Data Warehousing

**Transactional Data vs Analytical Data**

* **Transactional Data** refers to real-time data which is gathered from a company’s multiple transactions (Before ETL)
  + Airline reservations
  + ATM Withdrawal
  + Purchasing stock
  + Sales receipts
* **Analytical Data** refers to data which has already been processed and is used by an organization to make decisions (After ETL)
  + Transactional Information
  + \*Can be external information as well
  + Economic conditions
  + Industry
  + Market
  + Sales
  + Product statistics
* \* This is from the best of my understanding from the slides. If I’m wrong feel free to correct me

**Cost of poor information**

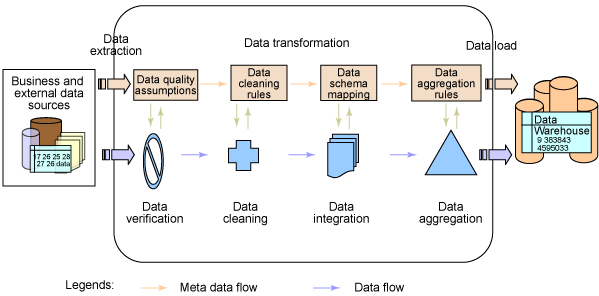
* Cannot accurately track customers
* Hard to identify valuable customers
* Cannot identify selling opportunities
* Difficulty tracking revenue
* Inability to build strong customer relationships

**Benefits of good information**

* Improves chances of making good decisions

**Data Warehousing**

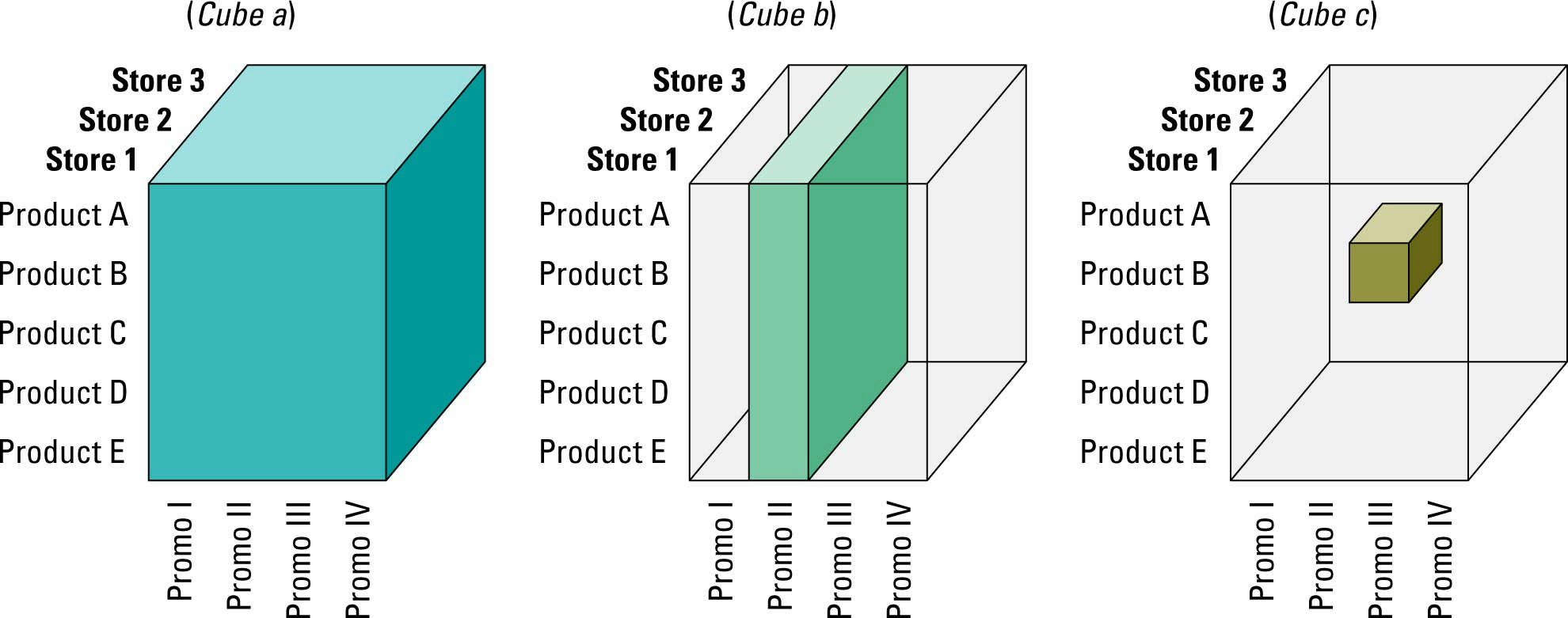
* **Definition:** Repository of historical data which is organized by subject to support decision makers
* **Purpose:** Aggregate information of an organization into a single repository for decision making
* Facilitates business intelligence
* **ETL(Extract, Transform, Load)**
  + **Extract:** Takes in data from internal and external databases.
  + **Transforms:** converts data using enterprise definitions.
  + **Loads:** Loads the data into the warehouse.

****

* Transforms data into information
* 1990s more occupied on overall business functions
* Support decision making without disrupting day-to-day operations
* Managers access corporate data to make decisions, must be able to respond quickly to market changes
* **Data Mart:** The mini-me of the warehouse. It focuses on a subset of information (i.e. Finance)
* **Characteristics:**
  + Subject-Oriented: organized by subjects
  + Consistent: data is encoded in a consistent manner
  + Historical: data kept for many years to see trends
  + Non-Volatile: data doesn’t change once in warehouse.
  + Multidimensional: (think of cube)

**Multi-dimensional Analysis**

* Databases contain information in a series of two-dimensional tables
* In a warehouse and data mart, it contains layers of columns and rows
* **Dimensions:** Characteristics of information. An attribute
* **Cube:** representation of multi-dimensional information
* **Historical Data:** Used to identify trends, forecasting, comparing



**Information Cleansing or Scrubbing**

* Process that weeds out and fixes inconsistent, incorrect or incomplete information.
* Sophisticated algorithms used by software tools to do this. (Oracle, SAS, IBM)
* Fixes
  + Redundant records
  + Missing records or attributes
  + Erroneous relationships
  + Inaccurate data
* Costs money. The more complete and accurate the more it costs.

**Building a Data Warehouse**

1. Extract data from various sources
2. Transform, integrate and load the data
3. Keep detailed and summary data in read-only environment
4. Access via query languages and analytical tools

**Data Warehousing Benefits**

* Can access data quickly and easily via web browsers
* Can conduct extensive analysis with data
* Users have a consolidated view of organized data (all in one place)

**Data Warehouse Issues**

* Expensive
* Incorporating data from obsolete mainframe is difficult
* Some departments may not want to share data with others
* Transferring data from another system can go through a cleansing process that changes the information

**Data Marts**

* Small data warehouse designed for users who seek specific information in a department or subset

**Data Mart Benefits**

* Less expensive
* They contain less information therefore they have a more rapid response time.
* Easier to learn and navigate
* Supports users locally instead of being centrally controlled by giving power or control to users

**OLAP vs OLTP**

* Online Analysis Processing (OLAP)
  + Analysis of accumulated data
  + Multi-dimensional analysis
  + Uses graphical software tools to provide complex analysis of data stored
  + Good to identify trends
  + Allows users to drill-down into data
* Online Transaction Processing (OLTP)
  + Involves a database, where data from business transactions are processed as soon as they occur
  + Real-time data
  + Immediate automated responses to the request of users
  + Designed to handle multiple concurrent transactions
  + Large role in E-commerce

**Business Intelligence**

* Using applications and technologies to consolidate, analyze, and provide access to large amounts of data to help make better business and strategic decisions.



* **Characteristics:** 
  + **Reliable:** information is approved beforehand so it can be used with confidence
  + **Consistent:** information is complete, predictable response times.
  + **Understandable:** information defined in business terms so it’s easy to understand.
  + **Easily Manipulated:** easy to change questions or set different parameters.

**Three forms of BI**

* **Strategic BI:** Helps with planning
* **Operational BI:** Helps with immediate actions
* **Tactical BI:** daily analysis

**Latency**

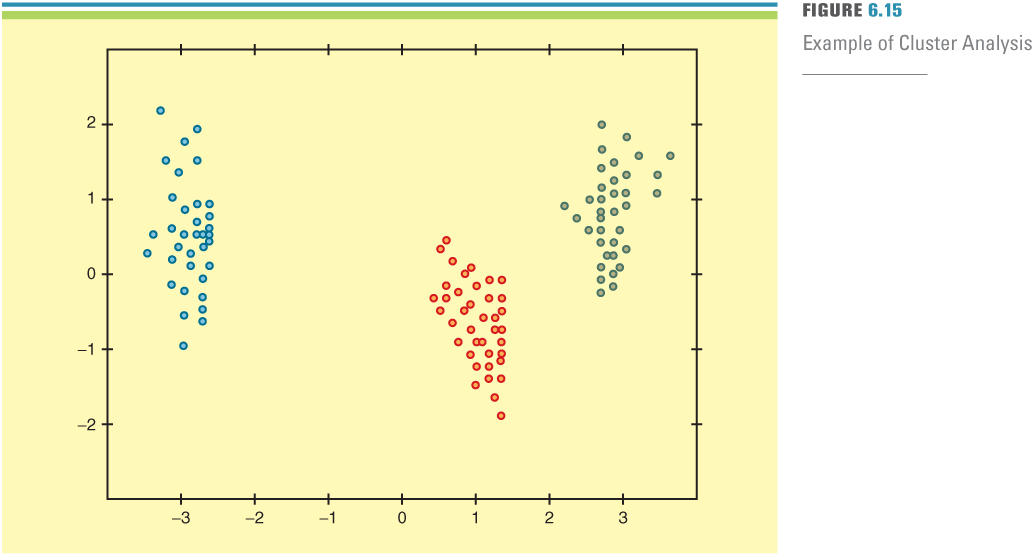
* **Data Latency:** Time needed to make transactional data ready for analysis
* **Analysis latency:** Time from which data is made available and complete analysis
* **Decision Latency:** Time it takes a human to comprehend analysis and make decisions.
* **\*Time is money so shorten these as much as possible to get the most value**

**Data Mining**

* Searching for valuable business information in a large database or warehouse
* **Two basic operations:** 
  + Predict trends
  + Identify previously unknown patterns
* **Example of data mining:** Detecting fraudulent credit card transactions

**Data Visualization**

* Analysts can visualize data with graphs
* **Common algorithms:**
  + Cluster analysis
    - Divides information sets into different groups
    - Example: postal code segmenting



* + Association detection
    - Reveals the relationship between variables along with the nature and frequency of the relationships
    - Example: Market basket analysis
  + Statistical analysis
    - Performs functions such as information correlations, distributions, calculations, etc
    - Forecast- prediction made on basis of time-series information
    - Time-series information – time-stamped information collected at a particular frequency

**Unsupervised Data Mining**

* Analysts do not create a model or hypothesis before running analysis
* Apply data mining techniques to data and observe results
* Create hypothesis after analysis
* i.e. Cluster Analysis

**Supervised Data Mining**

* Create model prior to analysis and apply statistical techniques to data to estimate parameters
* i.e. statistical analysis and association detection

**Data Mining – RFM analysis**

* Mining algorithm
* A way of analyzing and ranking customers according to purchase patterns
* **R** – How recently
* **F** – How frequent
* **M** – How much a customer spends

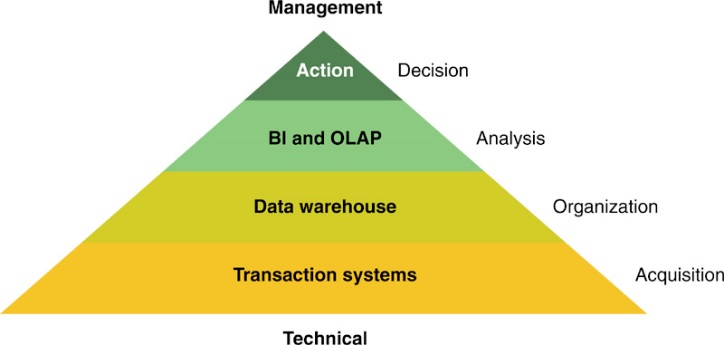
**Benefits of BI**

* Single point of access for all users
* Up-to-the-minute information for everyone (real time updates)
* BI across organizational departments (can be used in every step of value chain)
* **Categories of benefits:**
* **Direct quantifiable benefits:** amount of dollars saved, response time reduced, etc.
* **Indirect quantifiable benefits:** benefits that may be due to other factors
* **Unpredictable benefits:** analysis revealing a new target
* **Intangible benefits:** general expectations that change will improve bottom line but cannot be measured right away.

**Two types of BI systems:**

* **Those that provide data analysis tools**
  + Multidimensional data analysis
  + Data mining
  + Decision support systems (DSS refer to other notes on IS)
    - **What-if analysis:** study impact of a change in the input data on the proposed solution.
    - **Goal-seeking analysis:** attempts to find the value of the inputs necessary to achieve a goal.
    - **Sensitivity analysis:** Study of impact that changes in one part or more have on other parts
* **Those that provide information in structured format**
  + Dashboards: rapid access to timely information, direct access to management reports.

**The BI pyramid**

****