Chapter 1

Operational Vs Analytics DBs

Operational DBs

- What is current state ? Store up the last minute data
- Keep track of the on going state of process
- SEARCH

Analytical DBs

- Large table collection of data
- E.g. what change over past 5 years
- More static data set commutation of data set overtime or just collection of stable data
- ANALYSIS uncover new information

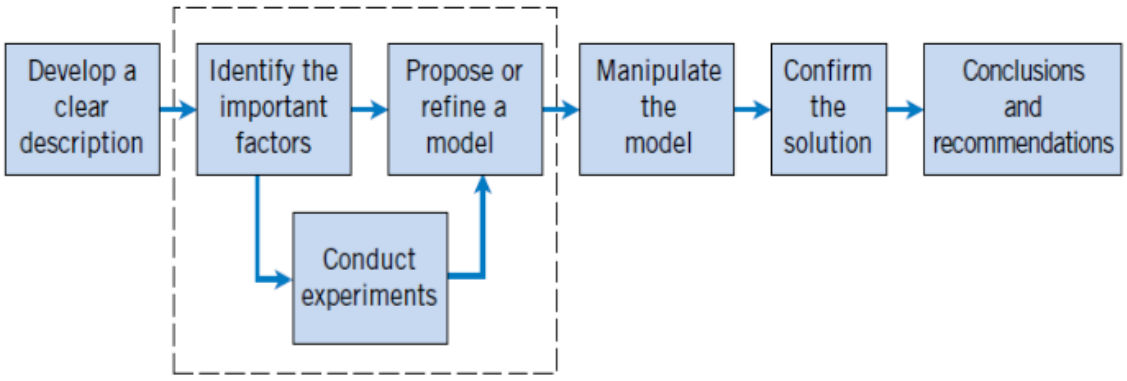

Fig 1.1 The Engineering Method

The steps in engineering methods towards solving problems

1. Develop a clear and concise description of the problem.
2. Identify, at least tentatively, the important factors that affect this problem or that may play a role in its solution.
3. Propose a model for the problem, using scientific or engineering knowledge of the phenomenon being studied. State any limitations or assumptions of the model.
4. Conduct appropriate experiments and collect data to test or validate the tentative model or conclusions made in steps 2 and 3.
5. Refine the model on the basis of the observed data.

6. Manipulate the model to assist in developing a solution to the problem.
7. Conduct an appropriate experiment to confirm that the proposed solution to the problem is both effective and efficient.
8. Draw conclusions or make recommendations based on the problem solution.

Statistics is a collection of methods which help us to describe, summarize, interpret, and analyze data.

Observation

The units on which we measure data—such as persons, cars, animals, or plants—are called observations.
These units/observations are represented by the Greek symbol ω. The collection of all units is called population and is represented by Ω.

Population

population refers to the total set of observations that can be made.

Example If we are studying the weight of adult women, the population is the set of weights of all the women in the world.

Sample

A sample is a set of data collected and/or selected from a population by a defined procedure.

Chapter 2

Business Intelligence(BI)

Business Intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. Chapter describes the problems, interconnections with other disciplines and components of Business intelligence

Data is a set of values of qualitative or quantitative variables. For example, for a retailer data refer to primary entities such as customers, points of sale and items, commercial transaction

Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain

Knowledge Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. knowledge - consisting of information put to work into a specific domain

The role of mathematical models

Adoption of a business intelligence system tends to promote a scientific and rational approach to the management of enterprises and complex organizations.

Business intelligence analysis can be summarized as

• First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
• Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.

• Finally, what-if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameter

Business Intelligence architecture

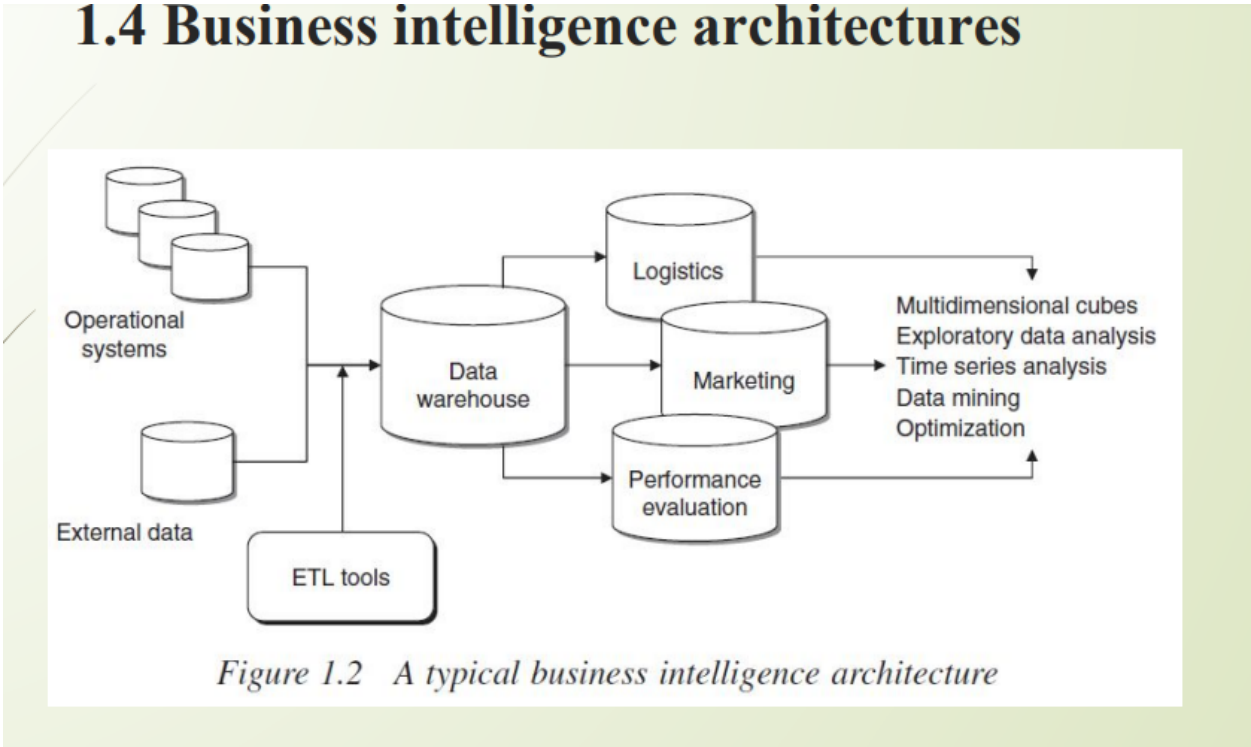

Figure 1.2 A typical business intelligence architecture

Ethics and business intelligence

The adoption of business intelligence methodologies, data mining methods and decision support systems raises some ethical problems that should not be over- looked.

Indeed, the progress toward the information and knowledge society opens up countless opportunities, but may also generate distortions and risks which should be prevented and avoided by using adequate control rules and mechanisms.

Usage of data by public and private organizations that is improper and does not respect the individuals' right to privacy should not be tolerate

1) **Decision support system :-**

A *decision support system (DSS) is an interactive computer-based application* that **combines data and mathematical models** to help decision makers solve complex problems faced in managing the public and private enterprises and organizations.

Keywords :- Info sys that supports decision making activities.

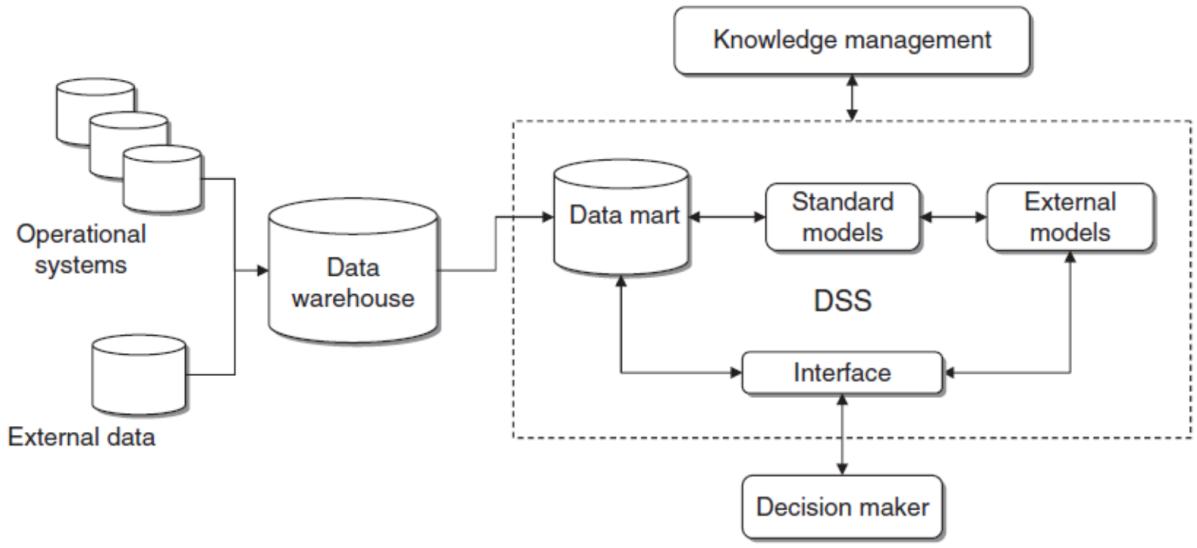

Figure 2.9 Extended structure of a decision support system

i) Data Management

This module includes a database designed to contain the data required by the decision-making processes to which the DSS is addressed. In most applications the database is a data mart.

The data management module of a DSS is usually connected with a company data warehouse, which represents the main repository of the data available to develop a business intelligence analysis.

ii) Model Management

The model management module provides end users with a collection of mathematical models derived from operations research, statistics and financial analysis.

These are usually relatively simple models that allow analytical investigations to be carried out that are very helpful during the decision-making process.

In certain applications such a module may be integrated with more complex models, referred to as **external models** in Figure 2.9, created to carry out specific analysis tasks. For example, a large-scale optimization model formulated to develop the annual logistic plan of a manufacturing company falls in this category

iii) Interaction

In most applications, knowledge workers use a DSS interactively to carry out their analyses. The module responsible for these interactions is expected to receive input data from users in the easiest and most intuitive way, usually through the graphic interface of a web browser, and then to return the extracted information and the knowledge generated by the system in an appropriate graphical form.

iv) Knowledge management

The knowledge management module is also interconnected with the company knowledge management integrated system. It allows decision makers to draw on the various forms of collective knowledge, usually unstructured, that represents the corporate culture.

2) **Types of Decisions :-**

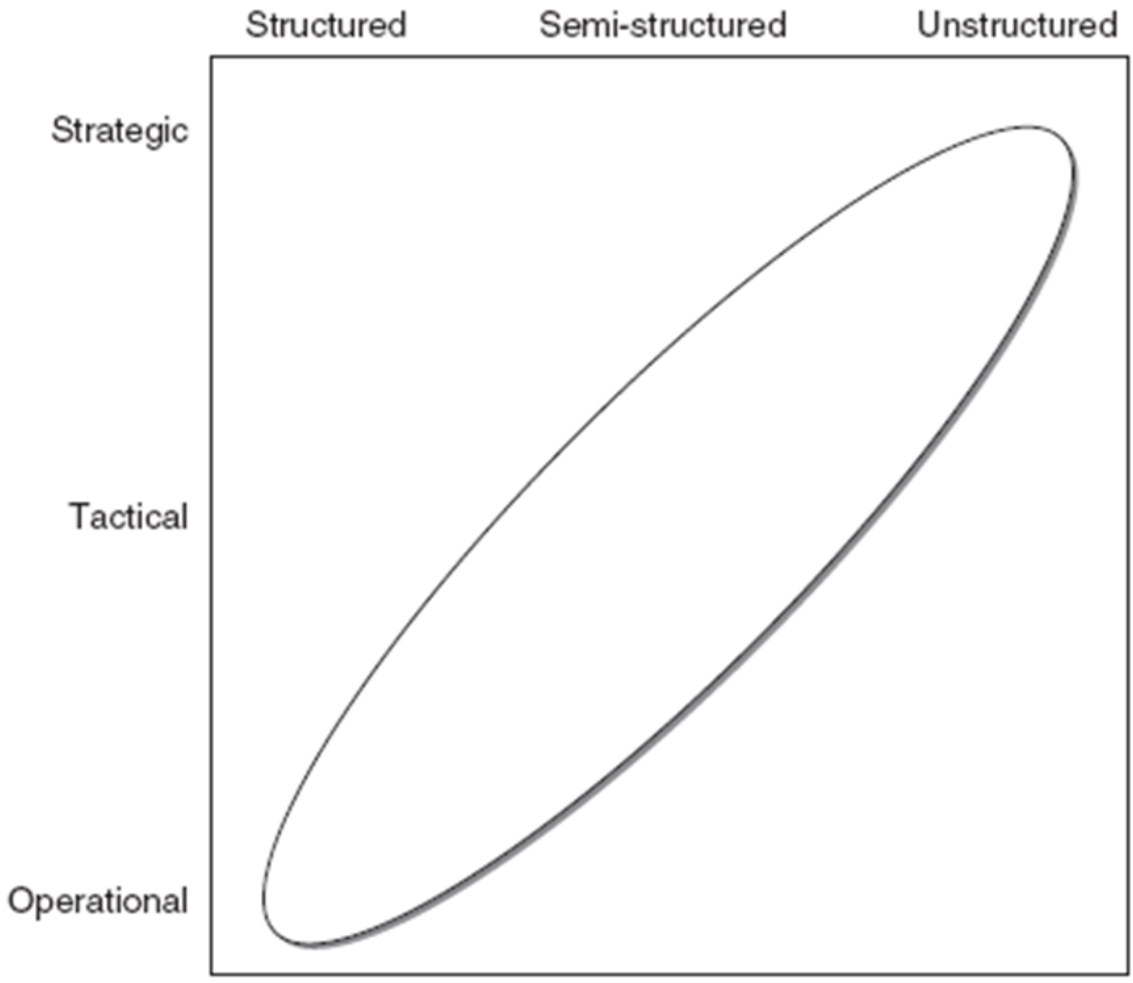

Figure 2.6 A taxonomy of decisions

1. **Structured decisions :-**

A decision is structured if it is based on a well-defined and recurring decision-making procedure. In most cases structured decisions can be traced back to an algorithm, which may be more or less explicit for decision makers, and are therefore better suited for automation.

are those in which the decision maker must provide judgment, evaluation, and insights into the problem definition

2. **Unstructured decisions :-**

A decision is said to be unstructured if the three phases of intelligence, design and choice are also unstructured. This means that for each phase there is at least one element in the system (input flows, output flows and the transformation processes) that cannot be described in detail and reduced to a predefined sequence of steps.

´are repetitive and routine, and decision makers can follow a definite procedure for handling them to be efficient.

3. **Semi-structured decisions :-**

when some phases are structured and others are not. Most decisions faced by knowledge workers in managing public or private enterprises or organizations are semi-structured. Hence, they can take advantage of DSSs and a business intelligence environment. Semi are those in which only part of the problem has a clear-cut answer provided by an accepted procedure.

Depending on their scope, decisions can be classified as:

´1. **Strategic decisions.** Decisions are strategic when they affect the entire organization or at least a substantial part of it for a long period of time.

Strategic decisions strongly influence the general objectives and policies of an enterprise. strategic decisions are taken at a **higher organizational level,** usually by the company top management.

2. ´**Tactical decisions.** Tactical decisions affect only parts of an enterprise and are usually restricted to a single department.

The time span is limited to a medium-term horizon, typically up to a year. Tactical decisions place themselves within the context determined by strategic decisions. In a company hierarchy, tactical decisions are made by **middle managers**, such as the heads of the company departments.

3. ´**Operational decisions.** Operational decisions refer to specific activities carried out within an organization and have a modest (uncertain) impact on the future.

Operational decisions are framed within the elements and conditions determined by strategic and tactical decisions. Therefore, they are usually made at a lower organizational level, by knowledge workers responsible for a single activity or task such as **sub-department heads, workshop foremen, back-office heads**.

Chapter 3

1. What is Big Data?

"Big Data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." -- Gartner

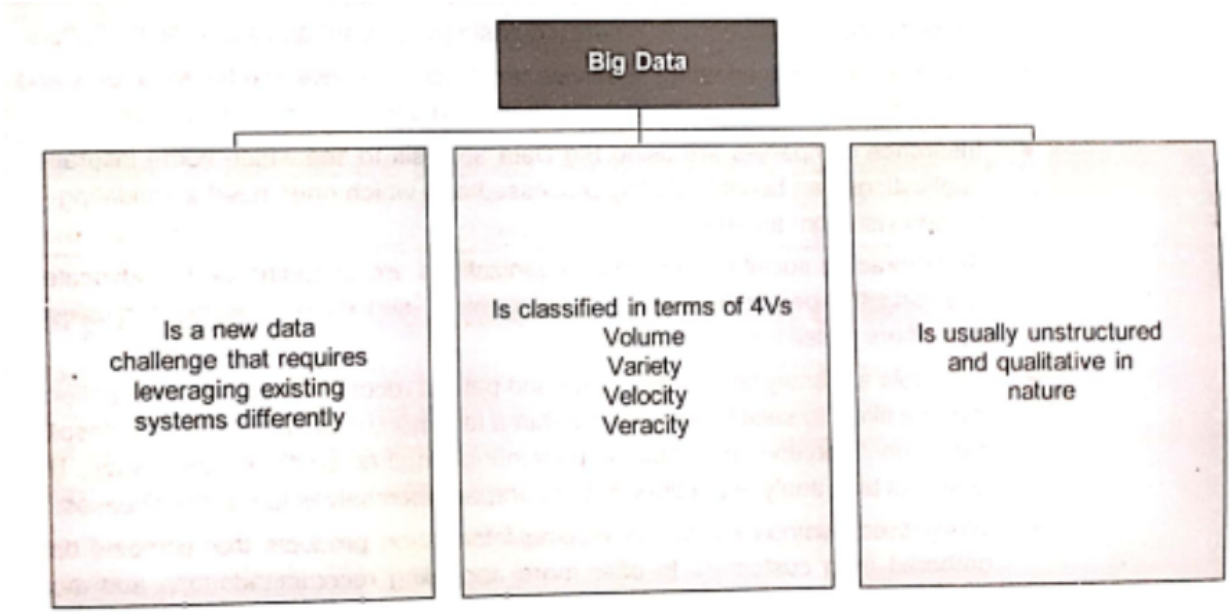

Figure 1.1: Features of Big Data

1. Volume:

- The name 'Big Data' itself is related to a size which is enormous.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.

2. Velocity:

- Velocity refers to the high speed of accumulation of data.
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Sampling data can help in dealing with the issue like ‘velocity’.
- *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

3. Variety:

- It refers to the nature of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.
    - **Structured data**: This data is basically an organized data. It generally refers to data that has defined the length and format of data.
    - **Semi- Structured data**: This data is basically a semi-organised data. It is generally a form of data that does not conform to the formal structure of data. Log files are examples of this type of data.

- **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are examples of unstructured data which can't be stored in the form of rows and columns.

4. Veracity:

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

5. Value:

- After having the 4 V's into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5 V's.

2. Evolution of Big Data

History of Data Management – Evolution of Big Data

Big Data is the new term of data evolution directed by the enormous velocity, variety, and volume of data. Velocity implies the speed with which the data flows in an organization; variety refers to the varied forms of data, such as structured, semi-structured, or unstructured; and volume defines the amount or quantity of data an organization has to deal with.

Figure 1.2 shows the challenges faced while handling data over the past few decades:

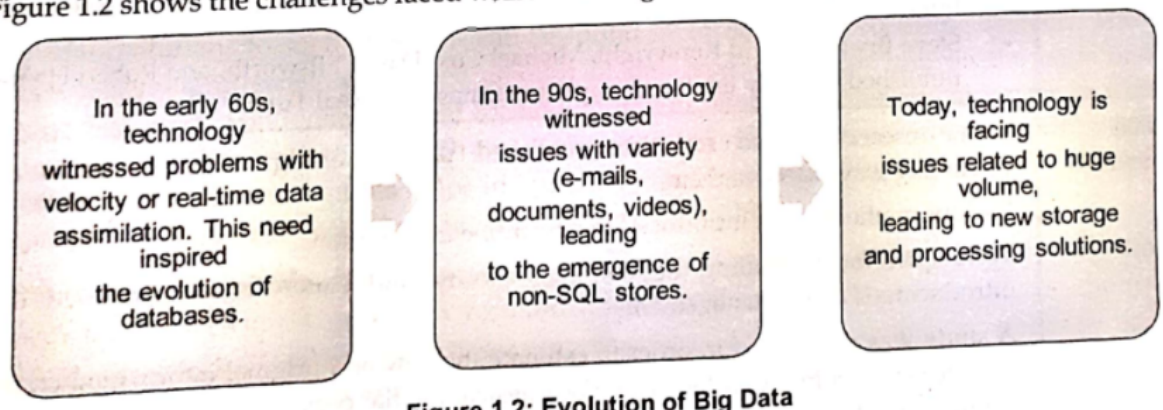

Figure 1.2: Evolution of Big Data

The advent of IT, the Internet, and globalization has facilitated increased volumes of data information generation at an exponential rate, which has led to 'information explosion.' This, in fueled the evolution of Big Data that started in 1940s and continues till date.

Information explosion is described as a continuous increase in the volume of the publi information or data and the effects of this abundant information.

Table 1.2 lists some major milestones in the evolution of Big Data:

Table 1.2: Evolution of Big Data

Year	Milestone
1940s	An American librarian speculated the potential shortfall of shelves and cataloging sta realizing the rapid increase in information and limited storage.
1960s	Automatic Data Compression was published in the Communications of the ACM, states that the explosion of information in the past few years makes it necessary tha requirements for storing information should be minimized. The paper described 'Automatic Data Compression' as a complete automatic and fas three-part compressor that can be used for any kind of information in order to reduce the slow external storage requirements and increase the rate of transmission from a computer system.
1970s	In Japan, the Ministry of Posts and Telecommunications initiated a project to study information flow in order to track the volume of information circulating in the country.
1980s	A research project was started by the Hungarian Central Statistics Office to account for the country's information industry. It measured the volume of information in bits.
1990s	Digital storage systems became more economical than paper storage. Challenges related to the amount of data and the presence of obsolete data became apparent. Some papers that discussed this concern are as follows: • Michael Lesk published How much information is there in the world? • John R. Masey presented a paper titled Big Data... and the Next Wave of InfraStress. • K.G. Coffman and Andrew Odlyzko published The Size and Growth Rate of the Internet. • Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes published Visually Exploring Gigabyte Datasets in Real Time.
2000 onwards	Many researchers and scientists published papers raising similar concerns and discussing ways to solve them. Various methods were introduced to streamline information. Techniques for controlling the Volume, Velocity, and Variety of data emerged, like introducing 3D data management. A study was carried out in order to estimate the new and original information created and stored worldwide in four types of physical media: paper, film, optical media and magnetic media.

3. Structure of Big Data :-

1. Structured :- organised data, have some repeated patterns, eg: relational database.
2. Unstructured :- data with different formats or patterns like audio, vdo, imgs, txts.
3. Semi-Structured :- kind of combination eg :- JSON

On the basis of the data received from the sources mentioned in Table 1.3, Big Data comprises:

- Structured data
- Unstructured data
- Semi-structured data

In a real-world scenario, typically, the unstructured data is larger in volume than the structured and semi-structured data, approximately 70% to 80% of data is in unstructured form. Figure 1.4 illustrates the types of data that comprise Big Data:

Figure 1.4: Types of Big Data

Let us discuss these types in detail in the following sections.

Structured Data

Structured data can be defined as the data that has a defined repeating pattern. This pattern makes it easier for any program to sort, read, and process the data. Processing structured data is much easier and faster than processing data without any specific repeating patterns.

Structured data:

- Is organized data in a predefined format
- Is stored in tabular form
- Is the data that resides in fixed fields within a record or file
- Is formatted data that has entities and their attributes mapped
- Is used to query and report against predetermined data types

Some sources of structured data include:

- Relational databases (in the form of tables)
- Flat files in the form of records (like comma separated values (csv) and tab-separated files)
- Multidimensional databases (majorly used in data warehouse technology)
- Legacy databases

Table 1.4 shows a sample of structured data in which the attribute data for every customer is stored in the defined fields:

Table 1.4: Sample of Structured Data

Customer ID	Name	Product ID	City	State
12365	Smith	241	Graz	Styria
23658	Jack	365	Wolfsberg	Carinthia
32456	Kady	421	Enns	Upper Austria

Unstructured Data

Unstructured data is a set of data that might or might not have any logical or repeating patterns.

SCENARIO

To better understand the concept of unstructured data, let us go back to the meeting of Mr. Smith. He explains that the publishing house also collects data from various blogs and websites. The data obtained from Web blogs or social media sites is considered as unstructured data because it does not follow any specific pattern and is inconsistent. The analysis of such data helps the organization to know more about customer preferences, feedback, and demands.

Unstructured data:

- Consists typically of metadata, i.e., the additional information related to data
- Comprises inconsistent data, such as data obtained from files, social media websites, satellites, etc.
- Consists of data in different formats such as e-mails, text, audio, video, or images

Some sources of unstructured data include:

- **Text both internal and external to an organization**—Documents, logs, survey results, feedbacks, and e-mails from both within and across the organization

- **Social media**—Data obtained from social networking platforms, including YouTube, Facebook, Twitter, LinkedIn, and Flickr
- **Mobile data**—Data such as text messages and location information

About 80 percent of enterprise data consists of unstructured content.

Semi-Structured Data

Semi-structured data, also known as having a schema-less or self-describing structure, refers to a form of structured data that contains tags or markup elements in order to separate elements and generate hierarchies of records and fields in the given data. Such type of data does not follow the proper structure of data models as in relational databases. In other words, data is stored inconsistently in rows and columns of a database.

Some sources for semi-structured data include:

- File systems such as Web data in the form of cookies
- Data exchange formats such as JavaScript Object Notation (JSON) data

SCENARIO

Mr. Smith also observes the presence of some semi-structured data saved in the database system of the publishing house. This data contains personal details of the authors working for the publishing house, as shown in Table 1.5:

Table 1.5: Semi-Structured Data

Sl. No	*Name*	*E-Mail*
1.	*Sam Jacobs*	*smj@xyz.com*
2.	*First Name: David* *Last Name: Brown*	*davidb@xyz.com*

As you can notice from Table 1.5, semi-structured data indicates that the entities belonging to the same class can have different attributes even if they are grouped together. In this case, different names and different e-mails are grouped under a common column name.

4. Elements of Big Data :- 5 V's

5. Big Data Analytics :-

Businesses, nowadays, rely heavily on big data to gain better knowledge about their customers. The process of extracting meaningful insights from such raw big data is known as big data analytics.

Three main types of Big data analytics :-

1. **Descriptive Analytics :-**

- It answers the question " What happened in the business?"
- Provide insights into what has occurred in the past and with the trends to dig into for more detail.
- This helps in creating reports like a company's revenue, profits, sales, and so on.
- Examples of descriptive analytics include summary statistics, clustering, and association rules used in market basket analysis.

2. **Predictive Analytics :-**

- It answers the question " What could happen?" by using statistical model and different forecast technique.
- Predictive Analytics, as can be discerned from the name itself, is concerned with predicting future incidents. These future incidents can be market trends, consumer trends, and many such market-related events.

- This type of analytics makes use of historical and present data to predict future events. This is the most commonly used form of analytics among businesses.
- Predictive analytics doesn’t only work for the service providers but also for the consumers. It keeps track of our past activities and based on them, predicts what we may do next.

- "The purpose of predictive analytics is NOT to tell you what will happen in the future. It cannot do that. In fact, no analytics can do that. Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature."

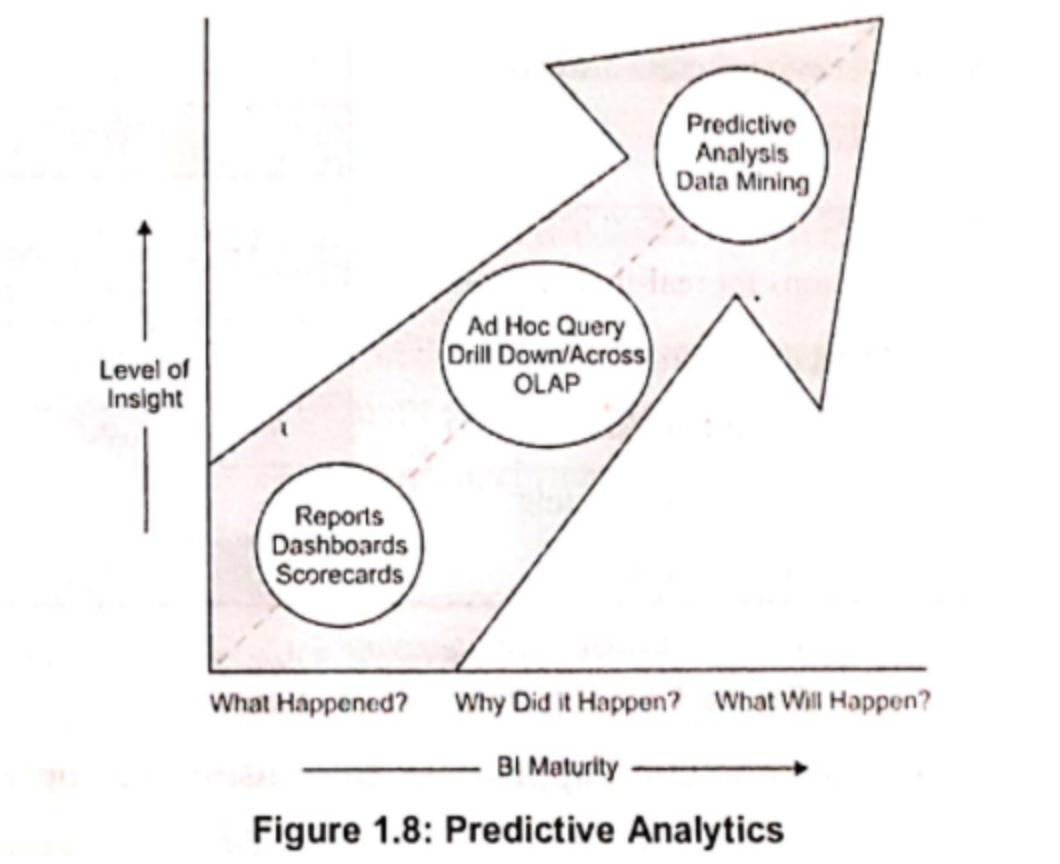

Figure 1.8: Predictive Analytics

3. Prescriptive Analytics :-

- It answers the question " What should we do?" on basis of complex data obtained from descriptive and predictive analysis.
- Prescriptive analytics is a combination of data and various business rules. The data of prescriptive analytics can be both internal (organizational inputs) and external (social media insights).

- Prescriptive analytics allows businesses to determine the best possible solution to a problem. When combined with predictive analytics, it adds the benefit of manipulating a future occurrence like mitigate future risk.
- Examples of prescriptive analytics for customer retention is the next best action and next best offer analysis.

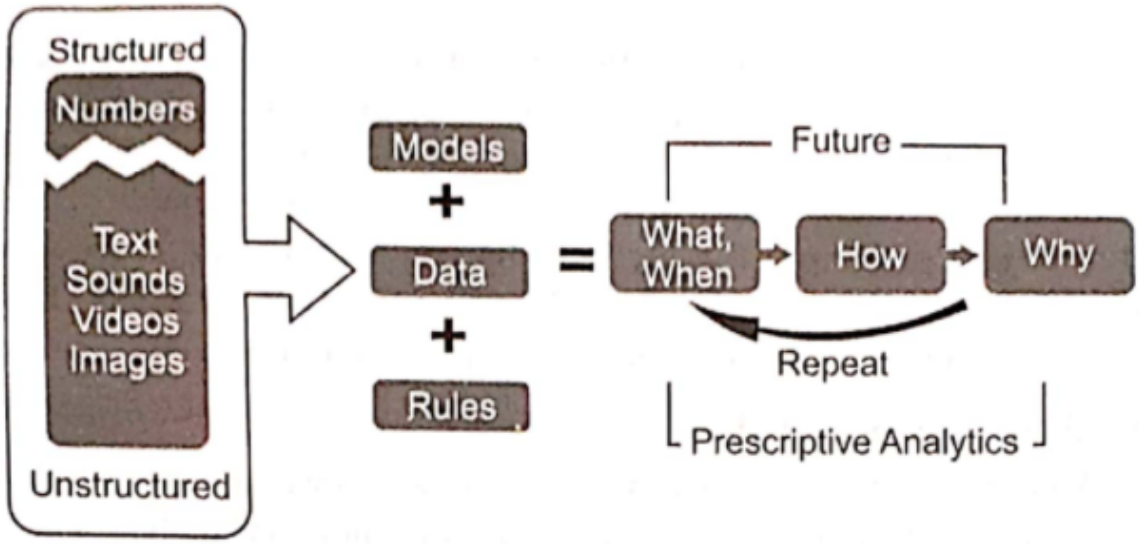

Figure 1.9: Prescriptive Analytics