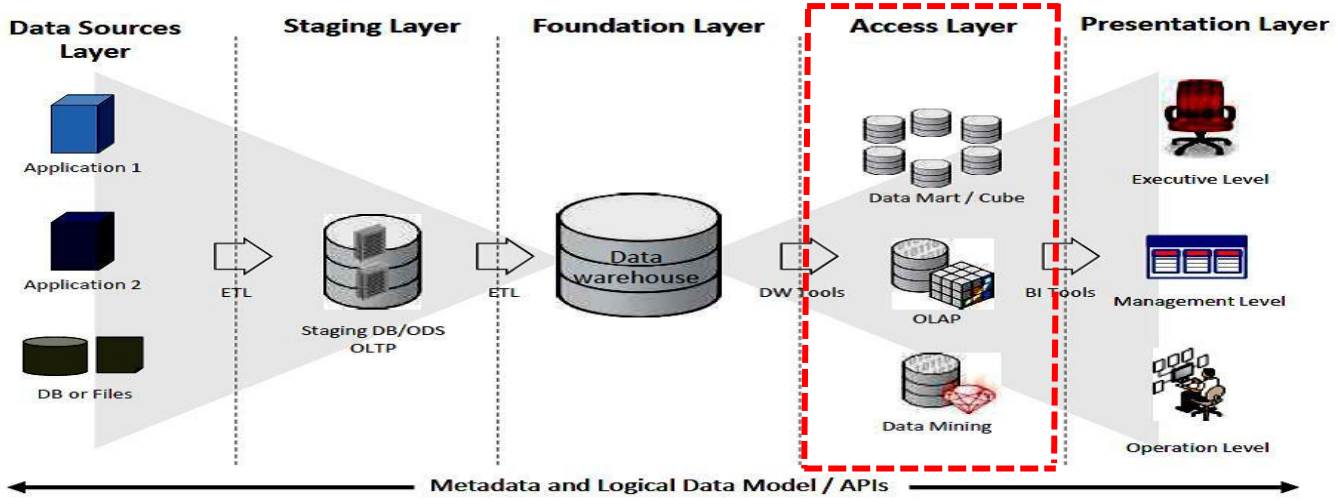
- **ETL (Extract, Transform, and Load)** and ***Staging Layer***:
  - **Extraction :** accessing and extracting the data from the source systems, including database, flat files, spreadsheets, etc.
  - **Transformation :** data cleanse, change the extracted data to a format and structure that conform to the destination data.
  - **Loading :** load the data to the destination database, and check for data integrity
    - Tools to move data to staging DB
    - Staging DB is a temporary storage to be loaded to DWH
    - Staging DB could be operational reporting tool/platform

Traditional Data Warehousing and Business Intelligence



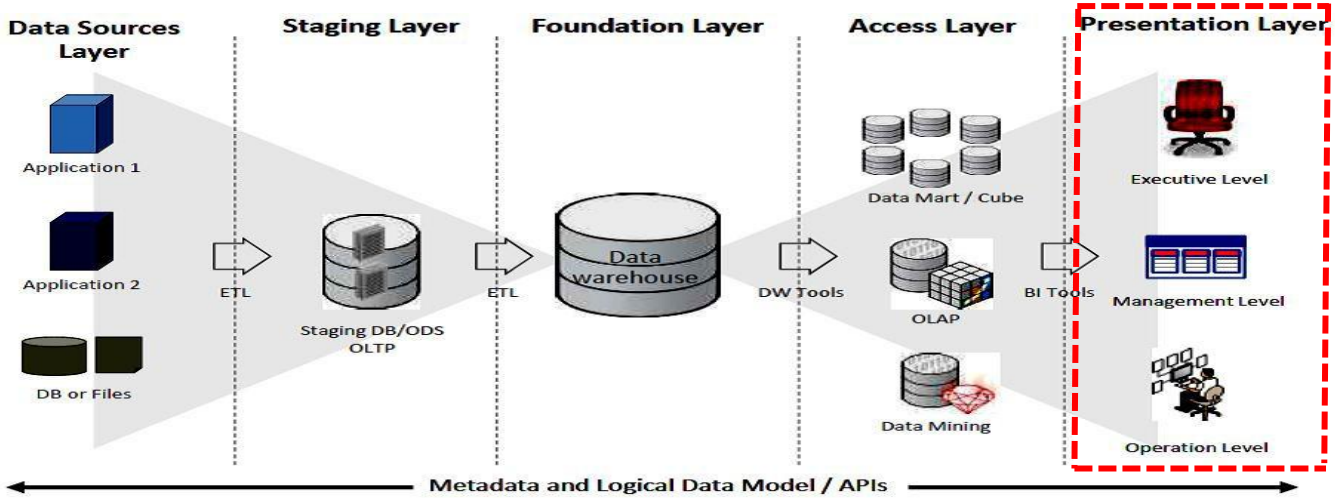
- **Data Warehouse:-**
  - Used for reporting
  - A scalable DB storing historical enterprise data
  - Online Analytical processing
  - Not used for transaction processing

Traditional Data Warehousing and Business Intelligence



- **Access Layer:-**
  - Data Mart for business fast query (Star Schema)
  - OLAP uses a multidimensional data model, allowing for complex
  - analytical and adhoc queries with a rapid execution time
  - Data mining for mostly in structured data format

Traditional Data Warehousing and Business Intelligence



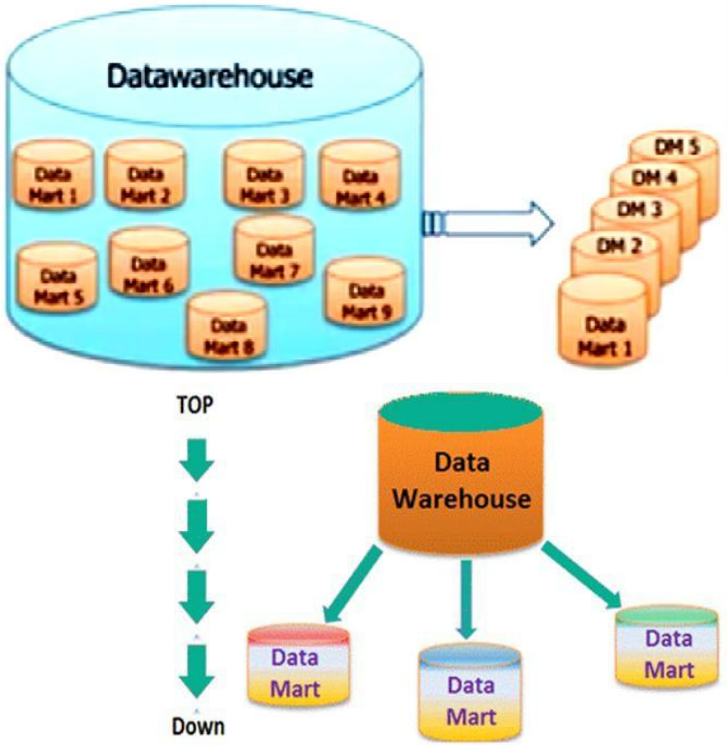
- **Presentation Layer:-**
  - Need to gather requirements from Business Units for Visualization and Touch points
  - Need to identify data sources and method to deliver results
  - Enterprise dashboards, reports and alerts that present findings from the
  - analysis

Data Warehouse Models

- Three Data Warehouse Models
  - **Enterprise warehouse**
    - ✓ collects all of the information about subjects spanning the entire organization
  - **Data Mart**
    - ✓ a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - ✓ Independent vs. dependent (directly from warehouse) data mart
  - **Virtual warehouse**
    - ✓ A set of views over operational databases
    - ✓ Only some of the possible summary views may be
    - ✓ materialized

Data Marts

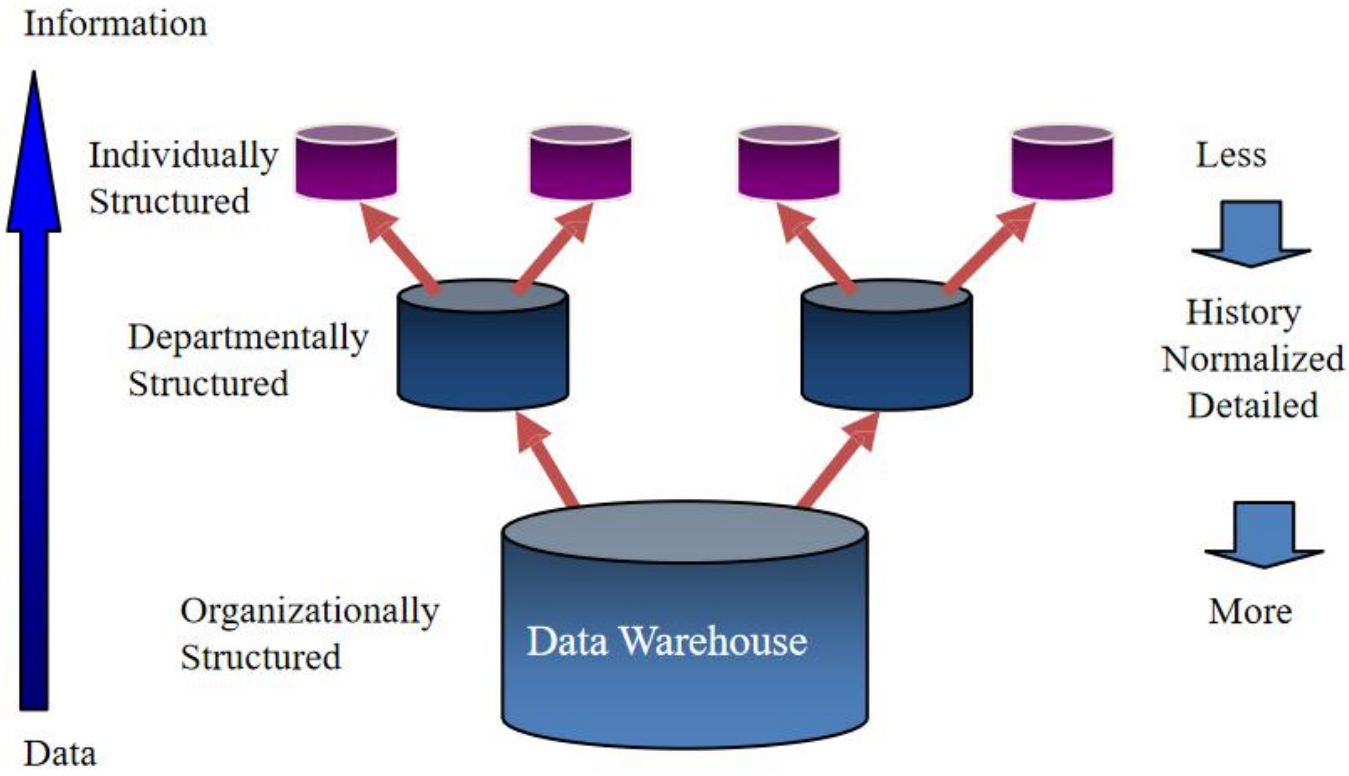
- Data mart is a smaller version of the Datawarehouse.
- Data mart deal with a signle subject.
- Data marts are focused on one area and they draw data from a limited number of sources.
- Time taken to build the data is very low compared to the time taken to build a Datawarehouse.
- Data Mart helps to enhance user's response time due to reduction in volume of data
- It provides easy access to frequently requested data.
- **Type of Data Mart**



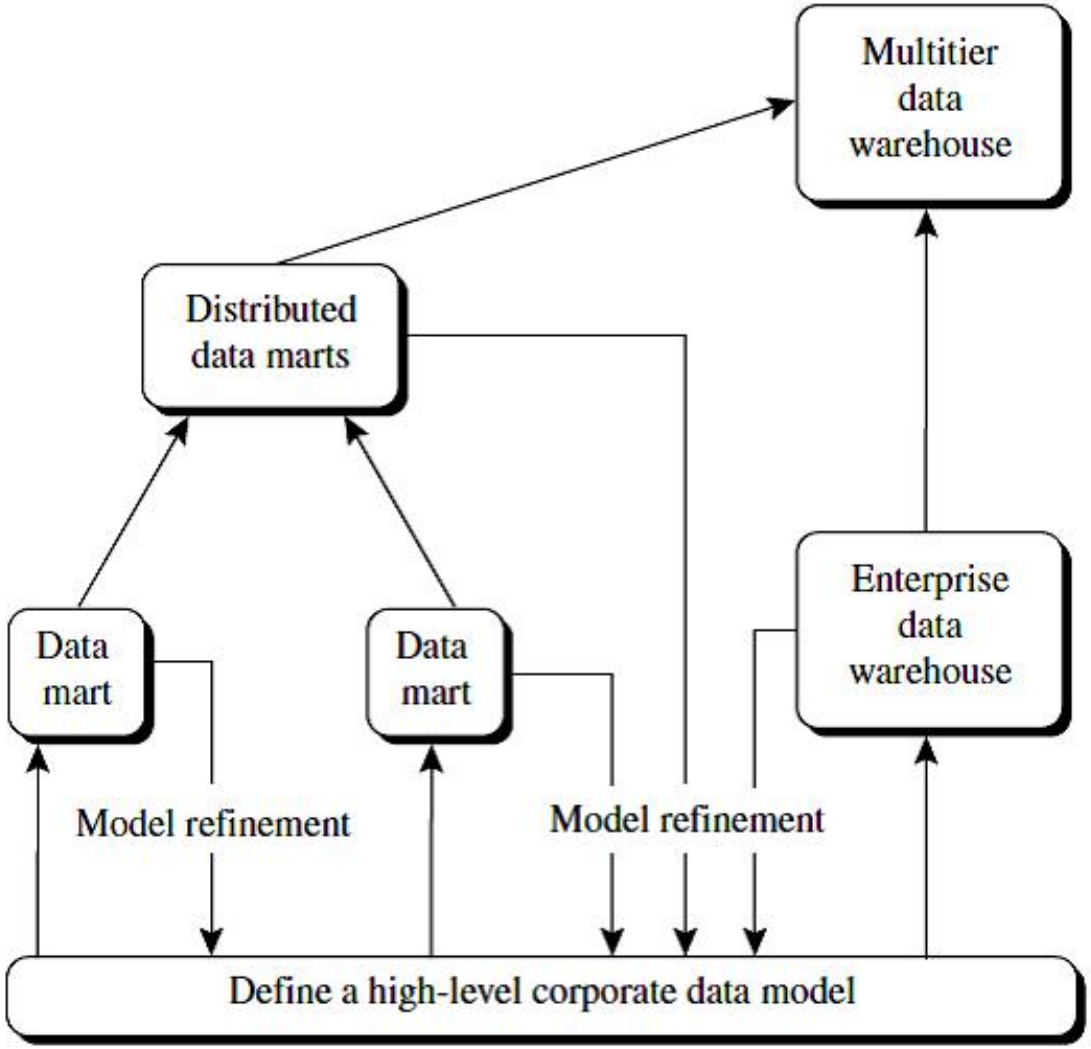
  - ***Dependent:*** Dependent data marts are created by drawing data directly from operational, external or both sources.
  - ***Independent:*** Independent data mart is created without the use of a central data warehouse.
  - ***Hybrid:*** This type of data marts can take data from data warehouses or operational systems.

From the Data Warehouse to Data Marts

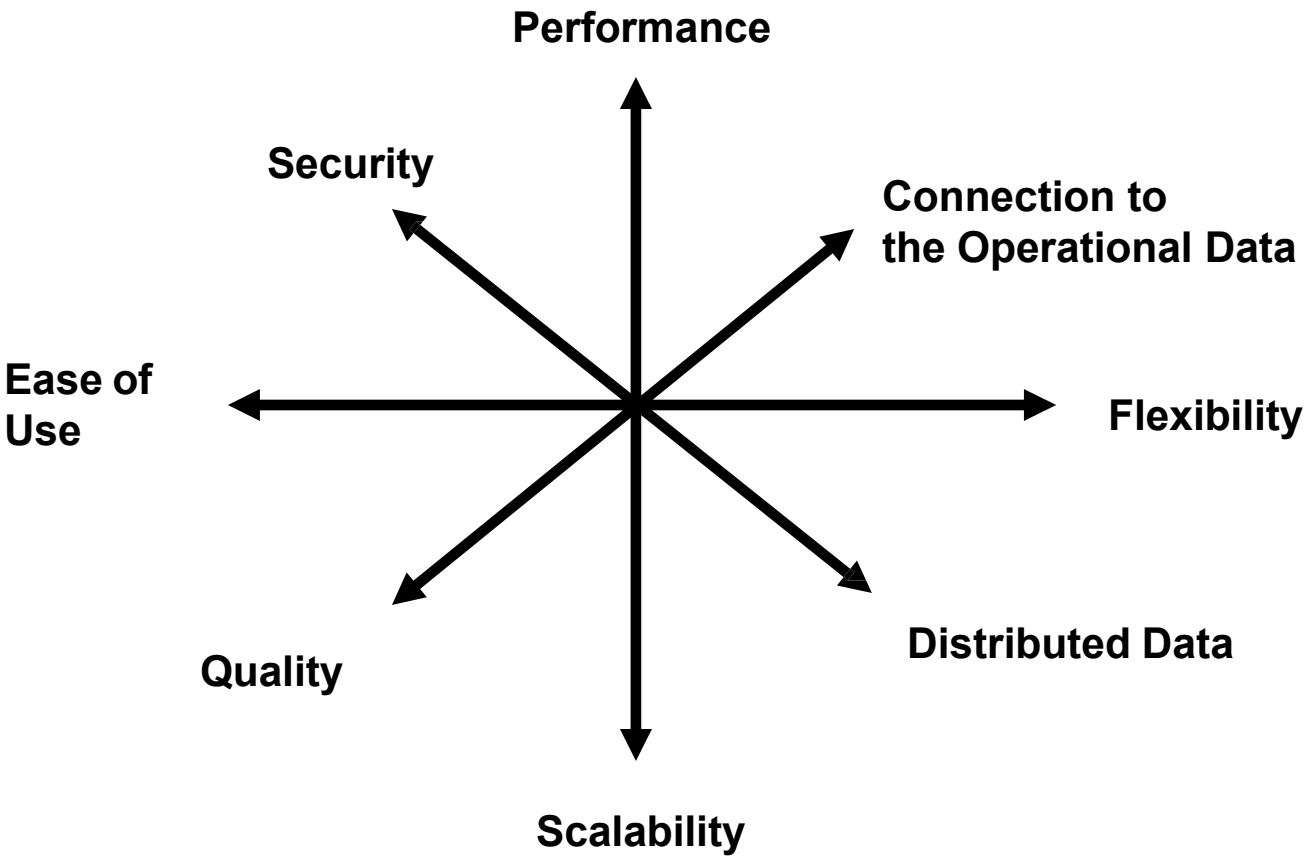
- **Data Mart** – A logical subset of the complete data warehouse. Often viewed as a restriction of the data warehouse to a single business process or to a group of related business processes targeted toward a particular business group.



A recommended approach for data warehouse development.

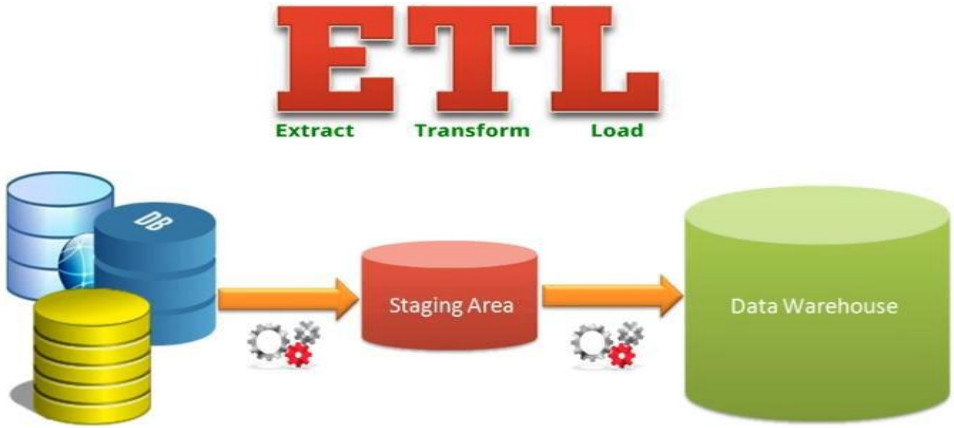


Dimensions of Data Warehousing



Extraction, Transformation, and Loading (ETL)

- Data extraction
  - get data from multiple, heterogeneous, and external sources
- Data cleaning
  - detect errors in the data and rectify them when possible
- Data transformation
  - convert data from legacy or host format to warehouse format
- Load
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse



Recommended Text and Reference Books

- **Text Book:**
  - J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- **Reference Books:**
  - H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2006.
  - I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2000.
  - D. Hand, H. Mannila and P. Smyth. Principles of Data Mining.Prentice-Hall. 2001.

THANK YOU!