

Business Intelligence

Title: **Business Intelligence**

Code: **IFT2433, IFT4202**

Credit Units: **2**

Programme: **BSC- IT / BCA / MCA**

Batch: **2019-2022**

2

# Business Intelligence (BI)

**Business intelligence** (BI) is a technology-driven process for analyzing data and presenting actionable information which helps executives, managers and other corporate end users make informed **business** decisions.

BI encompasses a wide variety of tools, applications and methodologies that enable organizations to collect data from internal systems and external sources, prepare it for analysis, develop and run queries against that data and create reports, dashboards and data visualizations to make the analytical results available to corporate decisionmakers, as well as operational workers.

Other potential benefits of business intelligence tools include:

* accelerating and improving decision-making;
* optimizing internal business processes;
* increasing operational efficiency;
* driving new revenues;
* gaining competitive advantage over business rivals;
* assisting companies in the identification of market trends; and
* spotting business problems that need to be addressed.

**Business intelligence software and systems**

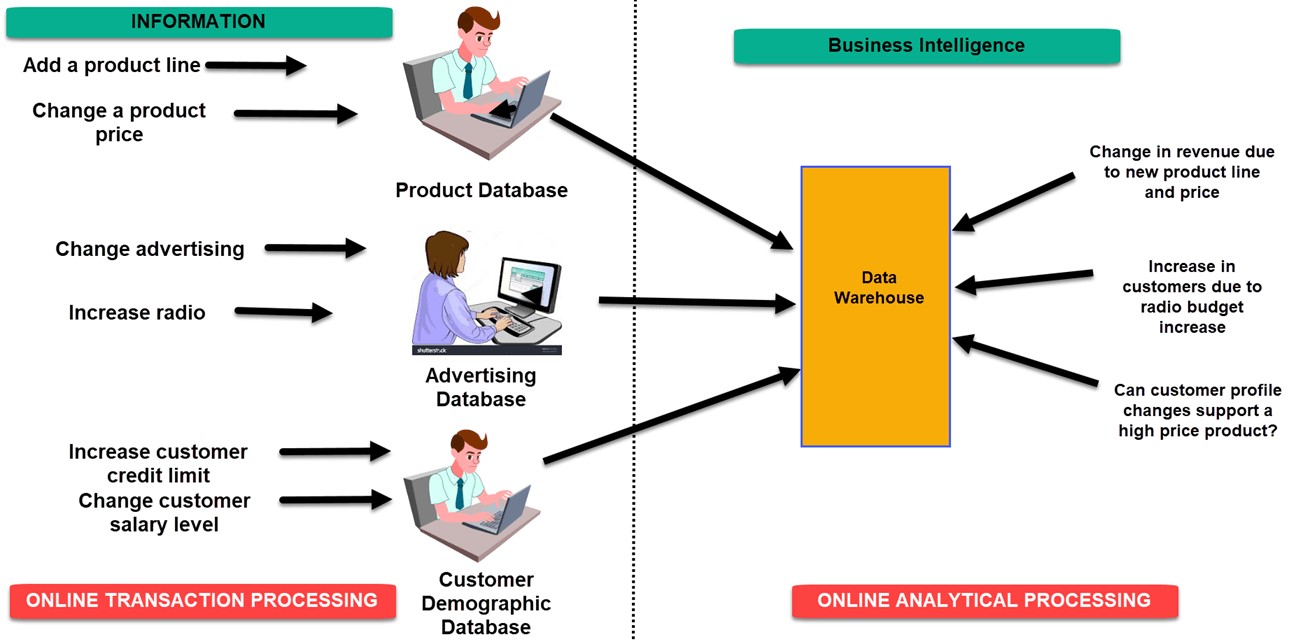
A variety of different types of tools fall under the business intelligence umbrella. Some of the most important categories and features:

* Dashboards
* Visualizations
* Reporting
* Data mining
* ETL (extract-transfer-load —tools that import data from one data store into another)
* OLAP (online analytical processing)

**How Business Intelligence systems are implemented?**

* Here are the steps:
* **Step 1**) Raw Data from corporate databases is extracted. The data could be spread across multiple systems heterogeneous systems.
* **Step 2)** The data is cleaned and transformed into the data warehouse. The table can be linked, and data cubes are formed.
* **Step 3)** Using BI system the user can ask queries, request ad-hoc reports or conduct any other analysis.

**Example 1**



**Example 2**

* A hotel owner uses BI analytical applications to gather statistical information regarding average occupancy and room rate. It helps to find aggregate revenue generated per room.
* It also collects statistics on market share and data from customer surveys from each hotel to decides its competitive position in various markets.
* By analyzing these trends year by year, month by month and day by day helps management to offer discounts on room rentals.

**Example 3**

* A bank gives branch managers access to BI applications. It helps branch manager to determine who are the most profitable customers and which customers they should work on.
* The use of BI tools frees information technology staff from the task of generating analytical reports for the departments. It also gives department personnel access to a richer data source.

## Business Intelligence

**Increasing**

**potential**

**to support**

**business**

**decisions**

**End**

**User**

**Bu**

**s**

**iness**

**Analyst**

**Data**

**Analyst**

**DBA**

**Making**

**Decisions**

**Data**

**Presentation**

***Visualization***

***Techniques***

**Data**

**Mining**

***Information***

***Discovery***

**Data**

**Exploration**

***Statistical***

***Analysis, Querying***

***and***

***Reporting***

**Data**

**Warehouses**

**/**

**Data**

**Marts**

***OLAP,***

***MDA***

**Data**

**Sources**

***Paper,***

***Files,***

***Information Providers,***

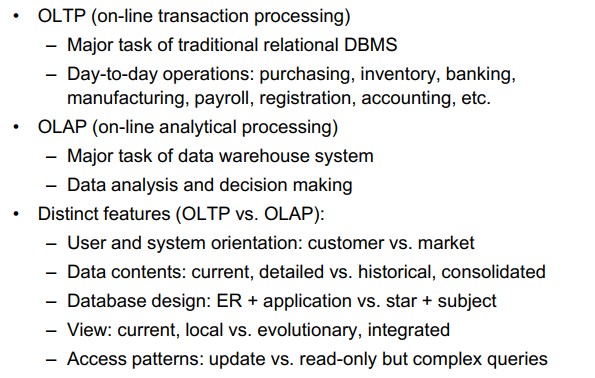
***Database Systems,***

***OLTP***

Top Business Intelligence Companies

* Microsoft • Tableau Software • Sisense • Pentaho • Domo • SAP • Oracle
* IBM…….

### OLTP vs. OLAP



### OLTP Systems

* OLTP Systems refers to a class of systems that manage transactions oriented applications.
* These applications are mainly concerned with entry, storage and retrieval of data.
* They are designed to cover most of the day to day operations of an organization such as purchasing, inventory management etc.
* They are characterized by large number of short on line transactions like INSERT,DELETE,UPDATE
* The data captured by OLTP systems are usually stored in commercial relation database.

### OLTP

• **Advantages –** • Simplicity

#### • Efficiency • Fast Query Processing

* **Disadvantages –** • Security –
* OLTP system data content not suitable for decision making

### OLAP Systems

* They differ from traditional databases in the way data is conceptualized and stored.
* In OLAP data is held in dimensional form rather than relational form.
* OLAP tools are based on multidimensional data models.
* The multidimensional data models views data in the form of data cubes.

* Advantages :
* Multidimensional data representation • Consistency of Information • What if Analysis can be carried out • Provides a single platform for all information and business needs – planning, budgeting, reporting and analysis • Fast and interactive adhoc exploration

## Architecture of OLAP

### • ROLAP • MOLAP • HOLAP

**Relational Online Analytical Processing (ROLAP) :**

* **Relational Online Analytical**  **Processing (ROLAP) :**

ROLAP servers are placed between relational backend server and client front-end tools. It uses relational or extended DBMS to store and manage warehouse data. ROLAP has basically 3 main components: Database Server, ROLAP server, and Front-end tool.

* **Advantages of ROLAP –**
* ROLAP is used for handle the large amount of data.

* ROLAP tools don’t use pre-calculated data cubes.
* Data can be stored efficiently.
* ROLAP can leverage functionalities inherent in the relational database.
* **Disadvantages of ROLAP**
* Performance of ROLAP can be slow.
* In ROLAP, difficult to maintain aggregate tables.
* Limited by SQL functionalities.

### **Multidimensional Online Analytical Processing (MOLAP)**

* **Multidimensional Online Analytical Processing (MOLAP) :**
* The storage is in proprietary formats and not in the relational databases.
* MOLAP does not uses relational database to storage. It stores in optimized multidimensional array storage. The storage utilization may be low With multidimensional data stores. Many MOLAP server handle dense and sparse data sets by using two levels of data storage representation. MOLAP has 3 components : Database Server, MOLAP server, and Front-end tool.

* **Advantages:**
* MOLAP is basically used for complex calculations.
* MOLAP is optimal for operation such as slice and dice.
* MOLAP allows fastest indexing to the pre-computed summarized data.

* **Disadvantages of MOLAP –**
* Limited in the amount of data that it can handle.
* In MOLAP, Requires additional investment.
* Without re-aggregation, difficult to change dimension.

### **Hybrid Online Analytical Processing (HOLAP)**

* **Hybrid Online Analytical Processing (HOLAP) :**

Hybrid is a combination of both ROLAP and MOLAP. It offers functionalities of both ROLAP and as well as MOLAP like faster computation of MOLAP and higher scalability of ROLAP. The aggregations are stored separately in HOLAP store. Its server allows storing the large data volumes of detailed information.

* **Advantages of HOLAP –**
* HOLAP provides the functionalities of both MOLAP and ROLAP.
* HOLAP provides fast access at all levels of aggregation.

* **Disadvantages of HOLAP –**

HOLAP architecture is very complex to understand because it supports both MOLAP and ROLAP.

### Digital Data - Introduction

* Data growth has seen exponential acceleration since the advent of computer and Internet.
* Computers and Internet have imparted the digital form to data.

Digital Data

Unstructured

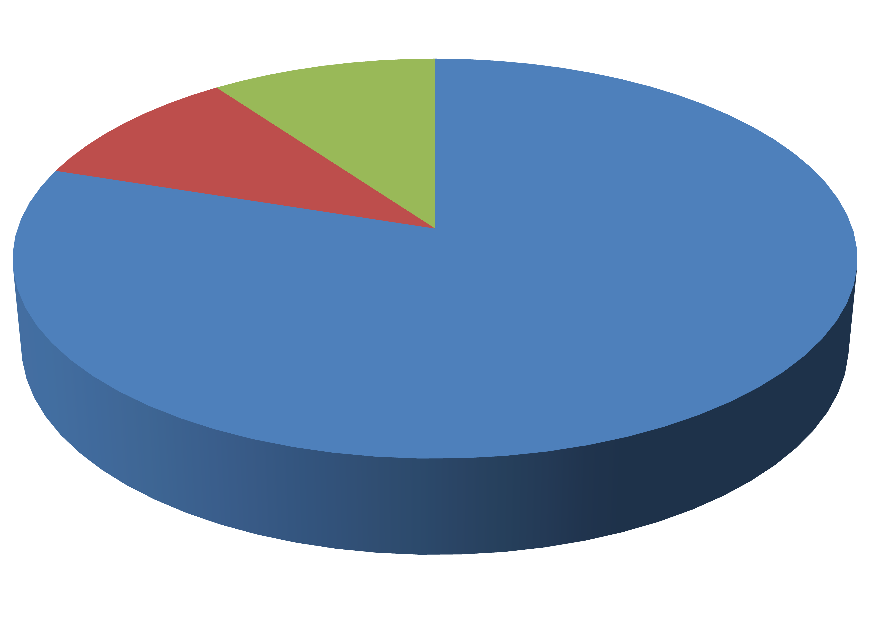
Semi

Structured

Structured

**Distribution of Data**

Unstructure Data



80

%

10

%

10

%

Semi Structured Data Structured Data

### Digital Data Types

Unstructured Data – This data does not conform to any particular data model or is not in a form that can be used easily by a computer program.

(e.g) Bitmap objects (Images, Videos, Audio files) , Textual Objects (Letters , body of an email, Memos , Power point Presentations, Survey, White Papers) etc.

Semi Structured Data - This data does not conform to any particular data model but has got some structure . However it is not in a form that can be used easily by a computer program. (e.g) XML, HTML etc. Metadata is available but not sufficient.

Structured Data – This data is organized (rows and columns) and can be used by the computer programs . Relationships exist between entities of data, such as classes and objects.(e.g) Data stored in databases.

### Characteristics of Structured Data

Similar entities

are grouped

Data Stored in

form of Rows

and Columns

Conforms to a

Data Model

Data resides in

fixed fields

within a record

or file

Attributes in a

group

are same

Definition,

format and

meaning of data

is explicitly

known

Characteristics of Unstructured Data

Not in any

particular format

or sequence

Data cannot be

stored in form of

Rows and

Columns

Does not

conforms to a

Data Model

Not easily usable

by a program

Does not follow

any rules or

semantics

Has no

identifiable

Structure

Characteristics of Semi structured Data

Attributes in a

groups may not

be the same

Data cannot be

stored in form of

Rows and

Columns

Does not

conforms to a

Data Model but

contains tags

and elements

(

metadata

)

Not sufficient

metadata

Similar entities

are grouped

Tags and

elements

describe data is

stored

### Evolution of BI -MIS

* To carry out managerial functions for various functional areas (production, marketing etc.) and integrate them with the external environment (government, customers etc.), managers need different types of information (quantitative and qualitative).
* Earlier, this information was provided by the accounting system which was limited and quantitative in nature.
* With introduction of computers, managers have access to huge quantity of data at very high speed. Computers help managers create data base and manipulate the information for taking various managerial decisions.
* It helps in storage of and retrieval of information.
* Computers enable managers to collect data at very short intervals of time; process, analyze and convert it into useful information and relate it to the external environment.
* This system of obtaining timely, relevant and accurate information based on computer technology is known as **management information system(MIS)**.
* The system helps in preparing reports for carrying out planning and controlling functions.

* **“MIS is a formal method of making available to management the accurate and timely information necessary to facilitate the decision-making process and enable the organization's planning, control and operational functions to be carried out effectively.”**

**Effective management information system has the following features:**

* 1. **Timeliness 4. Conciseness**
  2. **Accuracy 5. Completeness**

**3.Relevance**

### MIS

* MIS evaluates, analyzes and processes organizational data to produce useful information.
* It collects stores and distributes data as information, which helps to manage the organization properly. It also allows taking summarized reports.
* These reports help the management to monitor the organization, understand the current performance status and to take future business decisions.
* **Executive Information System (EIS) and Decision Support System (DSS)** are two types of MIS.
* The **main difference** between EIS and DSS is that the **EIS allows senior managers to take decisions to meet the strategic goals of the organization while the DSS allows senior managers to take non-routine decisions.**

### Evolution of BI -DSS

* It is also called knowledge based systems.
* It supports business decision making activities and supports decision making that is required to run day to day operations.
* A decision support system is an interactive, flexible and adaptable computer system that is easily accessible to, and operated by, non-computer specialists to assist them in planning and decision-making functions.
* It is especially developed for supporting the solution of a non-structured management problem for improved decision making.
* The emphasis is in on use of business graphics to present information from multiple sources.
* **Advantages :** • It saves time.
* Enhances efficiency.
* Reduces the cost.
* It improves personal efficiency.
* It increases the decision maker satisfaction.
* **Disadvantages :**
* Information Overload.
* Status reduction.
* Over-emphasize decision making.

### Evolution of BI -EIS

* EIS is defined as a system that comes with powerful reporting and analytical abilities.
* It supports decision making at senior management level.
* It is a structured, automated tracking system that operates continuously to keep everything managed.
* It provides exception and status reporting capabilities.
* It mainly focuses on metrics and KPIs that indicate the health of the functions/projects or business performance.
* It enables organizations to integrate and coordinate business processes and has support for metric based performance.

* **Advantages :**
* Easy to use.
* Ability to analyze the trends.
* Time management.
* Efficiency.
* Enhances business problem solving.
* **Disadvantages :**
* Functions are limited.
* Difficult to keep current data.
* System can run slow.
* Less reliable.

|  |  |
| --- | --- |
| DSS | EIS |
| It is used by professionals. | It is used by executives. |
| It is required for day-to-day operations. | It is required for strategic plans and procedures. |
| It deals with semi and unstructured data. | It deals with only unstructured data. |
| It consists of only internal information. | It consists of both internal and external information. |
| It allows taking non-routines decisions. | It allows taking decisions to meet the strategic goals of the organization. |
| It is used with mainframes, micro and distributed systems. | It is used with distributed systems. |

The Different Users of Business

Intelligence

* There are many different users who can benefit from business intelligence
  + Executives – Those who focus on the overall business
  + Business Decision Makers – Usually focused on single areas of the business (finance, HR, manufacturing, and so forth)
  + Information Workers – Typically managers or staff working in the back office
  + Line Workers – Employees who might use BI without knowing it
  + Analysts – Employees who will perform extensive data analysis

### Important Fact…

**Business intelligence** usually refers to the information that is available for the enterprise to make decisions on. A data warehouse (or data mart) system is the backend, or the infrastructural, component for achieving business intelligence. Business intelligence also includes the insight gained from doing data mining analysis, as well as unstructured data

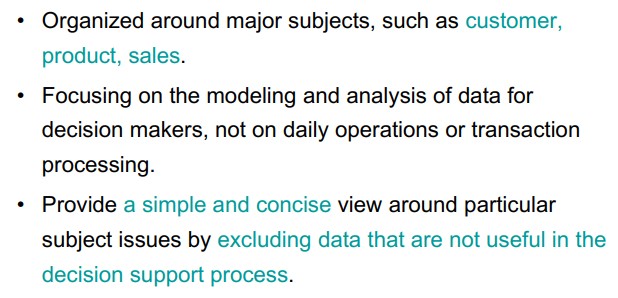
What is Data Warehouse?

Defined in many different ways, but not rigorously.

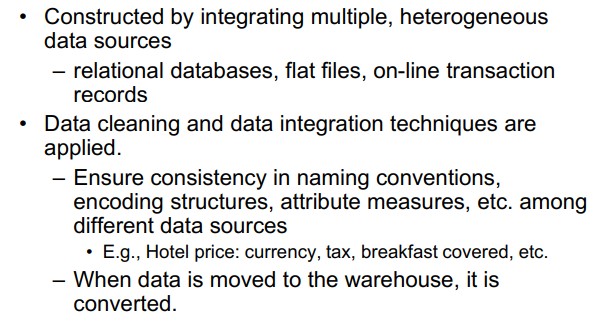
* A decision support database that is maintained separately from the organization’s operational database
* Support information processing by providing a solid platform of consolidated, historical data for analysis.

“A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”-W. H. Inmon

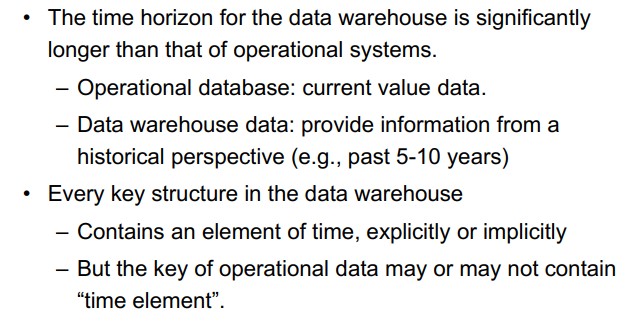
### Data Warehouse –Subject Oriented



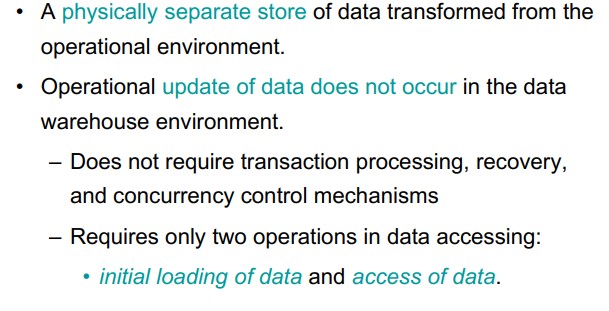
### Data Warehouse –Integrated



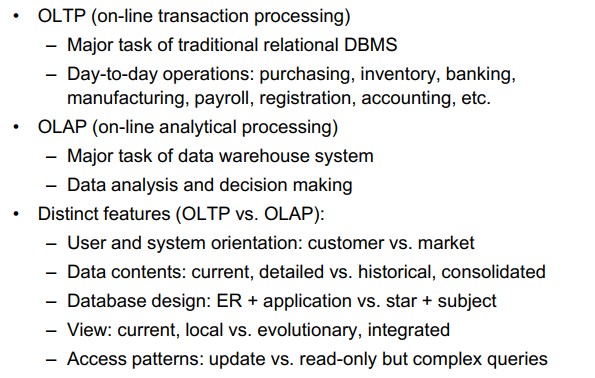
### Data Warehouse –Time Variant



### Data Warehouse –Non Volatile



### Data Warehouse vs. Operational DBMS



Data Marts

From a data warehouse data flows to various departments for their customized DSS usage.

* These individual department components are called as data marts.
* Data Mart is a body of DSS data for a department that has an architectural foundation of a data warehouse.

### Reasons for creating a Data Mart

#### • -Easy access to frequently needed data -Creates collective view by a group of users -Improves end-user response time • -Ease of creation -Lower cost than implementing a full data warehouse -Potential users are more clearly defined than in a full data warehouse -Contains only business essential data and is less cluttered. 41

Data Mart is a natural outgrowth of a data warehouse

Data Mart is much popular than a Data Warehouse because -

-As central database keeps growing it becomes more and more complex.

-Gradually the data becomes harder to customize, summarize and analyze for drawing analytical results easily.

-Cost of processing the data also increases as the volume of the data increases.

All the issues related to data marts are equally relevant for a data warehouse, since basically a data warehouse is a collection of data marts

We are drowning in data, but starving for knowledge!

Why do we have so much data ?

• Data explosion problem

– Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories

### Solution : Data Mining

Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Motivation: ―Necessity is the Mother of Invention‖

What Is Data Mining?

* Data mining (Knowledge Discovery in Databases KDD)
  + Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases



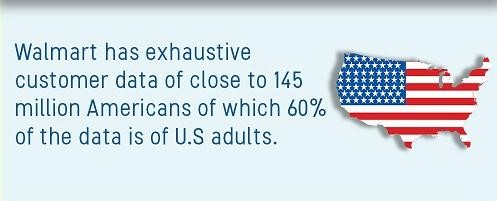
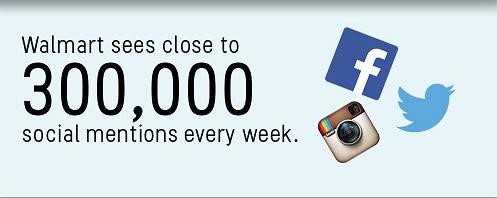
* Alternative names and their ―inside stories‖:



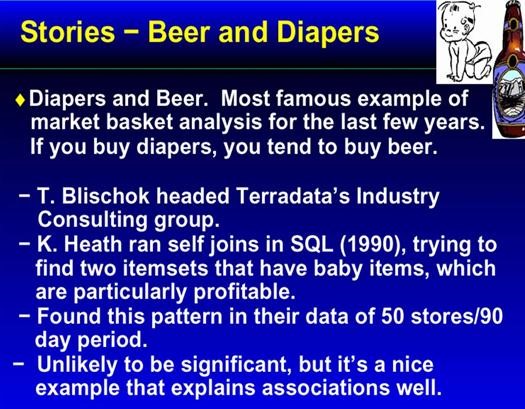
* + Data mining: a misnomer?
  + Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

* What is not data mining?
  + (Deductive) query processing.
  + Expert systems or small ML/statistical programs

### Example …Walmart



How can DM be used?

Probably mom was calling dad at work to buy diapers on way home and he decided to buy a six‐ pack as well.

The retailer could move diapers and beers to separate places and position high‐profit items of interest to young fathers along the path.

### Few More Examples

* Wal‐Mart customers who purchase Barbie dolls have a 60% likelihood of also purchasing one of three types of candy bars [*Forbes*]

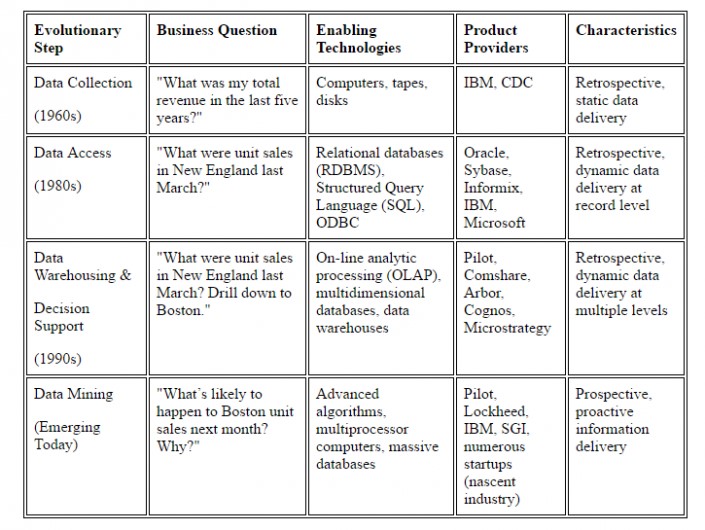
* Customers who purchase maintenance agreements are very likely to purchase large appliances (Linoff and Berry experience)

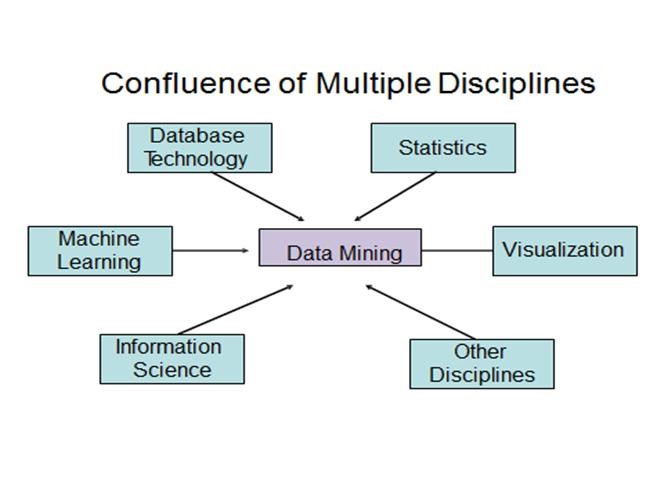
* When a new hardware store opens, one of the most commonly sold items is toilet bowl cleaners (Linoff and Berry experience)

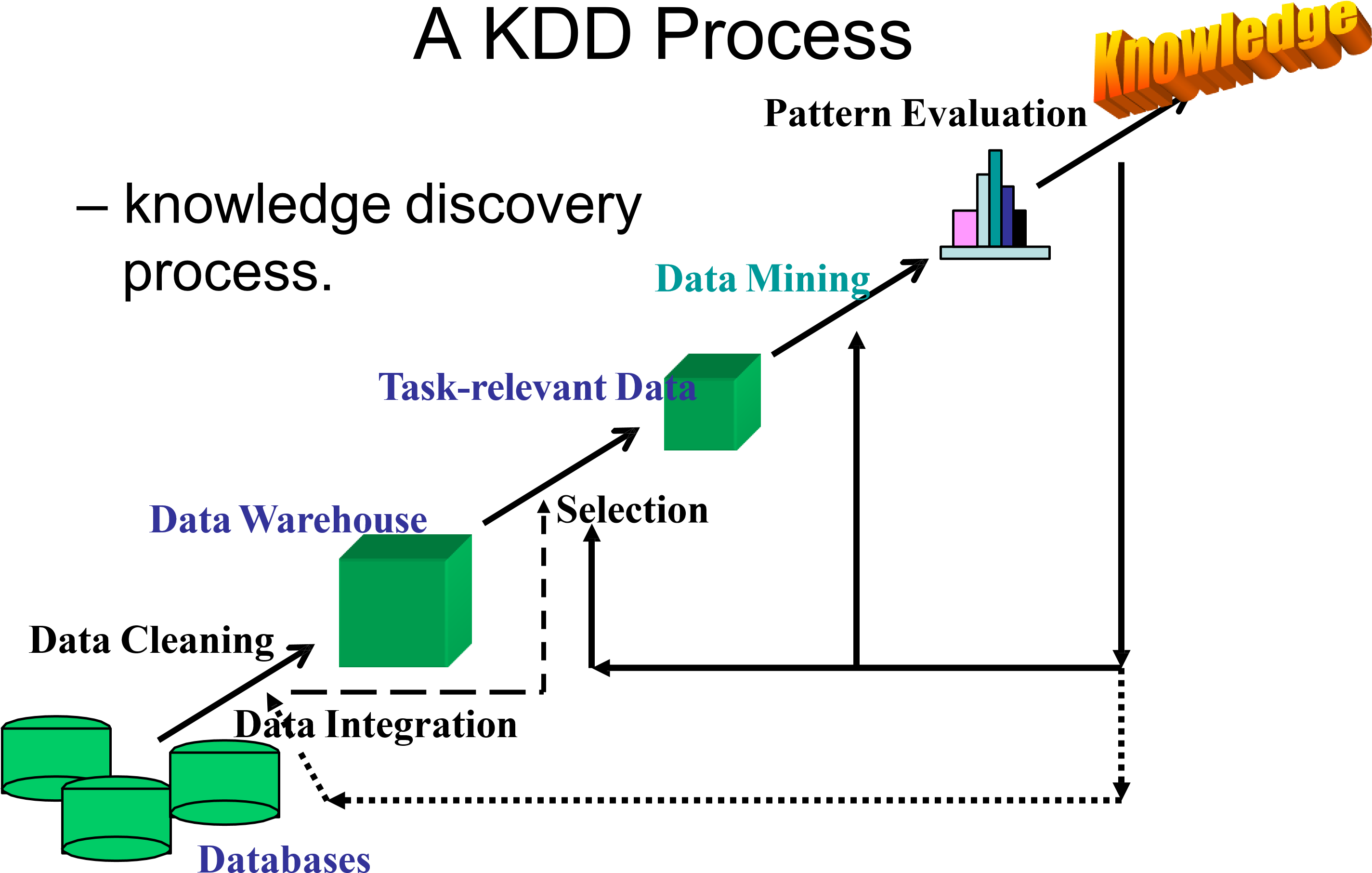
### Some Suggestions

* By increasing the price of Barbie doll and giving the type of candy bar free, walmart can reinforce the buying habits of that particular types of buyer
* Highest margin candy to be placed near dolls.
* Special promotions for Barbie dolls with candy at a slightly higher margin.
* Take a poorly selling product X and incorporate an offer on this which is based on buying Barbie and Candy. If the customer is likely to buy these two products anyway then why not try to increase sales on X?

Evolution of Database Technology







## Data Mining Applications

1. Financial Data Analysis
2. Retail Industry
3. Telecommunication Industry
4. Biological Data Analysis
5. Other Scientific Applications
6. Intrusion Detection

Data Mining : Market Trends

Data Mining Vs. Big Data

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Mining** | **Big Data** |
| **Focus** | It mainly focusses on lots of details of a data | It mainly focusses on lots of relationships between data |
| **View** | It is a close-up view of data | It is the Big Picture of data |
| **Data** | It expresses what about the data | It expresses Why of the data |
| **Volume** | It can be used for small data or big data | It refers to a large number of data sets |
| **Definition** | It is a technique for analyzing data | It is a concept than a precise term |
| **Data Types** | Structured data, relational and dimensional database. | Structured, Semi-Structured and Unstructured data (in NoSQL) |
| **Analysis** | Mainly Statistical Analysis, focus on prediction and discovery of business factors on small scale. | Mainly data analysis, focus on prediction and discovery of business factors on a large scale. |
| **Results** | Mainly for strategic decision making | Dashboards and predictive measures |

NOTE : We can say that Data Mining need not be depended on Big Data as it can be done on the small or large amount of data but big data surely depends on Data Mining because if we are not able to find the value/importance of a large amount of data then that data is of no use.

## General Statistics…2019

* 95 percent of businesses need to manage unstructured data. (**Forbes**, 2019)
* 40 percent of businesses say they need to manage unstructured data on a frequent basis. (**Forbes**, 2019)
* The big data industry will be worth an estimated $77 billion by 2023. (**Entrepreneur**, 2019)
* 90 percent of IT professionals plan to increase spending on BI tools. (**Forbes**, 2019)

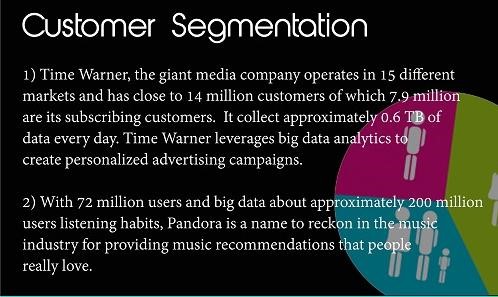
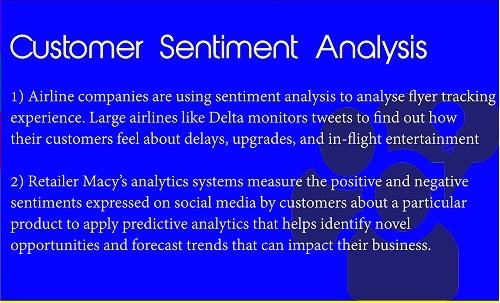
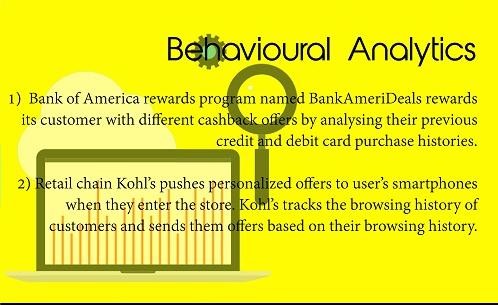
•

* The number of IT professionals using descriptive and predictive analytics grew from the mid-40th percentile to high 60th percentile between January 2018 and January 2019. (**Forbes**, 2019)
* Businesses that use big data saw a profit increase of 8–10 percent. (**Entrepreneur**, 2019)
* Businesses that use big data saw a 10 percent reduction in overall cost. (**Entrepreneur**, 2019)
* The amount of data generated each second in the banking sector will grow 700 percent by 2020. (**The Hill**, 2019)

Companies do data mining and make their business better

Data mining is used by companies to increase revenue, decrease costs, identify customers, provide better customer service, listen to what others are saying and do competitive intelligence.

### AMAZON DELTA MACY’s MCDONALD’S NETFLIX NIKE BANK OF AMERICA



### Some DM tools and popular open source initiatives

In big data mining and analysis, some tools and popular open source initiatives are as follows:

* Apache Mahout
* MOA
* R
* GraphLab
* Rapid-I Rapidminer



* KNIME
* Weka/Pentaho
* Orange

Some DM tools and popular

licensed initiatives

In big data mining and analysis, some tools and popular Licensed are as follows:

* Sisense
* SQL Server Data Tools
* Oracle Data Mining

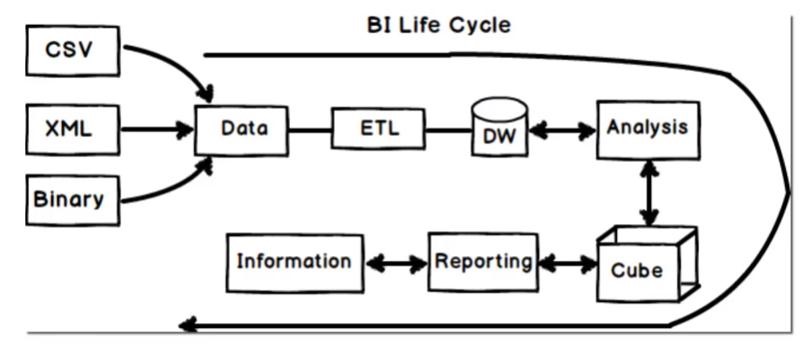
* IBM Cognos
* IBM SPSS Modeler
* SAS Data Mining
* Teradata
* Board

Microsoft SQL Server Data

Tools(MSDT)

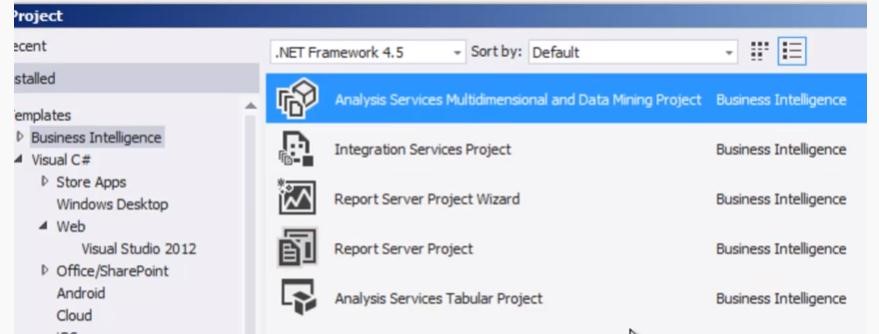
* Microsoft SQL Server Data Tools is introduced with SQL Server 2012 and is the successor of the Business Intelligence Development Studio (BIDS). It is a, Visual Studio based, integrated development environment, used for all your BI-related development.

### BI Life Cycle



Templates while creating a MSBI

Project



Each Template serves some section of the BI Life Cycle

• Download • SQL Server 2014 –Enterprise edition – Actual SQL Server RDBMS(Create D/B, creating tables) • SQL Server Data Tools 2013(will be used to create a BI Project)

### ETL Process

* **ETL** is short for ***e****xtract,* ***t****ransform,* ***l****oad*, three database functions that are combined into one tool to pull data out of one database and place it into another database.

* **Extract** is the process of *reading data* from a database. In this stage, the data is collected, often from multiple and different types of sources.

* **Transform** isthe process of *converting the extracted data* from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.

* **Load** is the process of *writing the data* into the target database.

#### **ETL and Business Intelligence**

• ETL is an important part of today's business intelligence (BI) processes and systems. It is the IT process from which data from disparate sources can be put in one place to programmatically analyze and discover business insights.

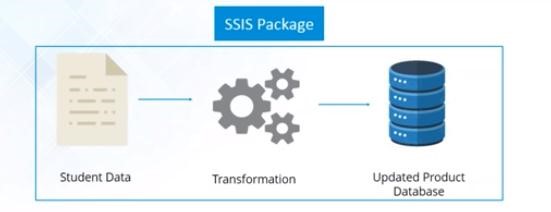
#### **How ETL Works**

• Data from one or more sources is *extracted and then copied* to the data warehouse. When dealing with large volumes of data and multiple source systems, the data is consolidated. ETL is used to migrate data from one database to another, and is often the specific process required to load data to and from data marts and data warehouses, but is a process that is also used to to large convert (transform) databases from one format or type to another.

### SSIS Package



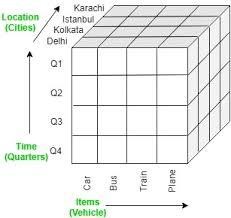
### Perform ETL using SSIS



From Tables and Spreadsheets to Data Cubes

* A data warehouse is based on a multidimensional data model which views data in the form of a data cube
* A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  + Dimension tables, such as item (item\_name, brand, type), or time(day, week, month, quarter, year)
  + Fact table contains measures (such as dollars\_sold) and keys to each of the related dimension tables
* In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

### Data Cube



## Cube: A Lattice of

Cuboids



a

l

l

t

i

m

e

i

t

em

loca

t

ion

sup

p

l

i

er

**time,location**

**ti**

**m**

**e,suppl**

**i**

**er**

**ite**

**m**

**,loc**

**at**

**ion**

**ite**

**m**

**,suppl**

**i**

**er**

**time,item,location**

**time,item,supplier**

**time,location,supplier**

**item,location,supplier**

**time, item,**

**location,**

**supplier**

0

-

D(apex)

cuboid

1

-

D

cuboids

**location,supplier**

2

-

D

cuboids

3

-

D

cuboids

4

-

D(base)

cuboid

**time,item**

### OLAP Operations

\*\*Roll up (drill-up): summarize data

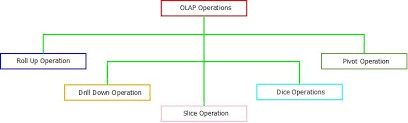
– *by climbing up hierarchy or by dimension reduction*

*\*\** Drill down (roll down): reverse of roll-up

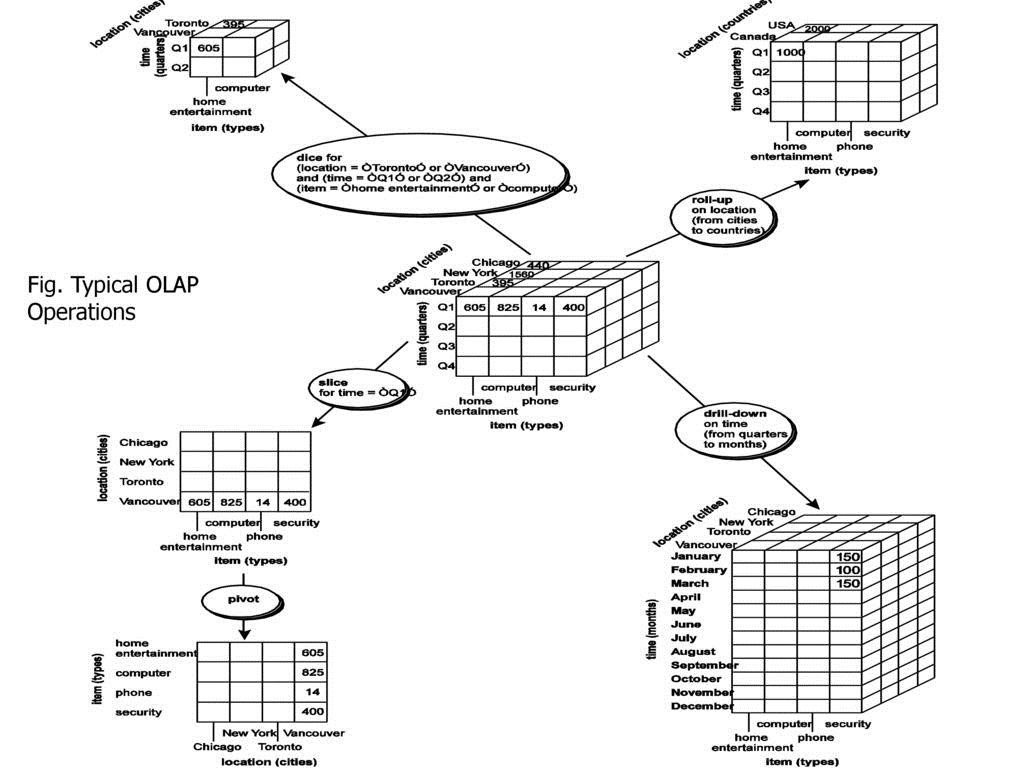
* *from higher level summary to lower level summary or detailed data, or introducing new dimensions*  \*\*Slice and dice: *project and select*  Pivot (rotate):
* *reorient the cube, visualization, 3D to series of 2D planes*

*\*\** Other operations

– *drill across: involving (across) more than one fact table* – *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*



### OLAP Operation



### Dimensional Modeling

* **Dimensional Modeling (DM)** is a technique optimized for data storage in a Data warehouse.
* The purpose of dimensional modeling is to optimize the database for faster retrieval of data.
* The concept of Dimensional Modelling was developed by Ralph Kimball and consists of “fact” and “dimension” tables.

Dimensional Model vs. Relational

Model

* A dimensional model in data warehouse is designed to read, summarize, analyze numeric information like values, balances, counts, weights, etc. in a data warehouse. In contrast, relation models are optimized for addition, updating and deletion of data in a real-time Online Transaction System.
* For instance, in the relational mode, normalization and ER models reduce redundancy in data. On the contrary, dimensional model in data warehouse arranges data in such a way that it is easier to retrieve information and generate reports.
* Dimensional models are used in data warehouse systems and not a good fit for relational systems.

### Few Terminologies

* **Fact :** Facts are the measurements/metrics or facts from business process. For a Sales business process, a measurement would be quarterly sales number.
* **Dimension :** It provides the context surrounding a business process event. In simple terms, they give who, what, where of a fact. In the

Sales business process, for the fact quarterly sales number, dimensions would be

* Who – Customer Names , Where – Location , What – Product Name
* **Attributes :** The Attributes are the various characteristics of the dimension in dimensional data modeling. In the Location dimension, the attributes can be State, Country, Zipcode etc.

### Fact Table vs. Dimensional Table

* **Fact Table**
* A fact table is a primary table in dimension modeling.
* A Fact Table contains
* Measurements/facts
* Foreign key to dimension table
* **Dimension Table**
* A dimension table contains dimensions of a fact.
* They are joined to fact table via a foreign key.
* Dimension tables are de-normalized tables.
* The Dimension Attributes are the various columns in a dimension table
* Dimensions offers descriptive characteristics of the facts with the help of their attributes
* No set limit is given for number of dimensions
* The dimension can also contain one or more hierarchical relationships

### Life Cycle of Dimensional Modeling

Selection of Business Process (

e.g

)

Marketing,Sales,HR

Declare the grain. It

describes the level of

detail for the business problem/solution

Identify the Dimension.

Dimensions are nouns

like date, store, inventory, etc.

Identify Facts.

Most of the fact table rows

are numerical values like price or cost per

unit,

etc

Build the Schema(Star/

SnowFlake

/Fact

Constellation)

### Conceptual Modeling of

Data Warehouses

* Modeling data warehouses: dimensions & measures
  + Star schema: A fact table in the middle connected to a set of dimension tables
  + Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  + Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

The Conceptual Data Model

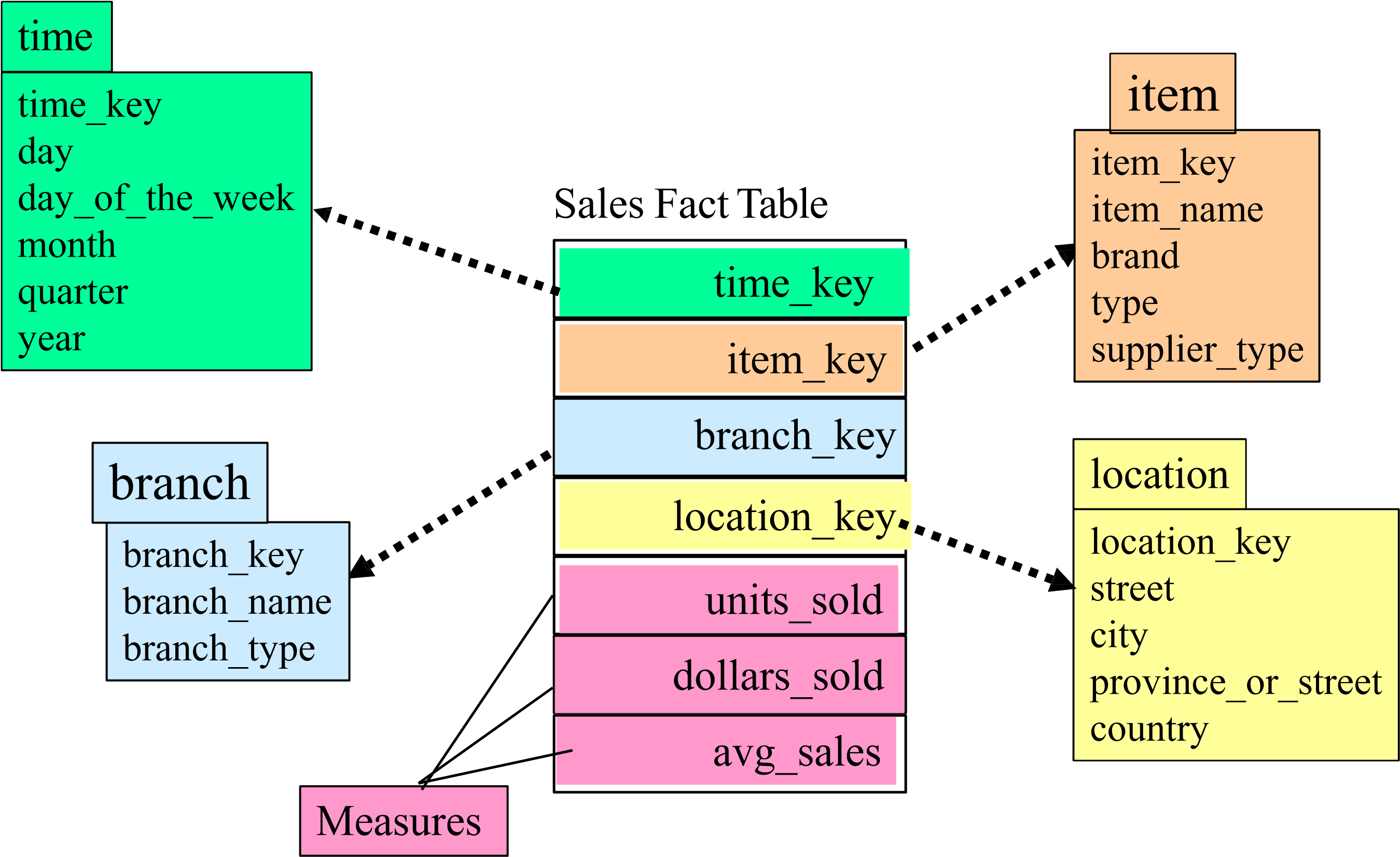
* A conceptual data model is used to describe entities and their relationships.
* High-level business users, such as executive managers, can comprehend the model diagram.

### Star Schema

* **Star Schema** in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables.
* It is known as star schema as its structure resembles a star.
* The Star Schema data model is the simplest type of Data Warehouse schema.
* It is also known as Star Join Schema and is optimized for querying large data sets.
* **Characteristics of Star Schema:**

* Every dimension in a star schema is represented with the only onedimension table.
* The dimension table should contain the set of attributes.
* The dimension table is joined to the fact table using a foreign key
* The dimension table are not joined to each other
* Fact table would contain key and measure
* The Star schema is easy to understand and provides optimal disk usage.
* The dimension tables are not normalized.
* The schema is widely supported by BI Tools

### Star Schema

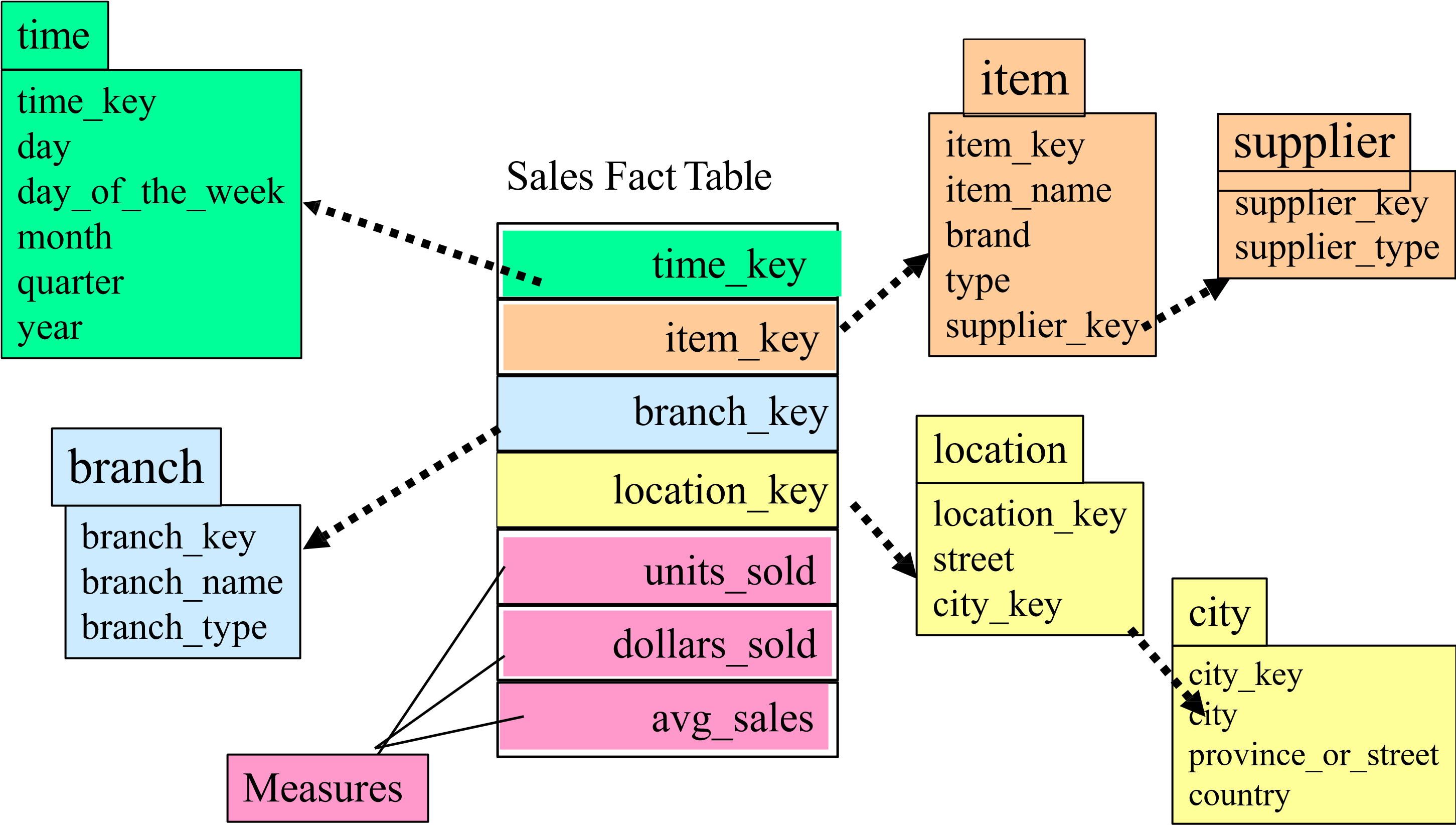


### Fact Constellation/ Galaxy Schema

* **Galaxy Schema** contains two fact table that share dimension tables between them. It is also called Fact Constellation Schema. The schema is viewed as a collection of stars hence the name Galaxy Schema.
* **Characteristics of Galaxy Schema:**
* The dimensions in this schema are separated into separate dimensions based on the various levels of hierarchy.
* For example, if geography has four levels of hierarchy like region, country, state, and city then Galaxy schema should have four dimensions.
* Moreover, it is possible to build this type of schema by splitting the onestar schema into more Star schemes.
* The dimensions are large in this schema which is needed to build based on the levels of hierarchy.
* This schema is helpful for aggregating fact tables for better understanding.

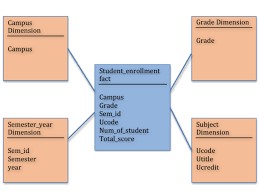
### Snowflake Schema

* **Snowflake Schema** in data warehouse is a logical arrangement of tables in a multidimensional database such that the ER diagram resembles a snowflake shape.
* A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions.
* The dimension tables are normalized which splits data into additional tables.
* **Characteristics of Snowflake Schema:**
* The main benefit of the snowflake schema it uses smaller disk space.
* Easier to implement a dimension is added to the Schema
* Due to multiple tables query performance is reduced
* The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

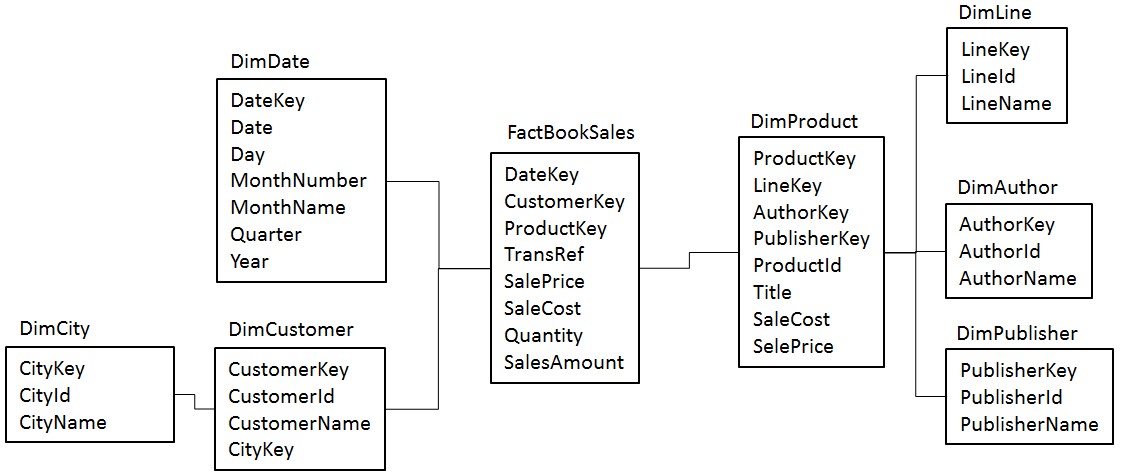
Snowflake Schema

### Fact Constellation

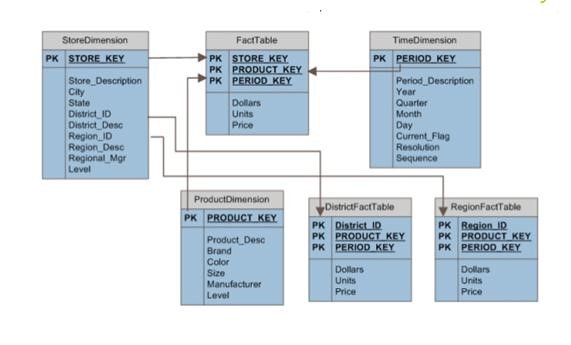
Example – Star Schema for a Students enrollment system



Example – Snowflake Schema for a Book Sales system



Example – Fact Constellation for a Store – Product system



### Types of Dimensional Tables

Dimensional

Tables

**Slowly**

**Changing**

**Dimensions**

**Degenerate**

**Dimension**

**Junk**

**Dimension**

**Conformed**

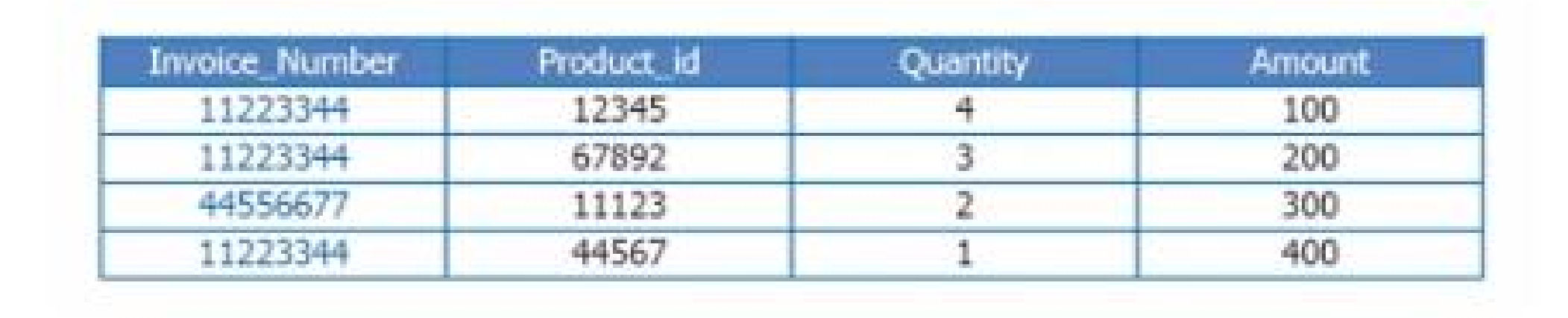
**Dimension**

**Role play**

**dimension**

* **Slowly Changing Dimensions**– Dimension attributes that change slowly over a period of time rather than changing regularly is grouped as SCDs. Attributes like name, address can change but not too often.
* These attributes can change over a period of time and that will get combined as a slowly changing dimension. Consider an example where a person is changing from one city to another. Now there are 3 ways to change the address;
* Example - Type 1 is to over write the old value, Type 2 is to add a new row and Type 3 is to create a new column.

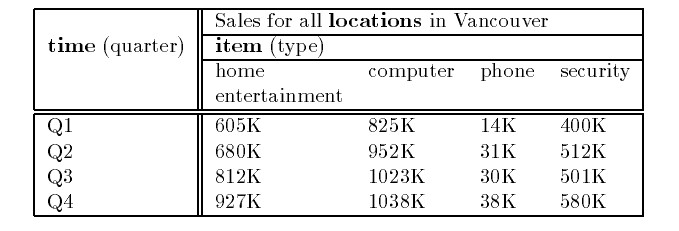


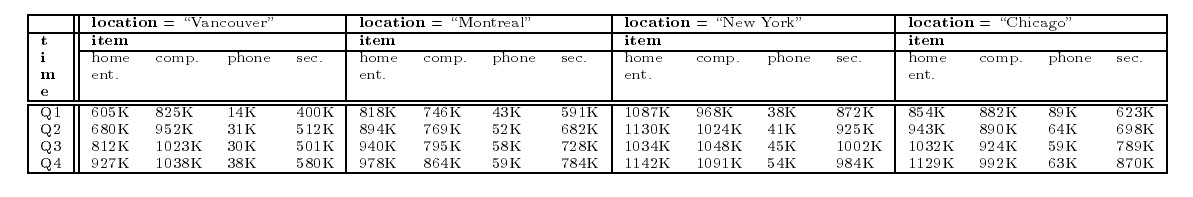
* **Conformed Dimension-** This is used in multiple locations**.** It helps in creating consistency so that the same can be maintained across the fact tables. Different tables can use the table across the fact table and it can help in creating different reports.
* For example, there are two fact tables. Fact table 1 is to determine the number of products sold by geography. This table will calculate just the number of products by geography and fact table 2 will determine the revenue generated by customer. Both are dependent on the product which contains product Id, name and source.
* There is the geography dimension and customer dimension which are being shared by two fact tables. The revenue fact gives the revenue generated by both the geography and the customer, while the product units fact gives number of units sold in the geography to a customer.
* **Degenerate Dimension**– A degenerate dimension is when the dimension attribute is stored as part of the fact table and not in a separate table. A degenerate dimension is when the dimension attribute is stored as part of fact table, and not in a separate dimension table. These are essentially dimension keys for which there are no other attributes. Product id comes from product dimension table. Invoice number is a standalone attribute and has no other attributes associated with it. An invoice number can be crucial since the business would want to know the quantity of the products.

* **Junk Dimension**– It is a single table with a combination of different and unrelated attributes to avoid having a large number of foreign keys in the fact table. They are often created to manage the foreign keys created by rapidly changing dimensions.

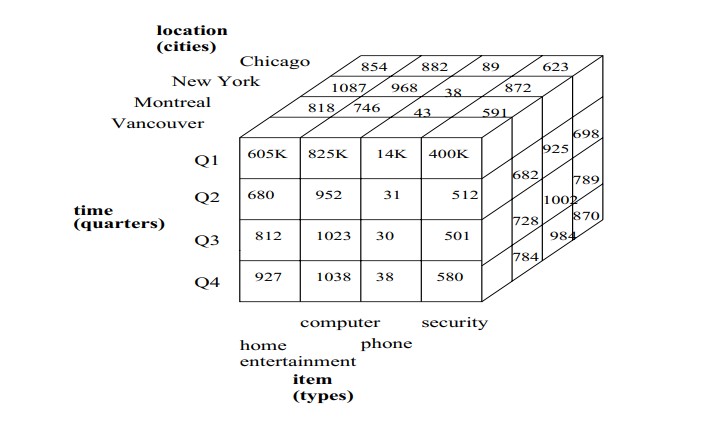
For example, attributes such as flags, weights, BMI (body mass index) etc.

* **Role play dimension**– A role-playing dimension is one where the same dimension key — along with its associated attributes — can be joined to more than one foreign key in the fact table. For example, a fact table may include foreign keys for both ship date and delivery date. But the same date dimension attributes apply to each foreign key, so you can join the same dimension table to both foreign keys. Here the date dimension is taking multiple roles to map ship date as well as delivery date, and hence the name of role playing dimension.
* A 2-D view of sales data for sales according to the dimensions time and item, where the sales are from branches located in the city of Vancouver. The measure displayed is dollars sold.



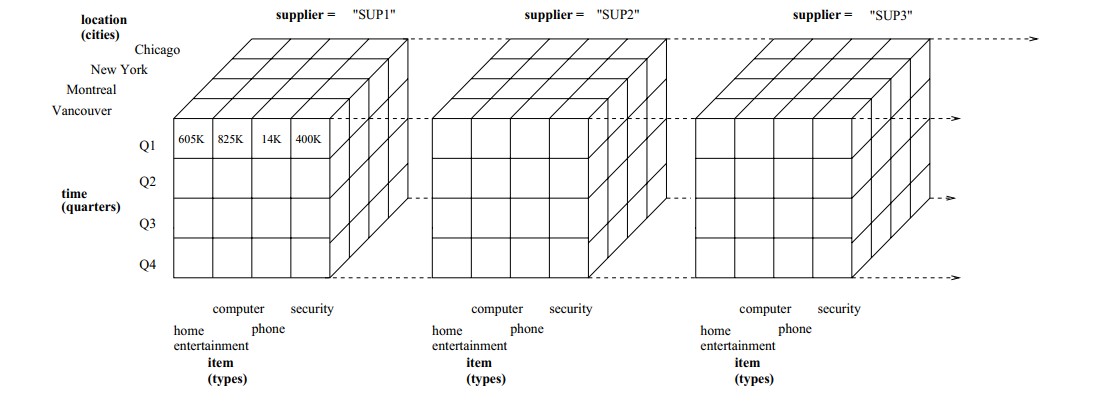
* A 3-D view of sales data for Sales, according to the dimensions time, item, and location. The measure displayed is dollars sold.

* A 3-D data cube representation of the data in Table of Example 2 according to the dimensions time, item, and location. The measure displayed is dollars sold.



### Examples -4

• A 4-D data cube representation of sales data, according to the dimensions time, item, location, and supplier. The measure displayed is dollars sold.



### Measurements in BI

* It is easy to claim that BI helps people make better decisions. For all but the simplest business process, improving decision support involves more than delivering a report with graphs and gauges.
* Measuring business performance has a long tradition in companies, and it can be useful in the case of BI to perform activities such as determining the actual value of BI to a company or to improve and manage existing BI processes.

Measurement system terminology

* Data – Collection of Facts which have similar attributes or characteristics.(e.g.) Phone Number , email id etc.
* Measure – Data with associated unit of measure(UOM).(e.g.) Lab hours per month, time duration.
* Metric – It is a system of measures based on standard UOM with a business context. (e.g) Product defect rate, employee attrition rate.
* Indicator – It is a business metric used to track business result or success/performance.(e.g.) Call drop frequency, earnings per share
* Index – It consists of a composite set of indicators used to address the overall health of the business operation.(e.g) customer satisfaction index rated on a scale of 1 to 5.

### 4 components of metric data

* Subject – This measure is about a customer, employee, supplier etc.
* Quantum – It is the value of the measure.(e.g) cost , frequency, duration, amount.
* Stratum – It is the grouping consideration expressed by location, customer, quarter etc.
* Application – Value compared with similar measurement.(e.g) Previous month forecast or target.
* Cell phone cost in Asia Pacific region is USD 100 against target of USD 75

### Key Performance Indicators (KPI)

* **Key performance indicators (KPIs) are a set of performance measurements that demonstrate how effectively an organization is achieving key objectives.**
* KPIs not only provide an organization with a focus for strategic and operational improvement, but a way to compare achievements to similar organizations.
* The KPIs in business intelligence are typically around major business areas such as financial metrics, marketing metrics, customer service metrics and HR metrics. KPIs also cover specific areas like project management metrics and retail metrics.
* To define your KPIs for successful business intelligence, you have to identify which aspects of your business you want to look into and have all the relevant data in your hand.
* Investing in a powerful business intelligence tool should let you calculate your KPIs faster, with many of them capable of diving deep into your data and generating novel insights to inform your next decisions and actions.
* Remember :
* Not everything that can be counted counts….and not everything that counts can be counted.
* To be effective, a KPI must be:

* Well-defined and quantifiable.
* Communicated throughout your organization and department.
* Crucial to achieving your goal. (Hence, key performance indicators.)
* Applicable to your Line of Business (LOB) or department.

* How Do I Determine Which KPIs To Use?

* There are a lot of KPIs to choose from, and you'll want to narrow down the list so you're only tracking the KPIs that will truly help drive your strategy forward.
* The right KPIs for you might not be the right KPIs for another organization.

### KPI usage in companies

* KPIs could be used in the company at strategic, tactical and operational level. There are techniques like Cause and effect modeling, Goal Question Metrics and measure (GQMM), Balanced

Scorecard(BSC), Six Sigma, Total Quality Management (TQM), Economic Value Add(EQA) Management framework to represent KPIs to define, align, track and communicate strategy.

Few Financial Metrics

* **Profit**: Analyze both gross and net profit margin to better understand how successful your organization is.
* **Cost**: Measure cost effectiveness and find the best ways to reduce and manage your costs.
* **LOB Revenue Vs. Target**: This is a comparison between your actual revenue and your projected revenue.
* **Day Sales Outstanding (DSO)**: Take your accounts receivable and divide them by the number of total credit sales. Take that number and multiply it by the number of days in the time frame you are examining.
* **Sales By Region**: Through analyzing which regions are meeting sales objectives, you can provide better feedback for underperforming regions.
* **LOB Expenses Vs. Budget**: Compare your actual overhead with your forecasted budget.

Few Financial Metrics

* **Cash Flow From Financing Activities**: This metric demonstrates an organization’s financial strength. Formula: (Cash Received from Issuing Stock or Debt) – (Cash Paid as Dividends and Reacquisition of Debt/Stock) = (Cash Flow from Financing Activities).
* **Average Annual Expenses To Serve One Customer**: This is the average amount needed to serve one customer. Formula: (Total Expenses) / (Total Customers) = (Average Annual Expenses to Serve One Customer).
* **(Customer Lifetime Value) / (Customer Acquisition Cost)**: The ratio of customer lifetime value to customer acquisition cost should ideally be greater than one, as a customer is not profitable if the cost to acquire is greater than the profit they will bring to a company. Formula: (Net Expected Lifetime Profit from Customer) / (Cost to Acquire Customer).

Few Process Metrics

* **Customer Support Tickets**: Analysis of the number of new tickets, the number of resolved tickets, and resolution time will help you create the best customer service department in your industry.
* **Percentage Of Product Defects**: Take the number of defective units and divide it by the total number of units produced in the time frame you’re examining. This will give you the percentage of defective products. Clearly, the lower you can get this number, the better.
* **LOB Efficiency Measure**: You can measure your organization’s efficiency by analyzing how many units you have produced every hour, and what percentage of time your plant was up and running.

Few Customer Metrics

* **Customer Lifetime Value (CLV)**:CLV helps you look at the value your organization is getting from a long-term customer relationship. Use this performance indicator to narrow down which channel helps you gain the best customers for the best price.
* **Customer Acquisition Cost (CAC)**: Divide your total acquisition costs by the number of new customers in the time frame you’re examining.
* **Customer Satisfaction & Retention**: You can use multiple performance indicators to measure CSR, including customer satisfaction scores and percentage of customers repeating a purchase.
* **Net Promoter Score (NPS)**: To determine your NPS score, send out quarterly surveys to your customers to see how likely it is that they’ll recommend your organization to someone they know.
* **Number Of Customers**: By determining the number of customers you’ve gained and lost, you can further understand whether or not you are meeting your customers’ needs.

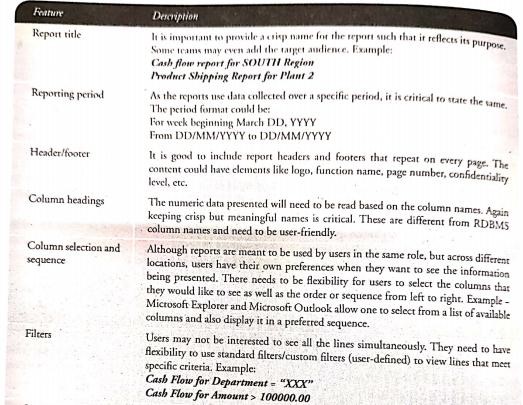
Few People Metrics

* **Employee Turnover Rate (ETR)**: To determine your ETR, take the number of employees who have departed the company and divide it by the average number of employees.
* **Percentage Of Response To Open Positions**: When you have a high percentage of qualified applicants apply for your open job positions, you know you are doing a good job maximizing exposure to the right job seekers.
* **Employee Satisfaction**: Happy employees are going to work harder—it’s as simple as that. Measuring your employee satisfaction through surveys and other metrics is vital to your departmental and organizational health.

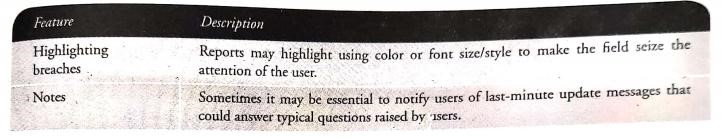
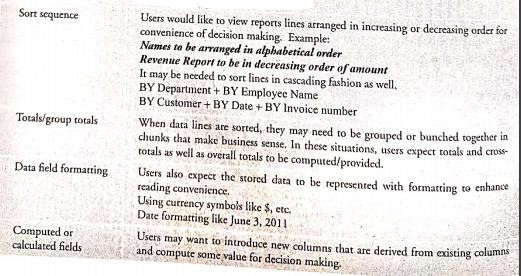
### Basics of Enterprise Reporting

* Reporting is the integral part of OLTP application.
* Our focus is on OLAP centric Reporting.
* This class of enterprise reporting will objectively communicate facts related to strategy, corporate/department performance against plan, status of critical initiatives and metrics that matter to the stakeholders.
* These reports help leaders align their business activities to the vision and strategies of their enterprise and to monitor their performance against the organizational goals.

### Features of Good Reporting



### Features of Good Reporting



The Balanced Scorecard

Definition:

The Balanced Scorecard is a management tool that provides stakeholders with a comprehensive measure of how the organization is progressing towards the achievement of its strategic goals.

It was originated by Dr. Robert Kaplan (Harvard Business School) and David Norton as a performance measurement framework that added strategic non-financial performance measures to traditional financial metrics to create a more ‘balanced’ view of organizational performance.

The Balanced Scorecard What is it?

The Balanced Scorecard:

* Balances financial and non-financial measures

* Balances short and long-term measures

* Balances performance drivers (leading indicators) with outcome measures (lagging indicators)

* Should contain just enough data to give a complete picture of organizational performance… and no more!

* Leads to strategic focus and organizational alignment.
* Four strategic perspectives are addressed within the Balanced Scorecard framework:
* Customer
* Financial
* Internal Processes – commonly includes technology, systems, etc.
* Learning and Growth (aka “Organization Capacity”) – commonly includes people, training, etc.

* Balanced Scorecard dashboards include both leading and lagging indicators. For example, customer and financial KPIs are traditionally lagging indicators – the numbers indicate *what has already happened.* KPIs for the two perspectives of internal processes and learning/growth are leading indicators. This is because positive results achieved with respect to internal processes and learning/growth initiatives should lead to a positive result in the customer and financial KPIs.

Balanced Scorecard Performance Measures

**Internal Business Process Perspective**



**Customer Perspective**

New Hire Termination %

Quality of Hire

Cycle Time to Hire

Voluntary Turnover %

Employee Satisfaction Score



**Financial Perspective**

Non

-

productive Labor Cost %

Average Cost per Claim

Cost per Hire

HR Expenses per 100 FTE

Overtime

Pay Ratio

Bonus Ratio %

**OUTCOME MEASURES**

HR Headcount per 100 FTE

Unscheduled Overtime %

Lost-time injuries per 100 FTE

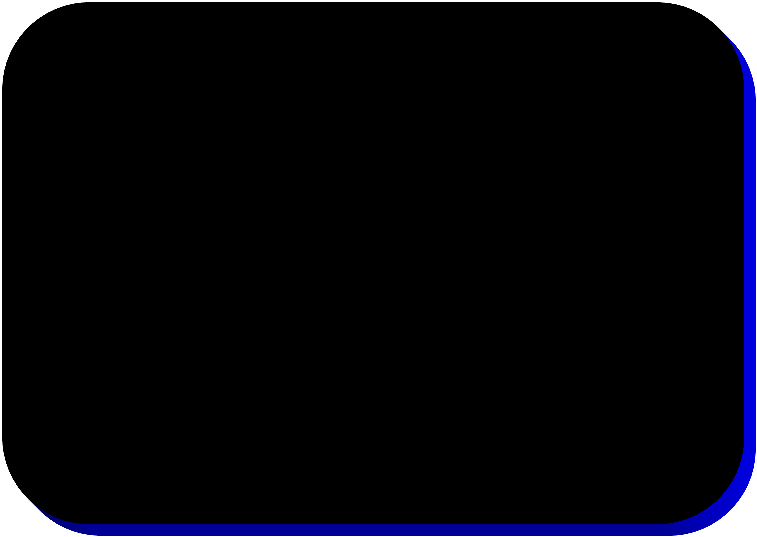
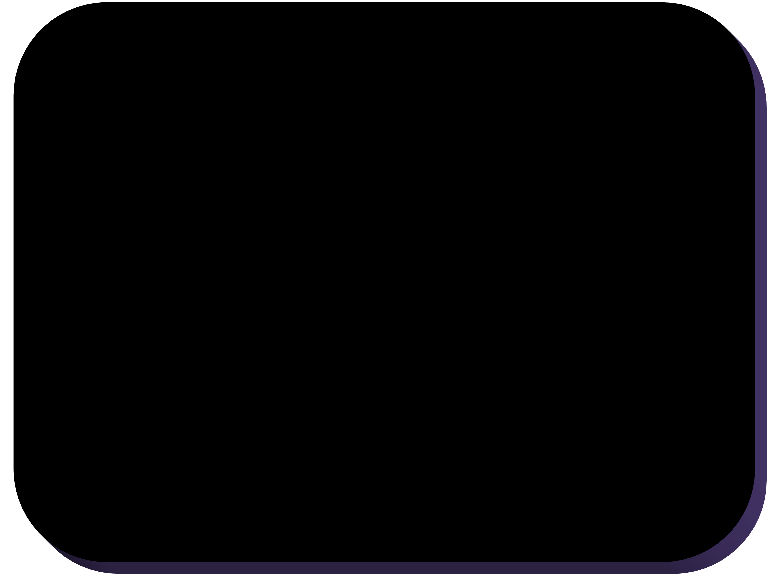
Self-Service Participation %

Scheduling Effectiveness %

**Learning & Growth Perspective**

Training Hours/FTE

Productivity Gain per Training Hour



**ACTIVITY MEASURES**

Career Path Ratio

Improving Performer Ratio

Position Tenure

Skill Match Ratio

FTE –Full Time Equivalent

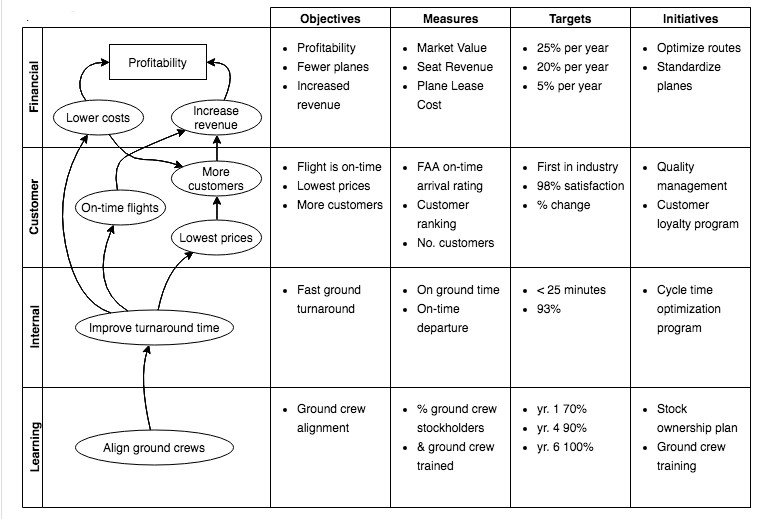
### The Contents of a Scorecard

* Scorecards usually contain some or all of the following elements:
  + Key Performance Indicators (KPIs)
  + KPI actual values compared to historical values (for trend analysis)
  + KPI actual values compared to a forecast or budget amount
  + Rankings of different departments, locations, products, and so forth

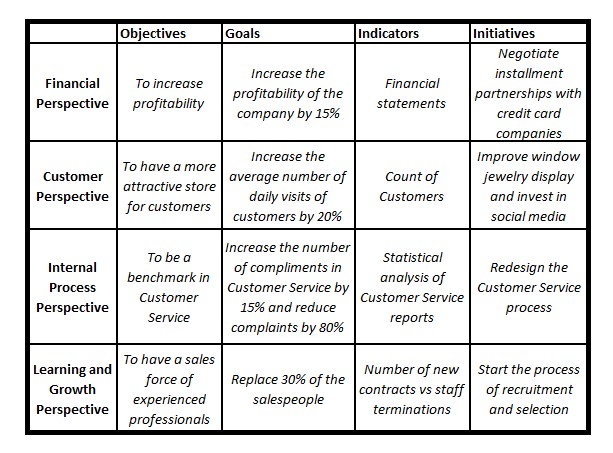
Strategy Map with Balanced Scorecard

* The Balance score card was first plotted as a four box model. This model has evolved since then and is now plotted as a strategy map.
* A strategy map is a diagram that is used to document the primary strategic goals being pursued by an organization or management team developed by Robert S. Kaplan and David P. Norton in 1996.
* The map places the four balanced scorecard perspective into a causal hierarchy. The causal relationship shows that the objectives support each other and that delivering the right performance in the lower perspective will help achieve the objectives in the upper perspective.
* A strategy map is typically an element of the documentation associated with the Balanced Scorecard by translating strategy into actions that models the relationship between the drivers and the desired outcomes.
* Strategy Map can help us to:
* Increase Focus On Strategy And Results
* Improve Organizational Performance By Measuring What Matters
* Align The Work People Do On A Day-To-Day Basis With Strategy
* Focus On The Drivers Of Future Performance
* Improve Communication Of The Organization’s Vision And Strategy
* Prioritize Action Items In The Implementation Roadmap In Tough Economic Times

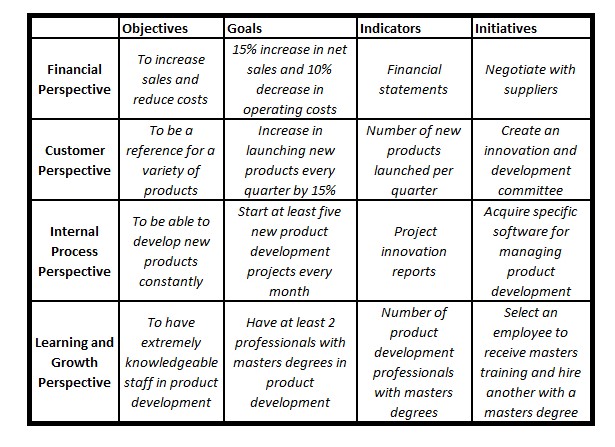
Combining the Power of Strategic Map and Balanced Scorecard



Balanced Scorecard example: Strategic map for a Jewelry store



Balanced Scorecard example: Strategic map for an E-Commerce Business



#### **Data Dashboard**

* A **data dashboard** is an information management tool that visually tracks, analyzes and displays **key performance indicators (KPI)**, metrics and key data points to monitor the health of a business, department or specific process. They are customizable to meet the specific needs of a department and company.
* Behind the scenes, a dashboard connects to your files, attachments, services and API’s, but on the surface displays all this data in the form of tables, line charts, bar charts and gauges.
* A data dashboard is the most efficient way to track multiple data sources because it provides a central location for businesses to monitor and analyze performance.
* Real-time monitoring reduces the hours of analyzing and long line of communication that previously challenged businesses.

### Examples



Importance of Dashboards for Enterprises

Improves

Corporate Dashboards  Accountability and

Transparency across organizations

Leads to

Improved

Better

Compliance Decision Making

#### **How are data dashboards used in business intelligence**

• Dashboards are a **data**  **visualization tool** that allow all users to understand the analytics that matter to their business, department or project. Even for non-technical users, dashboards allow them to participate and understand the analytics process by compiling data and visualizing trends and occurrences. Data dashboards provide an objective view of **performance metrics** and serve as an effective foundation for further dialogue. A dashboard is a **business intelligence tool** used to display data visualizations in a way that is immediately understood.

### Types of Dashboards

Enterprise

Performance

Dashboards

•

Sales Revenue

•

Compliance or regulatory data

•

Business Unit KPIs

Customer Support

Dashboards

•

Online trading

•

Utility Services

•

B2B SLA

Divisional

Dashboards

•

Purchasing Dashboards

•

Supply Chain Dashboards

•

Operational Dashboards

•

Marketing Dashboards

#### **How are balanced scorecards and dashboards different**

* The difference between scorecards and dashboards comes from the context in how they are applied. Hence Scorecards and dashboards are not contradictory.
* They both display measurements, but they serve different purposes.
* Scorecards are intended to be strategic. Dashboards are intended to be operational.
* KPIs should be displayed in a balanced scorecard, and PIs should be reported in dashboards.
* Dashboards do not communicate why something matters, why someone should care about the reported measure or what the impact may be if an undesirable declining measure continues. In short, dashboards report what can be measured.
* Scorecards do not start with the existing data, but rather they begin with identifying what strategic projects to complete and core processes to improve and excel in.

##### **Scorecards monitor the progress toward accomplishing strategic objectives**

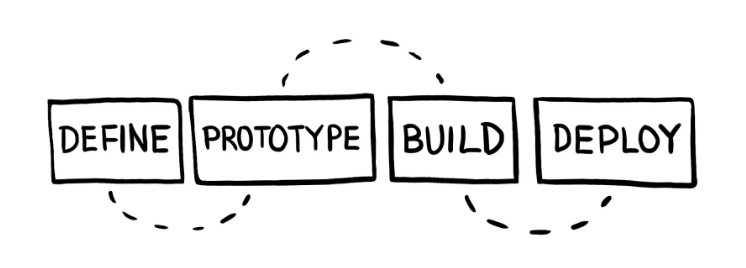
* A scorecard displays periodic snapshots of performance associated with an organisation’s strategic objectives and plans.
* Directionally upward from the employee-centric innovation, learning and growth perspectives, KPIs should reveal the cumulative build of potential to realised economic value.
* There are two key distinctions of scorecards: First, each KPI must require a predefined target measure. Second, KPIs should include both project-based KPIs such as milestones, progress percentage of completion and degree of planned versus accomplished outcome, as well as process-based KPIs such as customer satisfaction and per cent on-time delivery against customer promise dates.
* Process improvement is important, but innovation and change is even more important. Strategy is all about change and not just doing the same things better.

#### **Dashboards monitor and measure processes and outputs**

* A dashboard is operational and reports information typically more frequently than scorecards.
* The organisation’s traction and torque are reflected in the dashboard’s PI measures. PIs serve more to monitor trends across time or results against upper- or lower-threshold boundary limits.
* As PIs are monitored and responded to, the corrective actions contribute to achieving the KPI target levels with actual results.
* However, each dashboard measure is reported with little regard to its relationship to other dashboard measures.
* Dashboard measures do not directly reflect the context of strategic objectives.
* Dashboard information can be more real-time in nature, like an automobile dashboard that lets drivers check their current speed, fuel level and engine temperature at a glance.
* It follows that a dashboard should ideally be linked directly to systems that capture events as they happen, and it should warn users through alerts or exception notifications when performance against any number of metrics deviates from the norm or what is expected.

### Steps for creating Dashboards

• **Dashboard Design Process**



#### **1. Define**

* Having clarity about who this dashboard is for and what metrics matter to them is critical to creating a dashboard that will be used.

* **Stakeholders**
* There are 4 main stakeholders
* The designer (you)
* The audience (who will be viewing this dashboard)
* The point person (the member of the audience who has the most experience)
* The Data Gatekeeper (member of the data team who will help with the database)
* **Metrics**
* You will work with the point person to go from decisions that need to be made to metrics that can be queried and tracked.

#### **2. Prototype**

* We need to figure out how to best display them so that the dashboard is useful to the whole audience.
* **Visualizations**
* Use visualizations that present the metrics clearly and accurately. Even when sketching and prototyping graphs, making the right visualization decisions here will improve the prototype and the feedback loop.
* **Sketching and Iteration**
* At this stage, it is recommended that the visualizations and dashboards be sketched out on paper or using a tool that is not connected to any real data.

#### **3. Build**

* Once we are satisfied with the prototype we have to create the dashboard using real data.
* **Find the Data**
* Many challenges can arise at this point. Where is the data stored? Is the data messy? Do we even have the data available? Working with the data team and the Data Gatekeeper is critical to navigating this step.
* **Build Metrics/Dashboard**
* We need to create queries to power the metrics, create formulas, and transform the data into charts. Using a framework to log the metrics, formulas, and data sources makes creating queries much easier.

#### **4.Deploy**

* Finally, we have a fully functioning dashboard. Now we need to share it with the full audience. We should enhance the dashboard to make it more effective at scale and we need to make sure to maintain it as usage grows and changes.
* **Sharing**
* The audience will have varying levels of data literacy and context for the data presented in the dashboard. You need to verify that you have enough context within the dashboard and that you provide enough training so that people can get insights out of it easily.
* **Scaling**

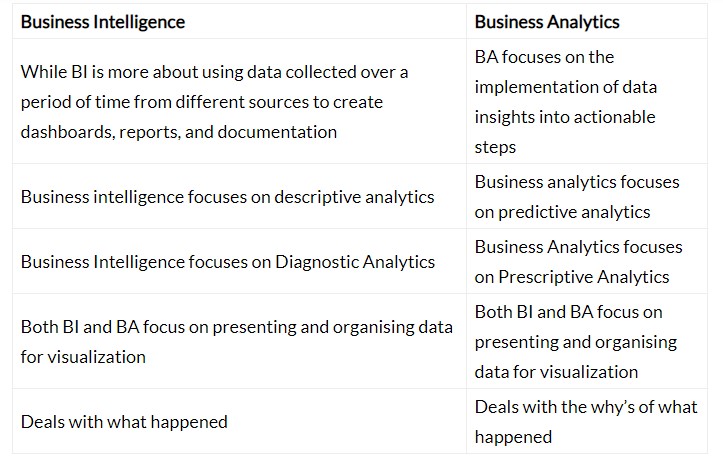
* If the dashboard is useful, the amount of views and number of viewers is likely to grow. Adding links, interactivity, and documentation to a dashboard helps it accommodate more use cases and inspire other dashboard creators. Also as the number of views and viewers increases, spending time optimizing queries becomes an important way of keeping the dashboard useful.
* **Maintenance**
* Datasources, tables, and fields change, and dashboards need to change with them. Setting up scheduled times to review dashboards is critical to keeping them relevant and functional. Providing a way for the audience to alert you about issues will allow you to make informed improvements to the dashboard.

#### **Business Analytics**

• Gartner says, “**Business analytics** is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states. Business analytics includes data mining, predictive analytics, applied analytics and statistics, and is delivered as an application suitable for a business user. These analytics solutions often come with prebuilt industry content that is targeted at an industry business process (for example, claims, underwriting or a specific regulatory requirement).”

##### **Relation Between Business Analytics and Business Intelligence**

* While BI is more about using data collected over a period of time from different sources to create dashboards, reports, and documentation; BA focuses on the implementation of data insights into actionable steps.
* Business intelligence is slightly more generic where you are using data from various sources and then you are analyzing it. That’s also a part of BA but in BI you finally finish the loop by implementing the insights generated from data. So in that sense, BI is slightly all-encompassing.



###### 4 Types of Business Analytics

* **Descriptive Analytics -** It summarizes an organization's existing data to understand what has happened in the past or is happening currently. It makes data more accessible to members of an organization such as the investors, shareholders, marketing executives, and sales managers.
* **Diagnostic Analytics -** This type of Analytics helps shift focus from past performance to the current events and determine which factors are influencing trends. Diagnostic analytics makes use of probabilities, and likelihoods to understand why events may occur. Techniques such as sensitivity analysis, and training algorithms are employed for classification and regression.

###### 4 Types of Business Analytics (contd.)

Predictive Analytics

This type of Analytics is used to forecast the possibility of a future event with the help of statistical models and ML techniques. It builds on the result of descriptive analytics to devise models to extrapolate the likelihood of items. To run predictive analysis, Machine Learning experts are employed. They can achieve a higher level of accuracy than by business intelligence alone.

One of the most common applications is sentiment analysis. Here, existing data collected from social media and is used to provide a comprehensive picture of an users opinion. This data is analyzed to predict their sentiment (positive, neutral or negative).

Prescriptive Analytics

Going a step beyond predictive analytics, it provides recommendations for the next best action to be taken. It suggests all favorable outcomes according to a specific course of action and also recommends the specific actions needed to deliver the most desired result. It mainly relies on two things, a strong feedback system and a constant iterative analysis.

It learns the relation between actions and their outcomes. One common use of this type of analytics is to create recommendation systems.

Classification and Prediction

* Classification is the process of finding a model that describes the data classes or concepts.
* The purpose is to be able to use this model to predict the class of objects whose class label is unknown.
* This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms −

* Classification (IF-THEN) Rules
* Decision Trees
* Mathematical Formulae
* Neural Networks

#### Example Use Cases

* **Use Case – 1**
* **Business Problem:** A bank loan officer wants to predict if the loan applicant will default on a loan, based attributes such as Loan amount, monthly payment installments, employment tenure, number of times delinquent, annual income, debt to income ratio etc. The target variable would be ‘past default status’ and the predicted class would contain values ‘yes or no’ representing ‘whether the applicant is likely to default/unlikely to default’.
* **Business Benefit:** Once classes are assigned, the bank will have a loan applicant dataset with each applicant labeled as “likely/unlikely to default”. Based on these labels, the bank can easily make a decision on whether to give loan to an applicant and how much credit to extend, as well as the interest rate each applicant is eligible for based on the amount of risk involved.

#### Example Use Cases

* **Use Case – 2**
* **Business Problem:** A doctor wants to predict the likelihood of successful treatment of a patient illness based on various attributes such as blood pressure, hemoglobin level, blood sugar, prescription medications, and current and previous treatments. The target variable would be ‘past cure status’ and predicted class would contain values ‘yes or no’ meaning ‘prone to cure/not prone to cure’ respectively.
* **Business Benefit:** Given the patient profile, and current and previous treatments and medications, the doctor can establish a probability of success and make changes in treatments/medications.

##### **Top Algorithms for Classification**

**Decision Trees**

**Logistic Regression**

**Naive Bayes Classification**

**k-nearest neighbors**

**Support Vector Machine**

Process (1): Model Construction

Training

Data

**NAME**

**RANK**

**YEARS**

**TENURED**

Mike

Assistant Prof

3

no

Mary

Assistant Prof

7

yes

Bill

Professor

2

yes

Jim

Associate Prof

7

yes

Dave

Assistant Prof

6

no

Anne

Associate Prof

3

no

Classification

Algorithms

IF rank = ‘professor’

OR years > 6

THEN tenured = ‘yes’

Classifier

)

Model

(

Process (2): Using the Model in Prediction

Classifier

Testing

Data

Unseen Data

(

Jeff, Professor,

4)

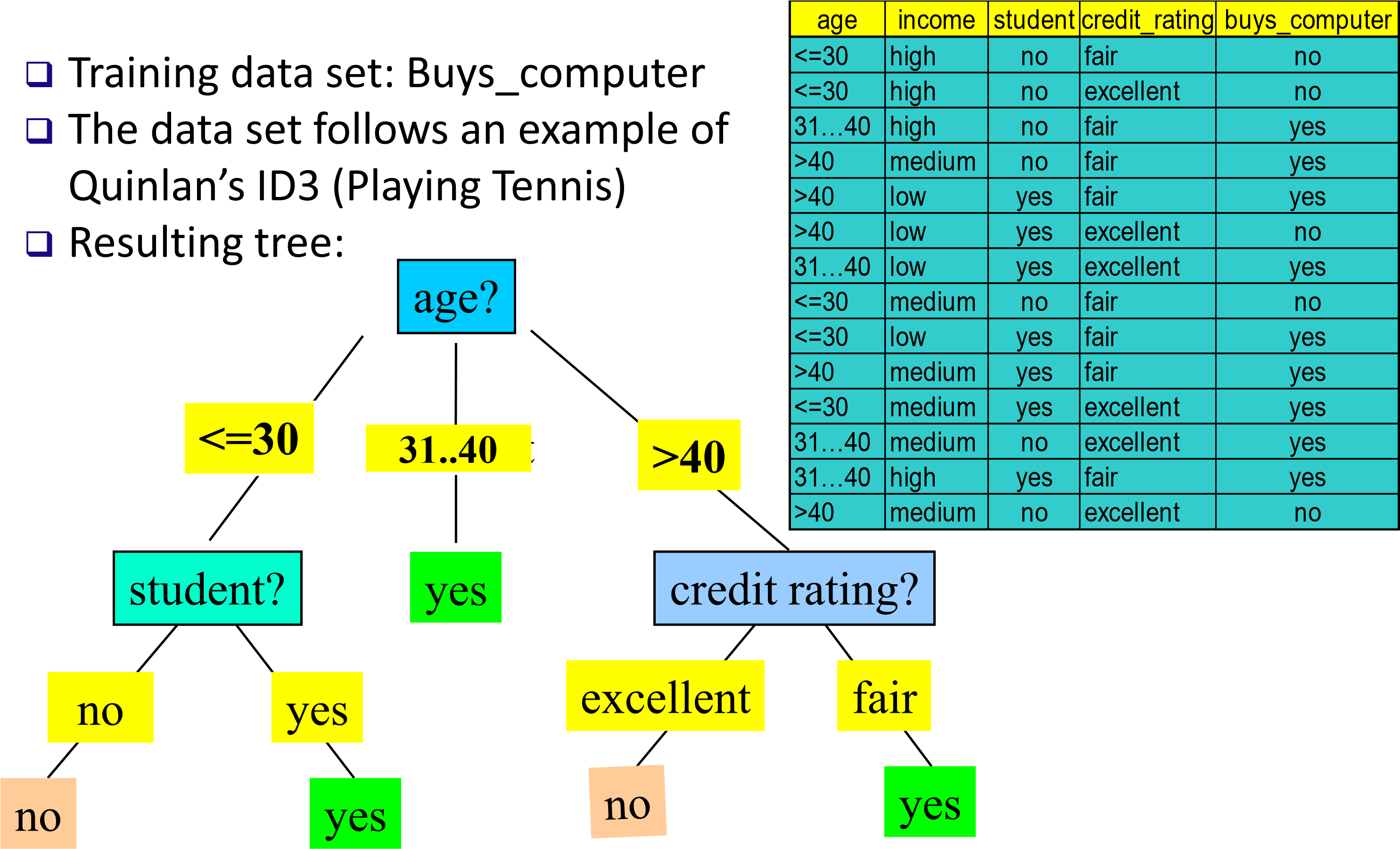
|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **RANK** | **YEARS** | **TENURED** |
| Tom | Assistant Prof | 2 | no |
| Merlisa | Associate Prof | 7 | no |
| George | Professor | 5 | yes |
| Joseph | Assistant Prof | 7 | yes |

Tenured?

**What is a Decision Tree?**

* A Supervised Machine Learning Algorithm, used to build classification and regression models in the form of a tree structure.
* *A decision tree is a tree where each -*

###### • Node - a feature(attribute) • Branch - a decision(rule) • Leaf - an outcome(categorical or continuous)

Decision Tree Induction: An Example

153

#### CART

* The CART or Classification & Regression Trees methodology was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer to the following types of decision trees.
* **Classification Trees**: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.
* **Regression Trees**: where the target variable is continuous and tree is used to predict it's value.
* The representation of a CART model is a binary tree.
* Each node represents a single input variable (x) and a split on that variable.
* The leaf node of the tree contains an output variable (y) which is used for making a prediction.

#### Gini Index

* Gini index is a metric for classification tasks in CART. It stores sum of squared probabilities of each class. We can formulate it as illustrated below.
* Gini = 1 – Σ (Pi)2 for i=1 to number of classes

#### Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temp.** | **Humidity** | **Wind** | **Decision** |
| **1** | Sunny | Hot | High | Weak | No |
| **2** | Sunny | Hot | High | Strong | No |
| **3** | Overcast | Hot | High | Weak | Yes |
| **4** | Rain | Mild | High | Weak | Yes |
| **5** | Rain | Cool | Normal | Weak | Yes |
| **6** | Rain | Cool | Normal | Strong | No |
| **7** | Overcast | Cool | Normal | Strong | Yes |
| **8** | Sunny | Mild | High | Weak | No |
| **9** | Sunny | Cool | Normal | Weak | Yes |
| **10** | Rain | Mild | Normal | Weak | Yes |
| **11** | Sunny | Mild | Normal | Strong | Yes |
| **12** | Overcast | Mild | High | Strong | Yes |
| **13** | Overcast | Hot | Normal | Weak | Yes |
| **14** | Rain | Mild | High | Strong | No |

Implement CART on 14 instances of golf playing decisions based on outlook, temperature, humidity and wind factors

#### For outlook

|  |  |  |  |
| --- | --- | --- | --- |
| **Outlook** | **Yes** | **No** | **Number of instances** |
| **Sunny** | 2 | 3 | 5 |
| **Overcast** | 4 | 0 | 4 |
| **Rain** | 3 | 2 | 5 |

Gini(Outlook=Sunny) = 1 – (2/5)2 – (3/5)2 = 1 – 0.16 – 0.36 = 0.48

Gini(Outlook=Overcast) = 1 – (4/4)2 – (0/4)2 = 0

Gini(Outlook=Rain) = 1 – (3/5)2 – (2/5)2 = 1 – 0.36 – 0.16 = 0.48

Then, we will calculate weighted sum of gini indexes for outlook feature.

Gini(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48 = 0.171 + 0 + 0.171 = **0.342**

#### For Temperature

|  |  |  |  |
| --- | --- | --- | --- |
| **Temperature** | **Yes** | **No** | **Number of instances** |
| **Hot** | 2 | 2 | 4 |
| **Cool** | 3 | 1 | 4 |
| **Mild** | 4 | 2 | 6 |

Gini(Temp=Hot) = 1 – (2/4)2 – (2/4)2 = 0.5

Gini(Temp=Cool) = 1 – (3/4)2 – (1/4)2 = 1 – 0.5625 – 0.0625 = 0.375

Gini(Temp=Mild) = 1 – (4/6)2 – (2/6)2 = 1 – 0.444 – 0.111 = 0.445

We’ll calculate weighted sum of gini index for temperature feature

Gini(Temp) = (4/14) x 0.5 + (4/14) x 0.375 + (6/14) x 0.445 = 0.142 + 0.107 + 0.190 =

0.439

#### For Humidity

|  |  |  |  |
| --- | --- | --- | --- |
| **Humidity** | **Yes** | **No** | **Number of instances** |
| **High** | 3 | 4 | 7 |
| **Normal** | 6 | 1 | 7 |

Gini(Humidity=High) = 1 – (3/7)2 – (4/7)2 = 1 – 0.183 – 0.326 = 0.489

Gini(Humidity=Normal) = 1 – (6/7)2 – (1/7)2 = 1 – 0.734 – 0.02 = 0.244

Weighted sum for humidity feature will be calculated next

Gini(Humidity) = (7/14) x 0.489 + (7/14) x 0.244 = **0.367**

#### For Wind

|  |  |  |  |
| --- | --- | --- | --- |
| **Wind** | **Yes** | **No** | **Number of instances** |
| **Weak** | 6 | 2 | 8 |
| **Strong** | 3 | 3 | 6 |

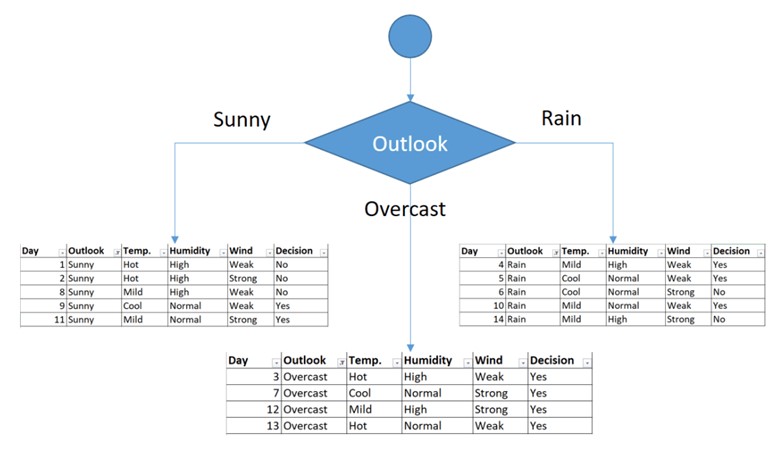
Gini(Wind=Weak) = 1 – (6/8)2 – (2/8)2 = 1 – 0.5625 – 0.062 = 0.375

Gini(Wind=Strong) = 1 – (3/6)2 – (3/6)2 = 1 – 0.25 – 0.25 = 0.5

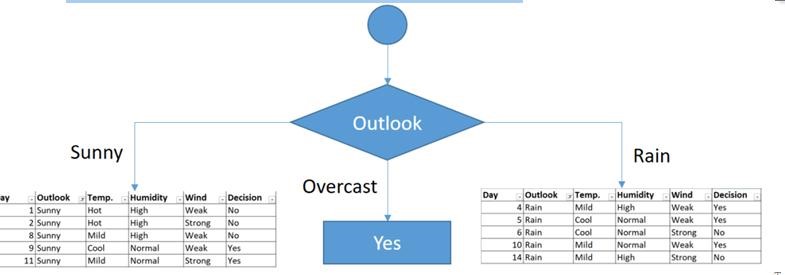
Gini(Wind) = (8/14) x 0.375 + (6/14) x 0.5 = **0.428**

The winner will be outlook feature because its cost is the lowest.

|  |  |
| --- | --- |
| **Feature** | **Gini index** |
| **Outlook** | 0.342 |
| **Temperature** | 0.439 |
| **Humidity** | 0.367 |
| **Wind** | 0.428 |



* The sub dataset in the overcast leaf has only yes decisions. This means that overcast leaf is over.



* We need to focus on the sub dataset for sunny outlook. We need to find the gini index scores for temperature, humidity and wind features respectively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temp.** | **Humidity** | **Wind** | **Decision** |
| **1** | Sunny | Hot | High | Weak | No |
| **2** | Sunny | Hot | High | Strong | No |
| **8** | Sunny | Mild | High | Weak | No |
| **9** | Sunny | Cool | Normal | Weak | Yes |
| **11** | Sunny | Mild | Normal | Strong | Yes |

**Gini of temperature for sunny outlook**

|  |  |  |  |
| --- | --- | --- | --- |
| **Temperature** | **Yes** | **No** | **Number of instances** |
| **Hot** | 0 | 2 | 2 |
| **Cool** | 1 | 0 | 1 |
| **Mild** | 1 | 1 | 2 |

##### • Gini(Outlook=Sunny and Temp.=Hot) = 1 – (0/2)2 – (2/2)2 = 0 • Gini(Outlook=Sunny and Temp.=Cool) = 1 – (1/1)2 – (0/1)2 = 0 • Gini(Outlook=Sunny and Temp.=Mild) = 1 – (1/2)2 – (1/2)2 = 1 – 0.25 – 0.25 = 0.5 • Gini(Outlook=Sunny and Temp.) = (2/5)x0 + (1/5)x0 + (2/5)x0.5 = 0.2

**Gini of humidity for sunny outlook**

|  |  |  |  |
| --- | --- | --- | --- |
| **Humidity** | **Yes** | **No** | **Number of instances** |
| **High** | 0 | 3 | 3 |
| **Normal** | 2 | 0 | 2 |

Gini(Outlook=Sunny and Humidity=High) = 1 – (0/3)2 – (3/3)2 = 0

Gini(Outlook=Sunny and Humidity=Normal) = 1 – (2/2)2 – (0/2)2 = 0 Gini(Outlook=Sunny and Humidity) = (3/5)x0 + (2/5)x0 = 0

|  |  |  |  |
| --- | --- | --- | --- |
| **Wind** | **Yes** | **No** | **Number of instances** |
| **Weak** | 1 | 2 | 3 |
| **Strong** | 1 | 1 | 2 |

**Gini of wind for sunny outlook**

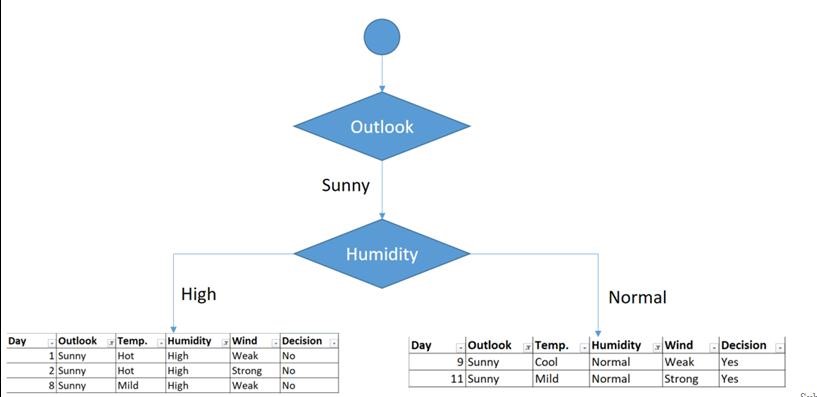
Gini(Outlook=Sunny and Wind=Weak) = 1 – (1/3)2 – (2/3)2 = 0.266

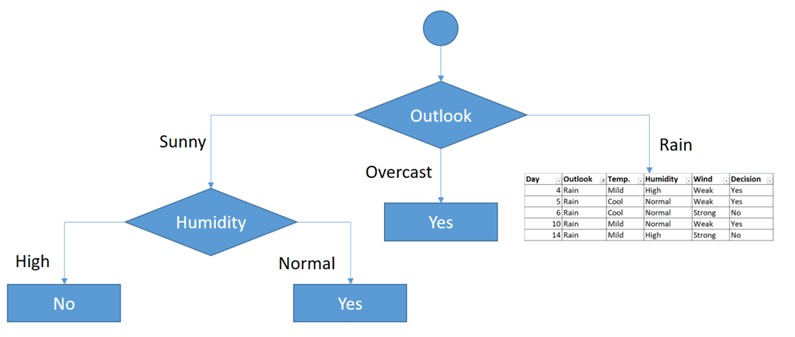
Gini(Outlook=Sunny and Wind=Strong) = 1- (1/2)2 – (1/2)2 = 0.2

Gini(Outlook=Sunny and Wind) = (3/5)x0.266 + (2/5)x0.2 = 0.466

* **Decision for sunny outlook**
* We’ve calculated gini index scores for feature when outlook is sunny. The winner is humidity because it has the lowest value.

|  |  |
| --- | --- |
| **Feature** | **Gini index** |
| **Temperature** | 0.2 |
| **Humidity** | 0 |
| **Wind** | 0.466 |





##### **Rain outlook**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temp.** | **Humidity** | **Wind** | **Decision** |
| **4** | Rain | Mild | High | Weak | Yes |
| **5** | Rain | Cool | Normal | Weak | Yes |
| **6** | Rain | Cool | Normal | Strong | No |
| **10** | Rain | Mild | Normal | Weak | Yes |
| **14** | Rain | Mild | High | Strong | No |

Calculate gini index scores for temperature, humidity and wind features when outlook is rain.

**Gini of temperature for rain outlook**

|  |  |  |  |
| --- | --- | --- | --- |
| **Temperature** | **Yes** | **No** | **Number of instances** |
| **Cool** | 1 | 1 | 2 |
| **Mild** | 2 | 1 | 3 |

Gini(Outlook=Rain and Temp.=Cool) = 1 – (1/2)2 – (1/2)2 = 0.5

Gini(Outlook=Rain and Temp.=Mild) = 1 – (2/3)2 – (1/3)2 = 0.444

Gini(Outlook=Rain and Temp.) = (2/5)x0.5 + (3/5)x0.444 = 0.466

##### **Gini of humidity for rain outlook**

|  |  |  |  |
| --- | --- | --- | --- |
| **Humidity** | **Yes** | **No** | **Number of instances** |
| **High** | 1 | 1 | 2 |
| **Normal** | 2 | 1 | 3 |

Gini(Outlook=Rain and Humidity=High) = 1 – (1/2)2 – (1/2)2 = 0.5

Gini(Outlook=Rain and Humidity=Normal) = 1 – (2/3)2 – (1/3)2 = 0.444 Gini(Outlook=Rain and Humidity) = (2/5)x0.5 + (3/5)x0.444 = 0.466

##### **Gini of wind for rain outlook**

|  |  |  |  |
| --- | --- | --- | --- |
| **Wind** | **Yes** | **No** | **Number of instances** |
| **Weak** | 3 | 0 | 3 |
| **Strong** | 0 | 2 | 2 |

Gini(Outlook=Rain and Wind=Weak) = 1 – (3/3)2 – (0/3)2 = 0

Gini(Outlook=Rain and Wind=Strong) = 1 – (0/2)2 – (2/2)2 = 0

Gini(Outlook=Rain and Wind) = (3/5)x0 + (2/5)x0 = 0

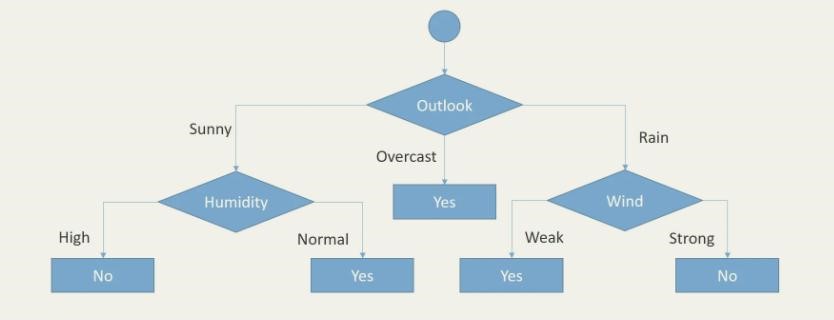
• **Decision for rain outlook -**The winner is wind feature for rain outlook because it has the minimum gini index score in features.

|  |  |
| --- | --- |
| **Feature** | **Gini index** |
| **Temperature** | 0.466 |
| **Humidity** | 0.466 |
| **Wind** | 0 |



Final form of the decision tree built by

CART



#### Naïve Bayes Analysis

* It is probability based classifier
* It uses the Bayes theorem and incorporates evidences and prior knowledge in its predictions.
* It assumes that instances are independent of each other which is a rather unrealistic assumption in real world.
* It can be trained very efficiently in supervised learning setting.
* P(C|A) = [P(A|C)P(C)]/P(A)

* P(C|A) = [P(A|C)P(C)]/P(A)
* P- prob of the variable written in parentheses.

##### C and A - Events P(C) –Prior prob / marginal prob of C

P(C|A) – conditional prob of C given A. Also called as posterior prob.

P(A|C) – conditional prob of A given C P(A) –Prior prob / marginal prob of A

#### Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Age Young ?** | **Income High ?** | **Student ?** | **Credit rating Good ?** | **Buy Car** |
| **1** | Yes | Yes | No | No | No |
| **2** | Yes | Yes | No | Yes | No |
| **3** | No | Yes | No | No | Yes |
| **4** | No | No | No | No | Yes |
| **5** | No | No | Yes | No | Yes |
| **6** | No | No | Yes | Yes | No |
| **7** | No | No | Yes | Yes | Yes |
| **8** | Yes | No | No | No | No |
| **9** | Yes | No | Yes | No | Yes |
| **10** | No | No | Yes | No | Yes |
| **11** | Yes | No | Yes | Yes | Yes |
| **12** | No | No | No | Yes | Yes |
| **13** | No | Yes | Yes | No | Yes |
| **14** | No | No | No | Yes | No |

#### Question

• Implement the Naive Bayes classification for the following credentials.

##### • Age Young ?= Yes • Income High ? = No • Student = No • Credit Rating Good ? = Yes

* Buy Car =?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age Young ?** | | **Income High?** | | **Student ?** | | **Credit Rating Good?** | |
| Buy Car | Yes | No | Yes | No | Yes | No | Yes | No |
| Yes | 2 | 3 | 2 | 2 | 6 | 1 | 3 | 3 |
| No | 7 | 2 | 7 | 3 | 3 | 4 | 6 | 2 |
| Ratio  Yes/Total | 2/9 | 3/5 | 2/9 | 2/5 | 6/9 | 1/5 | 3/9 | 3/5 |
| Ratio No/Total | 7/9 | 2/5 | 7/9 | 3/5 | 3/9 | 4/5 | 6/9 | 2/5 |

**P(Buy Car =Yes|A) = [P(A|P(Buy Car =Yes)P(P(Buy Car =Yes)] / P(A)**

* P(A|**Buy Car =Yes**) :
* P(Age Young ? = Yes |**Buy Car =Yes**) = 2/9
* P(Income High ? = No |**Buy Car =Yes**) = 7/9
* P(Student? =No |**Buy Car =Yes**) = 3/9
* P(Credit Rating Good ?= Yes |**Buy Car =Yes**) = 3/9
* P(A|**Buy Car =Yes**) = (2/9)(7/9)(3/9)(3/9) = 0.0192
* **P(Buy Car =Yes|A) = [P(A|Buy Car =Yes)P(Buy Car =Yes)] / P(A)**

* **P(Buy Car =Yes|A) = (0.0192\* 9/14 ) / P(A) = 0.0123 /P(A)**

* P(A|**Buy Car = No**) :
* P(Age Young ? = Yes |**Buy Car =No**) = 3/5
* P(Income High ? = No |**Buy Car = No**) = 3/5
* P(Student? =No |**Buy Car = No**) = 4/5
* P(Credit Rating Good ?= Yes |**Buy Car = No**) = 3/5
* P(A|**Buy Car = No**) = (3/5) (3/5) (4/5) (3/5) = 0.0357
* **P(Buy Car = No|A) = [P(A|Buy Car = No)P(Buy Car = No)] / P(A)**

* **P(Buy Car = No|A) = (x\*5/14)/P(A) = 0.00714 / P(A)**

* **Concluding remark - The person will buy a car.**

#### Linear Regression

* It is a statistical approach for modeling relationship between a dependent variable (responses) with a given set of independent variables(features).

**Simple Linear Regression**

* Simple linear regression is an approach for predicting a **response** using a **single feature**.
* It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).
* Let us consider a dataset where we have a value of response y for every feature x:
* For generality, we define:
* x as **feature vector**, i.e x = [x1, x2, …., xn],
* y as **response vector**, i.e y = [y1, y2, …., yn ]
* for **n** observations
* Now, the task is to find a **line which fits best** in above scatter plot so that we can predict the response for any new feature values. (i.e a value of x not present in dataset)
* This line is called **regression line**.
* The equation of regression line is represented as:
* h(𝑥𝑖)= β0 + β1 𝑥𝑖
* Here,
* h(𝑥𝑖) represents the **predicted response value** for ith observation.
* β0 and β1  are regression coefficients and represent  **y-intercept** and **slope** of regression line respectively.

##### • The mathematical technique to fit a regression line involves estimating the values of β0 , β1

This has to be done in such a way that the sum of the squared errors for all the data points is minimum. **This is known as the least squares criterion.**

### Linear regression (Numerical)

#### • The problem is to regress Y on X

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **State** | **No. of**  **Institutions (X)** | **Membership (Y)** | **X2** | **Y2** | **XY** |
| NSW | 17 | 5987 | 289 | 3.58442\*107 | 101779 |
| QLD | 11 | 5950 | 121 | 3.54025\*107 | 65450 |
| SA | 10 | 3588 | 100 | 1.28737\*107 | 35880 |
| TAS | 3 | 1356 | 9 | 1.83873\*106 | 4068 |
| VIC | 41 | 14127 | 1681 | 1.99572\*108 | 579207 |
| WA | 9 | 4847 | 81 | 2.34934\*107 | 43623 |
| Others | 11 | 3893 | 121 | 1.51554\*107 | 42823 |
| Total | 102 | 39748 | 2402 | 3.241799\*108 | 872830 |

If b0 and b1 are the estimates of β0 , β1

#### Clustering

* Clustering is the process of grouping observations of similar kinds into smaller groups within the larger population.
* It has widespread application in business analytics.
* One of the questions facing businesses is how to organize the huge amounts of available data into meaningful structures.
* Or break a large heterogeneous population into smaller homogeneous groups.
* Cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise.

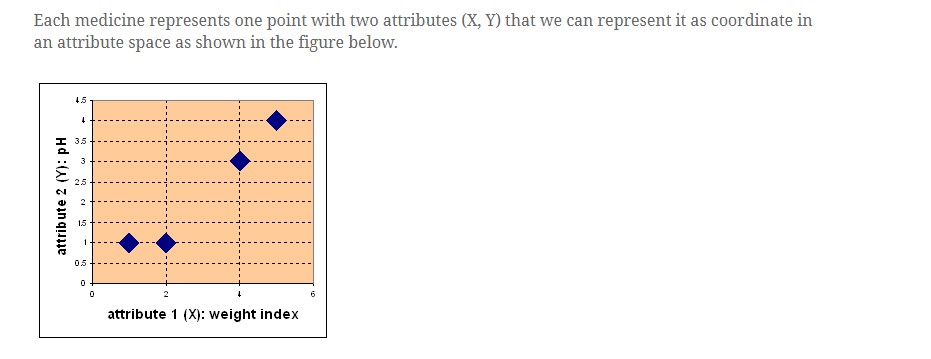
#### What is clustering

##### • Clustering: the process of grouping a set of objects into classes of similar objects

* Documents within a cluster should be similar.
* Documents from different clusters should be dissimilar.

###### • The commonest form of *unsupervised learning*

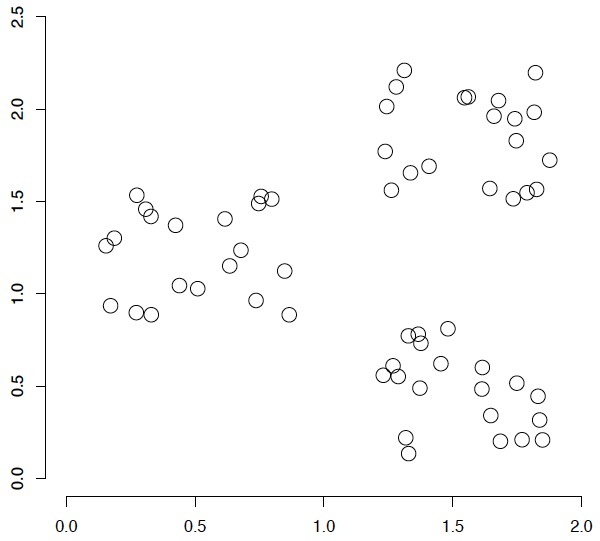
• Unsupervised learning = learning from raw data, as opposed to supervised data where a classification is an example that is given.



##### **Business application of clustering – Use Case**

* A grocer retailer used clustering to segment its 1.3MM loyalty card customers into 5 different groups based on their buying behavior. It then adopted customized marketing strategies for each of these segments in order to target them more effectively.
* One of the groups was called ‘Fresh food lovers’. This comprised of customers who purchase a high proportion of organic food, fresh vegetables, salads etc. A marketing campaign that emphasized the freshness of the fruits and vegetables and year-round availability of organic produce in the stores appealed to this customer group.
* Another cluster was called ‘Convenience junkies’. This comprised of people who shopped for cooked/semi-cooked, easy-to prepare meals. A marketing campaign focusing on the retailer’s in-house line of frozen meals as well as the speed of the check-out counters at the store worked well with this audience.
* In this way the retailer was able to deliver the right message to the right customer and maximize the effectiveness of its marketing.

A data set with clear cluster structure



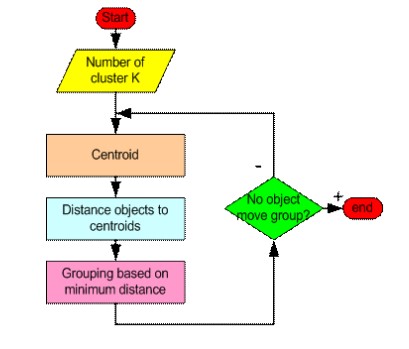
#### Types of Clustering

* Flat algorithms – Non Hierarchical algorithms

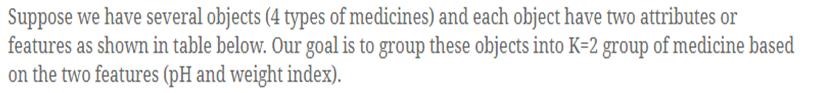
– Usually start with a random (partial) partitioning – Refine it iteratively

* *K* means clustering
* (Model based clustering)
* Hierarchical algorithms
* Agglomerative - Bottom-up
* Divisive - Top-down

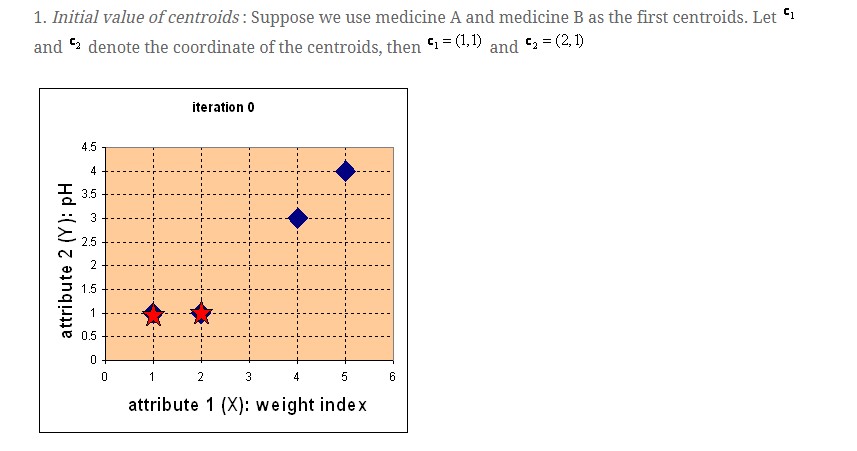
#### K Means

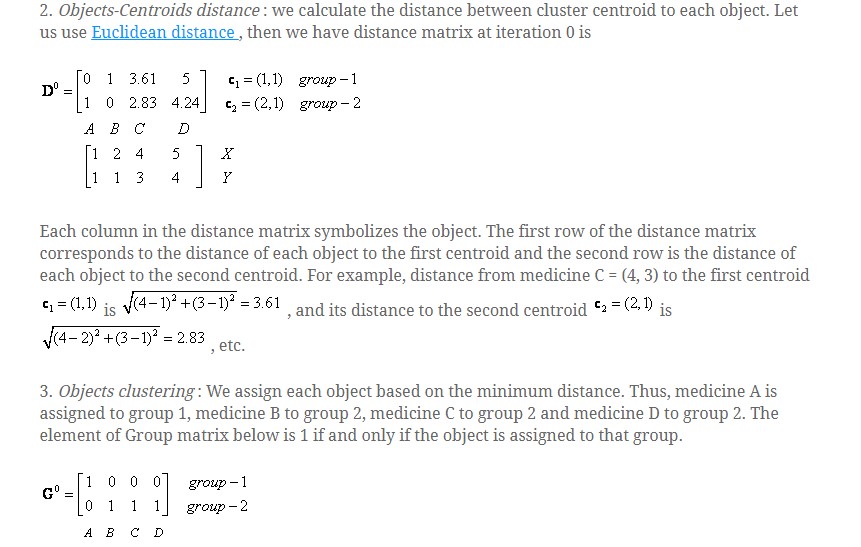


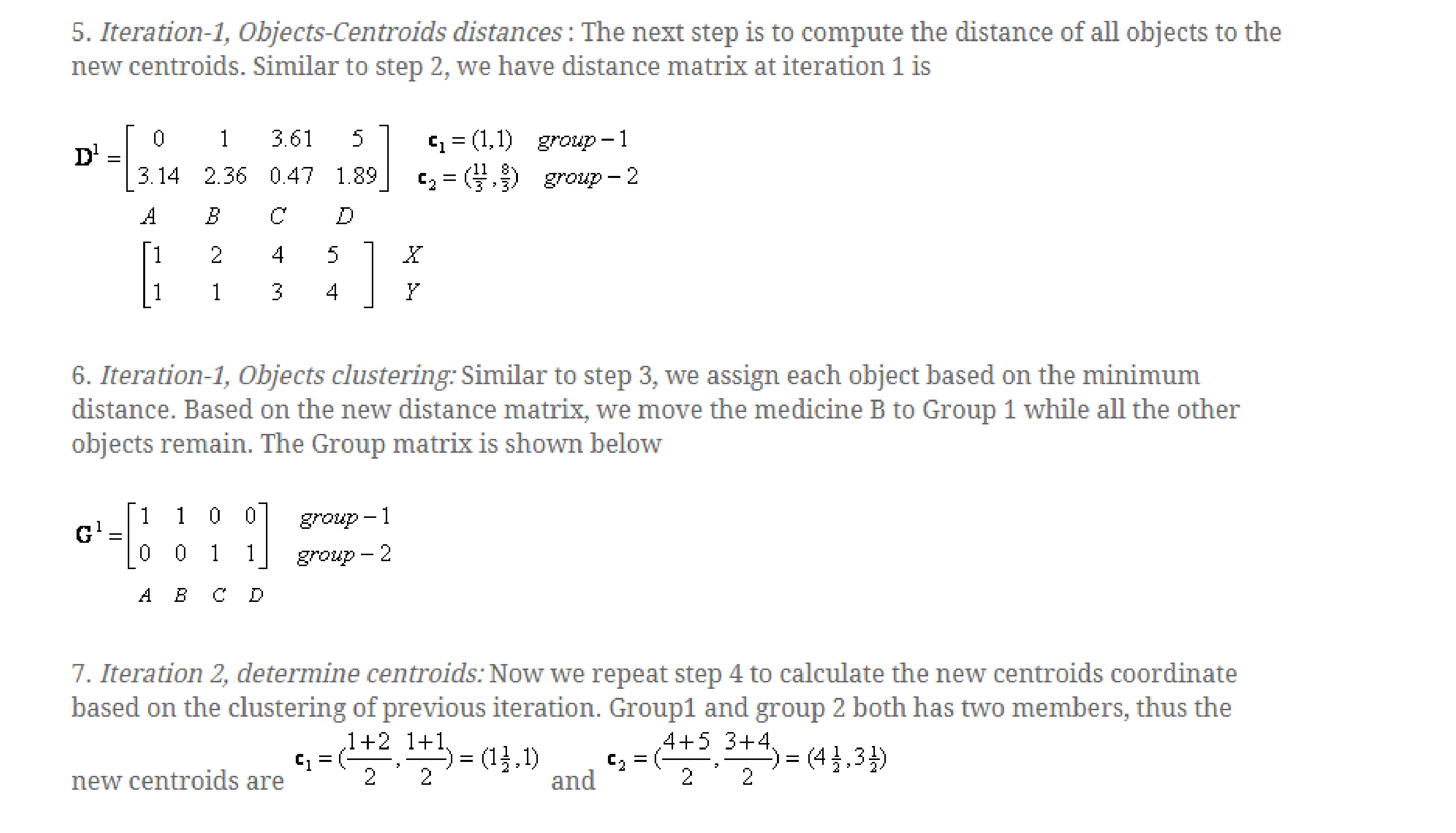
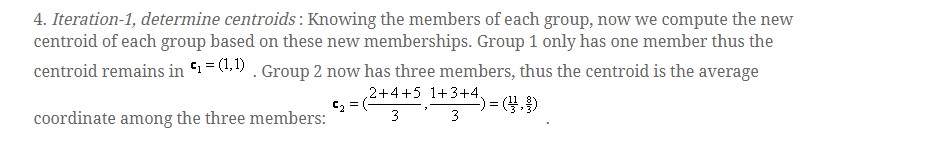
#### Example - KMeans



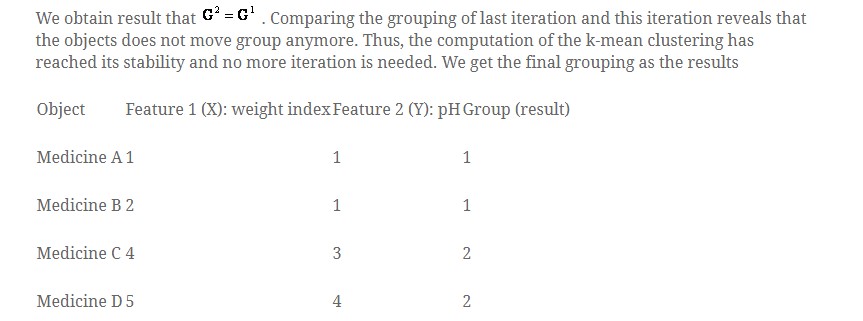
|  |  |  |
| --- | --- | --- |
| **Medicine** | **Attributes X Weight Index** | **Attributes Y pH** |
| Medicine A | 1 | 1 |
| Medicine B | 2 | 1 |
| Medicine C | 4 | 3 |
| Medicine D | 5 | 4 |











#### Association rule mining

* Proposed by Agrawal et al in 1993.
* It is an important data mining model studied extensively by the database and data mining community.
* It searches for interesting associations or relationships among items in the given dataset.
* With massive amounts of data continuously being collected and stored in databases many industries are becoming interested in mining association rules from the databases.
* Assume all data are categorical.
* No good algorithm for numeric data.
* Initially used for Market Basket Analysis to find how items purchased by customers are related.

Bread  Milk [sup = 5%, conf = 100%]

#### The model: data

* *I* = {*i*1, *i*2, …, *im*}: a set of *items*.
* Transaction *t* :

– *t* a set of items, and *t*  *I*.

* Transaction Database *T*: a set of transactions *T*

= {t1, t2, …, tn}.

#### Transaction data: supermarket data

##### • Market basket transactions: t1: {bread, cheese, milk}

t2: {apple, eggs, salt, yogurt}

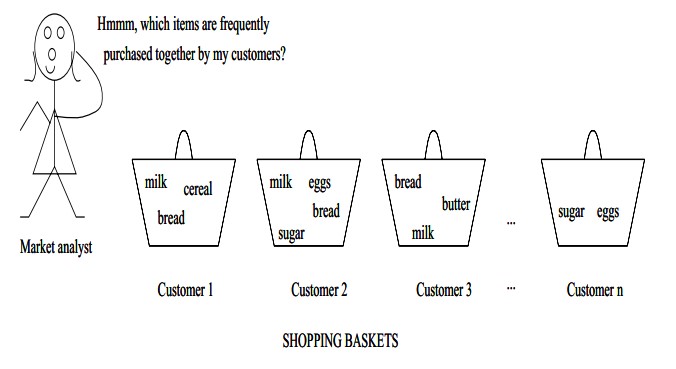
… …

tn: {biscuit, eggs, milk}

• Concepts:

* An *item*: an item/article in a basket
* *I*: the set of all items sold in the store
* A *transaction*: items purchased in a basket; it may have TID (transaction ID)
* A *transactional* *dataset*: A set of transactions

#### Market Basket Analysis



Transaction data: a set of documents

• **A text document data set. Each document is treated as a “bag” of keywords**

|  |  |
| --- | --- |
| doc1: | Student, Teach, School |
| doc2: | Student, School |
| doc3: | Teach, School, City, Game |
| doc4: | Baseball, Basketball |
| doc5: | Basketball, Player, Spectator |
| doc6: | Baseball, Coach, Game, Team |
| doc7: | Basketball, Team, City, Game |

#### The model: rules

* A transaction *t* contains *X*, a set of items (itemset) in *I*, if *X*  *t*.
* An association rule is an implication of the form:

*X*  *Y*, where *X*, *Y*  *I, and X* *Y* = 

* An itemset is a set of items.
  + E.g., X = {milk, bread, cereal} is an itemset.
* A *k*-itemset is an itemset with *k* items.
  + E.g., {milk, bread, cereal} is a 3-itemset

#### Rule strength measures

* Support: The rule holds with support *sup* in *T* (the transaction data set) if sup% of transactionscontain *X*  *Y*.
  + *sup* = Pr(*X*  *Y*)*.*
* Confidence: The rule holds in *T* with confidence *conf* if *conf*% of transactions that contain *X* also contain *Y.* 
  + *conf* = Pr(*Y* | *X*)
* An association rule is a pattern that states when *X* occurs, *Y* occurs with certain probability.

**Commonly Used Interest Measures for**

**Association Rules**

**-** Support , Confidence, Lift, Coverage, Leverage, Conviction

* Support count: The support count of an itemset *X*, denoted by *X.count*, in a data set *T* is the number of transactions in *T* that contain *X*. Assume *T* has *n* transactions.

##### • Then,

*support* (*X*  *Y*).*count*

*n*

* **Confidence (also called Strength)**

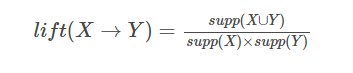
(*X*  *Y*).*count*

*confidence*

*X*.*count*

**Coverage -** A simple measure of how often a item set appears in the data set.

coverage(X) = P(X) = sup(X)

* **Lift(originally called Interest)**
* Lift is a ratio of observed support to expected support if X and Y were independent.

If lift(X→Y) = 1, then it would imply that probabilities of occurrences of itemset X and itemset Y are independent of each other, meaning that the ***rule*** doesn’t show any statistically proven relationship.

If lift(X→Y) > 1, then it would imply that probabilities of occurrences of the itemsets X and Y are positively dependent on each other. It will also tell us the magnitude of the level of dependence. The higher the lift value, the higher is the dependence, which can also be referred to as itemsets being complements to each other.

If lift(X→Y) < 1, then it would imply that the probabilities of occurrences of the itemsets X and Y are negatively dependent on each other. The lower the lift value, the lower is the dependence, which can also be referred to as itemsets being substitutes to each other.

* **Leverage**

Leverage(X -> Y) = P(X and Y) - (P(X)P(Y))

Leverage measures the difference of X and Y appearing together in the data set and what would be expected if X and Y where statistically dependent. The rational in a sales setting is to find out how many more units (items X and Y together) are sold than expected from the independent sells. **Conviction**

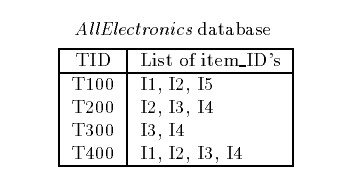
conviction(X -> Y) = P(X)P(not Y)/P(X and not Y)=(1-sup(Y))/(1-conf(X -> Y))

Conviction compares the probability that X appears without Y if they were dependent with the actual frequency of the appearance of X without Y.

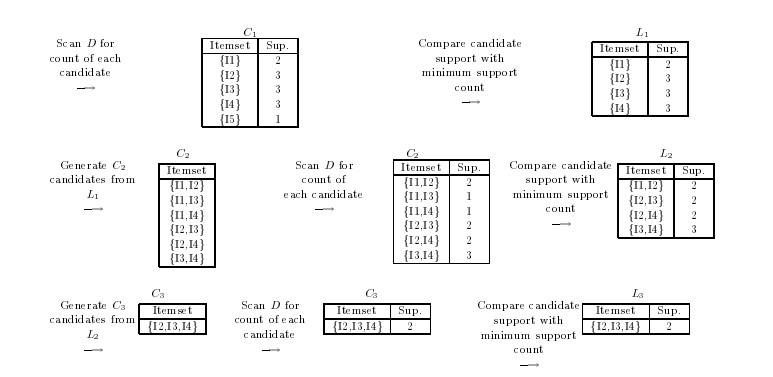
#### Goal

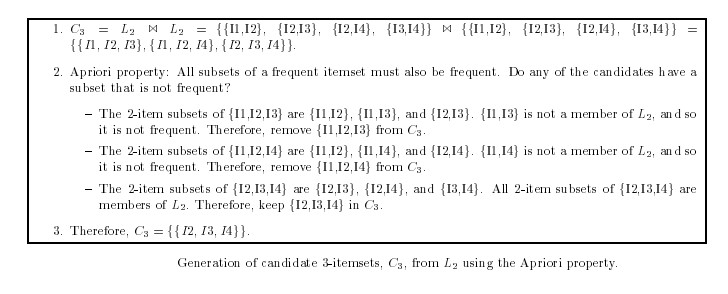
• **Goal:** Find all rules that satisfy the userspecified *minimum support* (minsup) and *minimum confidence* (minconf).

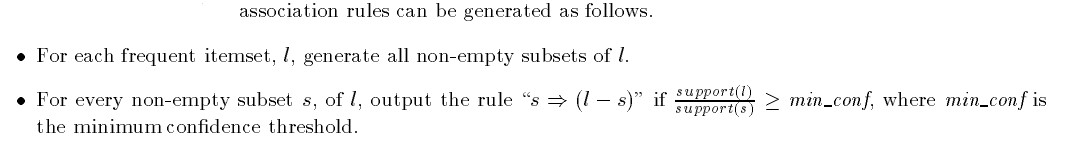
#### All Electronics Database

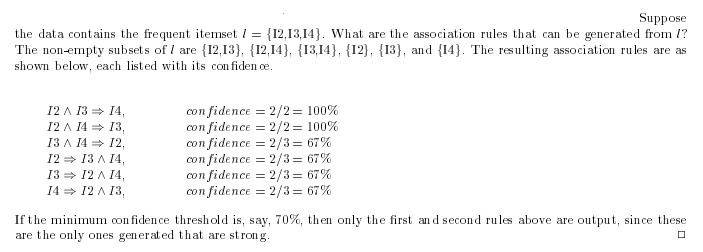


#### Apriori Algorithm









* Average of all the Y values = 5678 • Average of all the X values = 14.57 • Sxy = sum of cross product deviation = 293645 • Sxx = sum of squared deviation for X = 915.7 • b1 = 320.7 • b0 = 1005 • h(𝑥𝑖)= β0 + β1 𝑥𝑖
* Y= 1005+320.7\*X

#### BI Road Ahead

* BI as a concept is not new.
* Even the practice of storing data in databases, analyzing and revealing useful information in form of reports have found their way in organizations.

##### BI and Mobility

Mobility has two major offerings

* 24 X 7 connectivity – Ability to stay in contact with others.
* Mobile Workability – convenience of being able to work from any where.

##### Need of BI on the move

With the ever increasing volumes of enterprise data coupled with fast paced world of modern business, intelligent decisions are to be taken quickly.

So there is a need for better and faster transfer of information.

So management does not want to be limited to their office based desktops to access data and useful information through DSS and applications designed only for PCs.

##### BI Mobility timeline

* The Antediluvian Era
* Taking up the challenges
* The roadblocks
* Overcoming the shortcomings

##### The Present

* Exceptions and Alerts
* Push Reporting
* Pull Reporting

##### Examples of Small business mobile applications dealing with Local Data analysis

* Pedometer
* Music Suggestion
* Map Integration

##### Data Security Concerns

* Device Security
* Transmission Security
* Authorization, authentication and Network Security

#### BI and Cloud Computing

* Cloud computing and business intelligence are an ideal match.
* Business intelligence is about delivering the right information to the right people at the right time, and cloud computing provides a lightweight, agile way to access BI applications.
* The beauty of Cloud BI applications is that they are accessible on multiple devices and web browsers.
* This is circumventing traditional software barriers such as the requirement to access the application on-site.

#### Cloud Business Intelligence (BI)

* **Cloud Business Intelligence (BI)** applications are hosted on a virtual network, such as the internet. They are used to provide organizations access to BI-related data such as dashboards, KPIs and other business analytics.
* Enterprises are increasingly turning to cloud-based tools, like Customer Relationship Management (CRM) applications (Salesforce), online file collaboration and storage (Dropbox, Box) and help desk software (UserVoice, Zendesk).
* This trend includes business intelligence tools embracing the agility and accessibility of the Cloud.

Advantages of cloud computing for BI users

* **Ease of use**
* Cloud BI applications, like other cloud applications, tend to be easier for end-users to operate and set up. This means reduced IT involvement and costs.
* **Scalability and elasticity**
* Cloud applications can be rapidly scaled to accommodate for an increase in the number of users in an organization.
* **Accessibility**
* Cloud BI applications can be accessed on any web browser or on any mobile device.
* **Deployment speed**
* Cloud applications are very simple to deploy, since they require no additional hardware or software installations.

#### Social CRM and BI

* Customer relationship management (CRM) and business intelligence (BI) are two tools that deliver the same benefits: Understanding and analyzing data to make better business decisions, improve customer relationships, forecast and impact business performance and grow revenue.

* **Steps to integrate CRM and BI**
* Know your customers
* Identify data sources
* Test and clean data
* Measure and refine

#### Social CRM and BI

* **Connect with your customers via Twitter, LinkedIn, and Facebook**

Read their latest tweets and posts, view their business profiles, and more.

* **Identify key business opportunities and insights, and uncover new ones**

Identify your customer's interests to quickly identify key sales opportunities that you can use to grow your business.

* **Understand your customer's social interactions and relationships**

See your customer's connections and friends. See how these relationships interact.

BI and ERP(Enterprise Resource

Planning)

* An ERP system integrates all departments of a business, collates the data and makes it easier to access. It helps the users in taking the right decisions by providing an accurate overview of the whole business. ERP in itself is a very strong tool that is vital in streamlining processes and resources. However, if armed with powerful Business Intelligence, the capability of an ERP system increases ten-fold and it becomes more dependable.
* **ERP collects the enterprise data while Business Intelligence analyzes the enterprise data** and uses dashboards and other interfaces to present that data in ways that make it easily understood and helps identify actionable opportunities.
* With ERP software organizing your business data, Business Intelligence technologies can mine this data and present it in easily digestible and actionable formats. ERP and BI have a symbiotic relationship – **BI requires an ERP in order to function,** while the value of ERP software is significantly enhanced when it is paired with BI.

Advantages of bringing ERP and BI

together

1.Businesses today are moving away from historical reporting and getting more inclined towards forecasting. While ERP, through its integration of all departments, enables users to bring in efficiency in operations, it deals more with providing total control of business to the owners. On the other hand, Business Intelligence enables users to strategize for the future. With all the data of past and present activities of business, Business Intelligence can provide predictive analyses to users which can be used to plan operations. Thus BI plays a highly important role in overall strategy and future planning.

2. Earlier, when business owners derived data through traditional means, it took a massive amount of time and resources. By the time it got into their hands, the data would have become obsolete, rendering any decision making opportunities wasted. However, Business Intelligence provides data in real time, that too with highly reliable accuracy.

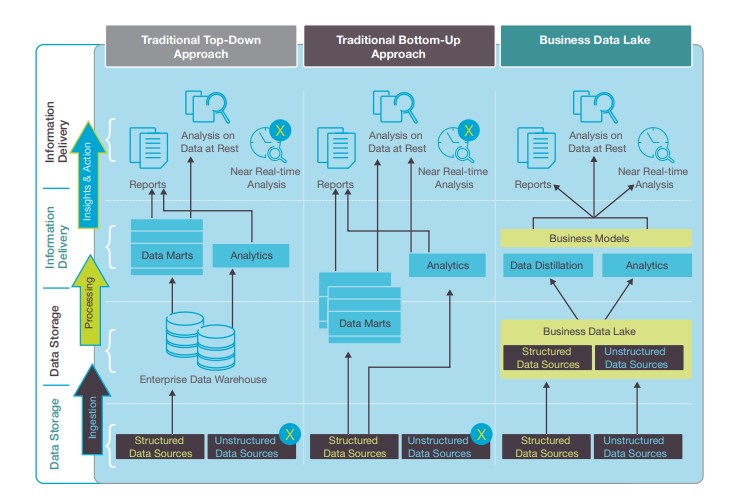
3.The reporting capabilities of BI enabled ERP systems are overwhelming.

4. Lastly, ERP is used more at the operational level where it brings clarity and smoothness in day to day activities of a business. Business Intelligence is used more on the decision making level, where it provides quick, yet deep insights as to where the business stands and where can it be taken to.

##### **Data lake - Introduction**

* A data lake is a large storage repository that holds a vast amount of raw data in its native format until it is needed. An *“enterprise data lake”* (EDL) is simply a data lake for enterprise-wide information storage and sharing.
* A Business Data Lake is a data repository that can store and handle massive amounts of structured, semi-structured and unstructured data in its raw form in low cost commodity storage as it arrives.
* A data lake architecture incorporating enterprise search and analytics techniques can help companies unlock actionable insights from the vast structured and unstructured data stored in their lakes.
* A Business Data Lake is a simple but powerful approach to solve business problems. It caters to ever-changing business needs by allowing for storage of all data and providing the capability of deriving actionable insights from any kind of data, yet working in a transparent and seamless fashion, in an enterprise-wide environment.

Architecture Comparison — Traditional BI and Business Data Lake



##### **Difference between Data lakes and Data warehouse**

|  |  |  |
| --- | --- | --- |
| **Storage** | Data lakes design for low-cost storage. | Expensive storage that give fast response times are used |
| **Security** | Offers lesser control. | Allows better control of the data. |
| **Replacement of EDW** | Data lake can be source for EDW | Complementary to EDW (not replacement) |
| **Schema** | Schema on reading (no predefined schemas) | Schema on write (predefined schemas) |
| **Data Processing** | Helps for fast ingestion of new data. | Time-consuming to introduce new content. |
| **Data Granularity** | Data at a low level of detail or granularity. | Data at the summary or aggregated level of detail. |
| **Tools** | Can use open source/tools like Hadoop/ Map  Mostly commercial tools. Reduce | |

**Parameters** **Data Lakes** **Data Warehouse**

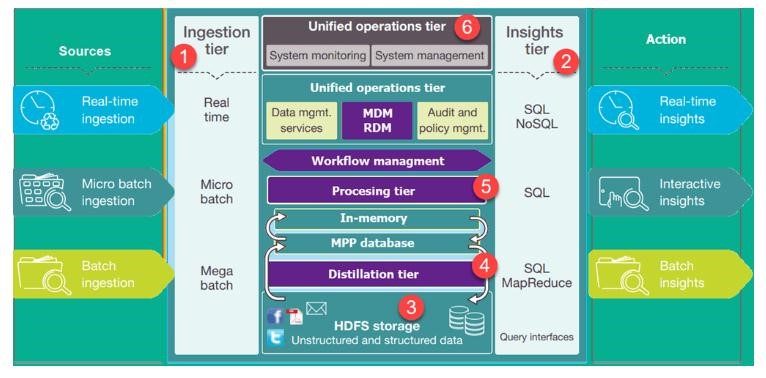
|  |  |  |
| --- | --- | --- |
| **Data** | Data lakes store everything. | Data Warehouse focuses only on Business Processes. |
| **Processing** | Data are mainly unprocessed | Highly processed data. |
| **Type of Data** | It can be Unstructured, semi-structured and structured. | It is mostly in tabular form & structure. |
| **Task** | Share data stewardship | Optimized for data retrieval |
| **Agility** | Highly agile, configure and reconfigure as needed. | Compare to Data lake it is less agile and has fixed configuration. |

Business professionals widely use

**Users** Data Lake is mostly used by Data Scientist

data Warehouse

##### **Data Lake Architecture**

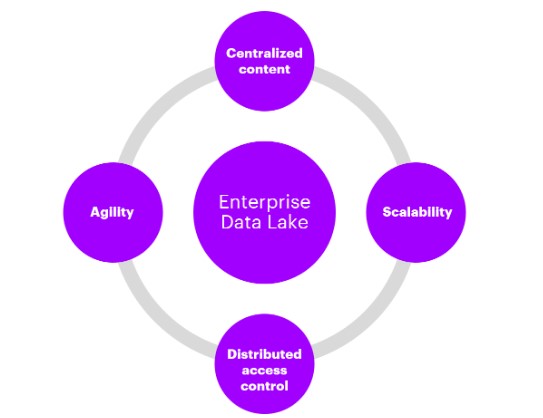


Important tiers in Data Lake

Architecture

* **Ingestion Tier**: The tiers on the left side depict the data sources. The data could be loaded into the data lake in batches or in real-time
* **Insights Tier:** The tiers on the right represent the research side where insights from the system are used. SQL, NoSQL queries, or even excel could be used for data analysis.
* **HDFS(Hadoop Distributed File System)** is a cost-effective solution for both structured and unstructured data. It is a landing zone for all data that is at rest in the system.
* **Distillation tier** takes data from the storage tire and converts it to structured data for easier analysis.
* **Processing tier** run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.
* **Unified operations tier** governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

#### Advantages of Data Lakes



* The main benefit of a data lake is the **centralization** of disparate content sources. Once gathered together (from their “information silos”), these sources can be combined and processed using big data, search and analytics techniques which would have otherwise been impossible.
* The **security** measures in the data lake may be assigned in a way that grants access to certain information to users of the data lake that do not have access to the original content source. These users are entitled to the information, yet unable to access it in its source for some reason.
* Once the content is in the data lake, it can be **normalized and enriched**. This can include metadata extraction, format conversion, augmentation, entity extraction, cross-linking, aggregation, de-normalization, or indexing.
* Users, from different departments, potentially scattered around the globe, can have **flexible access** to the data lake and its content from anywhere. This increases re-use of the content and helps the organization to more easily collect the data required to drive business decisions.

#### Case Study

• “Survey for You” is a service company that conducts surveys for gaining insights on matters important for companies.

#### Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Finance/Investment** | **Travel and Tour** | **Reading magazines** | **Health /Diet** | **Sex** |
| Yes | No | Yes | No | Male |
| Yes | Yes | No | No | Male |
| No | Yes | Yes | Yes | Female |
| No | Yes | No | Yes | Male |
| Yes | Yes | Yes | Yes | Female |
| No | No | Yes | No | Female |
| Yes | No | No | No | Male |
| Yes | Yes | No | No | Male |
| No | No | No | Yes | Female |
| Yes | No | No | No | Male |

### Question

* Finance – No
* Travel and Tour – Yes
* Reading Magazines – Yes • Health /Diet – No
* Sex = ?

#### • Classify this new instance

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Finance/ Investment** | | **Travel/Tour** | | **Reading Magazine** | | **Health/Diet** | |
| Sex | M | F | M | F | M | F | M | F |
| Yes | 5 | 1 | 3 | 2 | 1 | 3 | 1 | 3 |
| No | 1 | 3 | 3 | 2 | 5 | 1 | 5 | 1 |
| Ratio  Yes/Total | 5/6 | 1/4 | 3/6 | 2/4 | 1/6 | 3/4 | 1/6 | 3/4 |
| Ratio No/Total | 1/6 | 3/4 | 3/6 | 2/4 | 5/6 | 1/4 | 5/6 | 1/4 |

**P(Sex =Male|A) = [P(A|Sex =Male)P(Sex=Male)] / P(A)**

• P(A|Sex =Male) :

#### • P(Finance /Inv = No |Sex =Male ) = 1/6 • P(Travel/Tour = Yes |Sex =Male ) = 3/6 • P(Reading Magazines = Yes |Sex =Male ) = 1/6 • P(Health & Diet= No |Sex =Male ) = 5/6 • P(A|Sex =Male) = 1/6 \*3/6\*1/6\*5/6=5/432

* **P(Sex =Male|A) = [P(A|Sex =Male)P(Sex=Male)] / P(A)**
* **P(Sex =Male|A) = [**5/432 \* 6/10 ] / P(A) = 0.006944 /P(A)

* P(A|Sex =Female) :
* P(Finance /Inv = No |Sex =Female ) = 3/4
* P(Travel/Tour = Yes |Sex =Female ) = 2/4
* P(Reading Magazines = Yes |Sex =Female ) =3/4
* P(Health & Diet= No |Sex =Female ) = 1/4
* P(A|Sex =Female) = ¾ \* 2/4\*3/4\*1/4 = 9/128
* **P(Sex =Female|A) = [P(A|Sex =Female)P(Sex=Female)] / P(A)**
* **P(Sex =Female|A) = [**9/128 \* 4/10 ] / P(A) = 0.028125/P(A)

* **Hence the person is a female**

#### Thanks