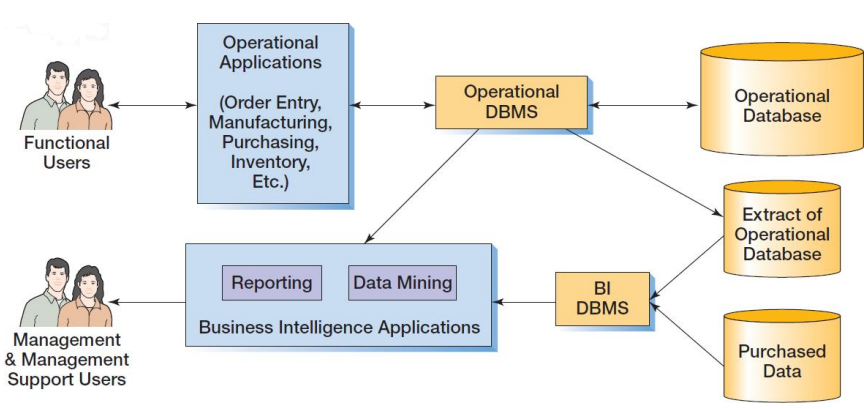
• Big Data—the current term for the enormous datasets generated by Web applications such as search tools (for example, Google and Bing) and Web 2.0 social networks (for example, Facebook, LinkedIn, and Twitter).  
• Although these new and very visible Web applications are highlighting the problems of dealing with large datasets, these problems were already present in other areas, such as scientific research and business operations.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Byte | Kilobyte | Megabyte | Gigabyte | Terabyte | Petabyte | Exabyte | Zettabyte | Yottabyte |
| Symbol |  | KB | MB | GB | TB | PB | EB | ZB | YB |
| Approx Value |  | 10^3 | 10^6 | 10^9 | 10^12 | 10^15 | 10^18 | 10^21 | 10^24 |
| Actual Value | 8 bits | 1024 bytes | 1024KB | 1024MB | 1024GB | 1024TB | 1024PB | 1024EB | 1024ZB |

3 characteristics of Big Data Volume, Velocity, Variety

• Business intelligence (BI) systems are information systems that assist managers and other professionals: – To analyze current and past activities. – To predict future events.   
• Two broad categories: – Reporting – Data mining

**Operational and BI systems relationship model**



• BI systems obtain data in three ways: – From the operational database • Read and process data only • DO NOT insert, modify or delete operational data – From extracts from the operational database • Data is in a BI DBMS • May be a different DBMS than the operations DBMS – From data purchased from data vendors  
  
• Reporting system applications: – Filter – Sort – Group – Make simple calculations – Classify entities • RFM Analysis – Can be performed using standard SQL – Extensions to SQL are sometimes used • online analytical processing (OLAP) – Summarize current business status – Compare current business status to past or future – Deal with critical report delivery  
  
• Data mining applications are used to: – Perform what-if analysis – Make predictions – Facilitate decision making • Data mining applications use sophisticated statistical and mathematical techniques. • Report delivery is not as critical.

**Characteristics of Business Intelligence Applications**  
Reporting:

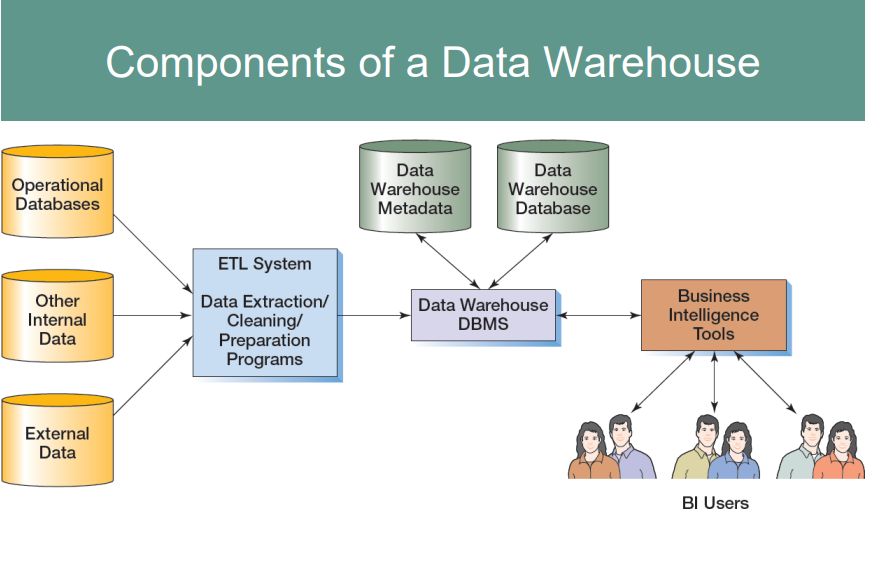
* Filter, sort, group, and make calculations
* Summarize current status
* Compare current status to past or predicted status
* Classify entities (customers, products, employees, etc…)
* Report delivery crucial

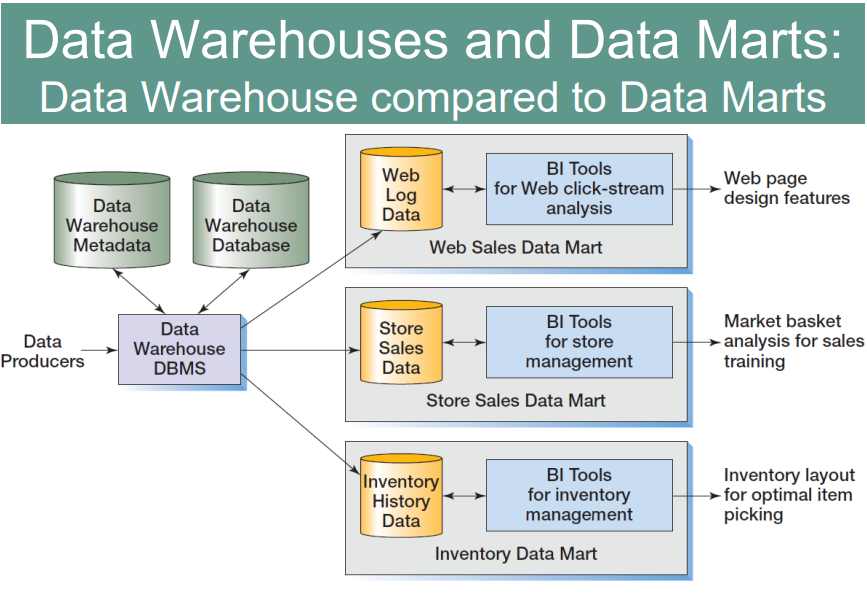
Data Mining

* Often employ sophisticated statistical and mathematical techniques
* Used for: What if analysis, Predictions, Decisions
* Results often incorporated into some other report or system

Amerilink Data Categories

* Name, Address, Phone
* Age, Gender
* Ethnicity, Religion
* Income
* Education
* Marital Status. Life Stage
* Height, Weight, Hair and Eye color
* Spouse’s name, Birth Date, etc…
* Kids’ names and birth dates
* Voter Registration
* Home Ownership
* Vehicles
* Magazine Subscriptions
* Catalog Orders
* Hobbies
* Attitudes





**Problems of using Transaction Data for Business Intelligence**

* Dirty Data
* Missing Values
* Inconsistent Data
* Data not integrated
* Wrong Format: Too fine, Not fine enough
* Too much data: Too many attributes, Too much volume

**Characteristics of Operational and Dimensional databases**

* Operational Database
  + Used for structured transaction data processing
  + Current data are used
  + Data are inserted, updated, and deleted by users
* Dimensional Database
  + Used for unstructured analytical data processing
  + Current and historical data are used
  + Data are loaded and updated systematically, not by users

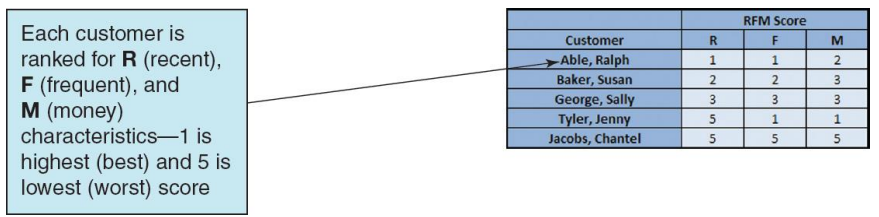
**Reporting Systems:**  
RFM Analysis analyzes and ranks customers according to purchasing patterns

R = recent (most recent order)

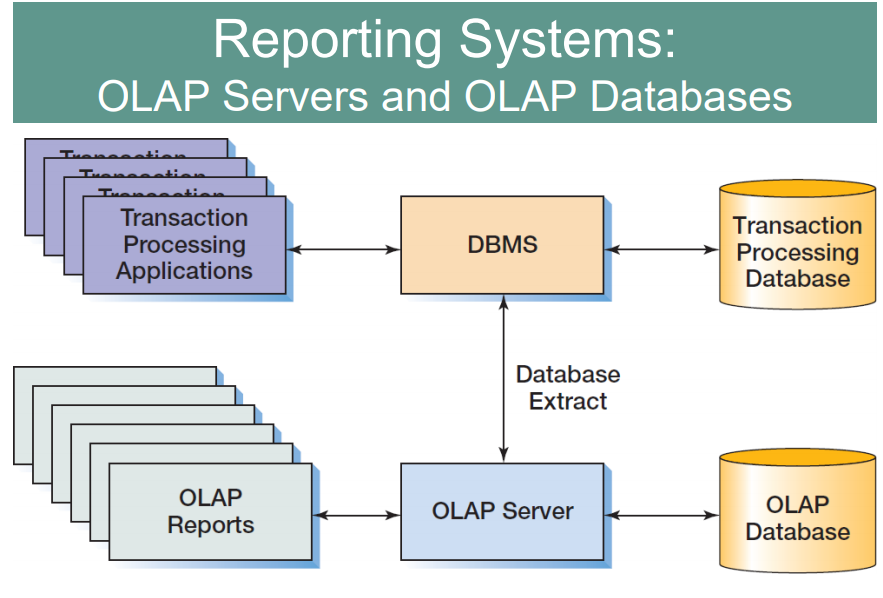
F = frequent (how often an order is made)

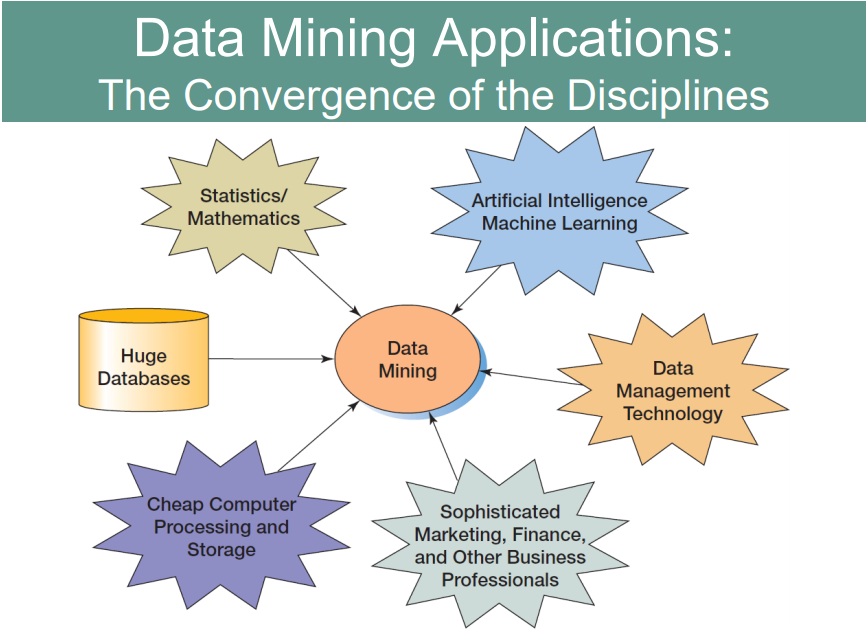
M = money (dollar amount of orders)

Customers are sorted into five groups, each containing 20% of the customers.   
Each group is given a numerical value: – 1 = top 20% – 2, 3, 4 = each 20% in between top and bottom 20% – 5 = bottom 20%

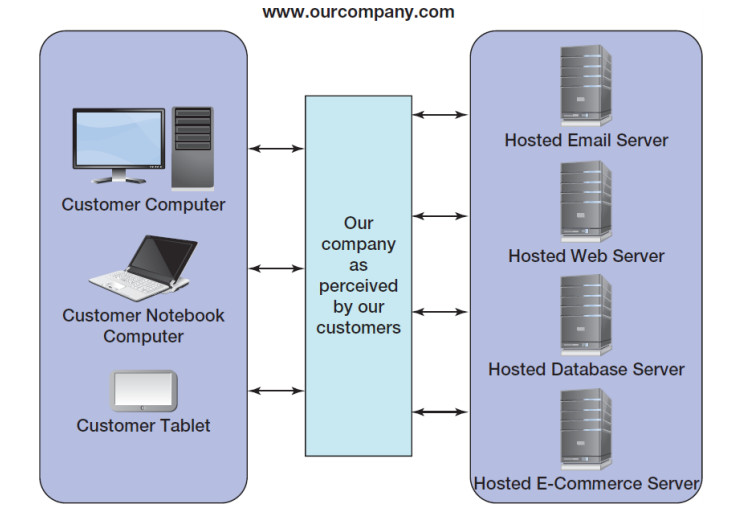


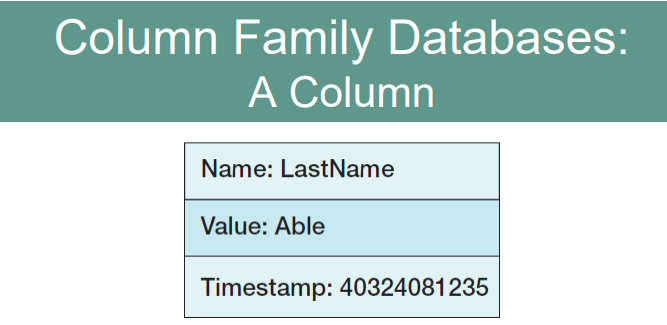
• An OLAP report has measures and dimensions:  
 – Measure—a data item of interest  
 – Dimension—a characteristic of a measure   
• OLAP cube—a presentation of a measure with associated dimensions.   
– An OLAP cube can have any number of axes.   
– The terms OLAP cube and OLAP report are synonymous.   
• OLAP allows drill-down—a further division of the data into more detail.





**Object Oriented Database**  
• Object-oriented programming (OOP) is a technique for designing and writing computer programs.   
• Objects are data structures that have:   
– Methods—computer programs that perform some task   
– Properties—data items particular to an object.  
•Storing the values of properties of an object is called object persistence.   
• Many different techniques have been used for object persistence.   
• One of them is to use some variation of database technology.  
• Special-purpose DBMS products for storing object data are called objectoriented DBMSs (OODBMSs) •Never achieved commercial success.   
– Billions of bytes of data were already stored in relational DBMS format   
– No organization wanted to convert their data to OODBMS format to be able to use an OODBMS.  
• The need for object persistence did not disappear.  
• Some vendors, most notably Oracle, added features and functions to their relational database DBMS products to create object-relational databases.   
• DBMS products for storing object data are called object-oriented DBMSs (OODBMSs).

**Cloud Computing**

**The NoSQL movement**  
• The NoSQL movement, better described as the Not only SQL movement, is a movement to using non-relational databases.   
• These databases are often distributed, replicated databases.   
• Used in many widely recognized Web applications – Used for Facebook – Used for Twitter  
• Four categories of NoSQL nonrelational databases:   
– key-value—examples are Dynamo and MemcacheDB  
– document—examples are Couchbase and MongoDB   
– column family—examples are Apache Cassandra and Hbase   
– graph—examples are Neo4j and AllegroGraph  


**HADOOP**

• Hadoop Distributed File System (HDFS) – provides standard file services to clustered servers so their file systems can function as one distributed file system.   
• The Hadoop family includes a full set of applications including:   
– Hbase – A nonrelational data store.  
– Pig – A query language.