2019

N. J. Sonecha management and Technical Institute

Prepared By: Jignesh J. Kariya

[BUSINESS ANALYTICS (BA)]

Subject Code: 4529201

MODULE – 1

Ch. 1 – Business Intelligence

Definitions

Business intelligence (BI) is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies. It is, like DSS, a content-free expression, so it means different things to different people.
It can also be defined as, *Business Intelligence is a set of concepts and methodologies to improve decision making in business through the use of facts and fact based IT systems.*

Meaning

BI uses a set of process technologies and tools to transform raw data into meaningful information. BI mainly focuses on *mining the information* to provide knowledge and uses the knowledge provided to beneficial insights. The insights then lead to impactful decision making which helps business and provide benefits like; increased productivity, increased profitability, reduced cost etc.

BI's major objective is to enable interactive access (sometimes in real time) to data, to enable manipulation of data, and to give business managers and analysts the ability to conduct appropriate analysis. By analysing historical and current data, situations, and performances, decision makers get valuable insights that enable them to make more informed and better decisions. The process of BI is based on the transformation of data to information, then to decisions, and finally to actions.
The goal of BI is improved business decisions. It is more than technologies. It compasses core concepts such as;

- **Extract-Transform-Load (ETL)**(Three database functions that are combined into one tool to pull data out of one database and place it into another database.)

- **Data Warehousing**
- **Data mart** (The data mart is a subset of the data warehouse and is usually oriented to a specific business line or team.)

- **Metrics and KPIs** (A key performance indicator –KPI– is used to measure performance and success. A metric is nothing more than a number within a **KPI** that helps track performance and progress.)

- **Scorecards** (Scorecard uses dashboards inside business intelligence platforms or business objects to measure how strategy and operational activities align.)

- **Dashboards**

- **OLAP reporting**

In addition to these, Methodologies specific to ETL, Data Warehousing, Master Data Management and Data Governance.

Examples

In business terms for example; A Human Capital Management data mart will get many IT applications holding employee specific data such as employee profile, payroll, training, compensation, project performance and analyse employee productivity by department, management level, years of experience, gender and qualification. Such data marts will need ETL and OLAP reporting tools to function.

In general term BI is a form of AI used in business functions. General examples of AI may be as under;
Virtual Personal Assistants - Cortana, Siri, and Google Now are some of the intelligent digital personal assistants

Video Games - The efficacy of AI has increased making video game characters to become skilled at your behaviors, take action to stimuli, and respond in volatile ways.

Smart Cars - Google's project and Tesla's autopilot functioning feature are two examples that have been in the latest news. The algorithms created by Google could enable self-driving cars driving in the similar ways that humans do by intelligence and experience.

Fraud Detection - AI is used to create systems that learn what types of transactions are fraudulent in banking transactions (For example: OTP).

A Brief History of BI

The term BI was coined by the Gartner Group in the mid-1990s. However, the concept is much older; it has its roots in the MIS reporting systems of the 1970s. During that period, reporting systems were static, two dimensional, and had no analytical capabilities. In the early 1980s, the concept of executive information systems (EIS) emerged.

This concept expanded the computerized support to top-level managers and executives. Some of the capabilities introduced were dynamic multidimensional (ad hoc or on-demand) reporting, forecasting and prediction, trend analysis, drill-down to details, status access, and critical success factors. These features appeared in dozens of commercial products until the mid-1990s. Then the same capabilities and some new ones appeared under the name BI.

In the late 1990s and early 2000s, BI services began providing simplified tools, allowing decision makers to become more self-sufficient. The tools were easier to use, provided the functionality needed, and were very efficient. Business people could now gather data and gain insights by working directly with the data.

Today, a good BI-based enterprise information system contains all the information executives need. So, the original concept of EIS was transformed into BI. By 2020, BI systems started to include artificial intelligence capabilities as well as powerful analytical capabilities, beyond our imagination.

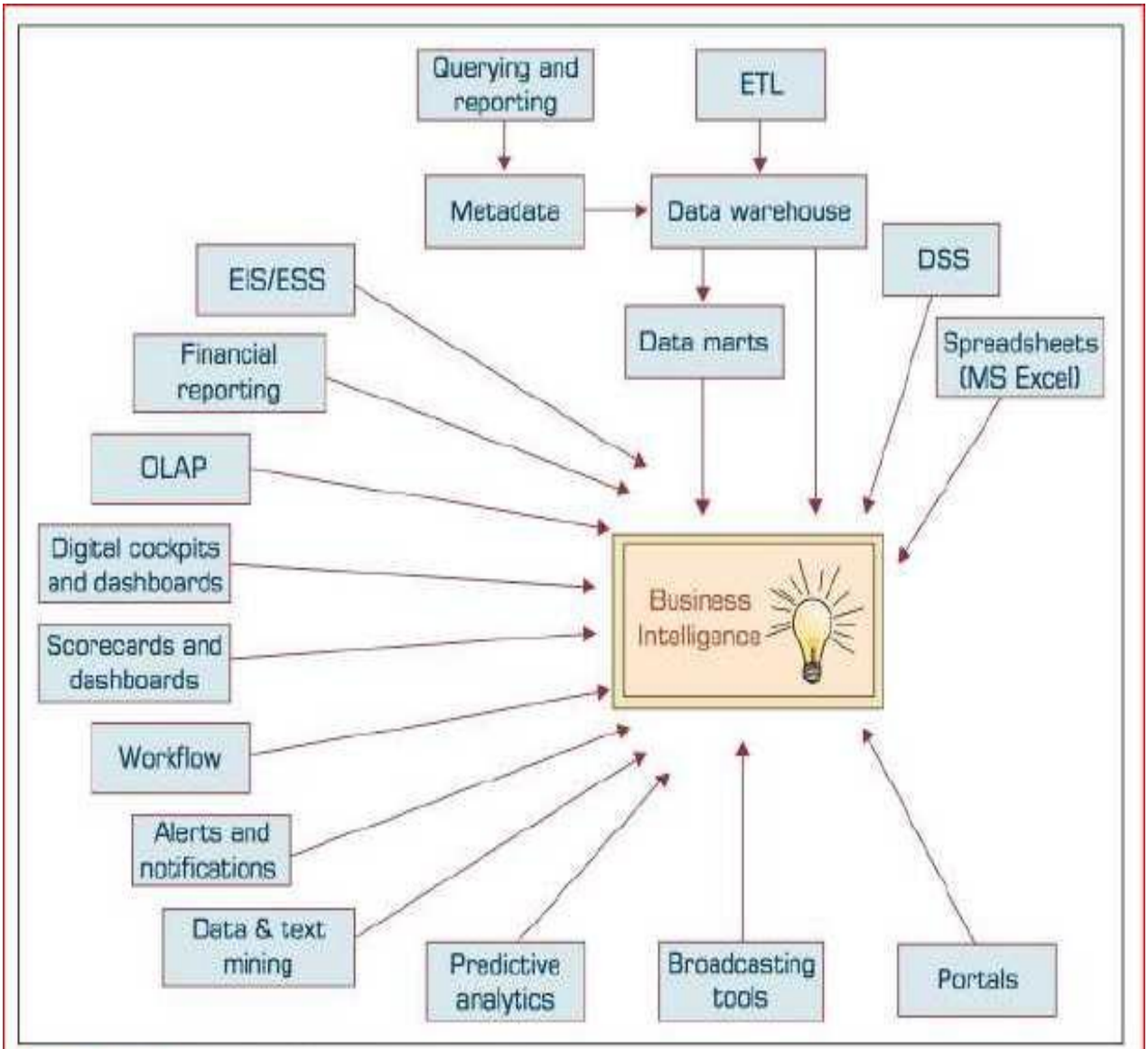


FIGURE I.9 Evolution of Business Intelligence (BI).

Need/Benefits:

1) To remove guess work: BI provides accurate data to update in real time and any other requirement to improve decision making in precise manner. Therefore, to remove guess work, BI is required.

2) For quick responses to business related queries: BI provides framework to get immediate answer to the business related query or question. Thus, BI provides tools to improve responses about business related query.

3) Valuable insights into the customer behaviour: BI tools help to predict market situation and consumer behaviour based on available data. Thus, for accurate purchasing pattern and consumer behaviour, BI is required.
4) Developing efficiency: BI helps regarding data collection, conversion and reporting same for communicate to the management regarding efficiency, performance and operation. Thus, to improve efficiency BI is required.

5) To identify and control costs: It's an important task for managers to identify various costs incurred during business operations. BI tools help to identify different costs regarding requiring a price statement on formulating pricing policy.
Moreover, BI tools help to control some costs which increase the overall profit.

6) For maintaining inventory / Inventory control: BI software helps to make right order for the inventory; it means the right quantity of right inventory at the right time.

7) For overall analysis of business: BI system helps to know or identify business statistics (*Profit & losses, Overall performance, Information regarding customer, employee, production etc.*) over a given period of time. This data related information help to take an advantage over the competition as well as sustain business.

Features/Characteristics:

1) Fact based decision making: Decision made through BI are purely based on facts and history. BI provides flow of data to the business system.

2) Single version of truth: It means some type of data or same data available at more than one place and all such data should agree completely and every respect.

3) 360 degree perspective of the business: BI allows looking at the business for various perspectives. Each person in the project team (BI system) will look at the data from his/her goal and will look for attributes that add value for decision making on his/her role.

4) Virtual team members on the same page: In today's business, team of people who work for common project but are spread across geographical location is long known as a 'virtual team'. Technologies like business intelligence bring them together and provide them some fact of the speed of life in personalized form.

5) Others: There are some common feature require in BI system like;
Data sources, Data filters/drill down, Security, Self-service, Data visualisation, Mobile application etc.

Uses

1) Business report categories: This includes the actual and how actual operations meet against the goals. Here, BI can be used for preparing standard weekly per month reports. For better decision making, BI can be used for smooth business operations.

2) Forecasting: Forecasting is an important tool for business decisions and operations. Without forecasting business cannot be performed well. Thus, BI can be used through estimated required tool as well as forecasting activities for the business functions.

3) For multidimensional Analysis: Such analysis offers good insights to the managers as per the requirement. Such analysis required sound data warehousing or data mart as well as constant flow of the data. BI provide all the required tools for multidimensional analysis.

4) To find correlation among different factors: For high level decision making, it is required to find correlation between different functions within and outside the business. Such analysis can only be programmed with the help of BI.

5) To manage business at all level and in all sectors: In today's competition it's required to manage business through large number of data, communicated many forms of media. BI can be used to handle such flow of data and analyse as per the requirements. BI is now a day applicable to every sector of business likes, retail, healthcare, transportation, insurance, banking and others.

BI Components

A BI system has four major components: a *data warehouse*, with its source data; *business analytics*, a collection of tools for manipulating, mining, and analysing the data in the data warehouse; *business performance management (BPM)* for monitoring and analysing performance; and a *user interface* (e.g., a dashboard).

i) Data Warehouse In simple terms, a data warehouse (DW) is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to managers throughout the organization. Data

are usually structured to be available in a form ready for analytical processing activities (i.e., online analytical processing [OLAP], data mining, querying, reporting, and other decision support applications). A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process.

ii) Business Analytics Business analytics (BA) refers to all the methods and techniques that are used by an organization to measure performance. Business analytics are made up of statistical methods that can be applied to a specific project, process or product.

iii) Business Performance Management Business performance management (BPM) is a form of business intelligence used to monitor and manage a company's performance. Key performance indicators (KPI) are used for this purpose.
These KPIs include revenue, return on investment, overhead and operational costs. Business performance management is also known as corporate performance management (CPM).

iv) User Interface User Interface (UI) is a broad term for any system, either physical or software based, that allows a user to connect with a given technology.
Many different kinds of user interfaces come with various devices and software programs.

A user interface, also sometimes called a human-computer interface, comprises both hardware and software components. It handles the interaction between the user and the system. There are different ways of interacting with computer systems which have evolved over the years. There are five main types of user interface:

- **Command Line Interface:** Command line interfaces are the oldest of the interfaces discussed here. It involves the computer responding to commands typed by the operator. It means interacting with a computer program where the user issues commands to the program in the form of successive lines of text.

- **Graphical UI** Graphical user interfaces (GUI) are sometimes also referred to as WIMP because they use *Windows*, *Icons*, *Menus* and *Pointers*. Operators use a pointing device (such as a mouse, touchpad or trackball) to control a pointer on the screen which then interacts with other on-screen elements. It

allows the user to interact with devices through graphical icons and visual indicators such as secondary notations.

- **Menu Driven** A menu driven interface is commonly used on cash machines (also known as automated teller machines (ATM's), ticket machines and information kiosks (for example in a museum). They provide a simple and easy to use interface comprised of a series of menus and sub-menus which the user accesses by pressing buttons, often on a touch-screen device.

- **Form Based** A form-based interface uses text-boxes, drop-down menus, text areas, check boxes, radio boxes and buttons to create an electronic form which a user completes in order to enter data into a system. This is commonly used on websites to gather data from a user, or in call centres to allow operators to quickly enter information gathered over the phone.

- **Natural Language** A natural language interface is a spoken interface where the user interacts with the computer by talking to it. Sometimes referred to as a 'conversational interface', this interface simulates having a conversation with a computer. Commonly used by telephone systems as an alternative to the user pressing numbered buttons the user can speak their responses instead. This is the kind of interface used by the popular iPhone application called Siri and Cortana in Windows.

Ch. 2 – Business Analytics

Introduction

Business analytics (BA) is the practice of iterative (interaction of mathematical o computational process), methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies committed to data-driven decision-making.
Business analytics (BA) refers to all the methods and techniques that are used by an organization to measure performance. Business analytics are made up of statistical method that can be applied to a specific project, process or product.

Business analytics can also be used to evaluate an entire company. Business analytics are performed in order to identify weaknesses in existing processes and highlight meaningful data that will help an organization prepare for future growth and challenges.

Successful business analytics depends on data quality, skilled analysts who understand the technologies and the business, and an organizational commitment.

Data Analysis and Data Analytics

There are two terms called, analysis and analytics. According to the Merriam-Webster dictionary, *analysis* is "a detailed examination of anything complex in order to understand its nature or to determine its essential features: a thorough study." *Analytics* is defined as "the method of logical analysis."

Data Analysis assesses the requirements of the business and sees how functions and processes can be used to improve performance and outcomes. Data analysis helps is breaking down the macro picture into a micro picture to rule out human bias with the help of statistical analysis.
Data Analytics is more exhaustive and detailed business practice that starts with identifying which data to analyze, collecting the right data, and then organizing that data into the right data sets using the right algorithms and statistical techniques.
Thus, the question that data analysis answers is, "What happened?" whereas data analytics answers, Why did it happen and what will happen next?

Need

1. To analyze data from multiple sources.
2. To Monitor KPIs (Key Performance Indicators) and react to changing trends in real-time.
3. To justify and revise decisions based on up-to-date information.
4. To reduce overall cost.
5. To sustain in competition.
6. To improve quality of decisions.
7. To make complete analysis of past and present data for the forecasting.

BA Components

There are 6 major components or categories in any analytics solution;

- **Data Mining –** Create models by uncovering previously unknown trends and pattern in vast amounts of data e.g. detect insurance claims frauds, Retail Market basket analysis. There are various statistical techniques through which data mining is achieved.

- **Classification** (when we know on which variables to classify the data e.g. age, demographics)

- **Regression**

- **Clustering** (when we don't know on which factors to classify data)

- **Associations & Sequencing Models**

- **Text Mining –** Discover and extract meaningful patterns and relationships from text collections e.g. understand sentiments of Customers on social media sites like Twitter, Face book, Blogs, Call centre scripts etc. which are used to improve the Product or Customer service or understand how competitors are doing.

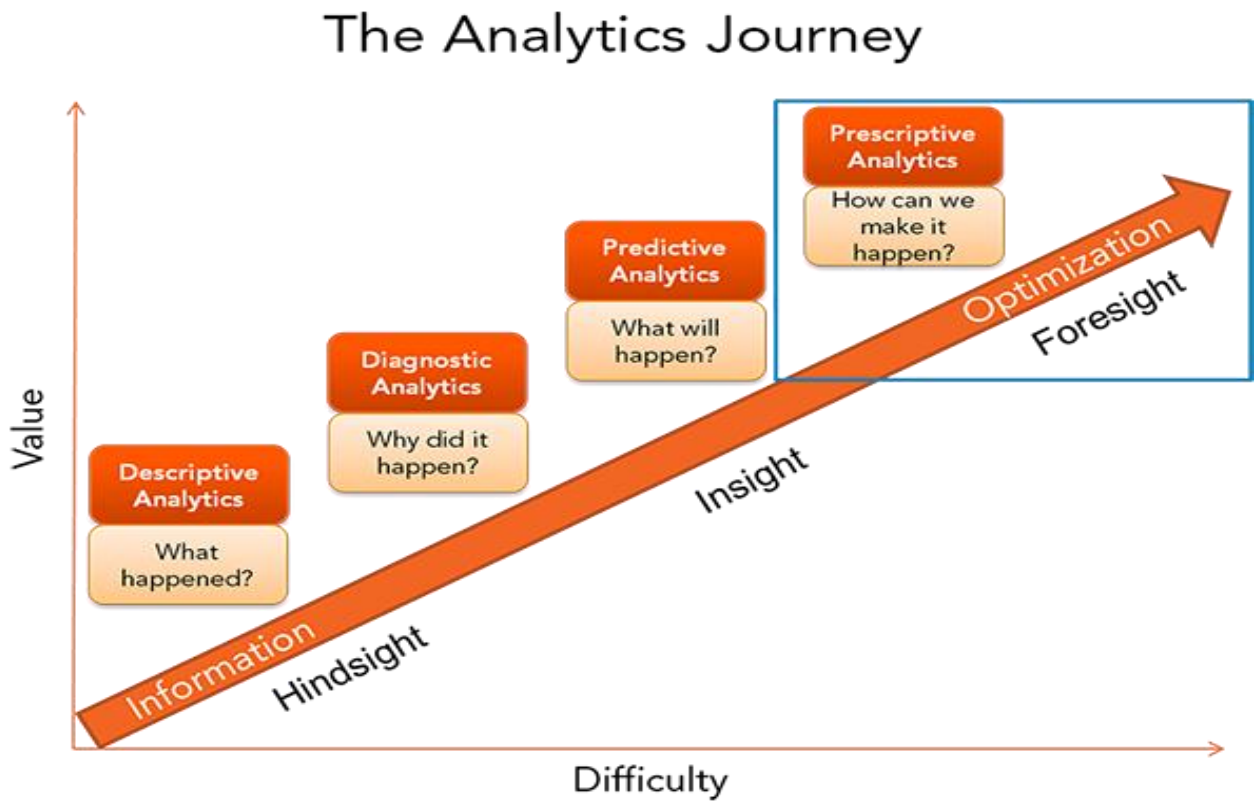
- **Forecasting –** Analyze & forecast processes that take place over the period of time e.g. predict seasonal energy demand using historical trends, predict how many ice creams cones are required considering demand

- **Predictive Analytics–** Create, manage and deploy predictive scoring models e.g. Customer churn &retention, Credit Scoring, predicting failure in shop floor machinery

- **Optimization –** Use of simulations techniques to identify scenarios which will produce best results e.g. Sale price optimization, identifying optimal Inventory for maximum fulfilment & avoid stock outs.

- **Visualization –** Enhanced exploratory data analysis & output of modelling results with highly interactive statistical graphics.

Types of BA



i) Descriptive Analytics

Descriptive analytics **answer the 'What happened?'** – As the name suggests, it analyzes the raw data of the past and gives it meaning. The past could refer to

something that occurred a week ago or two years ago. Through the use of data aggregation and data mining, it allows us to learn from past behaviors, and see how the past interactions can impact the future results.

Descriptive analytics looks at a variety of metrics like web traffic, monthly sales, and average rupees spent per customer, inventory levels, top product lines sold and provide insights around the company's production, finance, sales, operations, and customers.

In other words, it is the application of simple statistical techniques that describes what is contained in a data set or database. Example: An age bar chart is used to depict retail shoppers for a department store that wants to target advertising to customers by age.

Here, the purpose is to identify possible trends in large data sets or databases. The purpose is to get a rough picture of what generally the data looks like and what criteria might have potential for identifying trends or future business behavior.

Here, the methodology may include; Descriptive statistics, including measures of central tendency (mean, median, mode), measures of dispersion (standard deviation), charts, graphs, sorting methods, frequency distributions, probability distributions, and sampling methods.

ii) Predictive Analytics

Predictive analytics predicts the probability of something happening in the future or in other words answers the **'What is likely to happen'** question. It identifies past patterns and uncovers relationships between different data sets by using various techniques such as data mining, statistics, modeling, machine learning and artificial intelligence.

The outcome of predictive analytics is the probability and likelihood of future events, risks and opportunities. Examples of predictive analytics include forecasting customer behavior and purchasing patterns, inventory levels, sales activities etc.
In other words, it's an application of advanced statistical, information software, or operations research methods to identify predictive variables and build predictive models to identify trends and relationships not readily observed in a descriptive analysis.

Example: Multiple regressions are used to show the relationship (or lack of relationship) between age, weight, and exercise on diet food sales. Knowing that relationships exist helps explain why one set of independent variables influences dependent variables such as business performance.

Here, the purpose is to build predictive models designed to identify and predict future trends.

Here, the methodology may include; Statistical methods like multiple regression and ANOVA. Information system methods like data mining and sorting. Operations research methods like forecasting models.

iii) Perspective Analytics

Prescriptive analytics is a relatively new field of analytics which tries to answer the question **'What should we do about it'**. Prescriptive analytics uses extensive statistical methods and tools to prescribe (advise) a number of different potential actions and offers guidance on the best course of action. Companies get the 'advise' on how to optimize for future events.

Some of the techniques and tools used for prescriptive analytics are a combination of business rules, algorithms, machine learning and computational modelling. Prescriptive analytics are relatively harder to implement and most companies are not currently using them. However, when prescriptive analytics are applied and executed properly, they can have a very significant impact on the company's earnings and revenues.

In other words, it's an application of decision science, management science, and operations research methodologies (applied mathematical techniques) to make best use of allocable resources. Example: A department store has a limited advertising budget to target customers. Linear programming models can be used to optimally allocate the budget to various advertising media.

Here, the purpose is; to allocate resources optimally to take advantage of predicted trends or future opportunities.

Here, the methodology may include; Operations research methodologies like linear programming and decision theory.

Points	Business Intelligence	Business Analytics
Meaning	BI involves the process of collecting data from all sources and preparing it for Business Analytics (BA)	BA is the analysis of the answer required by the Business Intelligence. (BI)
Type of Questions/ Answers the Questions	✓ What Happened? ✓ When did it happened? ✓ Where did it happened? ✓ Who is responsible for what happened? ✓ How often?	✓ Why did it happened? ✓ Will it happen again? ✓ What will happen if we change X? ✓ What is the best that can happen?
	✓ How many?	
Makes Use of/ Methods	✓ Reporting (KPIs, Metrics) ✓ Automated monitoring/altering (threshold) ✓ Dashboards/Scorecards ✓ OLAP, etc.	✓ Statistical/Quantitative analysis ✓ Data Mining ✓ Predictive Modelling ✓ Text/Multi-media Mining, etc.
Data Types	Structured and sometime unstructured	Structured and unstructured both
Knowledge Generation	Manual	Automatic
Users	Business users only	✓ Data Scientists, ✓ Business Analysts, ✓ Business Users
Business Initiative	Reactive	Pro-active
Focus	✓ Information delivery and Reporting ✓ Data Visualisation ✓ Data Integration	✓ Business rules ✓ Data attributes (structures, quality and sources)

Transaction processing v/s Analytic Processing

OLTP

OLTP means On Line Transaction Processing. It supports transaction oriented application in the information systems or TPS. It administers day-to-day's transaction of an organisation. *The primary objective of OLTP is data processing and not data analysis.*

For example: ATM centre, online banking, online ticket booking, Order entry, sending text messages, online shopping etc.
Some advantages of OLTP may be;

It administers daily transactions of an organisation.

- It helps to increase the customers of an organisation by simplifying individual process.
- It is designed typically for use by clerks, cashiers, clients, etc. (Simplicity)
- It allows its users to read, write and delete data quickly. (Efficiency)
- It responds to user actions immediately and also supports transaction processing on demand. (Fast query processing)

Some challenges of OLTP may be;

- OLTP system faces hardware failures which affect the online transactions.
- It allows multiple users to access and change the same data at the same time, which creates unprecedented situation.
- Security – An OLTP system requires concurrency control (locking) and recovery mechanisms (logging).
- OLTP system data content not suitable for decision making – A typical OLTP system manages the current data within an enterprise/organization. This current data is far too detailed to be easily used for decision making.

OLAP

OLAP means On Line Analytical Processing. It is a category of software tool which provide analyses of data for the business decisions. OLAP system allows users to

analyse database information from multiple database system at one time. *The primary objective of OLAP is data analysis and not data processing.*

For example: any data warehouse system is an OLAP system, uses of OLAP are;

- A company might compare its mobile phone sales in September with the sales in October, and then compare those results with another location which may store in a separate database.
- Amazon analysis, purchase by its customers to come-up with a personalised home page with product which likely interest to their customers.

Types of OLAP Servers

We have four types of OLAP servers −

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)

(1) Relational OLAP

In ROLAP, data is stored in a relational database. In essence, each action of slicing and dicing is equivalent to adding a “WHERE” clause in the SQL statement.

Advantages:

- Can handle large amount of data
- Can leverage functionalities inherent in the relational database

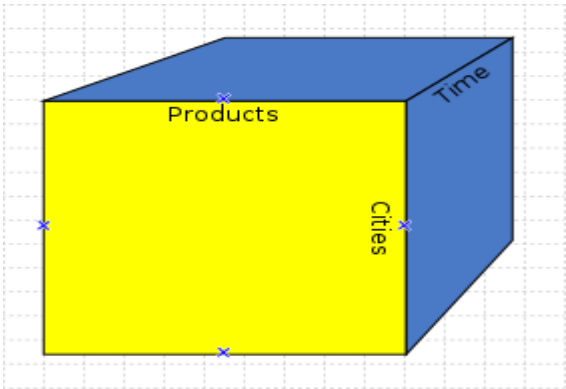
Disadvantages:

- Difficult to perform complex calculation using SQL.

Performance can be slow.

(2) Multidimensional OLAP

In MOLAP, data is stored in a multidimensional cube. The storage is in proprietary formats and not in the relational database.



Advantages:

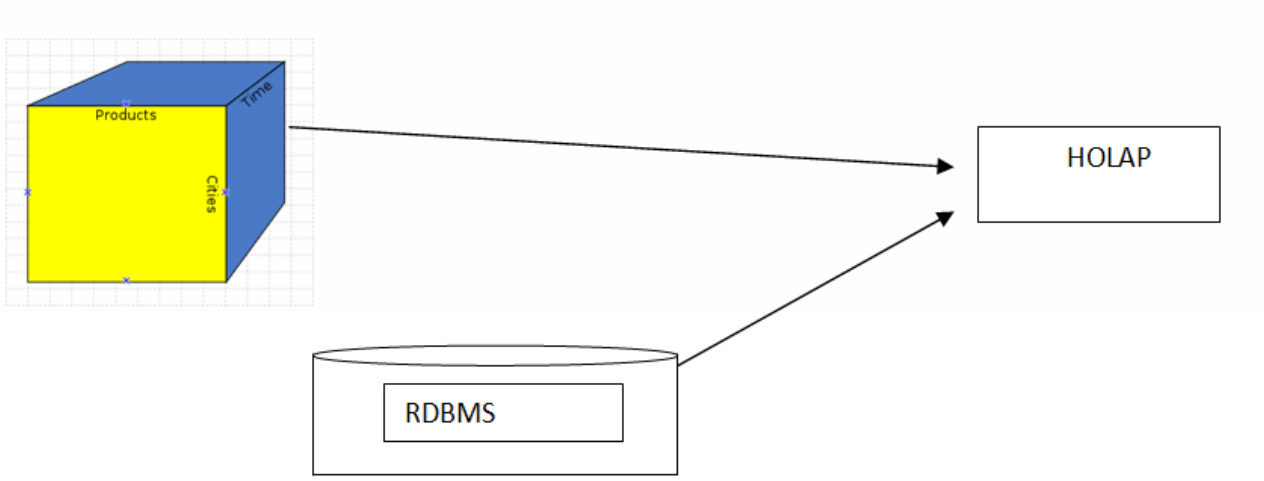
- Fast data retrieval
- Optimal for slicing and dicing
- Can perform complex calculations.

Disadvantages:

- Limited in the amount of data that it can handle. The reason being as all calculations are pre-generated when the cube is created, it is not possible to include a large amount of data.
- Additional investment in human and capital resources may be required as the cube technology is proprietary and might not exist in the enterprise.

Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allow storing the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.



OLAP Operations

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

Here is the list of OLAP operations −

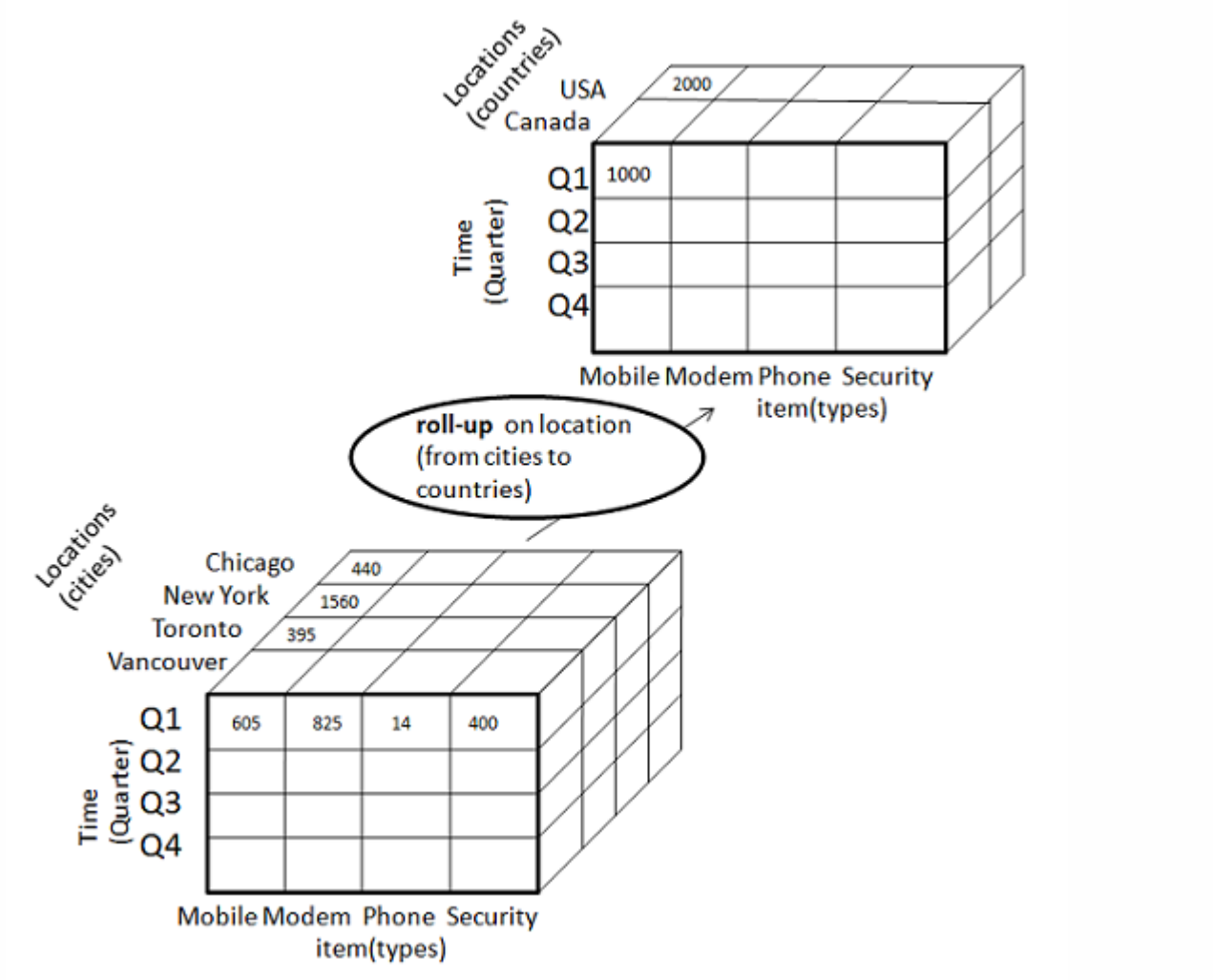
- Roll-up
- Drill-down
- Slice and dice
- Pivot rotate

(1) Roll-up

Roll-up performs aggregation on a data cube in any of the following ways −

- By climbing up a concept hierarchy for a dimension
- By dimension reduction

The following diagram illustrates how roll-up works.



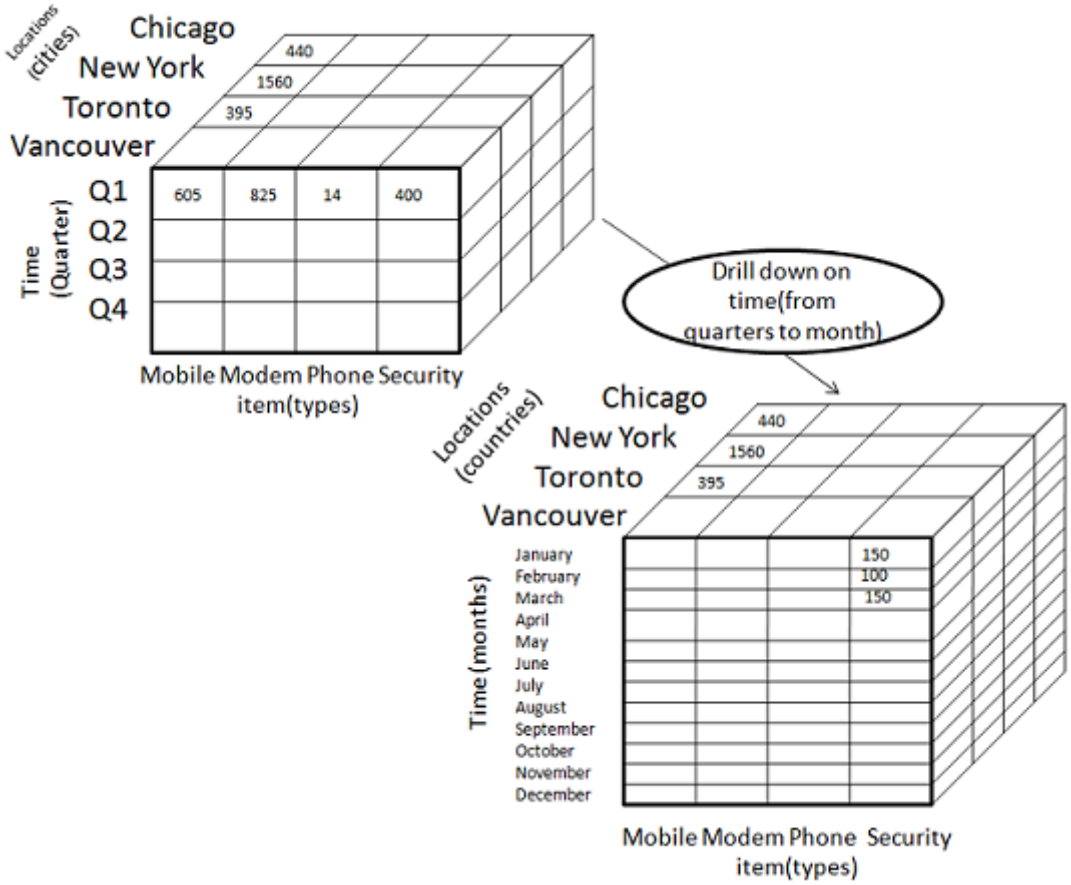
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

(2) Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways −

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

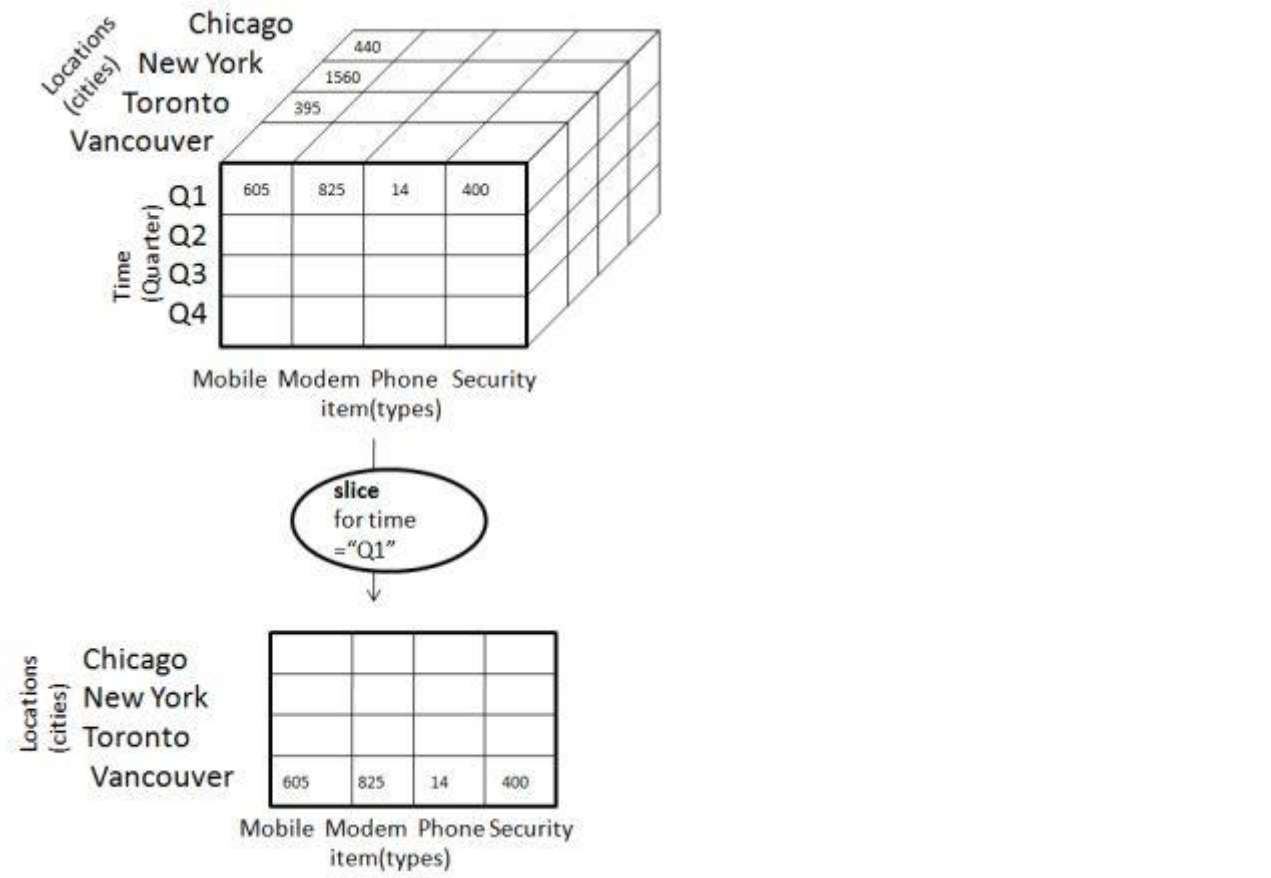
The following diagram illustrates how drill-down works −



- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

(3) Slice

The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



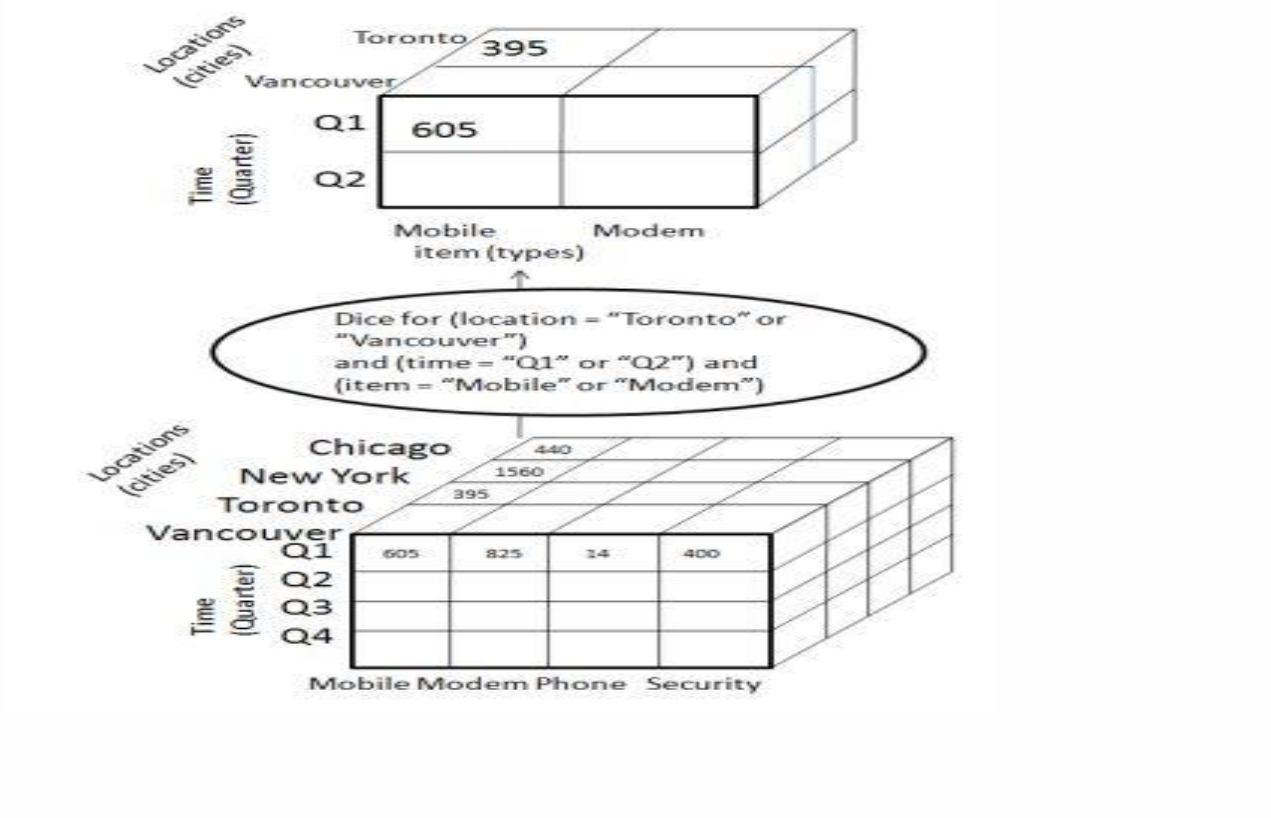
- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

(4) Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.

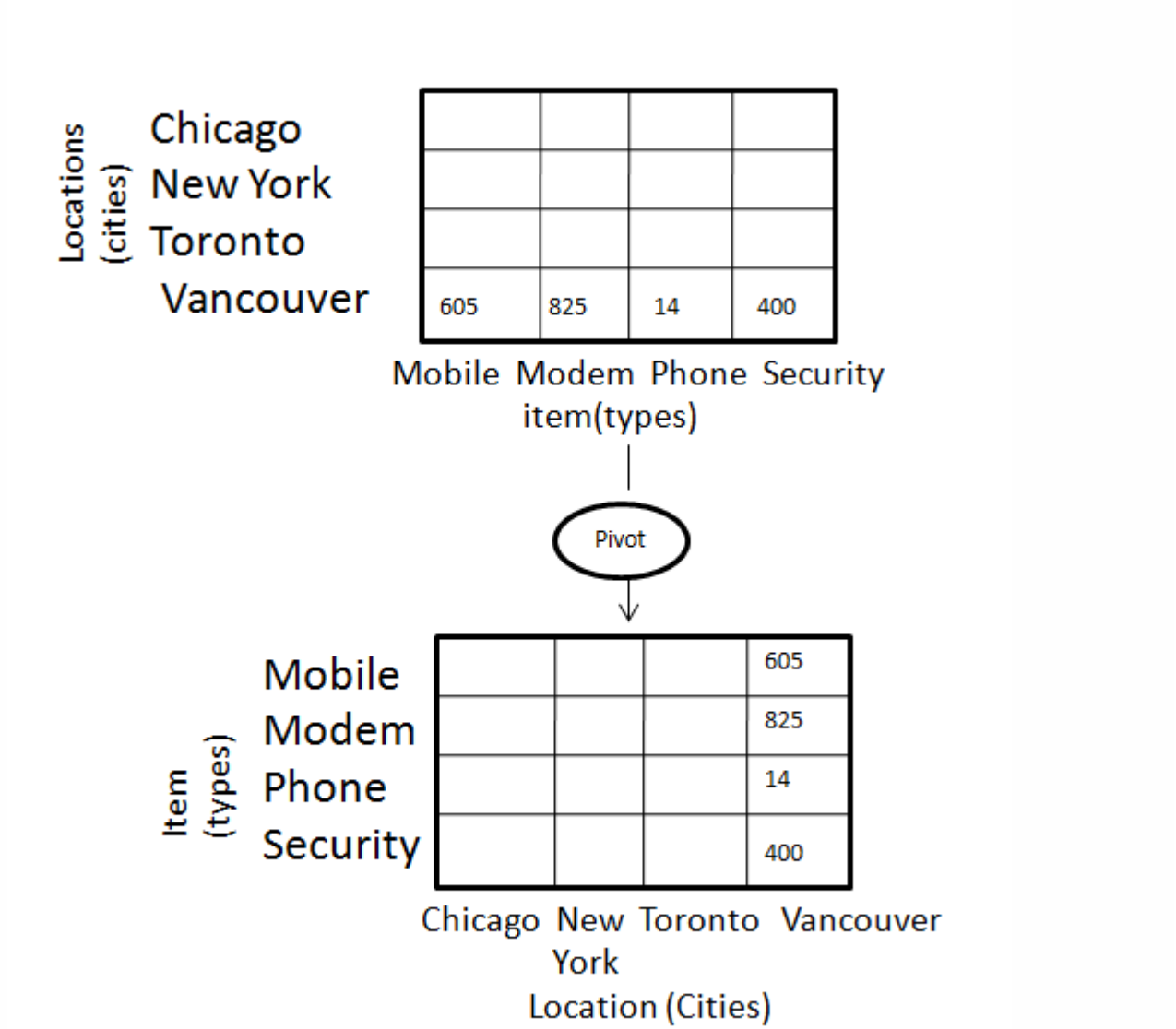
The dice operation on the cube based on the following selection criteria involves three dimensions.

- location="Toronto"or"Vancouver"location="Toronto"or"Vancouver"
- time="Q1"or"Q2"time="Q1"or"Q2"
- item="Mobile"or"Modem"item="Mobile"or"Modem"



(5) Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.



Advantages of OLAP System:

- Multidimensional data representation.
- Consistency of information.
- “What if” analysis.
- Provides a single platform for all information and business needs – planning, budgeting, forecasting, reporting, and analysis.
- Fast and interactive ad hoc exploration.

Drawbacks of OLAP may be;

- Implementation and maintenance are depending on IT progression which takes high costs.
- OLAP tools need cooperation between people of various departments to be effective which might not always be possible.

_____

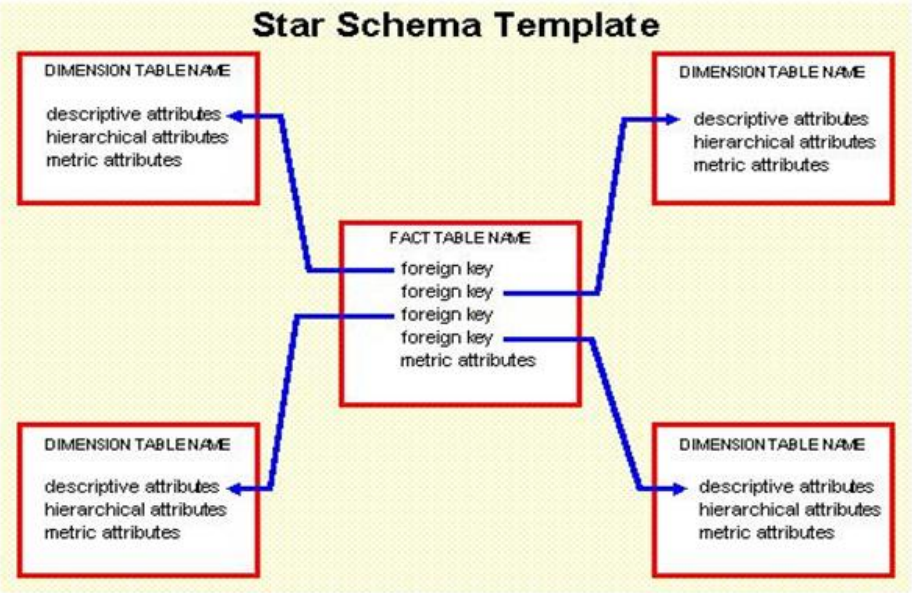
DATA MODEL OF OLAP

A multi dimensional can exist in the form of a star schema,

- A Snowflake schema
- A Star model
- A Fact Constellation/ Galaxy Schema

Star model:

It is the simplest of data warehousing schema. It consists of a large central table (called fact table) with no redundancy. The central table is being referred by a number of dimension tables. The schema graph looks like a starburst. The star schema is always very effective for handling queries.



In the star schema, the fact table is usually in 3NF of higher form of normalization. All the dimension tables are usually in de-normalized manner, and the highest form of normalization they are usually present in is 2NF.

The dimension tables are known as lookup or reference tables.

Following Figure shows the Star schema for "ElectronicsForAll".

Here sales are considered along four dimensions, i.e. Time, Product, Employee, and Customer. The schema diagram shows a central fact table for "Sales". The "Sales" central fact table has the keys to each of the four dimensions along with three measures- Total, Quantity and Discount. Each dimension is represented by only one table. Each table further has a set of attributes.

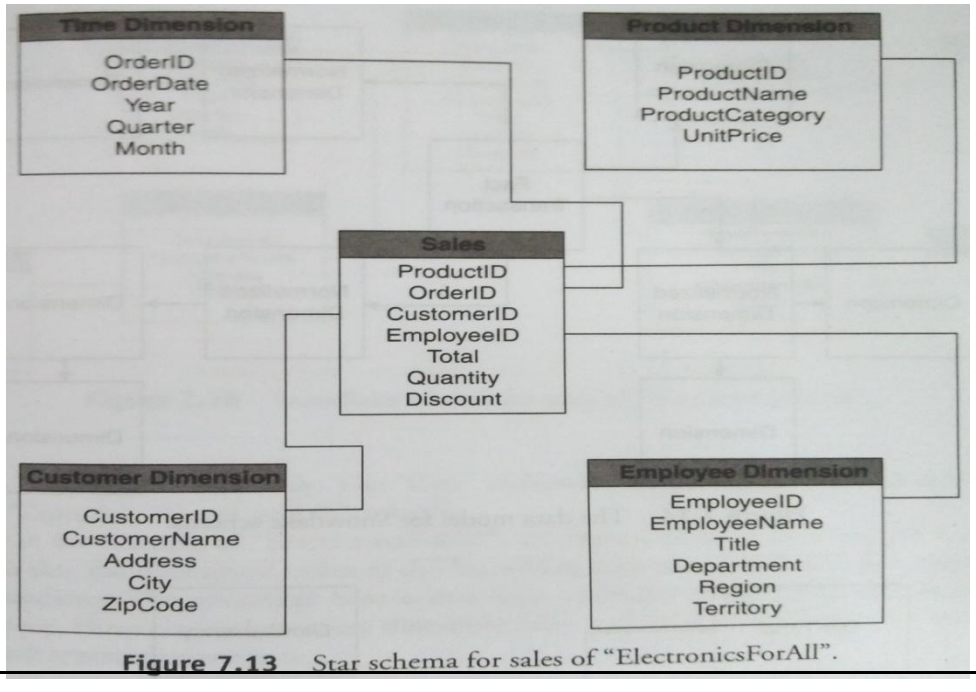
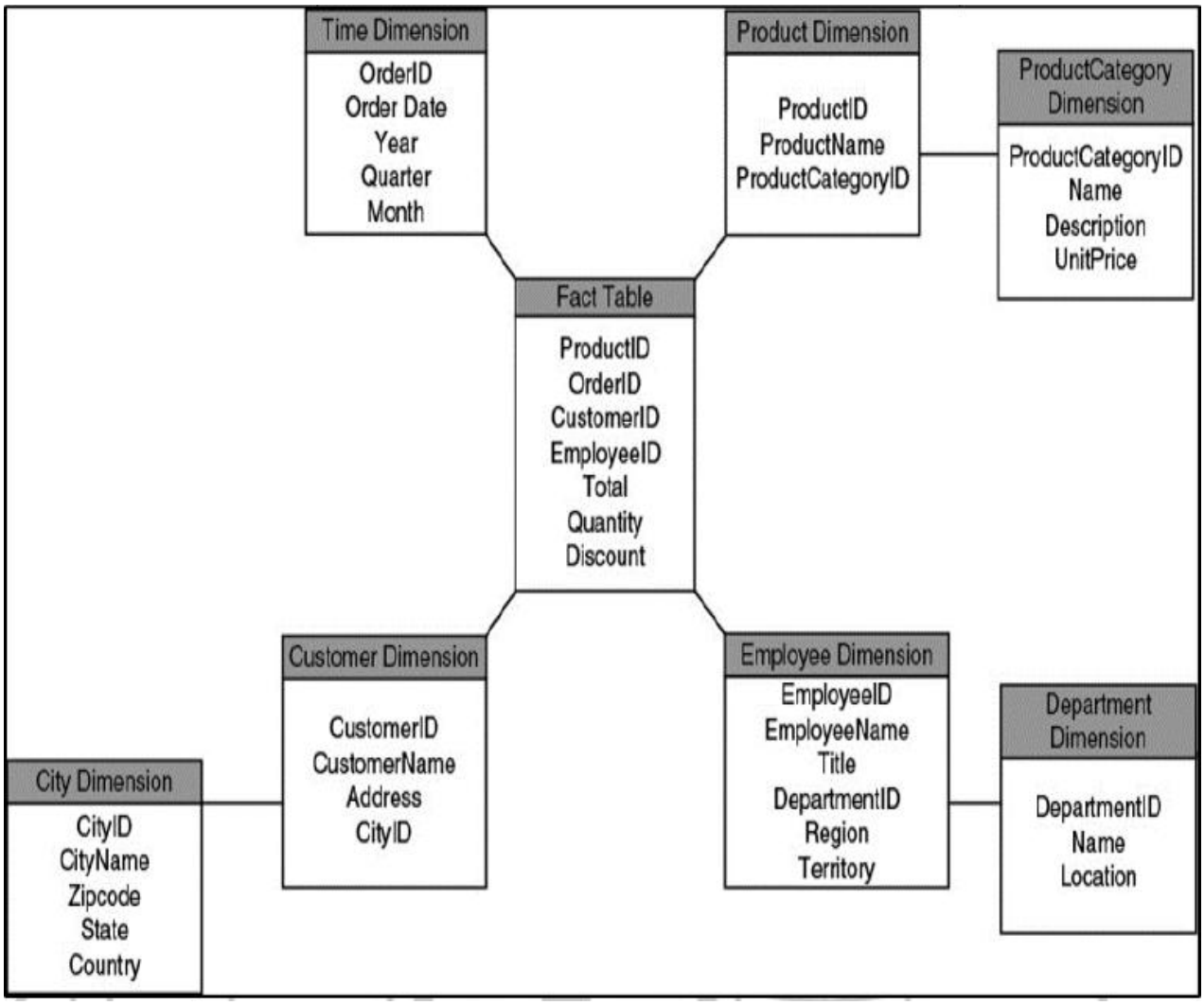


Figure 7.13 Star schema for sales of "ElectronicsForAll".

Snowflake Schema

The following figure shows the snowflake model. There is a central fact table connected to four dimensions. The 'product' dimension is further normalized to 'product category' dimension. Similarly, the 'employee' dimension is further normalized to the 'department' dimension. By now, you would have guessed that normalization of the dimension tables definitely helps in reducing redundancy; however it adversely impacts the performance as more joins will needed to execute a query.



Data Model for OLTP:

An OLTP system adopts an Entity Relationship (ER) model. **Entity Relationship (ER) Modelling** is a logical design technique whose main focus is to reduce data redundancy (Idleness). It is basically used for transaction capture and can

contribute in the initial stage of constructing a data warehouse. The reduction in the data redundancy solves the problems of inserting, deleting and updating data. The whole process ended up with creation of lots of tables and joins between these tables. It results a massive spider web of joins between tables.

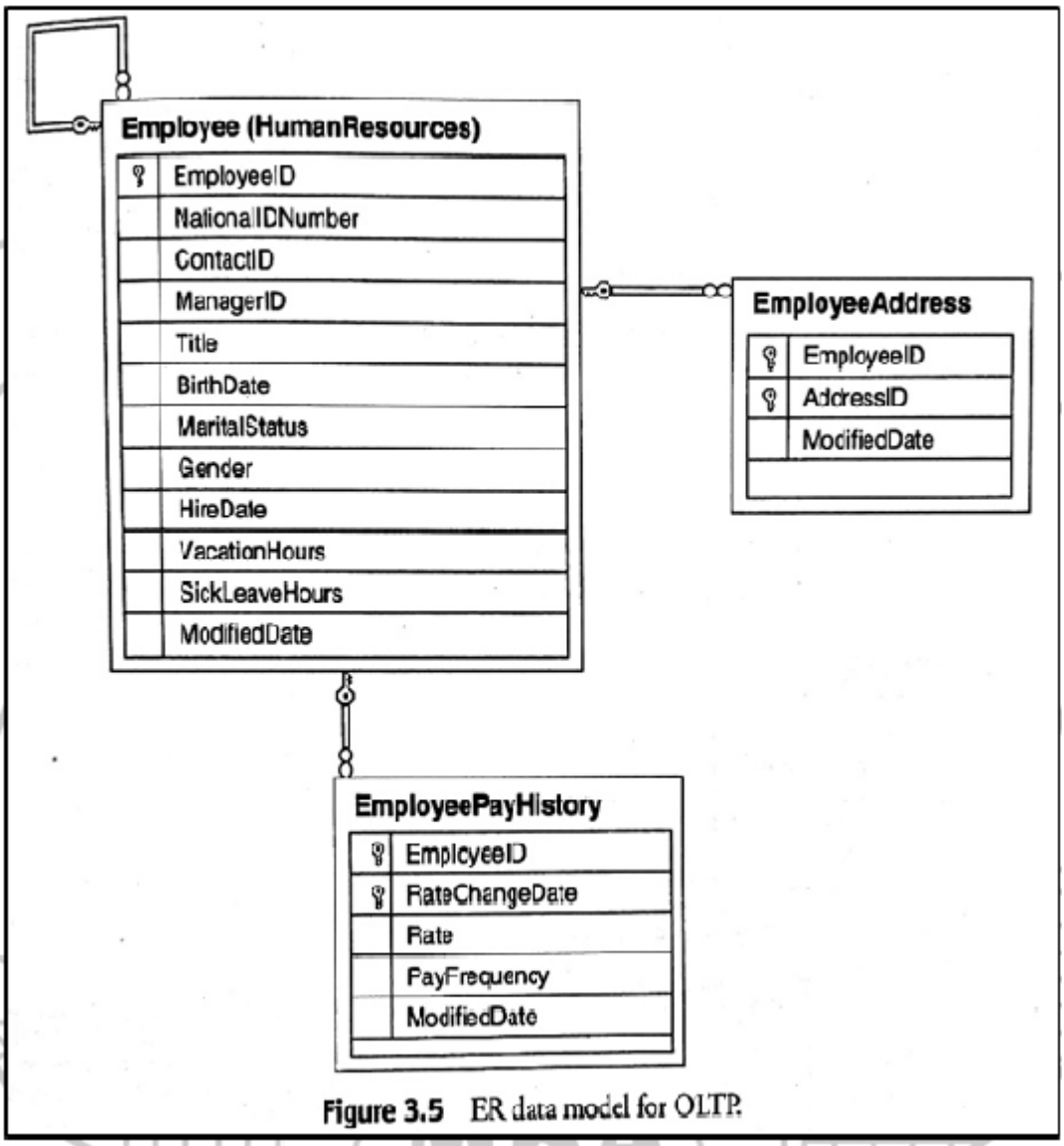


Figure 3.5 ER data model for OLTP

The figure shows an Entity Relationship (ER) data model for OLTP. We have considered following three Entities;

1. Employee (Employee ID is the primary key).

2. Employee Address (Employee ID is a foreign key referencing to the Employee ID attribute of Employee entity.)

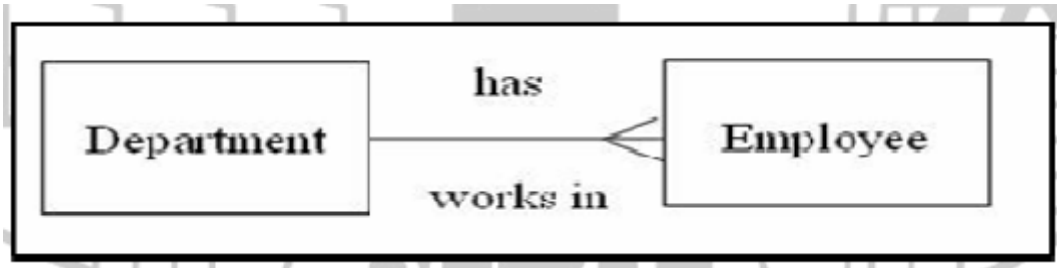
3. Employee Pay History (Employee ID is a foreign key referencing to the Employee ID attribute of Employee entity.)

For these entities, we see the following Relationships;

a. There is a (1: M cardinality) between Employee and Employee Address entities. This means that an instance of Employee entity can be related with multiple instances of Employee Address entity.
b. There is also (1: M cardinality) between Employee and Employee Pay History entities. This means that an instance of Employee entity can be related with multiple instances of Pay History entity.

The degree of relationship (also known as cardinality) is the number of occurrences in one entity which are associated (or linked) to the number of occurrences in another. There are three degrees of relationship, known as:
I. one-to-one (1:1)
II. one-to-many (1: M)
III. many-to-many (M: N)

One-to-Many (1: M): It is, where one occurrence in an entity relates to many occurrences in another entity. For example, taking the employee and department entities shown on the previous page, an employee works in one department but a department has many employees. Therefore, there is a one-to-many relationship between department and employee.

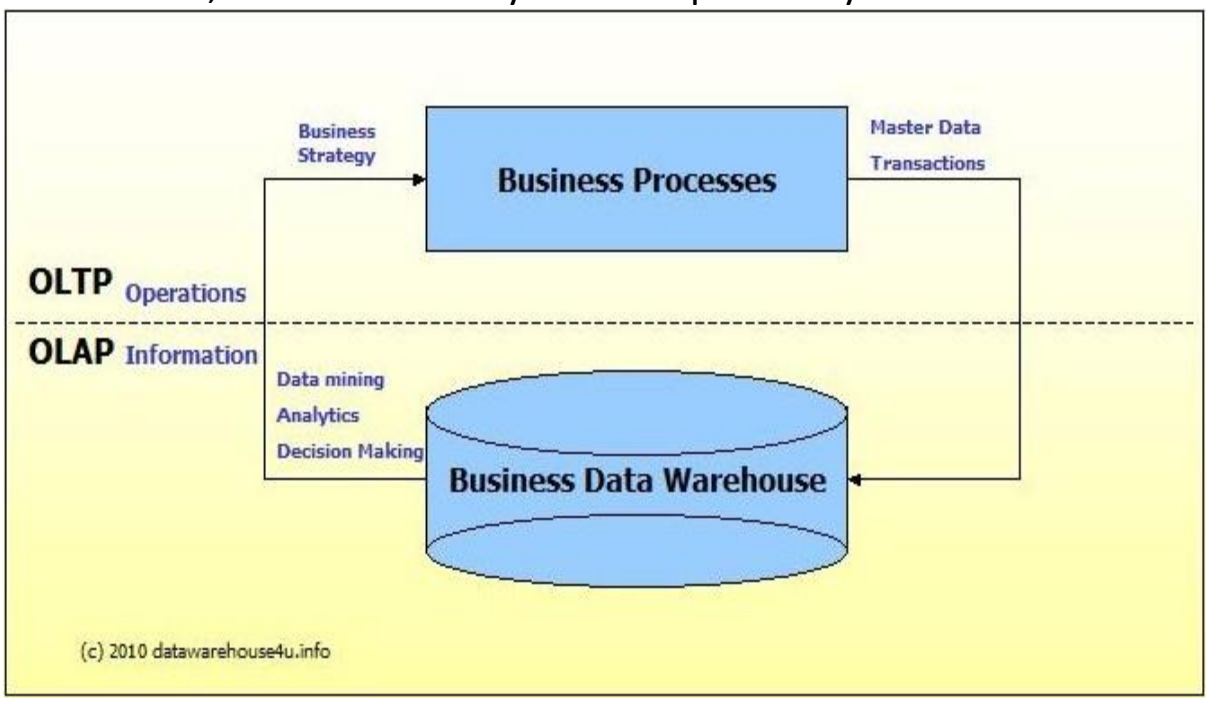


Differences between Star and Snowflake Schema

Property	Star Schema	Snowflake Schema
Ease of maintenance / change	Easy to maintain due to low redundancy.	Difficult to maintain due to high redundancy.
Facts and Dimension Properties	Dimension Tables are normalized, Fact tables are denormalized	Dimension Tables and tables are denormalized
Ease of Use	Difficult to understand to due to increased queries	Easier to understand due to simple queries
Query Performance	Poor, due to increased complexity in joins.(increased foreign keys)	Good, less complexity.(Less foreign keys).
Type of Data warehouse	Complex Relations (Many to Many)	Simple Relations (One to One/ One to Many)
When to use	Greater size of dimension tables, snowflake schema helps reduce space.	Smaller size of dimension tables.

OLTP vs. OLAP

We can divide IT systems into transactional (OLTP) and analytical (OLAP). In general we can assume that OLTP systems provide source data to data warehouses, whereas OLAP systems help to analyze it.



OLTP (On-line Transaction Processing) is characterized by a large number of short on-line transactions (INSERT, UPDATE, and DELETE). The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second. In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF).

OLAP (On-line Analytical Processing) is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

The following table summarizes the major differences between OLTP and OLAP system design.

FEATURES	OLTP SYSTEM ONLINE TRANSACTION PROCESSING (OPERATIONAL SYSTEM)	OLAP SYSTEM ONLINE ANALYTICAL PROCESSING (DATA WAREHOUSE)
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
Data Contents	Current data	Historical data. Stores and manages data at various levels of granularity, thereby suitable for decision making
Inserts and Updates	Very frequent updates and inserts	Periodic long-running updates to refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Queries usually take a long time to execute and return
Space Requirements	Can be relatively small if historical data is archived	Comparatively huge because of the existence of aggregation structure and historical data
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas

Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method
Access	Field level access	Typically aggregated access to data of business interest
Data Model	ER Model	Dimensional Model
Operations	Read/Write	Mostly read
Joins	Many	Few
Indexes	Few	Many
Data Structure	Complex	Multidimensional
Derived data and aggregates	Rare	Common

Sample Queries	• Search & locate student(s) record(s) • Print Students score • Filter records where student(s) have scored above 90% marks	• Which courses have productivity impact on-the-job? • How much training is needed on future technologies for non linear growth in BI?

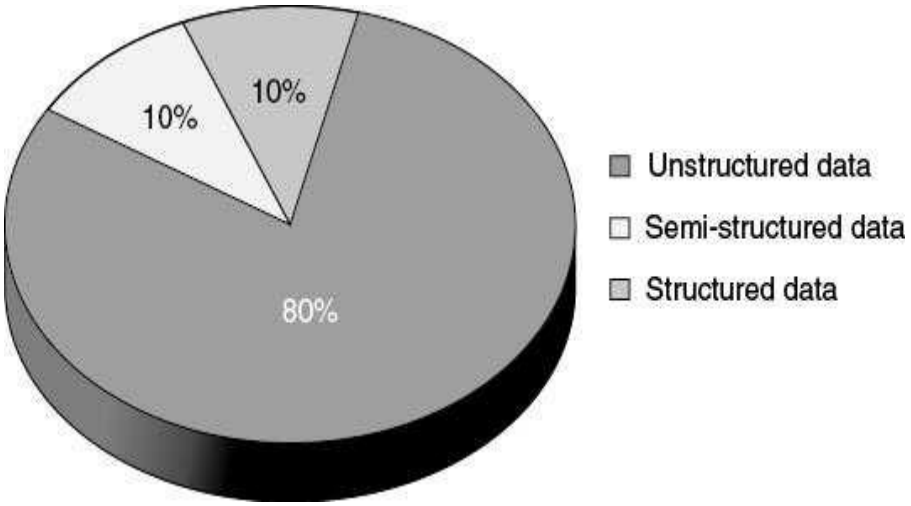
MODULE 2
CHAPTER 1: TYPES OF DIGITAL DATA

❖ Introduction:

Data refers to a collection of facts usually obtained as the result of experiences, observations, or experiments. Data may consist of numbers, letters, words, images, voice recording, and so on as measurements of a set of variables. Data are often viewed as the lowest level of abstraction from which information and then knowledge is derived. The need and growth of Data has been increased after advent and evolution of internet across the world. Today's battle for the corporate and government institutions are for acquiring and analyse maximum possible data. In digital form, there are mainly three types of data;

1) Structured Data;
2) Unstructured Data; and
3) Semi-structured Data.

Generally, every data is in the unstructured format and which makes collecting information from it quite difficult. According to a survey, 90% of the total business data is either unstructured or semi-structured form. Moreover, the unstructured data constitutes 80% of the whole business data. In short the percentage distribution of the forms of data can be viewed in the given image.



1) Structured Data

Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis. Structured data is easy to handle in terms of storage, scalability, update and delete.

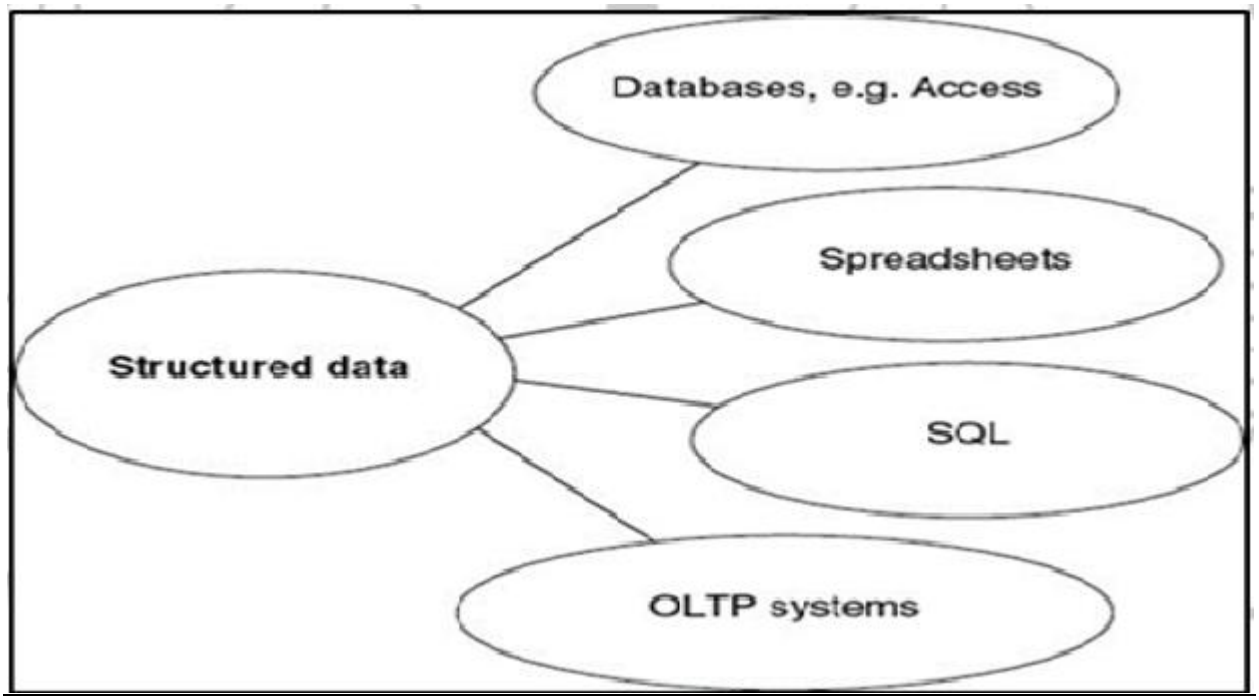
In short, structured data is organized in semantic chunks (entities), similar entities are grouped together (relations or classes), Entities in the same group have the same descriptions (attributes), Descriptions for all entities in a group (schema) have the same defined format, have a predefined length and follow the same order.

Definitions

Structured data refers to any data that resides in a fixed field within a record or file.
This includes data contained in relational databases and spreadsheets.

Structured data denotes the data in an organised form (in rows and columns) and can be easily used by computer programme. The data stored in data base may be the best example of the structured data. In other words, data stored and structure used in almost every research and business requirements.

Sources



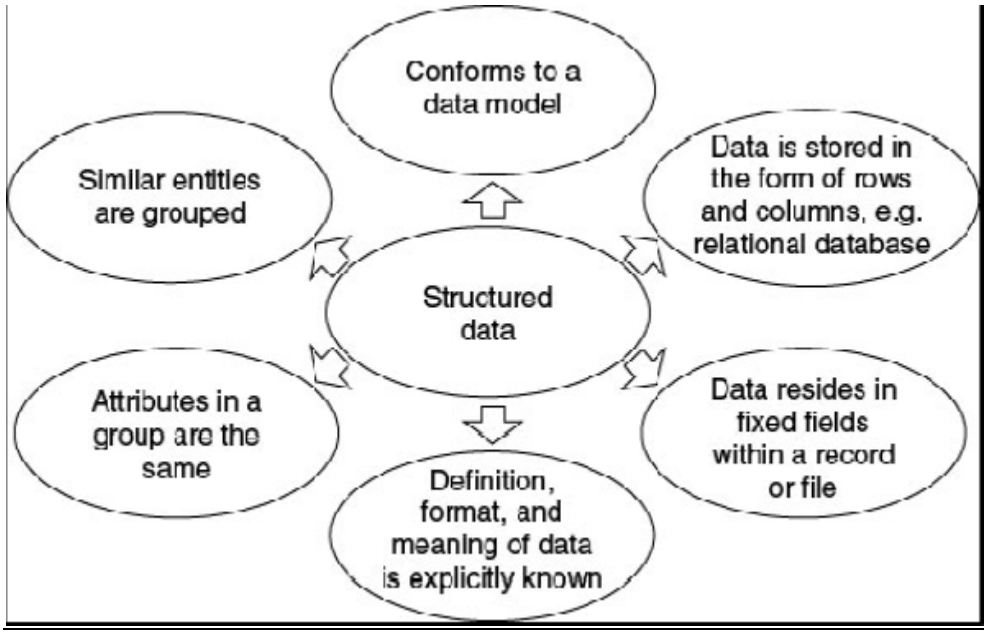
1) Databases: A database (DB), in the most general sense, is an organized collection of data. More specifically, a database is an electronic system that allows data to be easily accessed, manipulated and updated. In other words; a database is used by an organization as a method of storing, managing and retrieving information. One of the sources of structured data is database where structured data organised in rows and columns.

2) <u>Spreadsheets</u>: Spreadsheet is a computer application or programme where a file is made of rows and columns that help to short arrange and calculate data. It is also known as worksheet. Data stored in spreadsheets is structured data.

Some examples of spreadsheets are; MS Excel, Google Sheets (Google Docs), Numbers (iWork – by Apple Inc.)

3) <u>SQL</u>: ***S****tructured* ***Q****uery* ***L****anguage* is a standard language for storing, manipulating and restoring data in data bases. SQL is used to communicate with a data bases. It is also the standard language for **D**ata **B**ase **M**anagement **S**ystem. Query means a request for the data or information form a data base table or combination of tables.
SQL is particularly useful in handling structured data where there are relations between different entities/variables of the data.
4) <u>OLTP</u>: ***O****n* ***L****ine* ***T****ransaction* ***P****rocess* is an information system that helps in manages transaction oriented application for the data entry and other transaction process.
It administers day-to-day transactions of an organisation. For example; ATM, Online Ticket Booking etc.

Characteristics



2) Unstructured Data

Unstructured data does not confirm to a data model or information which can be used easily by a computer programme. Unstructured data cannot be stored in the form of rows and columns in a database. It is difficult to determine the meaning of such data. Unstructured data does not follow any rules. Such data can be of any time and therefore unpredictable and constitute 80% of the total business data in the world.

For example; Power Point Presentations, Images, Videos, Letters, memos etc.

Definitions

Unstructured data represents any data that does not have a recognizable structure. It is
Unorganized and raw and can be non-textual or textual. For example, email is a fine illustration of unstructured textual data. It includes time, date, recipient and sender details and subject, etc. but an email body remains unstructured.

Unstructured data also may be identified as loosely structured data, wherein the data sources include a structure, but not all data in a data set follow the same structure.

Sources:

Anything in a non-database form is unstructured data. It can be classified into two broad categories:

- Bitmap objects: For example, image, video, or audio files.
- Textual objects: For examples, Microsoft Word documents, emails. Or Microsoft Excel spreadsheets.

A Myth Demystified

Web pages are said to be unstructured data even though they are defined by HTML, a mark-up language which has a rich structure.

HTML is solely used for rendering and presentations. The tagged elements do not capture the meaning of the data the HTML page contains. This makes it difficult to automatically process the information in the HTML page.

Another characteristic that makes web pages unstructured data is that they usually carry links and references to external unstructured content such as images, XML files, etc.

Storage:

The challenges faced while storing unstructured data are as follows:

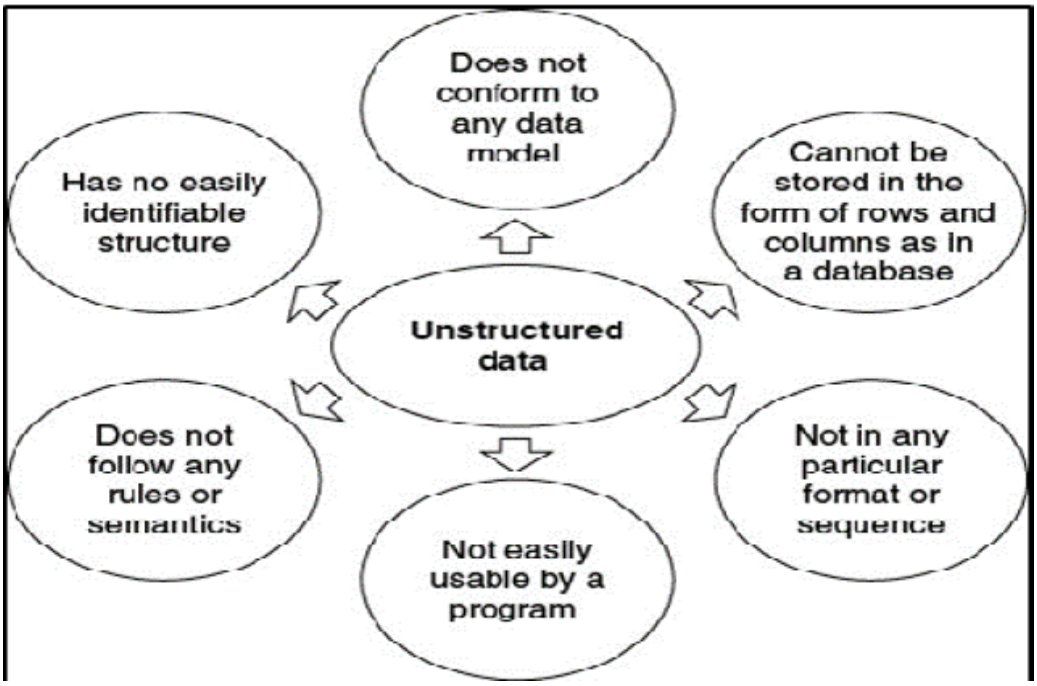
- **Storage space**: It is difficult to store and manage unstructured data. A lot of space is required to store such data. It is difficult to store images, videos, audios etc...
- **Scalability**: As the data grows, scalability becomes an issue and the cost of storing such data grows.
- **Retrieve information:** Even if unstructured data is stored, it is difficult to retrieve and recover from it.
- **Security**: Ensuring security is difficult due to varied sources of data, e.g. emails, web pages, etc.
- **Update and delete:** Updating and deleting unstructured data very difficult as retrieval is difficult due to no clear structure.
- **Indexing and searching:** Indexing unstructured data is difficult and error-prone as the structure is not clear and attributes are not pre-defined. As a result, the search results are not very accurate. Indexing becomes all the more difficult as the volume of data grows.

Now let's look at a few possible solutions described below:

- **Changing format**: Unstructured data may be converted to formats which are easily managed, stored and searched. For example, IBM is working on providing a solution which will convert audio, video, etc. to text.
- **Developing new hardware**: New hardware needs to be developed to support unstructured data. It may either complement the existing storage devices or may be a stand alone for unstructured data.
- **Storing in XML format**: Unstructured data may be stored in XML format which tries to give some structure to it by using tags and elements.

- **CAS(Content Addressable Storage)**: It organizes files based on their metadata and assigns a unique name to every object stored in it. The object is retrieved based on its content and not its location. It is used extensively to store emails, etc.

Characteristics



3) Semi-structured Data

Semi-structured data is data that is neither raw data, nor typed data in a conventional database system. It is structured data, but it is not organized in a rational model, like a table or an object-based graph. A lot of data found on the Web can be described as semi-structured.

Data integration especially makes use of semi-structured data. XML, other mark-up languages, email, and EDI (Electronic Data Interchange) are all forms of semi-structured data.

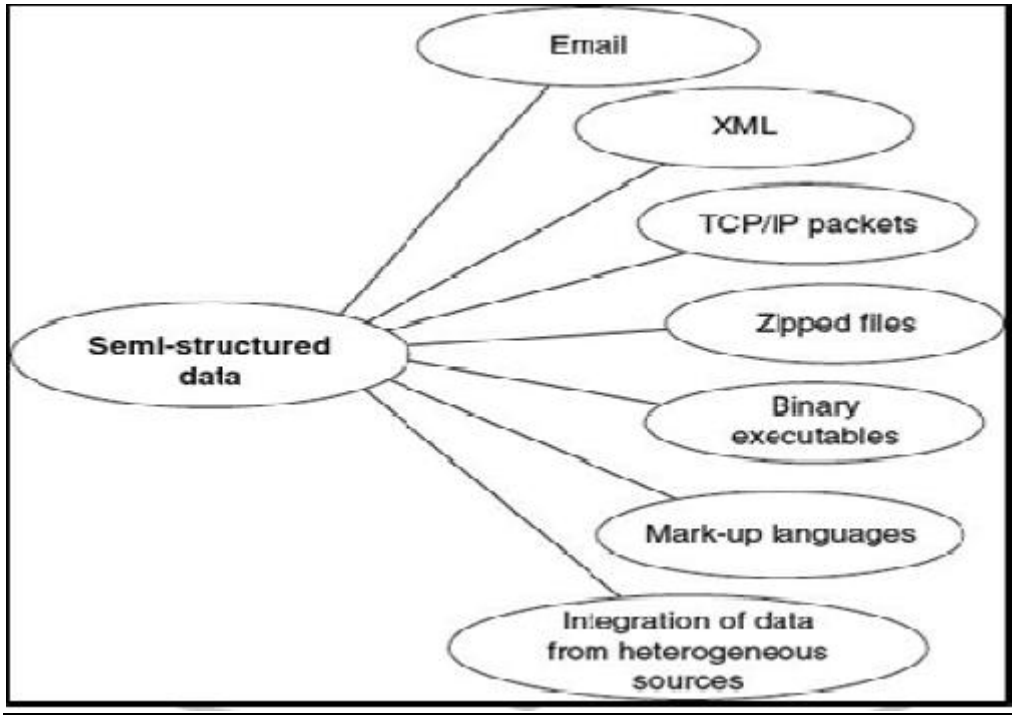
Definitions

Semi-structured data does not confirm to a data model but has some structure. It is difficult to determine a meaning of such data. This data cannot be stored in rows and columns as in database. Semi-structured data however has tags and markers which help grouping the data and describe how the data is stored.

Such data is not having features that sufficient for management. Here, similar entities are grouped and organised in a hierarchy. Semi-structured data is not in a form which can be easily used by a computer programme. About 10% of the data in the organisation stands for such data.

A blood test report may consider as Semi- structured data because it has s semi-structured data. fields like Date, Patient Name, Department, etc. data as well as unstructured fields like Diagnosis, conclusion etc.

Sources



1) E-mail: E-mail is having structured as well as unstructured data. Thus, e-mail is considered as one of the sources of semi-structured data.

2) XML: XML (E**x**tensible **M**ark-up **L**anguage) is also a source of semi - structured data as it allows data to be stored in a sequential manner.

3) TCP/IP Packets: TCP (***T**ransmission **C**ontrol **P**rotocol*) or IP (***I**nternet **P**rotocol*) is a communicate suite (*a set of rules and procedures*) used to transmit data on internet.
TCP/IP specifies how data is exchanged over the internet by providing end-to-end communications that identify how it should be broken into packets, addressed, transmitted, routed and received at the destination.

TCP defines how applications can create channels of communication across a network and IP defines how to address and route each packet to make sure it reaches the right destination. As in TCP/IP data is having varied versions, it is a source of semi-structured data.

4) Zipped Files: A zipped file is a computer file whose contents are compressed for the storage or transmission. A zipped file may contain one or more files with different characteristics. For example; documents, images, videos, etc. in one folder. Thus, it can be viewed as the source of semi-structured data.

5) **Binary Executables:** A binary file is a file stored in binary format. A binary file is computer-readable but not human-readable. All executable programs are stored in binary files, as are most numeric data files. In contrast, text files are stored in a form that is human-readable. Such files also one of the sources of semi-structured data.

6) Mark-up language: Mark-up language is a computer language that uses tags (*tag means a cod used to specify the formatting*) to define elements in a documents. Mark-up languages are designed for processing, definition and presentation of text. For example; XML.

7) Integration of data from heterogeneous sources: Heterogeneous means diverse in contents or mixed. Sometimes data are collected from heterogeneous/different sources where specific formats are not available. Thus, such data making of semi-structured data.

Characteristics:

- IT is organized into semantic entities.
- Similar entities are grouped together.
- Entities in the same group may not have same attributes.
- The order of attributes is not necessarily important.
- Not always all attributes are required.

- Size of the same attributes in a group may differ.
- Type of the same attributes in a group may differ.

Let us see with an example how entities can have different set of attributes. The attribute may also have different data types and most certainly can have different sizes. For example, names and emails of different people can be stored in more than one way as shown below.

One way is:

Name: Patrick Wood
Email: ptw@dcs.abc.ac.uk, p.wood@ymail.uk

Another way is:

First Name: Patrick
Last Name: Wood
Email: p.wood@ymail.uk

Integration of data from heterogeneous sources leads to the data being semi-structured. Because it is likely that data from one source may not have adequate structure while others may have information which is not required or the required information missing from them.

Challenges for Storage

There are various issues associated with storage of semi- structured data which are as;

1) Storage Cost: Storing data with their schemas (*organization or structure for a database*) increases cost as such data having mixed structures.
2) RDBMS: Semi-structured data cannot be stored in existing RDBMS (*Relational Data Base Management System*- is a system which prepares a database that stores data in structured format, using rows and columns) as data cannot be mapped into tables directly.

3) Irregular and Partial Structure: Some data elements may have extra information while others none at all, which create irregular and partial structure that lead to problems in storage.

4) Implicit Structure: In many cases the structure is implicit (*not clear/vague*). Interpreting relationships and correlations is very difficult in such structure.

5) Distinction between Schema and Data: Vague distinction between schema and data exists at times making it difficult to capture data

Ch. 2 Data Warehouse

1. Meaning:

Data warehouse is a centralised repository for all the data that various business systems of the organisation collect. Typically, the historical data of the enterprise is organised by subject area areas such as employees, marketing, customers, and products and so on.

The purpose of creating data warehouse is to guide management in decision making with integrated and summarised facts. Such data warehouse creation will also require ETL tools, multidimensional data storage and OLAP reporting or data mining tools.

Knowledge workers analyse data in the warehouse form multiple perspective like time, customer type, business location etc. to make business decision and stay ahead of competition.

2. Definition:

According to William H. Inmon, "*A data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process.*"

3. Characteristics:

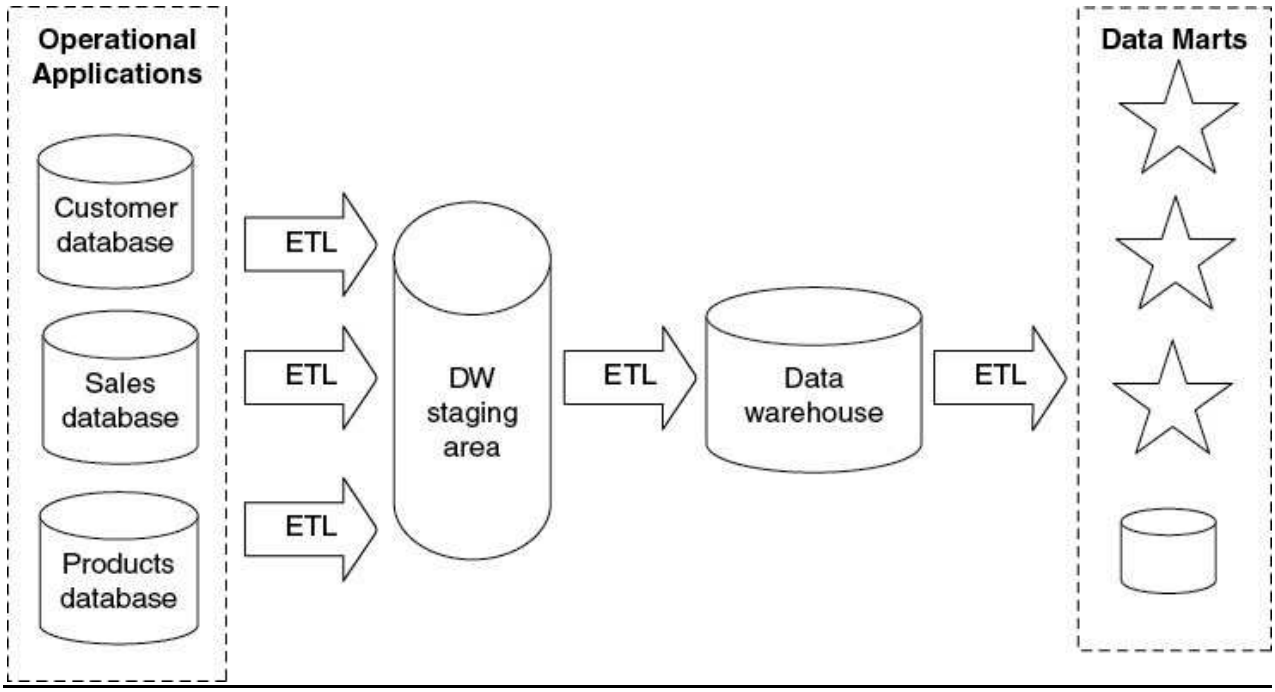
1) Subject oriented: A data warehouse collects data of subjects such as, customers, suppliers, partners, sales products, etc. spread across the enterprise or organisation.

2) Integrated: A data warehouse will serve to bring together the data from multiple disparate (different) sources after careful cleansing and transformation into a unified format to serve the information needs of the enterprise.

3) Time-variant: A data warehouse keeps historical data while and OLTP system will usually have the most up-to-date data. Form data warehouse, one can retrieve data that is 3 months, 6 months, 12 months or even older. For example; a transaction system may hold the most recent address of a customer, whereas a data warehouse can hold all address associated with a customer recorded, say; over the last five years.

4) Non-volatile: A data warehouse is a separate physical store of data transformed from the application data found in the operational environment.

4. Framework



ETL

ETL (***E**xtract **T**ransform and **L**oad*) is a three stage process in database usage, especially in data warehousing. It allows integration and analysis of data stored in different sources. After collecting the data from multiple varied sources (*extraction*), the data is reformatted and cleansed (to detect and rectify errors) to

meet the information needs (*transformation*) and then sorted, summarised and loaded into the desired end target (*loading*). In short; ETL is,

- Extracting data from different data sources.
- Transforming the extracted data into relevant format to fit information needs.
- Loading data into the final target database, usually a data warehouse.

Data Marts

Data marts can be defined as subset of data warehouse of an organisation which is limited to specific business units or group of users. It is a subject oriented data base and also **H**igh **P**erformance **Q**uery **S**tructure (HPQS). Data marts enable users to retrieve information for single departments or subjects, improving the user response time. Data marts exist within a single organizational data warehouse repository.

Data marts improve end-user response time by allowing users to have access to the specific type of data they need to view. Each data mart is dedicated to a specific business function or region. This subset of data may span across many or all of an enterprise's functional subject areas. It is common for multiple data marts to be used in order to serve the needs of each individual business unit (different data marts can be used to obtain specific information for various enterprise departments, such as accounting, marketing, sales, etc.).

Data Lake

A Data lake is a storage repository that holds a vast amount of collected raw data in its native format until it is needed for processing. Data lakes reduce data integration challenges.

The data lake supports the following capabilities to;

- Capture and store raw data at scale for low cost.
- Store many typed of data (structured to unstructured) in the same repository.
- Perform transformation on the data.
- Allow different communities of users like data scientists to gain access to raw data for their processing needs.

In other words, a data lake is a massive, easily accessible, centralized repository of large volumes of structured and unstructured data. A data lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, analytics and machine learning. A data lake can include structured data from relational databases (rows and columns), semi-structured data (XML), unstructured data (emails, documents, PDFs) and binary data (images, audio, and video).

Data Lakes allow you to import any amount of data that can come in real-time. Data is collected from multiple sources, and moved into the data lake in its original format. This process allows you to scale to data of any size, while saving time of defining data structures, schema, and transformations. The ability to harness more data, from more sources, in less time, and empowering users to collaborate and analyse data in different ways leads to better, faster decision making. Examples where Data Lakes have added value include; Improved customer interactions, Improve R&D innovation choices, Increase operational efficiencies etc.

A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose.

	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Data Structure	Raw	Processed
Price/ performance	Fastest query results using higher cost storage	Query 8results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

Ch. 3 – Business Reporting and Visual Analytics

Business Reporting

Definitions

A report is a communication article prepared with the specific intention of relaying information in a presentable form. If it concerns business matters, then it is called a business report.

A business report is a written document that contains information regarding business matters. Business reporting (also called enterprise reporting) is an essential part of the larger drive toward improved managerial decision making and organizational knowledge management. The foundation of these reports is various sources of data coming from both inside and outside the organization.

Business reporting or enterprise reporting refers to both "the public reporting of operating and financial data by a business enterprise," and "the regular provision of information to decision-makers within an organization to support them in their work."

Concept

Creation of these reports involves ETL (extract, transform, and load) procedures in coordination with a data warehouse and then using one or more reporting tools. Although reports can be distributed in print form or via e-mail, they are typically accessed via a corporate intranet.

Business reporting is an essential part of the business intelligence movement toward improving managerial decision making. Nowadays, these reports are more visually oriented, often using colours and graphical icons that collectively look like a dashboard to enhance the information content.

Due to the rapid expansion of information technology coupled with the need for improved competitiveness in business, there has been an increase in the use of computing power to produce unified reports that join different views of the enterprise in one place.

Usually, this reporting process involves querying structured data sources, most of which were created using different logical data models and data dictionaries, to produce a human readable, easily digestible report. These types of

business reports allow managers and co workers to stay informed and involved, review options and alternatives, and make informed decisions.

Types of Business Reports

Even though there are a wide variety of business reports, the ones that are often used for managerial purposes can be grouped into three major categories;

1) Metric Management Reports In many organizations, business performance is managed through outcome-oriented metrics. For external groups, these are service-level agreements (SLAs). For internal management, they are key performance indicators (KPIs). Typically, there are enterprise-wide agreed targets to be tracked against over a period of time. They may be used as part of other management strategies such as Six Sigma or Total Quality Management (TQM).

2) Dashboard-Type Reports A popular idea in business reporting in recent years has been to present a range of different performance indicators on one page, like a dashboard in a car. Typically, dashboard vendors would provide a set of predefined reports with static elements and fixed structure, but also allow for customization of the dashboard widgets, views, and set targets for various metrics. It's common to have color-coded traffic lights defined for performance (red, orange, green) to draw management attention to particular areas.

3) Balanced Scorecard–Type Reports This is a method developed by Kaplan and Norton that attempts to present an integrated view of success in an organization. In addition to financial performance, balanced scorecard–type reports also include customer, business process, and learning and growth perspectives.

Components of Business Reports

Following are the most common components of a business reporting system.

● **OLTP (online transaction processing):** A system that measures some aspect of the real world as events (e.g., transactions) and records them into enterprise databases. Examples include ERP systems, POS (Point of Sales) systems, Web servers, RFID (Radio Frequency Identification) readers, handheld inventory readers, and card readers.

● **Data supply:** A system that takes recorded events/transactions and delivers them reliably to the reporting system. The data access can be push or pull, depending on whether or not it is responsible for initiating the delivery process. It can also be polled (or batched) if the data are transferred periodically, or triggered (or online) if data are transferred in case of a specific event.

● **Business logic:** The explicit steps for how the recorded transactions/events are to be converted into metrics, scorecards, and dashboards.

● **ETL (extract, transform, and load):** The intermediate step where these recorded transactions/events are checked for quality, put into the appropriate format, and inserted into the desired data format.

● **Data storage:** The storage area for the data and metadata. It could be a flat file or a spreadsheet, but usually is a relational database management system (RDBMS) set up as a data mart, data warehouse, or operational data store (ODS). Data storage often employs online analytical processing (OLAP) functions likes cubes.

● **Publication:** The system that builds the various reports and hosts them (for users) or disseminates them (to users). These systems may also provide notification, annotation, collaboration, and other services.

● **Assurance:** Quality service offered to users by a good business reporting system. This includes determining if and when the right information is to be delivered to the right people in the right way/format.

Reporting Perspectives Common to all levels of Enterprise:

Typically enterprises have headquarters and several regional centres. Several geographic location focused operations may aggregate to the nearest regional centre. Each geographic location may have “revenue generating – customer facing units” and “support units”. There could be regional or corporate level support functions as well. IT could be leveraged at the local level or the regional level or corporate level. Hence, it is natural to expect IT – enabled reporting to occur at local, regional, or corporate levels.

- **Functional level**: Reports being generated at the functional level may be consumed by users within a department or geographic location or region or

by decision makers at the corporate level. Departments such as HR, marketing, production, purchase, accounting, etc. will need specific standard reports to handle operational, tactical, and strategic challenges. Reports could be produced in many languages to meet global user needs.

- **Internal/external**: Sometimes the consumers of reports may be external to the enterprise. We are very familiar with the annual reports of organizations. Correctness as well as attractive presentation of the report is of paramount importance.

- **Role based:** Today we are witnessing massive information overload. The trend is to provide standard format of report to similar roles across the enterprise, as they are likely to make similar information/ facts for decision making irrespective of the country/products he/she handles.
- **Strategic/operational:** Reports could also be classified based on the nature of the purpose they serve. Strategic reports inform the alignment with the goals, whereas operational reports present transaction facts. The quarterly revenue report indicates variance with regard to meeting targets, whereas the daily cash flow summary indicates summary of day's business transactions. When consolidated across several locations, regions, products/services, even this report will be of strategic importance.
- **Summary/detail:** As the name suggests, summary reports do not provide transaction-level information, whereas detailed reports list atomic facts. Even here several summaries could be aggregated to track enterprise-level performance.
- **Standard/ad hoc:** Departments tend to generate periodic reports, say weekly, monthly, or quarterly reports in standard formats. Executives many times need ad hoc or on demand reports for critical business decision making.
- **Purpose:** Enterprises classify reports as statutory that focus on business transparency and need to be shared with regulatory bodies. For example, a bank reporting to the Reserve Bank stipulated parameters of its operations. Audit reports that are produced to check the correctness and consistent application of business polices across global transactions. Analytical reports look into a particular area of operation like sales, production, and procurement, and they find patterns in historical data. These reports typically represent large data interpretations in the form of graphs.

- **Technology platform-centric:** Reporting in today's context need not use paper at all. Dashboards could be delivered on smart phones and tablets. Reports could be published in non editable (secure) form with watermarks. Reports could be protected to be used by a specific person, during specific hours from specific device! Reports could be delivered to the target user in user-preferred formats such as worksheets, word document, PowerPoint presentations, text file or HTML document, and so on. Reports could be generated once and shared with many users through an email link as well.

Features of good reporting

Feature	Description
Report Title	It is important to provide a crisp name for the report such that it reflects its purpose. Some teams may even add the target audience. Example: **Cash flow report for *SOUTH* Region** **Product Shipping Report for Plant 2**
Reporting period	As the reports use data collected over a specific period, it is critical to state the same. The period format could be: For week beginning March DD,YYYY From DD/MM/YYYY to DD/MM/YYYY
Header/footer	It is good to include report headers and footers that repeat on every page. The content could have elements like logo, function name, page number etc.
Column heading	The numeric data presented will need to be read based on the column names. Again keeping crisp but meaningful names is critical.
Column selection and sequence	Although reports are meant to be used by users in the same role, but across different locations, users have their own preferences when they want to see the information being presented. There needs to be flexibility for users to select the column that they would like to see as well as the order or sequence from left to right. Example – Microsoft Explorer and Microsoft Outlook allow one to select from a list of available columns and also display it in a preferred sequence.
Filters	Users may not be interested to see all the lines simultaneously. They need to have flexibility to use standard filters/customs filters to view lines that meet specific criteria. Example: **Cash Flow for Department = "XXX"** **Cash Flow for Amount> 100000.00**
Sort sequence	Users would like to view reports lines arranged in ascending or descending order for convenience of decision making. Example: **Names to be arranged in alphabetical order.** **Revenue Report to be in decreasing order of amount.** It may need to sort lines in cascading fashion as well. BY Department + BY Employee Name

	BY Customer + BY Date + BY Invoice Number
Totals/group totals	When data lines are sorted, they may need to be grouped or bunched together in chunks that make business sense. In this situation, users expect totals and cross totals as well as overall totals to be computed/provided.
Data filed formatting	Users also expect the stored data to be represented with formatting to enhance reading convenience. Using currency symbols like $, etc. Date formatting like June 5, 2018
Computed/calculated fields	Users may want to introduce new columns that are derived from existing columns and compute some value for decision making.
Highlighting breaches	Reports may highlight using color or font size/style to make the field seize the attention of the user.
Notes	Sometimes it may essential to notify users of last-minute update message that could answer typical question raised by user.

Common Report Layout Types

- **Tabular reports**: Tabular reports may have a finite number of columns. Typically representing the fields in a database. A tabular report has header and footer, and repeating detail rows. Data can be grouped on various fields. Each group can have its own header, footer, breaks, and subtotal. Table reports are typically used for logging detailed transactions.

Table 3: Average level of air components by section

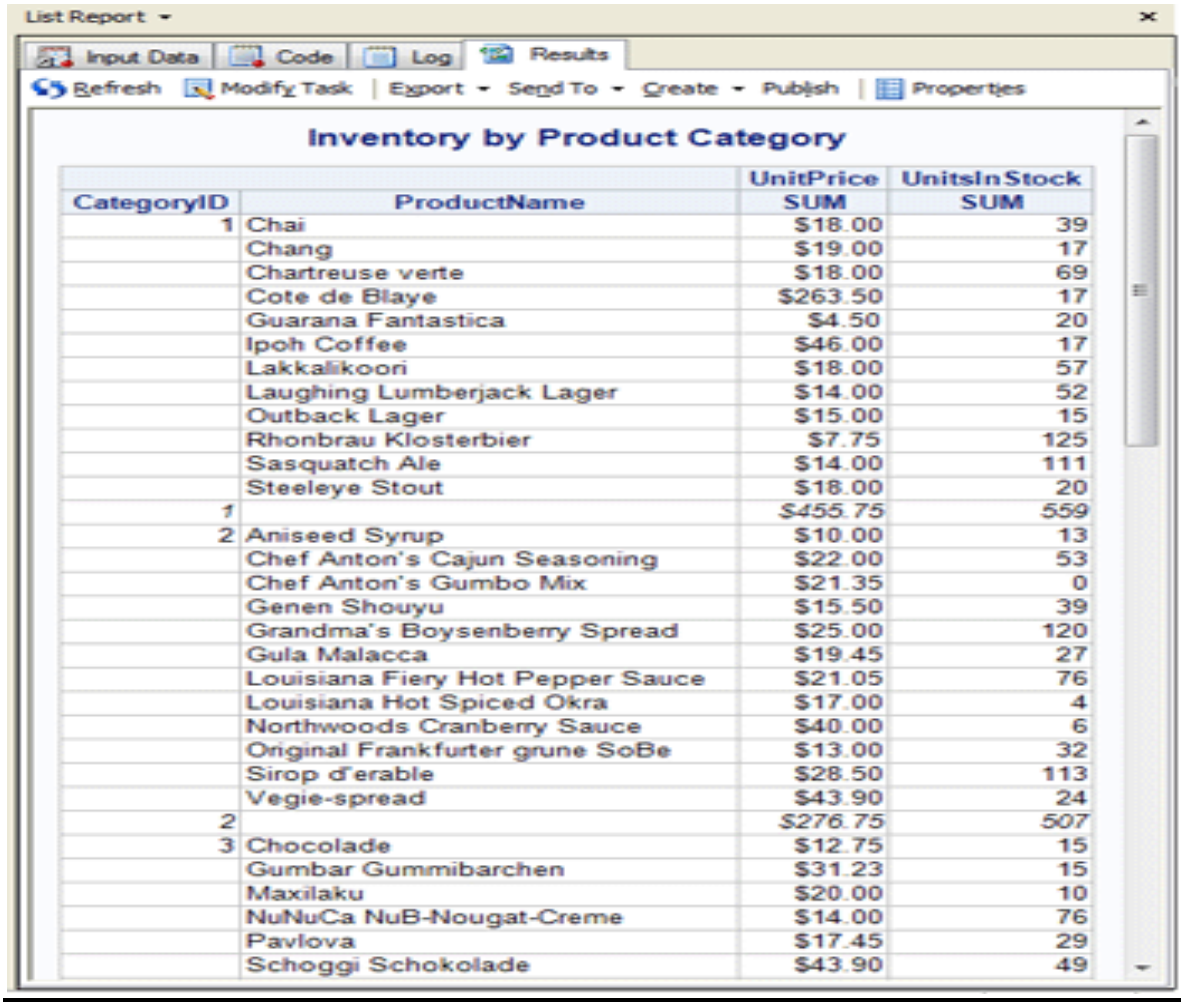
	Class and section							
	BNS		BS		TNS		TS	
RSP ($\mu g/m^3$)	60	(7)	250	(18)	160	(15)	220	(21)
TRP Average reading (mV)	0.6	(0.1)	1.9	(0.1)	1.3	(0.1)	1.7	(0.2)
TRP Maximum reading (mV)	4	(0.8)	29	(4.4)	7	(0.8)	18	(3.2)
Nicotine ($\mu g/m^3$)	5	(0.9)	41	(3.6)	21	(2.5)	32	(3.5)
Concentration of CO (ppm)	0.6	(0.06)	1.1	(0.08)	0.8	(0.07)	1.1	(0.11)
Concentration of CO_2 (ppm)	1310	(41)	1310	(42)	1270	(36)	1430	(36)
Mean Relative Humidity (%)	25	(0.7)	25	(0.8)	25	(0.6)	25	(1.0)

Cell content: - Weighted mean value, all flights
- Standard error for mean value in parenthesis

- **Matrix reports**: Business reporting is about summarizing information for analysis. A matrix, cross-tab, or pivot report aggregates data along the x-axis and y-axis of a grid to form a summarized table. Matrix report columns are not static but are based on group values.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Locations		AS OF Jan ,2012					AS OF Jan, 2013							
	Sub Loction	Net Revenue	%	Volume	%	Average	let Revenue	%	% Change	Volume	%	% Change	Average	% Change
1256														
afd	aa	15,910	12.84%	5	23.81%	3,182	0	0.00%	-100.00%	0	0.00%	-100.0%	nm	nm
	ab	0	0.00%	0	0.00%	nm	10,079	9.74%	nm	5	22.73%	nm	2,016	nm
	ac	108,001	87.16%	16	76.19%	6,750	93,421	90.26%	-13.50%	17	77.27%	6.3%	5,495	-18.6%
	ad	0	0.00%	0	0.00%	nm	0	0.00%	nm	0	0.00%	nm	nm	nm
	Total	123,911		21		5,901	103,500		-16.5%	22		4.8%	4,705	-20.3%
dfd	aa	9,626	10.20%	3	15.79%	3,209	6,630	6.11%	-31.12%	2	9.52%	-33.3%	3,315	3.3%
	ab	4,627	4.90%	3	15.79%	1,542	13,434	12.38%	190.36%	6	28.57%	100.0%	2,239	45.2%
	ac	79,482	84.21%	12	63.16%	6,623	88,437	81.51%	11.27%	13	61.90%	8.3%	6,803	2.7%
	ad	650	0.69%	1	5.26%	650	0	0.00%	-100.00%	0	0.00%	-100.0%	nm	nm
	Total	94,385		19		4,968	108,502		15.0%	21		10.5%	5,167	4.0%
Total		218,296		40		5,457	212,002		-2.88%	43		7.5%	4,930	-9.7%

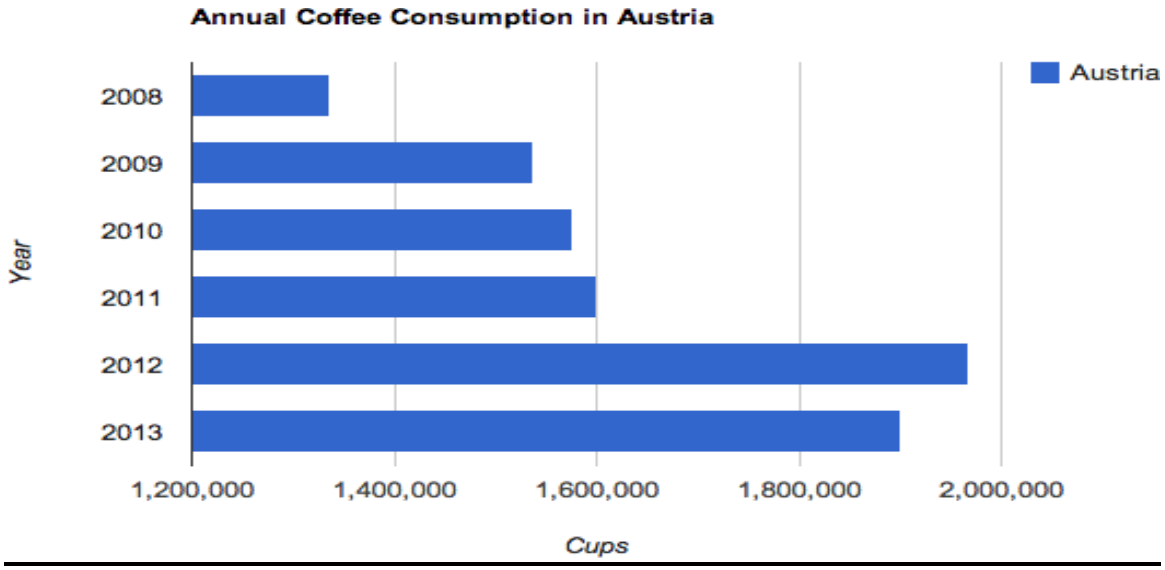
- **List reports**: A list report has a single, rectangular detail area that repeats for every record or group value in the underlying data set. Its main purpose is to contain other data regions and report items and to repeat them for a group of values.



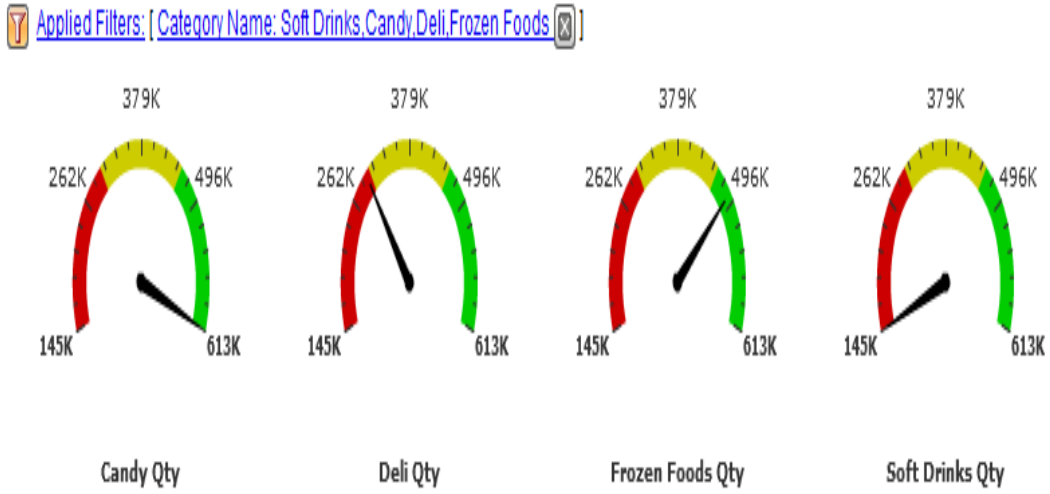
Inventory by Product Category

CategoryID	ProductName	UnitPrice SUM	UnitsInStock SUM
1	Chai	$18.00	39
	Chang	$19.00	17
	Chartreuse verte	$18.00	69
	Cote de Blaye	$263.50	17
	Guarana Fantastica	$4.50	20
	Ipoh Coffee	$46.00	17
	Lakkalikoori	$18.00	57
	Laughing Lumberjack Lager	$14.00	52
	Outback Lager	$15.00	15
	Rhonbrau Klosterbier	$7.75	125
	Sasquatch Ale	$14.00	111
	Steeleye Stout	$18.00	20
1		*$455.75*	*559*
2	Aniseed Syrup	$10.00	13
	Chef Anton's Cajun Seasoning	$22.00	53
	Chef Anton's Gumbo Mix	$21.35	0
	Genen Shouyu	$15.50	39
	Grandma's Boysenberry Spread	$25.00	120
	Gula Malacca	$19.45	27
	Louisiana Fiery Hot Pepper Sauce	$21.05	76
	Louisiana Hot Spiced Okra	$17.00	4
	Northwoods Cranberry Sauce	$40.00	6
	Original Frankfurter grune SoBe	$13.00	32
	Sirop d'erable	$28.50	113
	Vegie-spread	$43.90	24
2		*$276.75*	*507*
3	Chocolade	$12.75	15
	Gumbar Gummibarchen	$31.23	15
	Maxilaku	$20.00	10
	NuNuCa NuB-Nougat-Creme	$14.00	76
	Pavlova	$17.45	29
	Schoggi Schokolade	$43.90	49

- **Chart reports:** Chart reports provide a visual context for a lot of different kinds of data. There are several chart forms that can be used in the chart report such as bar chart, line chart, column chart, scatter plot, pie chart etc.



- **Gauge reports**: These are the reports with gauge controls. If gauge controls are appropriately designed one look at the gauge and you can say whether the enterprise is doing well, require attention, or is in a bad state.



Visual Analytics

Definition

Visual Analytics is the combination of visualisation and predictive analytic. It is aiming at answering, 'why is it happening?', 'what is more likely to happen?' and is usually associated with Business Analytics (*Forecasting, segmentation and correlation analysis*).

Visual analytics are a group of measuring systems and processes that combine analytical reasoning with information visualization.

Concept

Visual analytics works towards representing data in an easily understandable format but combines automated analysis techniques with interactive visualizations. This helps in the easier understanding of complex data and facilitates reasoning and decision-making based on large and complex data sets.

Due to the increasing demand for visual analytics with fast growing data volumes, there a movement toward investing in highly efficient visualisation systems. The companies like SAS investing in such concepts of high-powered visual

analytics, the new product of this company called, SAS Visual Analytics, is a very high performance computing in memory, solution for exploring massive amounts of data in a very short time. It empowers users to spot patterns, identify opportunities for further analysis and convey visual results via web reports or a mobile platform such as tablets and smart phones.

Different Types of Charts and Graphs

Basic Charts and Graphs

What follows are the basic charts and graphs that are commonly used for information visualisation.

1) Line Chart Line charts are perhaps the most frequently used graphical visuals for time-series data. Line charts (or a line graphs) show the relationship between two variables; they are most often used to track changes or trends over time (having one of the variables set to time on the x-axis). Line charts sequentially connect individual data points to help infer changing trends over a period of time. Line charts are often used to show time-dependent changes in the values of some measure, such as changes on a specific stock price over a 5-year period or changes on the number of daily customer service calls over a month.

2) Bar Chart Bar charts are among the most basic visuals used for data representation. Bar charts are effective when you have nominal data or numerical data that splits nicely into different categories so you can quickly see comparative results and trends within your data. Bar charts are often used to compare data across multiple categories such as percent of advertising spending by departments or by product categories. Bar charts can be vertically or horizontally oriented. They can also be stacked on top of each other to show multiple dimensions in a single chart.

3) Pie Chart Pie charts are visually appealing, as the name implies pie-looking charts. Because they are so visually attractive, they are often incorrectly used. Pie charts should only be used to illustrate relative proportions of a specific measure. For instance, they can be used to show the relative percentage of advertising budget spent on different product lines or they can show relative proportions of majors declared by college students in their sophomore year. If the number of categories to show is more than just a few (say more than 4), one should seriously consider using a bar chart instead of a pie chart.

4) Scatter Plot Scatter plots are often used to explore the relationship between two or three variables (in 2D or 2D visuals). Since they are visual exploration tools, having more than three variables, translating them into more than three dimensions is not easily achievable. Scatter plots are an effective way to explore the existence of trends, concentrations, and outliers. For instance, in a two-variable (two-axis) graph, a scatter plot can be used to illustrate the co-relationship between age and weight of heart disease patients or it can illustrate the relationship between number of customer care representatives and number of open customer service claims. Often, a trend line is superimposed on a two-dimensional scatter plot to illustrate the nature of the relationship.

5) Bubble Chart Bubble charts are often enhanced versions of scatter plots. Bubble charts, though, are not a new visualisation type; instead, they should be viewed as a technique to enrich data illustrated in scatter plots (or even geographic maps). By varying the size and/or colour of the circles, one can add additional data dimensions, offering more enriched meaning about the data. For instance, a bubble chart can be used to show a competitive view of college-level class attendance by major and by time of the day or it can be used to show profit margin by product type and by geographic region.

Specialized Charts and Graphs

The graphs and charts that we review in this section are either derived from the basic charts as special cases or they are relatively new and are specific to a problem type and/or an application area.

1) Histogram Graphically speaking, a histogram looks just like a bar chart. The difference between histograms and generic bar charts is the information that is portrayed. Histograms are used to show the frequency distribution of a variable or several variables. In a histogram, the x-axis is often used to show the categories or ranges, and the y-axis is used to show the measures/values/frequencies. Histograms show the distributional shape of the data. That way, one can visually examine if the data is normally or exponentially distributed. For instance, one can use a histogram to illustrate the exam performance of a class, where distribution of the grades as well as comparative analysis of individual results can be shown, or one can use a histogram to show age distribution of the customer base.

2) Gantt Chart Gantt charts are a special case of horizontal bar charts that are used to portray project timelines, project tasks/activity durations, and overlap amongst the tasks/activities. By showing start and end dates/times of tasks/activities and

the overlapping relationships, Gantt charts make an invaluable aid for management and control of projects. For instance, Gantt charts are often used to show project timelines, talk overlaps, relative task completions (a partial bar illustrating the completion percentage inside a bar that shows the actual task duration), resources assigned to each task, milestones, and deliverables.

3) PERT Chart PERT charts (also called network diagrams) are developed primarily to simplify the planning and scheduling of large and complex projects. They show precedence relationships among the project activities/tasks. A PERT chart is composed of nodes (represented as circles or rectangles) and edges (represented with directed arrows). Based on the selected PERT chart convention, either nodes or the edges may be used to represent the project activities/tasks (activity-on-node versus activity-on-arrow representation schema).Here, PERT states for *Programme Evaluation and Review Technique.*

4) Geographic Map When the data set includes any kind of location data (e.g., physical addresses, postal codes, state names or abbreviations, country names, latitude/longitude, or some type of custom geographic encoding), it is better and more informative to see the data on a map. Maps usually are used in conjunction with other charts and graphs, as opposed to by themselves. For instance, one can use maps to show distribution of customer service requests by product type (depicted in pie charts) by geographic locations. Often a large variety of information (e.g., age distribution, income distribution, education, economic growth, or population changes) can be portrayed in a geographic map to help decide where to open a new restaurant or a new service station. These types of systems are often called geographic information systems (GIS).

5) Bullet: Bullet graphs are often used to show progress toward a goal. A bullet graph is essentially a variation of a bar chart. Often they are used in place of gauges, meters, and thermometers in a dashboard to more intuitively convey the meaning with in a much smaller space. Bullet graphs compare a primary measure (e.g., year-to-date revenue) to one or more other measures (e.g., annual revenue target) and present this in the context of defined performance metrics (e.g., sales quota).A bullet graph can intuitively illustrate how the primary measure is performing against overall goals (e.g., how close a sales representative is to achieving his/her annual quota).

6) Heat Map Heat maps are great visuals to illustrate the comparison of continuous values across two categories using colour. The goal is to help the user quickly see where the intersection of the categories is strongest and weakest in

terms of numerical values of the measure being analysed. For instance, one can use heat maps to show segmentation analysis of target markets where the measure (colour gradient would be the purchase amount) and the dimensions would be age and income distribution.

7) Highlight Table Highlight tables are intended to take heat maps one step further. In addition to showing how data intersects by using colour, highlight tables add a number on top to provide additional detail. That is, they are two-dimensional tables with cells populated with numerical values and gradients of colours. For instance, one can show sales representative performance by product type and by sales volume.

8) Tree Map Tree maps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. A leaf node's rectangle has an area proportional to a specified dimension on the data. Often the leaf nodes are colored to show a separate dimension of the data. When the color and size dimensions are correlated in some way with the tree structure, one can often easily see patterns that would be difficult to spot in other ways, such as if a certain color is particularly relevant. A second advantage of tree maps is that, by construction, they make efficient use of space. As a result, they can legibly display thousands of items on the screen simultaneously.
Even though these charts and graphs cover a major part of what is commonly used in information visualisation, they by no means cover it all. Nowadays, one can find many other specialized graphs and charts that serve a specific purpose. Furthermore, the current trend is to combine/hybridize and animate these charts for better-looking and more intuitive visualisation of today's complex and volatile data sources. For instance, the interactive, animated, bubble charts available at the Gap minder Web site (gapminder.org) provide an intriguing way of exploring world health, wealth, and population data from a multidimensional perspective.

Emergence of Data Visualisation and Visual Analytics

Data visualization, quite simply, is the art of placing data in a visual context. The aim of data visualization is to identify trends, patterns, and contexts that usually go unrecognized in text-based data. Data visualization tools help in representing data beyond the typical spreadsheets, charts, and graphs and display it in more sophisticated formats using info graphics, maps, detailed bars, pie and communicate relationships between data values.

The images used in data visualizations can also have interactive capabilities which allow the users to manipulate data for query and analysis.

Right now, most of the activity in the business intelligence and analytics platform market is from organizations that are trying to mature their visualization capabilities and to move from descriptive to diagnostic (i.e., predictive and prescriptive)analytics.

Most of the IT solution providers are either relatively recently founded information visualization companies (e.g., Tableau Software, QlikTech, etc.) or well-established large analytics companies (e.g., SAS, IBM, Microsoft, SAP, etc.) that are increasingly focusing their efforts in information visualization and visual analytics.

Cheap hardware sensors and do-it-yourself frameworks for building your own system and the Internet have served as a fantastic distribution channel for visualizations. The future of data/information visualization is very hard to predict. There is a pretty good chance that we will see something that we have never seen in the information visualization realm invented before the end of this decade.

MODULE – 3
CH. 1 – DATA MINING

Concept

Simply defined, *data mining is a term used to describe discovering or "mining" knowledge from large amounts of data.*

Technically speaking, *data mining is a process that uses statistical, mathematical, and artificial intelligence techniques to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data.* These patternscan be in the form of business rules, affinities, correlations, trends, or prediction models.

The term data mining was originally used to *describe the process through which previously unknown patterns in data were discovered.* This definition has since been stretchedbeyond those limits by some software vendors to include most forms of data analysis in order to increase sales with thepopularity of the data mining label.

In other words, data mining is the computational process of sifting through existing business data to identify new patterns and establish relationships that will help in strategic decision making. This process of data mining will require pre-processing data by ETL or getting data form data warehouse as well as post-processing for displaying the new insights found.

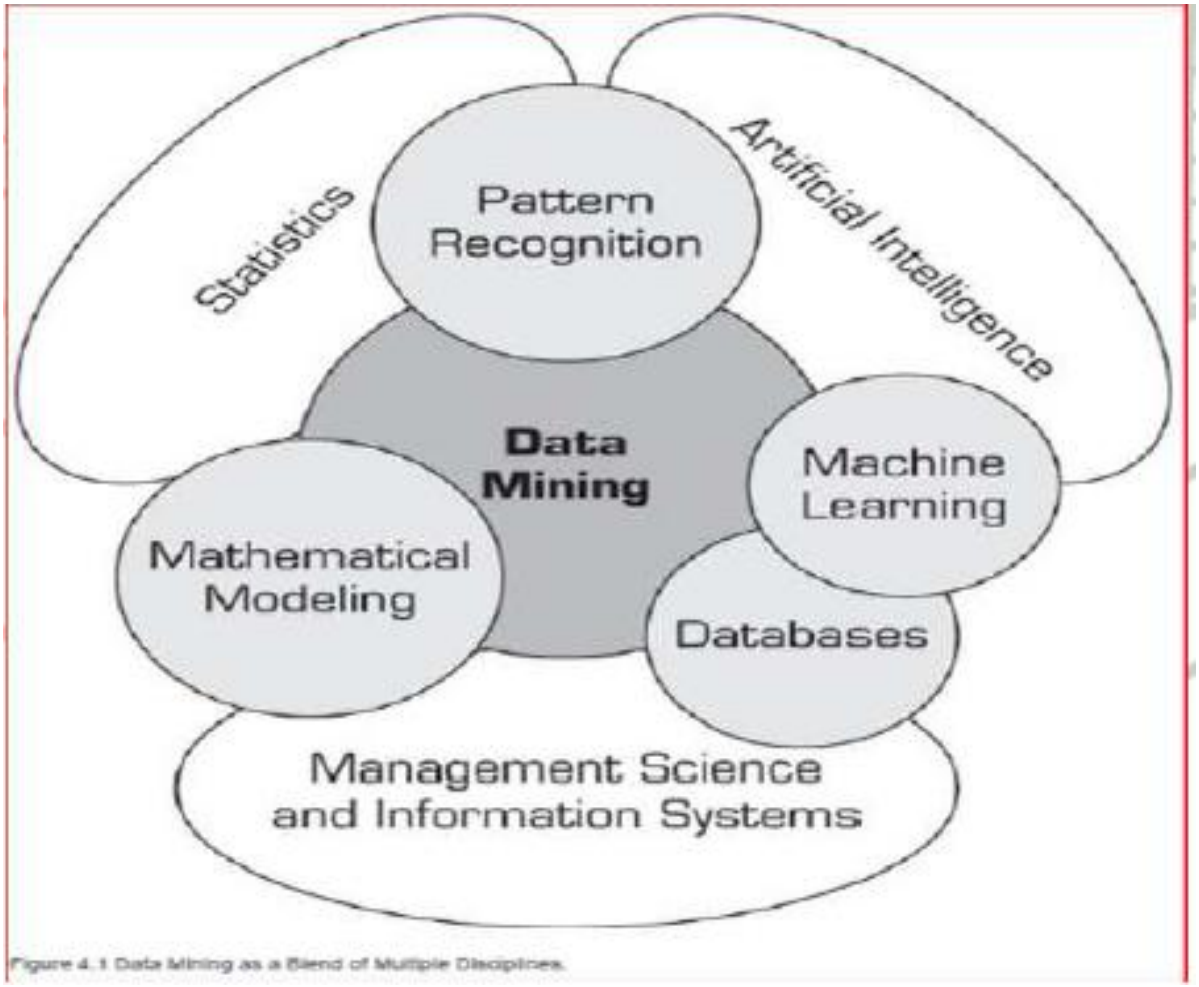


Figure 4.1 Data Mining as a Blend of Multiple Disciplines

Many other names that are associated with data mining include: *knowledge extraction, pattern analysis, data archaeology, information harvesting, pattern searching, and data dredging*

Data mining uses following techniques to discover new patterns in data;

Association: looking for patterns where one event is connected to another event.
Sequence Analysis: looking for patterns where one an event leads to another later event.
Classification: looking for new patterns and groups that could be of business value.
Clustering: **F**inding and visually documenting groups of facts not previously known.
Forecasting: **D**iscovering patterns in data that can lead to reasonable predictions about the future business events, conditions etc.

Applications

Data generated by the Internet is increasing rapidly in both volume and complexity.

Large amounts of genomic data are being generated and accumulated all over the world.
Medical and pharmaceutical researchers constantly generate and store data that can then be used in data mining applications to identify better ways to accurately diagnose and treat illnesses and to discover new and improved drugs.

On the commercial side, perhaps the most common use of data mining has been in the finance, retail, and healthcare sectors. Data mining is used to detect and reduce fraudulent activities, especially in insurance claims and credit card use, to identify customer buying patterns, to reclaim profitable customers, to identify trading rules from historical data; and to aid in increased profitability using market-basket analysis. Data mining is already widely used to better target clients, and with the widespread development of e-commerce, this can only become more imperative with time.

1) Customer relationship management Customer relationship management
(CRM) is the extension of traditional marketing. The goal of CRM is to create one-on-one relationships with customers by developing an intimate understanding of their needs and wants. As businesses build relationships with their customers over time through a variety of interactions (e.g., product inquiries, sales, service requests, warranty calls, product reviews, social media connections), they accumulate tremendous amounts of data.
When combined with demographic and socioeconomic attributes, this information-rich data can be used to;
(1) identify most likely responders/buyers of new products/services (i.e., customer profiling); (2) understand the root causes of customer attrition in order to improve customer retention (i.e., churn analysis); (3) discover time-variant associations between products and services to maximize sales and customer value; and (4) identify the most profitable customers and their preferential needs to strengthen relationships and to maximize sales.

2) Banking Data mining can help banks with the following:
(1) Automating the loan application process by accurately predicting the most probable defaulters; (2) detecting fraudulent credit card and online-banking transactions; (3) identifying ways to maximize customer value by selling them products and services that they are most likely to buy; and (4) optimizing the cash return by accurately forecasting the cash flow on banking entities (e.g., ATM machines, banking branches).

3) Retailing and logistics In the retailing industry, data mining can be used to;
(1) Predict accurate sales volumes at specific retail locations in order to determine correct inventory levels; (2) identify sales relationships between different products (with
market-basket analysis) to improve the store layout and optimize sales promotions;
(3) forecast consumption levels of different product types (based on seasonal and environmental conditions) to optimize logistics and, hence, maximize sales; and (4) discover interesting patterns in the movement of products (especially for the products that have a limited shelf life because they are prone to expiration, perish ability, and contamination) in a supply chain by analysing sensory and RFID data.
4) Manufacturing and production Manufacturers can use data mining to;
(1) predict machinery failures before they occur through the use of sensory data (enabling what is called condition-based maintenance); (2) identify anomalies and commonalities in production systems to optimize manufacturing capacity; and (3) discover novel patterns to identify and improve product quality.

5) Brokerage and securities trading Brokers and traders use data mining to;
(1) Predict when and how much certain bond prices will change; (2) forecast the range and direction of stock fluctuations; (3) assess the effect of particular issues and events on overall market movements; and (4) identify and prevent fraudulent activities in securities trading.

6) Insurance The insurance industry uses data mining techniques to;
(1) Forecast claim amounts for property and medical coverage costs for better
business planning; (2) determine optimal rate plans based on the analysis of claims and customer data; (3) predict which customers are more likely to buy new policies with special features; and (4) identify and prevent incorrect claim payments and fraudulent activities.
7) Computer hardware and software Data mining can be used to;
(1) predict disk drive failures well before they actually occur; (2) identify and filter unwanted Web content and e-mail messages; (3) detect and prevent computer network security bridges; and (4) identify potentially unsecure software products.
8) Travel industry (airlines, hotels/resorts, rental car companies) Data mining has a variety of uses in the travel industry. It is successfully used to;
(1) Predict sales of different services (seat types in airplanes, room types in hotels/resorts, car types in rental car companies) in order to optimally price services to
maximize revenues as a function of time-varying transactions (commonly referred to as yield management); (2) forecast demand at different locations to better

allocate limited organizational resources; (3) identify the most profitable customers and provide them with personalized services to maintain their repeat business; and (4) retain valuable employees by identifying and acting on the root causes for attrition.

9) Healthcare Data mining has a number of healthcare applications. It can be used to :

(1) Identify people without health insurance and the factors underlying this undesired phenomenon; (2) identify novel cost–benefit relationships between different treatments to develop more effective strategies; (3) forecast the level and the time of demand at different service locations to optimally allocate organizational resources; and (4) understand the underlying reasons for customer and employee attrition.

10) Entertainment industry: Data mining is successfully used by the entertainment industry to:

(1) Analyse viewer data to decide what programs to show during prime time and how to maximize returns by knowing where to insert advertisements; (2) predict the financial success of movies before they are produced to make investment decisions and to optimize the returns; (3) forecast the demand at different locations and different times to better schedule entertainment events and to optimally allocate resources; and (4) develop optimal pricing policies to maximize revenues.

Data mining process

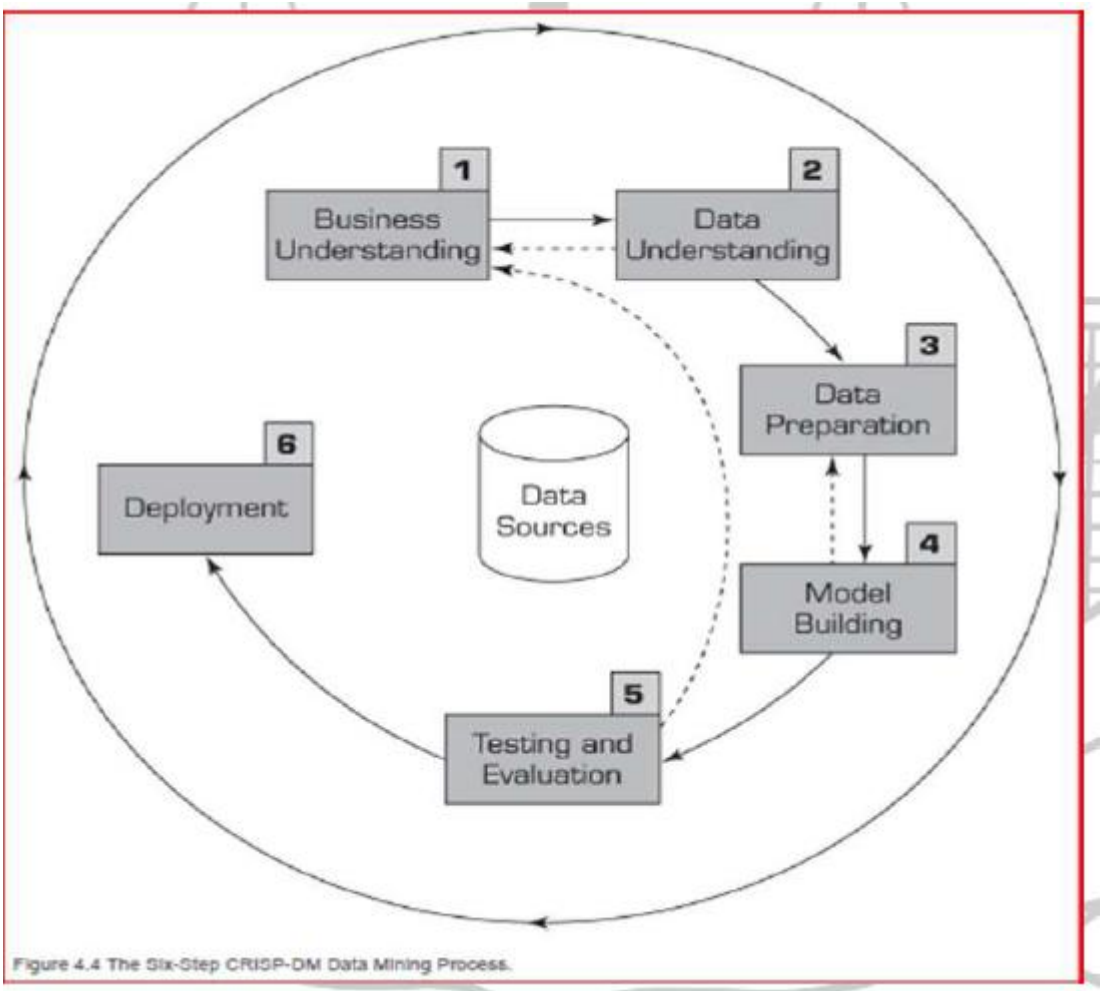


Figure 4.4 The Six-Step CRISP-DM Data Mining Process.

In order to systematically carry out data mining projects, a general process is usually followed. One such standardized process, arguably the most popular one, *Cross-Industry Standard Process for Data Mining- CRISP-DM* —was proposed in the mid - 1990s by a European group of companies to serve as a non - proprietary standard methodology for data mining. Figure 4.4 illustrates this proposed process, which is a sequence of six steps that starts with a good understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfied the specific business need.

Step-1: Business Understanding In the business understanding phase:

- First, it is required to understand business objectives clearly and find out what are the business's needs.
- Next, we have to assess the current situation by finding the resources, assumptions, constraints and other important factors which should be considered.

- Then, from the business objectives and current situations, we need to create data mining goals to achieve the business objectives within the current situation.
- Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible.

Step-2: Data Understanding

- First, the data understanding phase starts with initial data collection, which we collect from available data sources, to help us get familiar with the data. Some important activities must be performed including data load and data integration in order to make the data collection successfully.
- Next, the "gross" or "surface" properties of acquired data need to be examined carefully and reported.
- Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting, and visualization.
- Finally, the data quality must be examined by answering some important questions such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"

Step-3: Data Preparation

The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form.

The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.

Step-4: Model Building

- First, modelling techniques have to be selected to be used for the prepared dataset.
- Next, the test scenario must be generated to validate the quality and validity of the model.
- Then, one or more models are created by running the modelling tool on the prepared dataset.
- Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives.

Step-5: Testing and Evaluation

In step 5, the developed models are assessed and evaluated for their accuracy and generality. This step assesses the degree to which the selected model (or models)

meets the business objectives and, if so, to what extent (i.e., do more models need to be developed and assessed). Another option is to test the developed model(s) in a real-world scenario if time and budget constraints permit.

In this phase, new business requirements may be raised due to the new patterns that have been discovered in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

Step-6: Testing

Development and assessment of the models is not the end of the data mining project.

The knowledge or information, which we gain through data mining process, needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization.

In the deployment phase, the plans for deployment, maintenance, and monitoring have to be created for implementation and also future supports. From the project point of view, the final report of the project needs to summary the project experiences and reviews the project to see what need to improved created learned lessons.

MODULE – 3

CH. 2 - TEXT AND WEB ANALYTICS

Text Analytics

Text analytics is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms) as well as information extraction, data mining, and Web mining, whereas text mining is primarily focused on discovering new and useful knowledge from the textual data sources. Based on this definition of text analytics and text mining, one could simply formulate the difference between the two as follows:

Text Analytics = Information Retrieval + Information Extraction + Data Mining + Web Mining

Or

Text Analytics = Information Retrieval + Text Mining

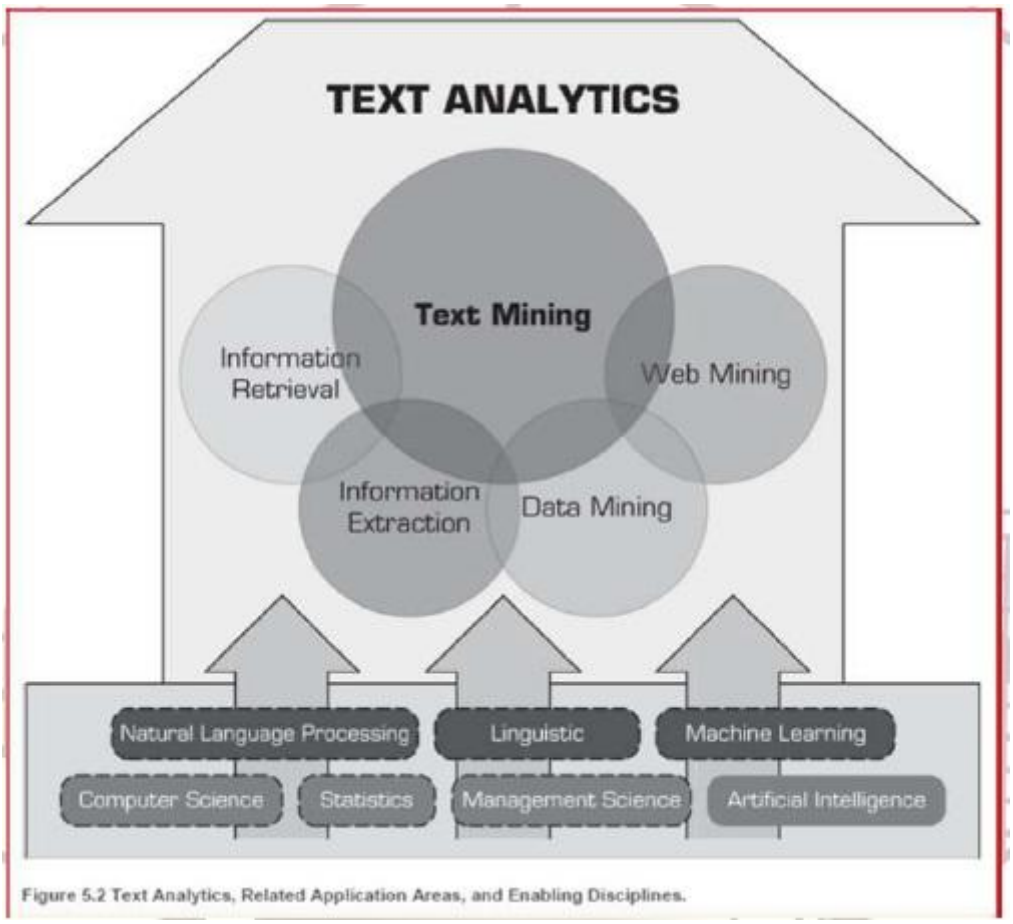


Figure 5.2 Text Analytics, Related Application Areas, and Enabling Disciplines.

Compared to text mining, text analytics is a relatively new term. With the recent emphasis on analytics, as has been the case in many other related technical application areas (e.g., consumer analytics, completive analytics, visual analytics, social analytics, etc.), the field of text has also wanted to get on the analytics bandwagon.
While the term text analytics is more commonly used in a business application context, text mining is frequently used in academic research circles. Even though they may be defined somewhat differently at times, text analytics and text mining are usually used synonymously.

Text mining

Overview
Text mining (*also known as text data mining or knowledge discovery in textual databases*) is the semi-automated process of extracting patterns (*useful information and knowledge*) from large amounts of unstructured data sources. Text mining is the same as data mining in that it has the same purpose and uses the same processes, but with text mining the input to the process is a collection of unstructured (or less structured) data files such as Word Documents, PDF files, text excerpts, XML files, and so on.
In essence, text mining can be thought of as a process (with two main steps) that starts with imposing structure on the text-based data sources followed by extracting relevant information and knowledge from this structured text-based data using data mining techniques and tools.

Applications

Following are among the most popular application areas of text mining;

Information extraction Identification of key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching.
Topic tracking Based on a user profile and documents that a user views, text mining can predict other documents of interest to the user.
Summarization Summarizing a document to save time on the part of the reader.
Categorization Identifying the main themes of a document and then placing the document into a predefined set of categories based on those themes.
Clustering: Grouping similar documents without having a predefined set of categories.

Concept linking Connects related documents by identifying their shared concepts and, by doing so, helps users find information that they perhaps would not have found using traditional search methods.
Question answering Finding the best answer to a given question through knowledge-driven pattern matching.

Web Mining and Web Analytics

Web mining (or Web data mining) is the process of discovering intrinsic relationships (i.e., interesting and useful information) from Web data, which are expressed in the form of textual, linkage, or usage information. The term Web mining was first used by Etzioni (1996); today, many conferences, journals, and books focus on Web data mining. It is a continually evolving area of technology and business practice. Web mining is essentially the same as data mining that uses data generated over the Web.

The Web is perhaps the world's largest data and text repository, and the amount of information on the Web is growing rapidly. A lot of interesting information can be found online: whose homepage is linked to which other pages, how many people have links to a specific Web page, and how a particular site is organized.

The goal of Web Mining is to turn vast repositories of business transactions, customer interactions, and Web site usage data into actionable information (i.e., knowledge) to promote better decision making throughout the enterprise. Because of the increased popularity of the term analytics, nowadays many have started to refer to Web mining as Web analytics. However, these two terms are not the same.

While Web analytics is primarily Web site usage data focused, Web mining is inclusive of all data generated via the Internet including transaction, social, and usage data. While Web analytics aims to describe what has happened on the Web site (employing a predefined, metrics-driven descriptive analytics methodology), Web mining aims to discover previously unknown patterns and relationships (employing a novel predictive or prescriptive analytics methodology). From a big-picture perspective, Web analytics can be considered to be a part of Web mining.

Social Media Analytics

Social media refers to the enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks. It is a group of Internet-based software applications.
Social media depends on mobile and other Web-based technologies to create highly interactive platforms for individuals and communities to share, co-create, discuss, and modify user-generated content.
By applying a set of theories in the field of media research (social presence, media richness) and social processes (self-presentation, self-disclosure), Kaplan and Haenlein (2010) created a classification scheme with six different types of social media;

- Collaborative projects (e.g., Wikipedia),
- Blogs and micro blogs (e.g., Twitter),
- Content communities (e.g., YouTube),
- Social networking sites (e.g., Facebook),
- Virtual game worlds (e.g., World of Warcraft), and
- Virtual social worlds (e.g. Second Life).

Thus, Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness. Social media analytics is rapidly becoming a new force in organizations around the world, allowing them to reach out to and understand consumers as never before. In many companies, it is becoming the tool for integrated marketing and communications strategies.
The exponential growths of social media outlets, from blogs, Facebook, and Twitter to LinkedIn and YouTube, and analytics tools that tap into these rich data sources offer organizations the chance to join a conversation with millions of customers around the globe every day.

Sentiment Analysis

As a human being, we rely on others' opinions to make better decisions, especially in an area where we don't have a lot of knowledge or experience. Thanks to the growing availability and popularity of opinion-rich Internet resources such as social media outlets (e.g., Twitter, Facebook, etc.), online review sites, and personal

blogs, it is now easier than ever to find opinions of others (thousands of them, as a matter of fact) on everything from the latest gadgets to political and public figures. Sentiment is a difficult word to define. It is often linked to or confused with other terms like belief, view, opinion, and conviction. Sentiment suggests a settled opinion reflective of one's feelings. Sentiment has some unique properties that set it apart from other concepts that we may want to identify in text.

As a field of research, sentiment analysis is closely related to computational linguistics, natural language processing, and text mining. Sentiment analysis has many names.
It's often referred to as *opinion mining, subjectivity analysis,* and *appraisal extraction*. Every opinion put on the Internet by an individual or a company will be accredited to the originator
(Good or bad) and will be retrieved and mined by others (often automatically by computer programs).

Sentiment analysis is trying to answer the question "what do people feel about a certain topic?" by digging into opinions of many using a variety of automated tools. In a business, especially in marketing and customer relationship management, sentiment analysis seeks to detect favourable and unfavourable opinions toward specific products and/or services using a large numbers of textual data sources (customer feedback in the form of Web postings, tweets, blogs, etc.).Timely collection and analysis of textual data, which may be coming from a variety of sources—ranging from customer call centre transcripts to social media postings—is a crucial part of the capabilities of proactive and customer focused companies, nowadays.

Sentiment analytics applied in fields like; *Brand management, politics, financial markets, government intelligence, e-commerce applications or websites, e-mails, etc.*

Ch. 3 – Big Data Analytics

Definition of Big Data

Big data is a high volume, high velocity, high variety information asset that demands cost-effective and innovative forms of information processing for enhanced business insight and decision making.
Hence, big data involves homogeneous voluminous data that could be structured (as in RDBMS) or unstructured (as in blogs, tweets, Facebook comments, emails) and the content is in different varieties (audio, picture, large text). Handling this type of data will need newer and innovative technologies for capturing, storing, searching, integrating, analysing and presenting newly found insights.
Example: A big data application in telecom industry could be to analyse millions of calls data, billing data, marketing data, competitive data, data usage and customer profiles to accurately recommend a service that will meet the needs of the customer. In this situation, the volume and split second response by the big data analytics application is critical to engage the customer on call and close deals.

Characteristics of Big Data

1) Volume: Volume is obviously the most common trait of Big Data. Many factors contributed to the exponential increase in data volume, such as transaction-based data stored through the years, text data constantly streaming in from social media, increasing amounts of sensor data being collected, automatically generated GPS data, and so on.
With the staggering increase in data volume, even the naming of the next Big Data echelon has been a challenge. The highest mass of data that used to be called peta bytes (PB) has left its place to zeta bytes (ZB), which is a terabytes (TB). (*1 Terabyte can hold 200,000 songs or 17,000 hours of music / 500 hours of movies*)

2) Variety: Data today comes in all types of formats formats—ranging from traditional databases to hierarchical data stores created by the end users and OLAP systems to text documents, e-mail, XML, meter-collected, and sensor-captured data, to video, audio, and stock ticker data. By some estimates, 80 to 85 percent of all organizations' data is in some sort of unstructured or semi - structured format (a format that is not suitable for traditional databases schemas). But there is no denying its value, and, hence, it must be included in the analyses to support decision making.

3) Velocity: In simple words, velocity means the speed of something in a given direction. According to Gartner, velocity means both how fast data is being produced and how fast the data must be processed (i.e., captured, stored, and analysed) to meet the need or demand. Velocity is perhaps the most overlooked characteristic of Big Data. Reacting quickly enough to deal with velocity is a challenge to most organizations.

4) Veracity: Veracity is a term coined by IBM that is being used as the fourth "V" to describe Big Data. It refers to conformity to facts: accuracy, quality, truthfulness, or trustworthiness of the data. Tools and techniques are often used to handle Big Data's veracity by transforming the data into quality and trustworthy insights.

5) Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal, and event-triggered peak data loads can be challenging to manage—especially with social media involved.

Fundamentals of Big Data Analytics

Critical Success Factors

Following are the most critical success factors for Big Data analytics:

1) A clear business need Business investments ought to be made for the good of the business, not for the sake of mere technology advancements. Therefore, the main driver for Big Data analytics should be the needs of the business, at any level— strategic, tactical, and operations.

2) Strong, committed sponsorship (executive champion) It is a well-known fact that if you don't have strong, committed executive sponsorship, it is difficult to succeed. If the scope is a single or a few analytical applications, the sponsorship can be at the departmental level. However, if the target is enterprise-wide organizational transformation, which is often the case for Big Data initiatives, sponsorship needs to be at the highest levels and organization wide.

3) Alignment between the business and IT strategy It is essential to make sure that the analytics work is always supporting the business strategy, and not other way around. Analytics should play the enabling role in successfully executing the business strategy.

4) A fact-based decision-making culture In a fact-based decision-making culture, the numbers rather than intuition, gut feeling, or supposition drive decision making.
There is also a culture of experimentation to see what works and what doesn't. To create a fact-based decision-making culture, senior management needs to:

- Recognize that some people can't or won't adjust
- Stress that outdated methods must be discontinued
- Ask to see what analytics went into decisions
- Link incentives and compensation to desired behaviours

5) A strong data infrastructure Data warehouses have provided the data infrastructure for analytics. This infrastructure is changing and being enhanced in the Big Data era with new technologies. Success requires marrying the old with the new for a holistic infrastructure that works synergistically.

Challenges

Following are the most common challenges for adopting Big Data analytics:

1) Data volume The ability to capture, store, and process the huge volume of data at an acceptable speed so that the latest information is available to decision makers when they need it.

2) Data integration The ability to combine data that is not similar in structure or source and to do so quickly and at reasonable cost.

3) Processing capabilities The ability to process the data quickly, as it is captured. The traditional way of collecting and then processing the data may not work. In many situations data need to be analysed as soon as it is captured to leverage the most value.

4) Data governance The ability to keep up with the security, privacy, ownership, and quality issues of Big Data. As the volume, variety (format and source), and velocity of data change, so should the capabilities of governance practices.

5) Skills availability Big Data is being harnessed with new tools and is being looked at in different ways. There is a shortage of people (often called data scientists) with the skills to do the job.

6) Solution cost Since Big Data has opened up a world of possible business improvements; a great deal of experimentation and discovery is taking place to determine the patterns that matter and the insights that turn to value. To ensure a positive ROI on a Big Data project, therefore, it is crucial to reduce the cost of the solutions used to find that value.

Problems addressed by Big Data Analytics

Here is a list of problems that can be addressed using Big Data analytics:

- Process efficiency and cost reduction
- Brand management
- Revenue maximization
- Enhanced customer experience
- Churn identification, customer recruiting
- Improved customer service
- Identifying new products and market opportunities
- Risk management
- Regulatory compliance
- Enhanced security capabilities, etc.

Benefits Big Data Analytics

Here is a list of advantages that can be achieved by using Big Data analytics:

- Understanding and Targeting Customers
- Understanding and Optimizing Business Processes
- Re-develop your products
- Personal Quantification and Performance Optimization
- Helps in Fraud Detection & improving Security
- Perform Risk Analysis
- Customize your website in real time
- Optimizing Machine and Device Performance

MODULE – 4
Ch. 1 – Business Performance Management

1. Concept

In the business and trade literature, business performance management (BPM) has a number of names, including corporate performance management (CPM), enterprise performance management (EPM), and strategic enterprise management (SEM).
The term business performance management (BPM) refers to the business processes, methodologies, metrics, and technologies used by enterprises to measure, monitor, and manage business performance. It encompasses three key components;
1. A set of integrated, closed-loop (an automatic control system in which an operation, process, or mechanism is regulated by feedback) management and analytic processes (supported by technology) that addresses financial as well as operational activities.
2. Tools for businesses to define strategic goals and then measure and manage performance against those goals.
3. A core set of processes, including financial and operational planning, consolidation (merging) and reporting, modelling, analysis, and monitoring of key performance indicators (KPIs), linked to organizational strategy

2. Business Performance Management Cycle

1) Strategize: Where do we want to go? Strategy, in general terms, is a high-level plan of action, encompassing a long period of time (often several years) to achieve a defined goal. It is especially necessary in a situation where there are numerous constraints (driven by market conditions, resource availabilities, and legal/political alterations) to deal with on the way to achieving the goal. In a business setting, strategy is the art and the science of crafting decisions that help businesses achieve their goals. More specifically, it is the process of identifying and stating the organization's mission, vision, and objectives, developing plans (at different levels of granularity—strategic, tactical, and operational) to achieve these objectives. Business strategies are normally planned and created by a team of corporate executives (often led by the CEO), approved and authorized by the board of directors, and then implemented by the company's management team under the supervision of the senior executives. Business strategy provides an overall

direction to the enterprise and is the first and foremost important process in the BPM methodology.

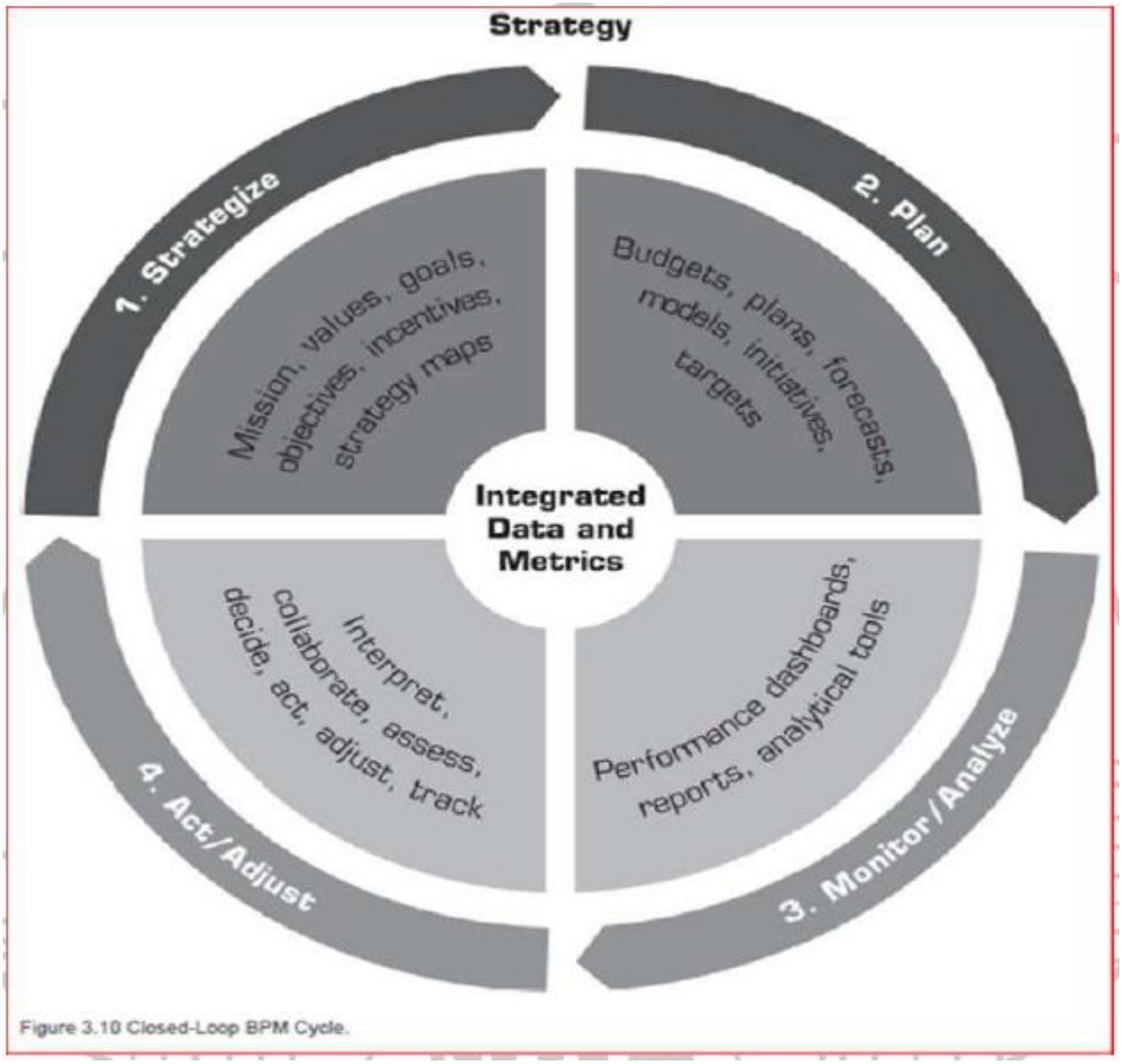


Figure 3.10 Closed-Loop BPM Cycle.

2) Plan: How do we get there? When operational managers know and understand the what (i.e., the organizational objectives and goals), they will be able to come up with the how (i.e., detailed operational and financial plans). Operational and financial plans answer two questions: What tactics and initiatives will be pursued to meet the performance targets established by the strategic plan? What are the expected financial results of executing the tactics?

An operational plan translates an organization's strategic objectives and goals into a set of well-defined tactics and initiatives, resource requirements, and expected results for some future time period. In essence, an operational plan is like a project

plan that is designed to ensure that an organization's strategy is realized. Most operational plans encompass a portfolio of tactics and initiatives. The key to successful operational planning is integration. Strategy drives tactics, and tactics drive results.

The financial planning and budgeting process has a logical structure that typically starts with those tactics that generate some form of revenue or income. In organizations that sell goods or services, the ability to generate revenue is based on either the ability to directly produce goods and services or acquire the right amount of goods and services to sell. After a revenue figure has been established, the associated costs of delivering that level of revenue can be generated.

In addition to the collaborative input, the organization also needs to add various overhead costs, as well as the costs of the capital required. This information, once consolidated, shows the cost by tactic as well as the cash and funding requirements to put the plan into operation.

3) Monitor/Analyze: How are we doing? When the operational and financial plans are under way, it is imperative that the performance of the organization be monitored. A comprehensive framework for monitoring performance should address two key issues: what to monitor and how to monitor. Because it is impossible to look at everything, an organization needs to focus on monitoring specific issues. After the organization has identified the indicators or measures to look at, it needs to develop a strategy for monitoring those factors and responding effectively. These measures are most often called key performance indicators (or KPI in short).

4) Act and Adjust: What do we need to do differently? Whether a company is interested in growing its business or simply improving its operations, virtually all strategies depend on new projects—creating new products, entering new markets, acquiring new customers or businesses, or streamlining some processes. Most companies approach these new projects with a spirit of optimism rather than objectivity, ignoring the fact that most new projects and ventures fail.

What is the chance of failure? Obviously, it depends on the type of project. Hollywood movies have around a 60percent chance of failure. The same is true for mergers and acquisitions. Large IT projects fail at the rate of 70 percent. For new food products, the failure rate is 80 percent. For new pharmaceutical products, it is even higher, around 90 percent. Overall, the rate of failure for most new projects or ventures runs between 60 and 80 percent. Given these numbers, the answer to the question of "what do we need to do differently?" becomes a vital issue.

_____

3. KPI – Key Performance Indicator

A KPI represents a strategic objective and measures performance against a goal. According to Eckerson (2009), KPIs are multidimensional. Loosely translated, this means that KPIs have a variety of distinguishing features, including:

- ***Strategy*:** KPIs embody a strategic objective.
- ***Targets*:** KPIs measure performance against specific targets. Targets are defined in strategy, planning, or budget sessions and can take different forms (e.g., achievement targets, reduction targets, absolute targets).
- ***Ranges***: Targets have performance ranges (e.g., above, on, or below target).
- ***Encodings***: Ranges are encoded in software, enabling the visual display of performance (e.g., green, yellow, red). Encodings can be based on percentages or more complex rules.
- ***Time frames***: Targets are assigned time frames by which they must be accomplished. A time frame is often divided into smaller intervals to provide performance mileposts.
- ***Benchmarks*:** Targets are measured against a baseline or benchmark. The previous year's results often serve as a benchmark, but arbitrary numbers or external benchmarks may also be used.
- **Relevance and functionality**: The KPIs chosen should be directly related to business results that the company is trying to produce in the specific business function. Like, your body temperature measurement can only indicate whether you have fever or not, but can say nothing about your blood pressure!
- **Understandable**: Chosen KPIs must be defined unambiguously. A KPI needs to be understood in one and only one way by all stakeholders. It must be documented, and its definition must be easily accessible to all users.
- **Reliability and creditability**: The value of KPIs needs to be authenticated and should be validated as "trusted" or "dependable". Someone is going to base an important decision on the chosen metric. Adequate checks are needed to declare data as trustworthy. This also means that the data must represent the "single version truth".
- **Abuse-proof**: An abuse-proof measure is unlikely to be used against intended purpose or individual(s) involved in the measurement process.

We may say that KPIs are objective, measurable attributes of business performance, which assist in informed decision making. They are means of assessing the business functions' health and a means of assisting in the prediction of business success and potential failure.

KPIs are quantitative or qualitative measures which reflect the business performance of a company in achieving its goals and strategies. KPIs reflect strategic value drivers rather than just measuring mom-critical business activities and strategies.

Let's look at a few sample KPIs used by the Human Capital and Training Management division of "GoodFood Restaurant Inc."

- Average time to recruit.
- Average open time of job positions.
- # of responses to open job positions.
- # of interviews to fill up open job positions.
- # of offers that made.
- # of responses to the offers made.
- % of vacancies that were filled with in x time.
- % of new employees that remained after x time.
- % of new employee satisfaction rate.

Few sample KPIs employees by the Employee Training Management Division of "GoodFood Restaurant" are as follows:

- % of employees who underwent training.
- Average training cost per employee.
- % of employees satisfied with training.
- Average training hours per employee.
- Ratio of internal vs. External training.
- % of budget spent on employees training.
- ROI of training.

Likewise, lets us look at a few sample KPIs likely in use by the Help Desk of "GoodFood Restaturant":

- Average no. of calls by customer in a day.
- % of complaints serviced in a day.

- % of customers satisfied by the services offered.
- % of complaints serviced well before the SLA (service-level agreement) time.

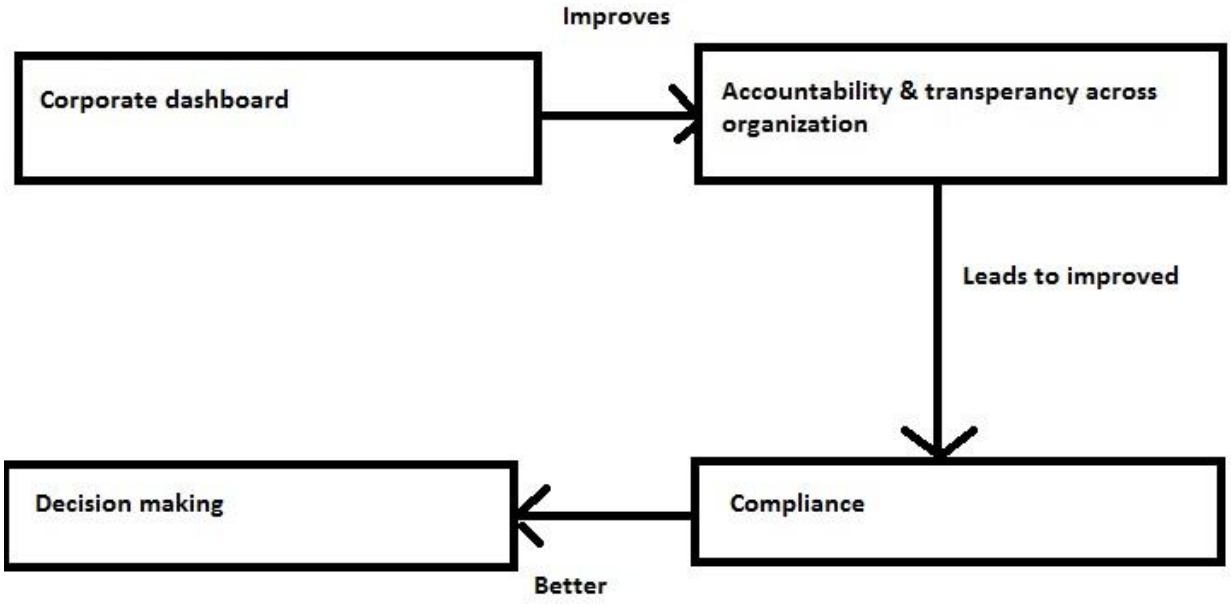
---

4. Dashboard

A dashboard is a graphical user interface that organizer and presents information in a way that is easy to read. It provides at-a-glance insight to what is actually happening in an organization. Dashboards have the following attributes:

- They display data relevant to their own objectives.
- They throw light on KPIs and metrics used to measure and monitor the organization's performance.
- Since dashboards are designed to serve a specific purpose, they inherently contain pre-defined conclusions that help the end-user analyze his or her own performance.

Following figure describes the benefits accruing to enterprises through dashboards.



Types of Dashboard

- ✓ ***Enterprise Performance Dashboards***
- ✓ ***Customer Support Dashboards***
- ✓ ***Divisional Dashboards***

1. Enterprise Performance Dashboards

These dashboards provide an overall view of the entire enterprise, rather than of specific business functions/process. Typical portlets in an enterprise performance dashboard include:

- Corporate financials.
- Sales revenue.
- Business Unit KPIs.
- Supply chain information.
- Compliance or regulatory data.
- Balanced scorecard information.

2. Customer Support Dashboards:

Organizations provide this type of dashboard to its customers as a value-add service. A customer support dashboard provides customers their personal account information pertaining to the business relationship, such as

- Online trading.
- Utility services.
- Entertainment.
- B2B SLA (service-level agreement) monitoring.

3. Divisional Dashboards:

These are one of the most popular dashboards used to provide at-a-glance actionable information to division heads, operational managers, and department managers. Each division has its own set of KPIs which can be visually displayed on the enterprise dashboard. Typical divisional dashboards include:

- Purchasing dashboards.
- Supply chain dashboards.
- Operations dashboards.
- Manufacturing dashboards.
- Quality control dashboards.
- Marketing dashboards.
- Sales dashboards.
- Finance dashboards.
- HR dashboards.

HOW DO YOU CREATE DASHBOARDS?

Most dashboards are created around a set of measures or KPIs. KPI is an indicator of the performance of a task, and it reveals the performance that is below the normal range so that corrective action can be taken. It draws attention to problem areas. The measures used in the dashboard should be relevant and support the initial purpose of the dashboard.

Steps for creating Dashboards

First Step:

Understand/identify the data that will go into an enterprise dashboard. Enterprise dashboard can contain either/both of the following mentioned data:

- Quantitative data.
- Non-quantitative data

Quantitative data is the data that gives an idea of what is currently going on. Example of quantitative data for an Educational dashboard:

- No. of students batches.
- No. of learning programs.
- No. of students who have successfully qualified the internal certification.
- No. of students being trained on the various learning programs.

Examples of non-qualitative data for an Education program.

- Salient features of the foundation learning program.
- Challenges faced by the instructor in classroom training.
- Users' comments on the effectiveness of the learning program.

Second Step

Decide on the timeframes. The various timeframes could be

- This month to date.
- This quarter to date.
- This year to date.
- Today so far.

Third Step

Decide on the comparative measures. The comparative measures could be

- The same measure at the same point in time in the past.
- The same measure at some other point in time in the past.
- A distinct yet relative measure.
- A competitor's measure.

Last Step

Decide on the evaluation mechanisms. The evaluation can be performed as follows:

- Using visual objects, e.g. traffic lights.
- Using visual attributes, e.g. color for the measure to alert a serious condition.

_____

Ch. 2 – Analytics in Business Support Functions

1. Sales and Marketing Analytics

Sales and marketing analytics are essential to unlocking commercially relevant insights, increasing revenue and profitability, and improving brand perception. With the help of the right analytics, one can uncover new markets, new audience niches, areas for future development and much more. The best and most important sales and marketing analytics that can help any business grow and succeed would be;

1) Unmet Need Analytics Business is all about meeting the needs of customers.
Unmet need analytics is the process of uncovering whether there are any unmet needs around your product or service or within your market which you could meet to increase customer satisfaction and revenue. Useful tools for unmet need analytics include; *product reviews, qualitative surveys, focus groups and interviews*. You could also use tools like *Google Trends* to help identify what customers are searching for.

2) Market Size Analytics If you don't understand the size and potential of your market you can easily jump to conclusions about how viable your business proposition is.
Market size analytics is the process of working out how large the market is for your products and services, and whether there is sufficient growth potential.
The size of the market is measured in terms of volume (how many units sold), value (money spent in that market) or frequency (how often a product or service is sold).

Useful data includes government data, trade association data, financial data from competitors, and customer surveys.

3) Demand Forecasting Understanding demand is essential in order to remain competitive. Demand forecasting is an area of predictive analytics that seeks to estimate the quantity of a product or service your consumers are likely to buy. It goes beyond educated guesses and looks at historical sales data or current data from test markets.
Analytic techniques such as time series analysis can be very useful here.

4) Market Trend Analytics Every business needs to know the direction its market is heading in. Market trend analytics is a process of establishing whether a market is growing, stagnant or in decline and how fast that movement is occurring. Understanding market size is important but knowing whether that market is trending up or down is also vital.

To monitor market trends you can run business experiments or scenario analysis to see what the market would look like and how it would impact your business in either a growing, stagnating or growth market. Customer surveys and focus groups can also help.

5) Non-customer Analytics Traditionally we've been told that we need to understand our customers so that we know what they look like and can find more people like them. And even as that makes sense, there is another group that could be even more important – the non-customer!
Non-customer analytics is about understanding what people who are currently not your customers think about your product, services or brand. By identifying who is not buying from you (and why), you can expand your market to include those individuals. If you want to know why people are not buying your product or service, you need to ask them: interviews, questionnaires and focus groups can help. It can be remarkably easy to get feedback from people who are not your customers using the power of social media.

6) Competitor Analytics Competitor analytics is important for marketing and strategic planning by identifying who your real competitors are, and how they are positioned in the market and in relation to your business. By understanding their strengths and weaknesses you can identify opportunities to exploit and threats to navigate.

There are many ways of gathering competitor data, such as business journals and newspapers, annual reports, product brochures and marketing activity. You could even have an employee, friend or family member buy a product or service from your key competitors and assess their experience.

7) Pricing Analytics What if you could find out exactly how many your customers would pay for your product ahead of time? Pricing analytics is the process that delivers that outcome. In short, it involves analysing price sensitivity in market segments and is especially useful in highly competitive markets where everything that can be done has been done.
Pricing analytics requires data mining and the development of forecasting models and algorithms. It also often involves multiple, concurrent business experiments that can be run quickly and easily so you can measure what is likely to happen with each price change.

8) Marketing And Sales Channel Analytics There are literally hundreds of possible channels and ways to market and sell your products and services. Marketing and sales channel analytics allows you to assess the different channels available to you and establish which are the most effective. It is likely you will reach different segments of your market via different channels but is it still good to know which ones are working and which are less effective. For each of your current marketing and sales channels and any potential as yet unused channels you will need to set some conversion rate goals so you know what you want that channel to deliver.

9) Brand Analytics Brands matter. Brand analytics seeks to determine the strength of your brand compared to your competitors. Your brand is more than just your logo and your commercial livery – it's the look and feel of your products and what they represent to your customers. It's important to really understand how customers perceive your brand as this will impact your decision making and strategic direction.
You can source this sort of data anywhere your customers and potential customers are discussing your brand, such as *customer service conversations, sales conversations, online forums, blogs, review sites, and social media*.

---

2. HR Analytics

HR departments are generating more data than ever before but at the same time they often struggle to turn their data into valuable insights. Some of the most

important analytics managers can use to better understand the people-related side of their business are;

1) Capability Analytics The success of the business depends on the level of expertise and skill of workforce. Capability analytics is a talent management process that allows manager to identify the capabilities or core competencies he want and need in his business. Once the manager knows what those capabilities that he can compare with the capabilities he has in place at the moment.
Here, Capabilities are not just about qualifications and skills; they can also include capabilities that may not be formally recognized, such as the ability to develop and maintain relationships.

2) Competency Acquisition Analytics Talent matters, and the acquisition and management of talent is often a critical factor in business growth. Competency acquisition analytics is the process of assessing how well or otherwise your business acquires the desired competencies. You need to start by identifying the core competencies your business requires now and in the future. Then assess the current levels of these competencies within your business and identify any gaps.
You can then monitor how effective you are at developing these competencies in house or spotting and recruiting candidates with those competencies. Key to effective competency acquisition analytics is focusing on a small set of core competencies.

3) Capacity Analytics Capacity affects revenue. Capacity analytics seeks to establish how operationally efficient people are in a business, e.g. are people spending too much time on admin and not enough on more profitable work, or are individuals stretched far too thin? It also allows businesses to establish of how much capacity they have to grow? The tricky part is establishing a system to track capacity without creating huge administrative burdens and without separating employees with a 'big-brother' approach. Big data and sensor system can be very effective here.

4) Employee Churn Analytics Hiring employees, training them and then integrating them into the business costs time and money. Employee churn analytics is the process of assessing your staff turnover rates in an attempt to predict the future and reduce employee churn. Historical employee churn can be identified through traditional KPIs such as *the employee satisfaction index, employee engagement level* etc. Surveys and exit interviews are also useful tools.

5) Corporate Culture Analytics Culture is difficult to pin point and even harder to change. It is essentially the collective (often unspoken) rules, systems and patterns of behaviour that represent your business. Corporate culture analytics is therefore the process of assessing and understanding more about your corporate culture or the different cultures that exists across your organization.
This then allows you to track changes in culture you would like to make, understand how the culture is changing, create early warning systems to detect toxic cultures in their development and ensure you are recruiting people that don't clash with the corporate culture. One way to assess culture is through the analysis of customer service conversations, which can provide a rich vein of data to assess corporate culture.

6) Recruitment Channel Analytics Employees represent the greatest cost and greatest opportunity in most businesses. Recruitment channel analytics is the process of working out where your best employees come from and what recruitment channels are most effective. Recruitment channel analytics will involve some historical assessment of employee value using KPIs such as *human capital value added and return per employee*.
Surveys and entry interviews are also useful sources of data.

7) Leadership Analytics Poor leadership, whether of a business, division or team costs money and prevents a business from fulfilling its potential. Leadership analytics unpacks the various dimensions of leadership performance via data to uncover the good, the bad and the ugly. Data about leadership performance can be gained through the use of *surveys, focus groups or employee interviews*.
8) Employee Performance Analytics Your business needs capable high performing employees to survive and thrive. Employee performance analytics seeks to assess individual employee performance. The resulting insights can identify who is performing well and who may need some additional training or support in order to raise their game. Today, we have many innovative ways of collecting and analysing performance, from crowd sourced performance assessments to big data analytics.

---

3. Financial Analytics

In today's data-filled world, analytics is an essential part of staying competitive.
Financial analytics help businesses understand current and past performance, predict future performance and make smarter decisions. Some of the key financial analytics that any business, regardless of size, should be using are;

1) Predictive Sales Analytics Sales revenue is the lifeblood of any business so knowing how much you can expect to receive has important tactical and strategic implications. Predictive sales analytics involves figuring out how successful your sales forecast is and improving your sales predictions in the future. There are many ways to predict sales, such as looking for trends in past data or using predictive techniques like correlation analysis.

2) Customer Profitability Analytics It's important to differentiate between the customers that make you money and the customers that lose you money. Customer profitability usually falls within the 80/20 rule, whereby 20% of your customers account for 80% of your profit, and 20% of your customers account for 80% of your customer related costs. Knowing which is which is important.

By understanding the profitability of certain groups of customers you can also analyse each group and extract useful insights. For example, you may discover that your very best customers made their first purchase from a particular advertisement in a particular magazine. That knowledge can help direct your future marketing efforts.

3) Product Profitability Analytics In order to stay competitive, businesses need to know where money is being made and lost. Product profitability analytics is a way of discovering profitability by individual product, rather than looking at the business as a whole. To do this you need to assess each product and its costs individually. This can be tricky because your products may well share production processes or cost bases.

Therefore, you need to find a reliable and fair way to apportion costs to your various products.

Product profitability analytics helps businesses uncover profitability insights across the product range so better decisions are made and profit is protected and grown over time. For example, if you discover that one product makes more profit than all the others then you may want to promote that product more heavily.

4) Cash Flow Analytics The day-to-day running of a business requires a certain amount of cash to keep the lights on, wages paid, etc. Knowing how money is moving in and out of your business is essential for gauging the health of your business. Cash flow analytics involves using retrospective or real-time indicators such as the Cash Conversion Cycle and Working Capital Ratio. You can also use tools like regression analysis to predict future cash flow.

Cash flow analytics can also support a variety of corporate functions. For

example, analytic software can help accounts receivable personnel to increase cash flow by prioritising which customers are contacted by collection staff and when.

5) Value Driver Analytics Most businesses have a sense of where they are heading and what they are trying to achieve. Often these goals are formalized on a strategy map that identifies the value drivers in the business. These value drivers are the key levers that the business needs to pull in order to meet its strategic objectives. Value driver analytics is the assessment of these levers to ensure they actually deliver the expected outcome.
Value drivers are often based on assumptions which need to be tested to check they are correct. For example, you may use price as one of your value drivers and assume that price influences sales and revenue, but you need to test that hypothesis so you can establish if you are right or not.

6) Shareholder Value Analytics The results and interpretation of the results by investors, analysts and the media will determine how successful your business is on the stock market. Shareholder value analytics is a calculation of the value of a company made by looking at the returns the business provides to its shareholders. It effectively measures the financial consequences of strategy and assesses how much value the business's strategy is actually delivering to the shareholders.
Shareholder value analytics should be used frequently alongside profit and revenue analytics. To measure shareholder value analytics, you can use a metric called
Economic Value Added (EVA). This calculates the profit of a business when the cost of equity finance has been removed.

---

4. Production and Operation Analytics

Production / Operations Management is defined as, the process which transforms the inputs/resources of an organization into final goods (or services) through a set of defined, controlled and repeatable policies. Production and operations management are more similar than different: if manufacturing products is a prime concern then it is called production management, whereas management of services is somewhat broader in scope and called operations management.
Aim of Production function is to add value to product or service which will create a strong and long lasting customer relationship or association. This can be achieved by healthy and productive association between Marketing and Production people.

Operation Management deals with; identify the customer needs and convert that into a specific product or service, Based on product requirement do back-ward working to identify raw material requirements as well as engage internal and external vendors to create supply chain for raw material and finished goods between vendor → production facility → customers.

There is also a term called *Industrial Analytics (IA)*, describes the collection, analysis and usage of data generated in industrial operations and throughout the entire product lifecycle, applicable to any company that is manufacturing and selling physical products. It involves traditional methods of data capture and statistical modelling.

In short, analytics help in Production and Operation functions in following ways;

1) Production Scheduling

As Scheduling is all about arranging, controlling and optimising work in production process, analytics help in many ways to improve this function. The problem like having a limited production capacity to make all of its offerings; as well as fluctuation in demand for those offerings, resolved by analytics. This can be done by prepare a forecasting model to predict demand for each product and visualization of the demand forecast via a custom application.
This will help by 90% reduction in planning time, reduced reliance on expertise of a single individual, confidence in planning by reducing human error or bias as well as continuous application relevance by updating the underlying model with new data.

2) **Variable Manufacturing Overhead Cost Analytics**

Manufacturing overhead costs are always varying. *Examples of Overhead Costs are; accounting and legal expenses, Administrative salaries, Depreciation, Insurance, Licenses and government fees, Rent etc*. To be profitable on any custom manufacturing project, it is required to monitor and closely control manufacturing overhead costs.
Analytics help to solve the problems by a custom built real-time cost tracking application as well as an interactive dashboard allowing custom views of individual & aggregate data (for cross-referencing).

This will help as; more than 3x (3 times) reduction in effort required to accurately measure and track costs, early identification of the impact of various factors, *e.g. overtime, new employee hire, etc. on overhead*, spot the reasons for a hike in

overhead costs as well as focus more cost/time on resolving cost issues rather than identifying causes for overhead fluctuations.

3) Supply Chain Analytics

Supply chain simply means, *the sequence of processes involved in the production and distribution of a commodity*. Supply chain management means, *the management of the flow of goods and services, involves the movement and storage of raw materials, of work-in-process inventory, and of finished goods from point of origin to point of consumption*.

In supply chain it is required, the acquisition of market intelligence to help minimize production costs. For that managers need to frequently access commodities Market data and related news; compare budgets and contract prices with that market data; forecast pricing changes; and update cost models of products to accurately reflect cost of goods. If done manually, this may be time/labour intensive.

So the goals would be; Make system that automatically track market prices and commodities data of raw materials and prepare model and forecast the cost of goods (to be sold) on a regular basis. Here, analytics help, by an application that tracks daily market data for commodities used in production, provide an intuitive tool that builds cost models of finished goods, an interactive dashboard that allows the end user to review the costs of each product based on underlying commodity prices.

4) Transportation Cost Forecasting Analytics

Sometimes, when product ships across the country via multiple distributions centers (DCs), and incurs different costs at different DCs at different times it is required to conduct proactive transportation budget planning with a high level of confidence. In short, the main objectives should be to understand, why different distribution centers (DCs) have different costs? Identify which costs are most influential to each DC and Predict aggregate transportation costs with confidence.

The said issues can be solved with the help of analytics by making an analysis of historical data to identify key cost factors and prepare a budget forecasting model to predict costs. This may be beneficial by Confidence in proactive planning by eliminating human error and bias and so on.

5) Customer Segmentation Analytics

Using big data and advanced analytics, manufacturers are able to view product quality and delivery accuracy in real-time as well as making transactions on which

suppliers receive the most time-sensitive orders. Managing to quality metrics becomes the priority over measuring delivery schedule performance alone.

Customer segmentation is required for scheduling manufacturing activities and planning as per the customer demand. Here, the task of the mangers would be to profile and segment prospective customers based on quantitative and qualitative data as well as automate the process of segmentation.
Analytics helps by making a high accuracy classification model that identifies the technological adoption category of any prospective firm based on its survey response, and quantifies the confidence of that result as well as prepare an interactive dashboard that allows managers to graphically view results, including the overall distribution of prospects across categories.

Ch. 3 – Analytics in Industries

Here we look how different industries apply analytics for business benefits.

One needs to have some understanding of industry domain basics, trends, common current challenges in order to develop industry specific analytics solutions.

- Analytics in Telecom
- Analytics in Retail
- Analytics in Healthcare
- Analytics in Financial services

1 Analytics in Telecom

In world of telecom industry, people and devices generate data 24*7, globally. Whether we are speaking with our friends, browsing a website, streaming a video, playing the latest game with friends, or making in-app purchases, user activity generates data about our needs, preferences, spending, complaints and so on.

Traditionally, communication service providers (CSPs) have leveraged this tsunami of data they generate to make decisions in areas of improving financial performance, increasing operational efficiency or managing subscriber relationship.

They have adopted advanced reporting and BI tools to bring facts and trends to decision makers.

Let us look at the role of analytics in CSP business:

Operational analytics:

1. **Network performance**: CSPs need to understand the bottlenecks in the network performance and optimize network utilization. They can use analytics to model capacity plans needed to meet service levels.

2. **Service Analytics:** This domain deals with analysis of customer problems, speed of resolution and identification of priority customers and ensures their satisfaction. Advanced analytics will deliver customer sentiment through social media analysis.

3. **Regulatory Analytics**: CSPs collaborate with other carriers and patterns to support roaming, sharing infrastructure, etc. They need to track regulatory compliance as per agreed contract norms and handle deviations through anomalies (something that deviates from what is standard, normal, or expected.) detections.

4. **Product Analysis**: It involves analysis of data to enhance revenue, launch promotions, create campaigns, and create new segments strategize pricing and study churn.

Subscriber analytics:

1. Fraud Detection: Analytics helps to detect billing and device theft, cloned SIMs, and related frauds as well as misuse of credentials.

2. **Subscriber Acquisition:** CSPs study customer behavior to identify the most suitable channels and sales strategy for each product.

3. **Churn Analytics**: This helps CSPs to not only model the loyalty programs but also predict churn and destination CSP.

4. **Value Segment Prediction**: Here CSPs will be able to enhance revenue by defining new subscriber base ahead of competition by matching their profitable offerings to subscribers needing them.

Financial analytics:

1. **Infrastructure Analytics:** CSPs study CAPEX and optimize investments in infrastructure and save money by considering utilization options.

2. Product portfolio Analytics: This area provides information into the profitable products and services and helps to exit from loss making products.

3. Channel Analytics: Helps CSPs to optimize the commercial terms with patterns to optimize distributor margins.

4. Cost Reduction: This area focuses on reducing service management cost, operations cost, compliance risks related cost, etc.

2. Analytics in Retail:

Only some industries have greater access to data around consumers, products, they buy and use, and different channels that sell and service products – and the lucky vertical is retail industry. Data coupled with insights are at the heart of what drives the retail business.

Technologies like Point of Sale (PoS), CRM, SCM, Big Data, Mobility and Social Media offer a means to understand shoppers via numerous digital touch points ranging from their online purchases, to their presence on social networks, to their visits to brick and mortar stores as well as tweets, images, video and more.

Value of analytics can come from three sources:

1. Gaining insight to improve process and resources optimization
2. Personalizing and localizing offers.
3. Creating community for branding and customer engagement.

1. Gaining insight to improve process and resource optimization:

Supply Chain Analytics: Every retailer needs to optimize the vendors of products, its cost and quality. They need to constantly track the performance of supply chain and initiate proactive actions for competitive advantage.

***Pricing Analytics*:** Helps retailers to optimize the product pricing, special offers, merchandizing, loyalty programs and campaigns that attract maximum numbers of consumers both from physical store and online store perspective.

***Buying Experience Analytics*:** Retailers can gain insight into the path taken to purchase, complaints registered, help provided by store personnel, store layout/item search time, product details availability, pricing, etc. and enhance the buying experience and train personnel for enhancing consumer loyalty.

2. Personalizing and localizing offers:

***Inventory Analytics**:* Retailers aim to fulfill consumer demand by optimizing stocks and ability to replenish when consumer demand increases due to seasonal effect or as a result of powerful campaigns. This area of analytics will alert store managers about the potential need for stocking highly moving items and reduce slow moving items.

Consumer Analytics: Every region around the world has people with different taste for goods and service levels. The purpose of consumer analytics is to equip store managers with insights to customize their products and services to the local consumer profile.

Campaign Analytics: All retailers will have digital marketing programs to entice consumers with value offers. Retailers invest in this area of analytics to design most effective campaigns that convert maximum number of consumers into buyers.

***Fraud Detection*:** All retailers strive to eliminate fraud relating to payments, shipping, and change of price tags and so on. Analytics can study transactions in real time to detect fraud and alert store personnel or online commerce teams.

3. Creating community for branding and customer engagement:

Web Analytics: Here the different perspective of each consumer's online behavior such as surfacing traffic, visitor and conversation trends, location of smart devices, and access to kiosk will be analyzed to recommend the best sales approach in response to each of the customer's real-time actions.

***Market Basket Analytics*:** The promotion, price offer and loyalty dimension of shopping behaviors will be used to understand sales patterns, customer preferences, and buying patterns to create targeted and profitable product promotions, customer offers and shelf arrangements.

Social Media Analytics: Listening and learning from the social community dimension of each consumer's online behavior is the scope of this area of analytics. Here store taps into customer-generated content with sentiment and behavioral analysis to answer key merchandise, and marketing strategy questions.

Consumer Behavioral Analytics: The focus area is consumer preferences such as channels, categories, brands and product attributes; return and exchange patterns; usage level of service programs; and partition in loyalty programs.

3. Analytics in health care (hospitals of healthcare providers)

Health care is very complex eco-system of multiple industries interconnected to achieve the health care goals of a country. These entities include healthcare providers, physicians, insurance companies, pharmaceutical companies, laboratories, healthcare volunteers, regulatory bodies, retail medicine distributors and so on centered on a patient.

You can imagine the complexity, variety, volume, velocity of data that gets generated in each of these independent enterprises and multitude of interconnected heterogeneous IT applications.

Analytics is applicable for all these enterprises, viz. insurance companies, pharmaceutical manufacturers, hospitals, etc. Here we will focus on how hospitals, that is, healthcare providers, can leverage analytics for goals like:

Hospital Management Analytics: It focuses on cost reduction, enhancing quality of care, improving patient satisfaction, improving outcomes (Performance of diagnosis, testing and treatment), providing secure access to patient data (Electronic Health Records - EHR). Analytics in this area can support fact-based

decisions in area of reduction of medical errors, manage diseases, understand physician performance and retain patients.

Compliance Analytics: Provide healthcare compliance metrics to regulatory authorities and bench mark against world-class hospitals using Baldridge criteria. Improvement in widespread use of digital data will support audits, analytics and improve hospital process needed for regulatory compliance.

Financial Analytics: This area of analytics will lead to enhance ROI, improved utilization of hospital infrastructure and human resources, optimize capital management, optimize supply chain and reduce fraud.

Predictive Models: They can help health care professionals go beyond traditional search and analysis of unstructured data by applying predictive root cause analysis, natural language and built-in medical terminology support to identify trends and patterns to achieve clinical and operational insights. Healthcare predictive analytics can help healthcare organizations get to know their patients better, so they can understand their individual patient's needs, while delivering quality, cost effective life saving services.

Social Analytics: It can help hospitals listen to patient sentiments, requirements, affordability, and insurance to model care and wellness programs customizing services by localization of needs.

Clinical Analytics: A number of other critical clinical situations can be detected by analytics applied to EHR such as:

1. Detecting postoperative complications.
2. Predicting 30 days risk of readmission.
3. Risk-adjusting hospital mortality rates.
4. Detecting potential delay in diagnosis.
5. Predicting out of intensive care unit each

4. Analytics in Financial Services Industry

Today's financial institutions have been compelled to deploy analytics and data driven capabilities to increase growth and profitability, to lower costs and improve efficiencies, to drive digital transformation, and to support risk and regulatory compliance priorities.

(A)How analytics can benefit Insurance companies?

1) Better Product Design and Marketing Insurers can take advantage of new sources of data to better target intended customers with specific – and potentially more suitable – products, making it possible to design offers based on what people need in the future, and to combine these with improvements in technology and regulation.

2) Risk Assessment and Pricing Processes Data analytics allow insurers to assess the risk profiles of their applicants in much greater detail, which should mean better informed underwriting decisions as well as premium calculations that will be more accurate in their alignment with the corresponding levels of risk.

3) Customer Reward Policy There is potential to reward policy holders with lower premiums, if their risk profile improves: this can be indicated by Smartphone apps that can monitor lifestyles of customers. The reward of a lower premium could also encourage policy holders to improve their lifestyle.

4) Better Claims Management Data analytics can be used to prioritize claims, and to set straight forward claims apart from complex cases. This can result in faster settlements for the straight forward claims, and more attention for the complex cases.

(B)How analytics can benefit Banking companies?

1) Better Customer Targeting By understanding clients more fully, and by using analytics of their transactions and trading activities, bank scan be sure that they are delivering the best services for what their customers need, resulting in higher levels of retention and acquisition.

2) Enhancing Risk Assessment As banks will be able to assess the risk profiles of their credit applicants in much greater detail, they will also be able to improve their credit assessments. Data analytics will advance the early-warning systems and data collection as well. All of these features will help banks to lower their risk costs, and to become aware of fraud more quickly.

3) More Business opportunities By collecting data from customers, data analytics Will enable banks to develop new business models and new sources of income: for

example, by sharing data with other companies, when the customer has agreed to this beforehand.

4) Improved Productivity and Decision Making With the advantage of advanced analytics, banks will be able to provide faster and more accurate responses to regulatory requests. Data will also enable better decisions for everyday activities: for example, better placement of ATMs and counters, and how much cash is required at each ATM.

5) Providing Efficient Digital Banking In today's society, most people conduct their transactions online, through their smart phones or their computers. By analysing real-time data, we can advance the customer experience and understand our customers much better.
