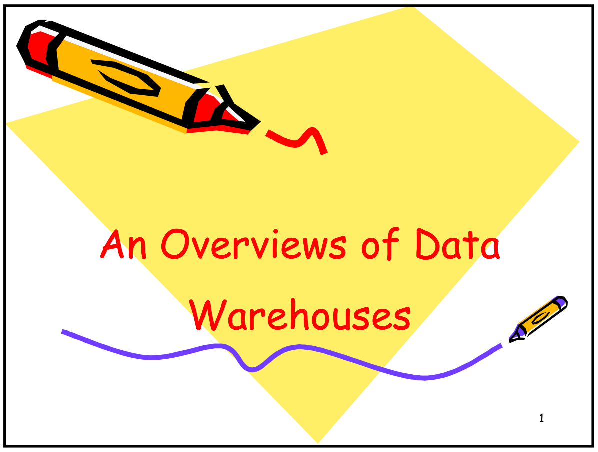
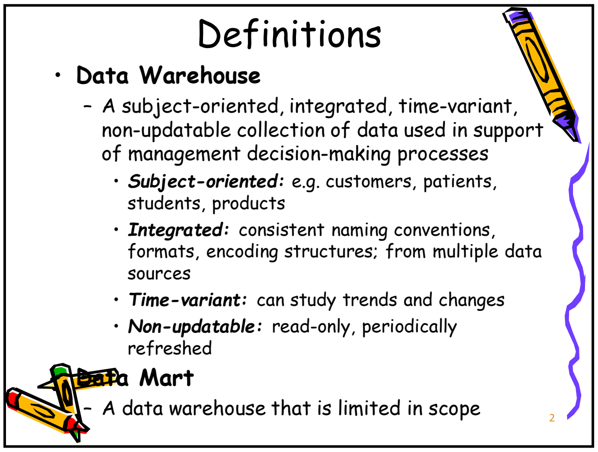
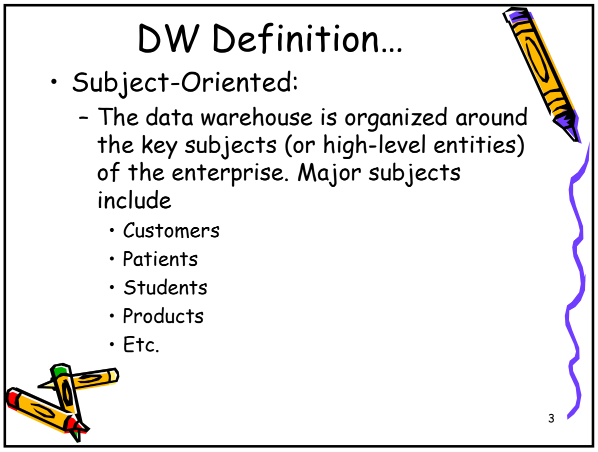


An Overviews of Data Warehouses



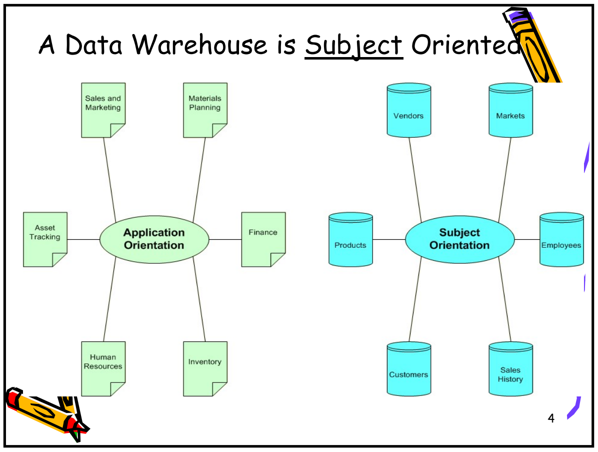
Definitions

- **Data Warehouse**
  - A subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making processes
    - ***Subject-oriented:*** e.g. customers, patients, students, products
    - ***Integrated:*** consistent naming conventions, formats, encoding structures; from multiple data sources
    - ***Time-variant:*** can study trends and changes
    - ***Non-updatable:*** read-only, periodically refreshed
- **Data Mart**
  - A data warehouse that is limited in scope

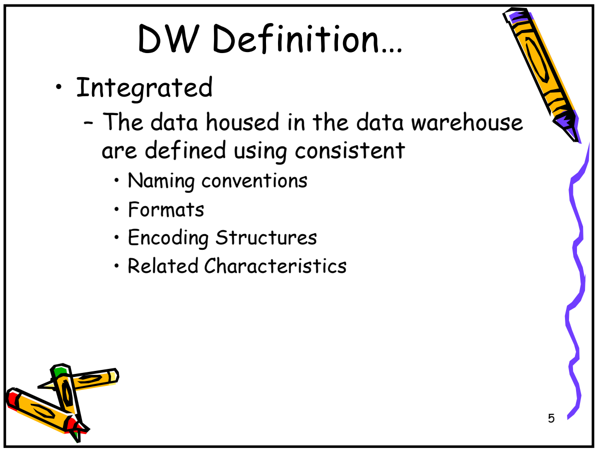


DW Definition…

- Subject-Oriented:
  - The data warehouse is organized around the key subjects (or high-level entities) of the enterprise. Major subjects include
    - Customers
    - Patients
    - Students
    - Products
    - Etc.

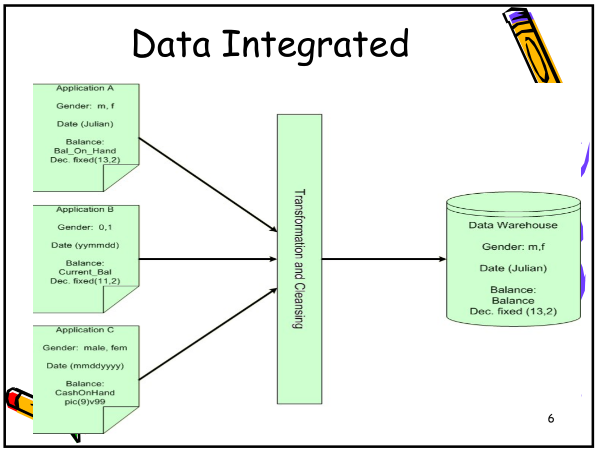


A Data Warehouse is <u>Subject</u> Oriented



DW Definition…

- Integrated
  - The data housed in the data warehouse are defined using consistent
    - Naming conventions
    - Formats
    - Encoding Structures
    - Related Characteristics



Data Integrated

Application A
Gender: m, f
Date (Julian)
Balance: Bal_On_Hand Dec. fixed(13,2)
Application B
Gender: 0,1
Date (yymmdd)
Balance: Current_Bal Dec. fixed(11,2)
Application C
Gender: male, fem
Date (mmddyyyy)
Balance: CashOnHand pic(9)v99
Transformation and Cleansing
Data Warehouse
Gender: m,f
Date (Julian)
Balance: Balance Dec. fixed (13,2)

DW Definition…

- Time-variant
  - The data in the warehouse contain a time dimension so that they may be used as a historical record of the business

Time Variant

- In an operational application system, the expectation is that all data within the database are accurate as of the moment of access. In the DW data are simply assumed to be accurate as of some moment in time and not necessarily right now.
- One of the places where DW data display time variance is in the structure of the record key. Every primary key contained within the DW must contain, either implicitly or explicitly an element of time(day, week, month, etc)

Time Variant

- Every piece of data contained within the warehouse must be associated with a particular point in time if any useful analysis is to be conducted with it.
- Another aspect of time variance in DW data is that, once recorded, data within the warehouse cannot be updated or changed.

DW Definition…

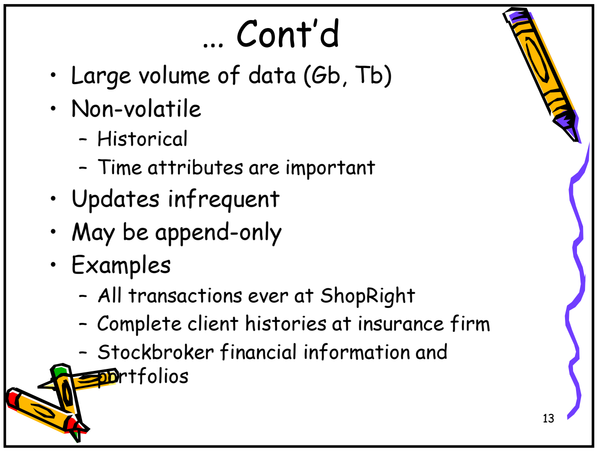
- Non-volatile
  - Data in the data warehouse are loaded and refreshed from operational systems, but cannot be updated by end-users

Nonvolatility

- Typical activities such as deletes, inserts, and changes that are performed in an operational application environment are completely nonexistent in a DW environment.
- Only two data operations are ever performed in the DW: data loading and data access

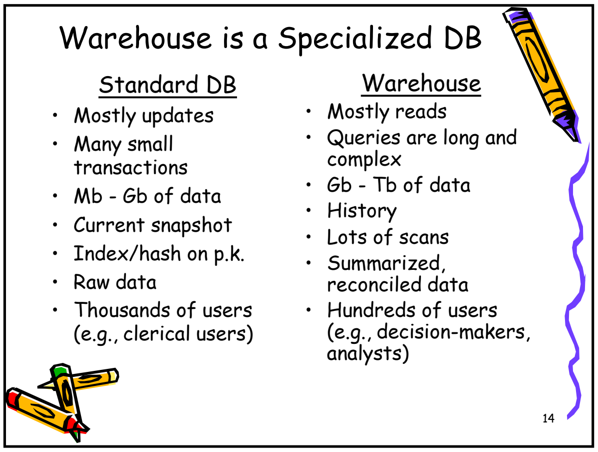
A Data Warehouse is...

- Stored collection of diverse data
  - A solution to data integration problem
  - Single repository of information
- Subject-oriented
  - Organized by subject, not by application
  - Used for analysis, data mining, etc.
- Optimized differently from transaction-oriented db
- User interface aimed at executive decision makers and analysts



… Cont'd

- Large volume of data (Gb, Tb)
- Non-volatile
  - Historical
  - Time attributes are important
- Updates infrequent
- May be append-only
- Examples
  - All transactions ever at ShopRight
  - Complete client histories at insurance firm
  - Stockbroker financial information and portfolios



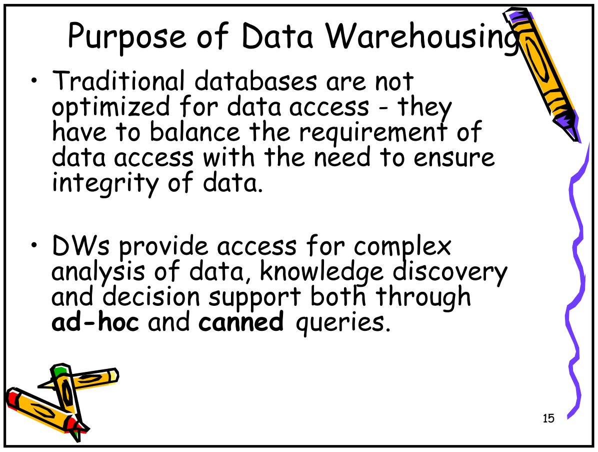
Warehouse is a Specialized DB

Standard DB

- Mostly updates
- Many small transactions
- Mb - Gb of data
- Current snapshot
- Index/hash on p.k.
- Raw data
- Thousands of users (e.g., clerical users)

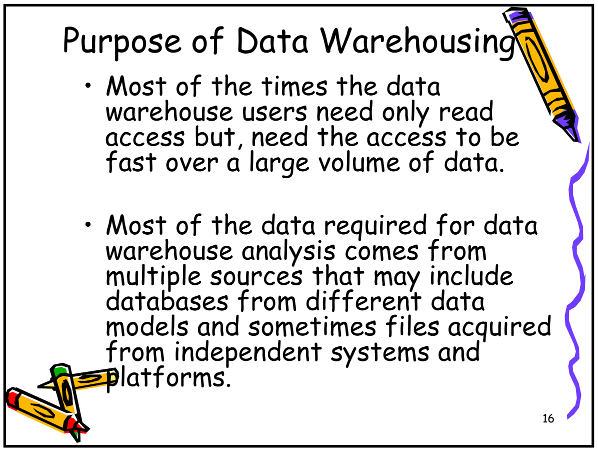
Warehouse

- Mostly reads
- Queries are long and complex
- Gb - Tb of data
- History
- Lots of scans
- Summarized, reconciled data
- Hundreds of users (e.g., decision-makers, analysts)



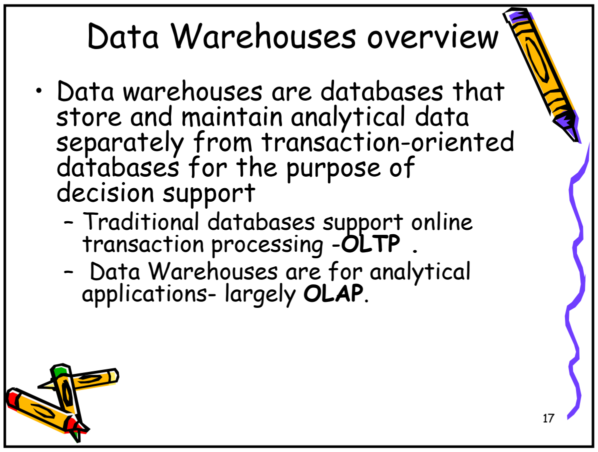
Purpose of Data Warehousing

- Traditional databases are not optimized for data access - they have to balance the requirement of data access with the need to ensure integrity of data.
- DWs provide access for complex analysis of data, knowledge discovery and decision support both through **ad-hoc** and **canned** queries.



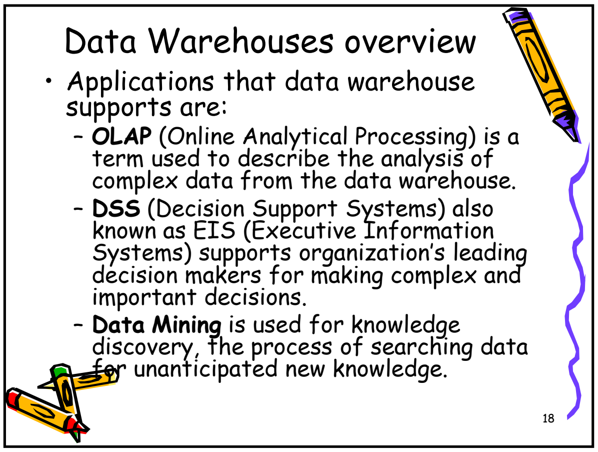
Purpose of Data Warehousing

- Most of the times the data warehouse users need only read access but, need the access to be fast over a large volume of data.
- Most of the data required for data warehouse analysis comes from multiple sources that may include databases from different data models and sometimes files acquired from independent systems and platforms.



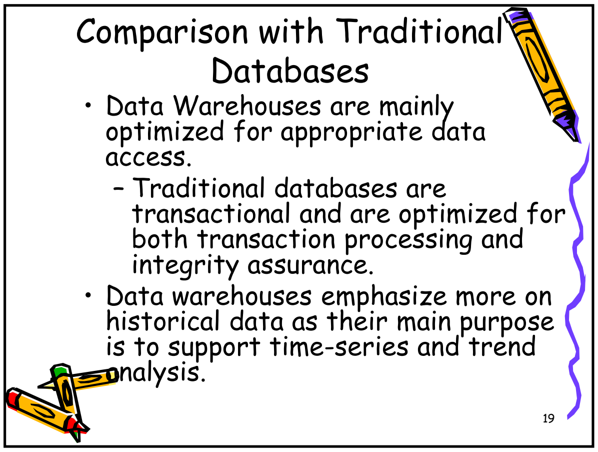
Data Warehouses overview

- Data warehouses are databases that store and maintain analytical data separately from transaction-oriented databases for the purpose of decision support
  - Traditional databases support online transaction processing -**OLTP** .
  - Data Warehouses are for analytical applications- largely **OLAP**.



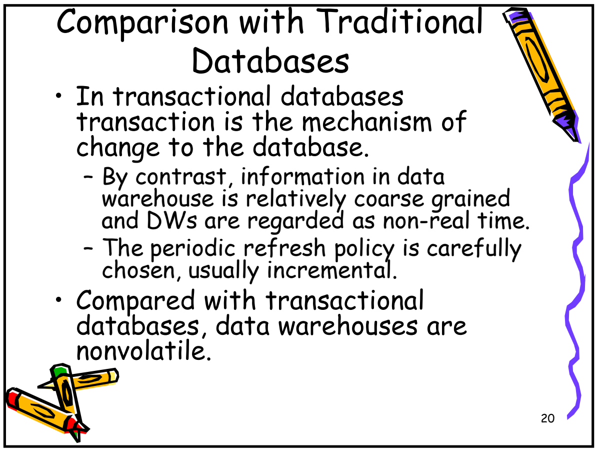
Data Warehouses overview

- Applications that data warehouse supports are:
  - **OLAP** (Online Analytical Processing) is a term used to describe the analysis of complex data from the data warehouse.
  - **DSS** (Decision Support Systems) also known as EIS (Executive Information Systems) supports organization's leading decision makers for making complex and important decisions.
  - **Data Mining** is used for knowledge discovery, the process of searching data for unanticipated new knowledge.



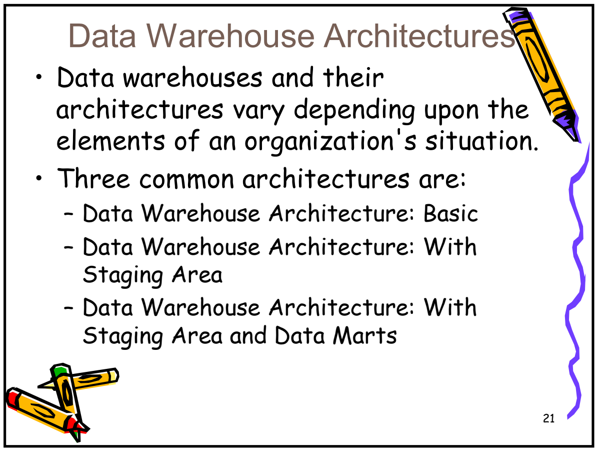
Comparison with Traditional Databases

- Data Warehouses are mainly optimized for appropriate data access.
  - Traditional databases are transactional and are optimized for both transaction processing and integrity assurance.
- Data warehouses emphasize more on historical data as their main purpose is to support time-series and trend analysis.



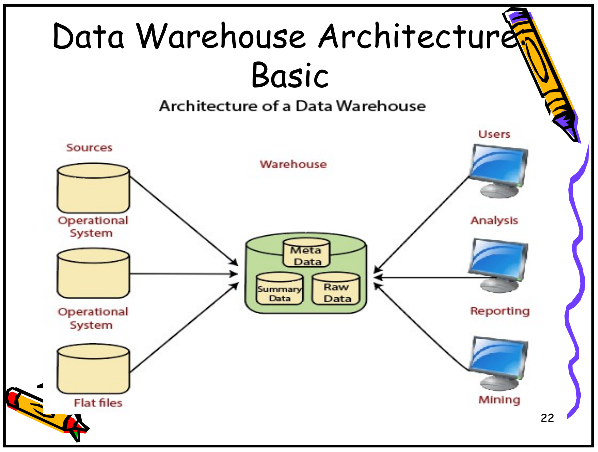
Comparison with Traditional Databases

- In transactional databases transaction is the mechanism of change to the database.
  - By contrast, information in data warehouse is relatively coarse grained and DWs are regarded as non-real time.
  - The periodic refresh policy is carefully chosen, usually incremental.
- Compared with transactional databases, data warehouses are nonvolatile.



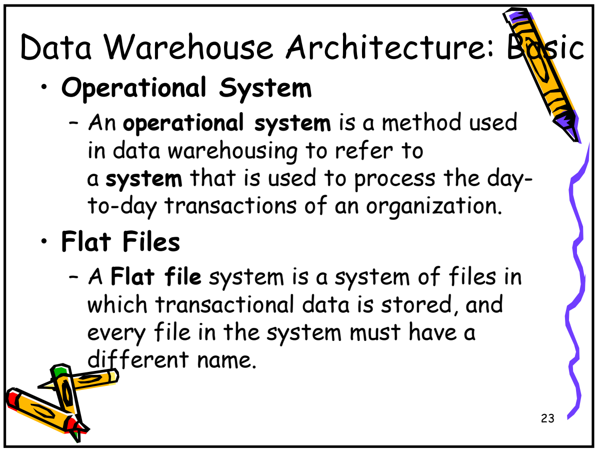
Data Warehouse Architectures

- Data warehouses and their architectures vary depending upon the elements of an organization's situation.
- Three common architectures are:
  - Data Warehouse Architecture: Basic
  - Data Warehouse Architecture: With Staging Area
  - Data Warehouse Architecture: With Staging Area and Data Marts



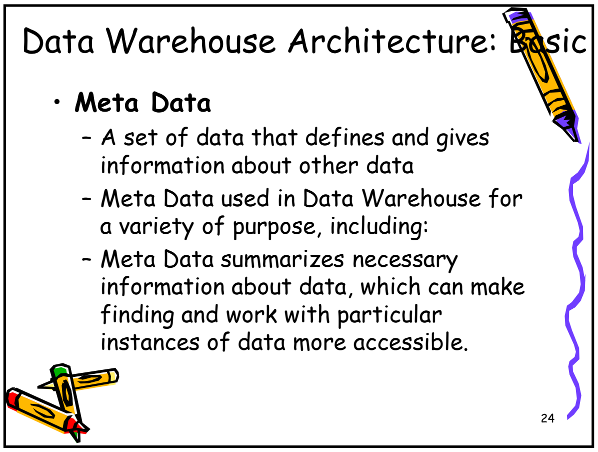
Data Warehouse Architecture: Basic

Architecture of a Data Warehouse



Data Warehouse Architecture: Basic

- **Operational System**
  - An **operational system** is a method used in data warehousing to refer to a **system** that is used to process the day-to-day transactions of an organization.
- **Flat Files**
  - A **Flat file** system is a system of files in which transactional data is stored, and every file in the system must have a different name.



Data Warehouse Architecture: Basic

- **Meta Data**
  - A set of data that defines and gives information about other data
  - Meta Data used in Data Warehouse for a variety of purpose, including:
  - Meta Data summarizes necessary information about data, which can make finding and work with particular instances of data more accessible.

Data Warehouse Architecture: Basic

- For example, author, data build, and data changed, and file size are examples of very basic document metadata.
- Metadata is used to direct a query to the most appropriate data source.

- **Lightly and highly summarized data**
  - The area of the data warehouse saves all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.

Data Warehouse Architecture: Basic

- The goals of the summarized information are to speed up query performance.
- The summarized record is updated continuously as new information is loaded into the warehouse.

- **End-User access Tools**
  - The principal purpose of a data warehouse is to provide information to the business managers for strategic decision-making. These customers interact with the warehouse using end-client access tools.

Data Warehouse Architecture: Basic

- The examples of some of the end-user access tools can be:
  - Reporting and Query Tools
  - Application Development Tools
  - Executive Information Systems Tools
  - Online Analytical Processing Tools
  - Data Mining Tools

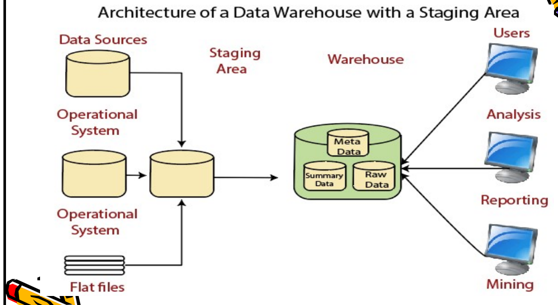
Data Warehouse Architecture: With Staging Area

- We must clean and process the operational information before putting it into the warehouse.
- We can do this programmatically, although data warehouses uses a **staging area** (A place where data is processed before entering the warehouse).

Data Warehouse Architecture: With Staging Area

- A staging area simplifies data cleansing and consolidation for operational method coming from multiple source systems, especially for enterprise data warehouses where all relevant data of an enterprise is consolidated.

Data Warehouse Architecture: With Staging Area



Data Warehouse Architecture: With Staging Area and Data Marts

- We may want to customize our warehouse's architecture for multiple groups within our organization.
- We can do this by adding **data marts**.
  - A data mart is a segment of a data warehouses that can provided information for reporting and analysis on a section, unit, department or operation in the company, e.g., sales, payroll, production, etc.

31

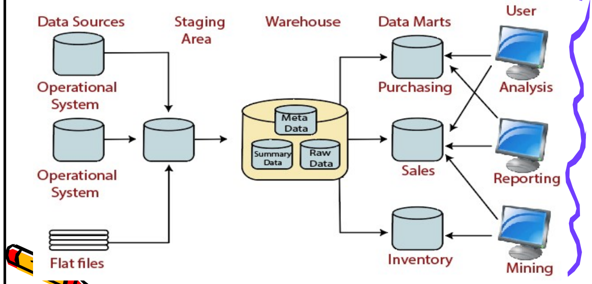
Data Warehouse Architecture: With Staging Area and Data Marts

- The figure illustrates an example where purchasing, sales, and inventory are separated.
- In this example, a financial analyst wants to analyze historical data for purchases and sales or mine historical information to make predictions about customer behavior.

32

Data Warehouse Architecture: With Staging Area and Data Marts

Architecture of a Data Warehouse with a Staging Area and Data Marts



Questions