INTRODUCTION TO DATA ANALYTICS

UNIT1-SYLLABUS

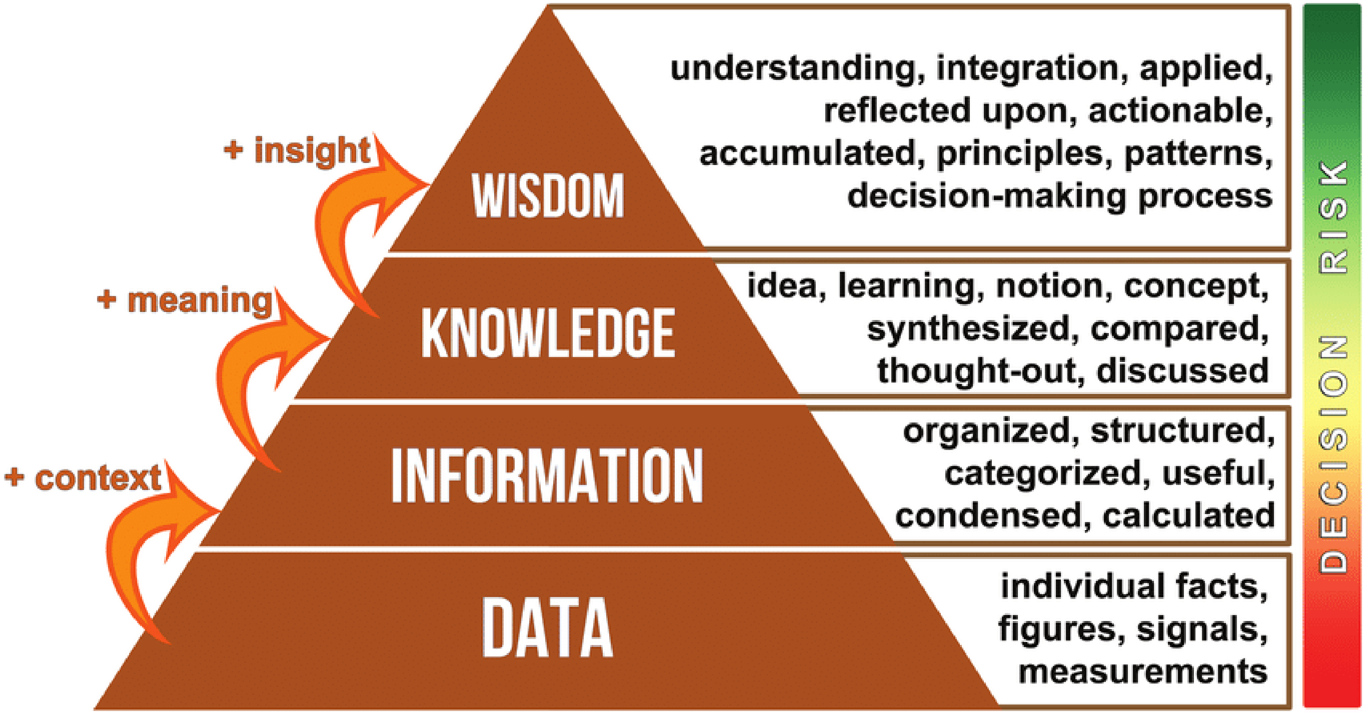
Introduction To Data Analytics

- Sources and nature of data
- Classification of data (structured, semi-structured, unstructured)
- Characteristics of data
- Introduction to Big Data platform
- Need of data analytics
- Evolution of analytic scalability
- Analytic process and tools
- Analysis vs reporting
- Modern data analytic tools
- Applications of data analytics

Data Analytics Life-Cycle

- Need of data analytics
- key roles for successful analytic projects
- Various phases of data analytics lifecycle
- Discovery
- Data preparation
- Model planning
- Model building,
- Communicating results
- Operationalization

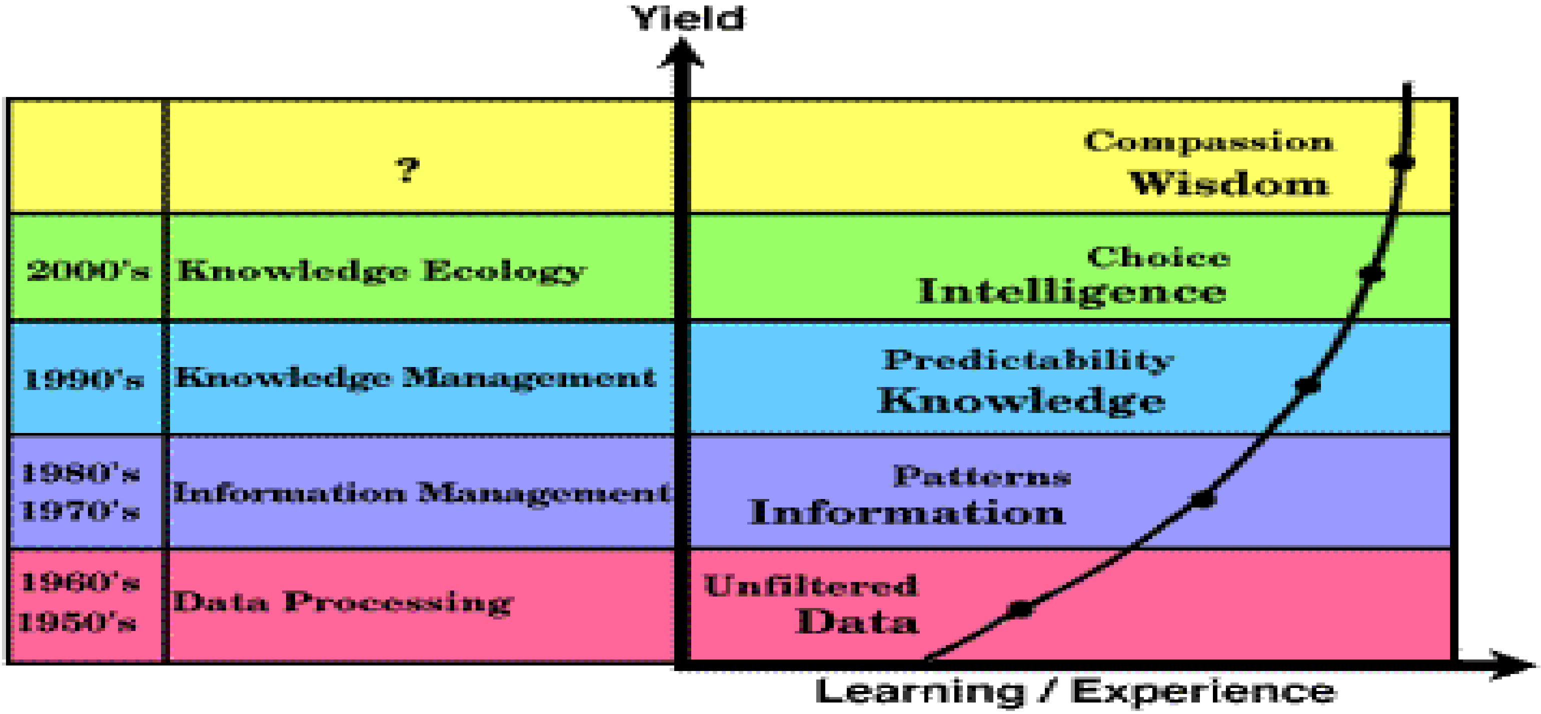
Data Analytics Needs Data



+ insight
+ meaning
+ context
WISDOM
KNOWLEDGE
INFORMATION
DATA
understanding, integration, applied, reflected upon, actionable, accumulated, principles, patterns, decision-making process
idea, learning, notion, concept, synthesized, compared, thought-out, discussed
organized, structured, categorized, useful, condensed, calculated
individual facts, figures, signals, measurements
DECISION RISK

Hierarchy of Visual Understanding

- Data are the pure and simple facts without any particular structure or organization, the basic atoms of information,
- Information is structured data, which adds meaning to the data and gives it context and significance,
- Knowledge is the ability to use information strategically to achieve one's objectives,
- Wisdom is the capacity to choose objectives consistent with one's values within a larger social context.



Yield = intellectual dividends per measure of effort invested.
Examples: increased clarity, deeper understanding.

Let’s be specific:

Data has been the buzzword for ages now. Either the data being generated from large-scale enterprises or the data generated from an individual, each and every aspect of data needs to be analyzed to benefit yourself from it. But how do we do it? Well, that’s where the term ‘**Data Analytics**’ comes in.

What is Data Analytics?

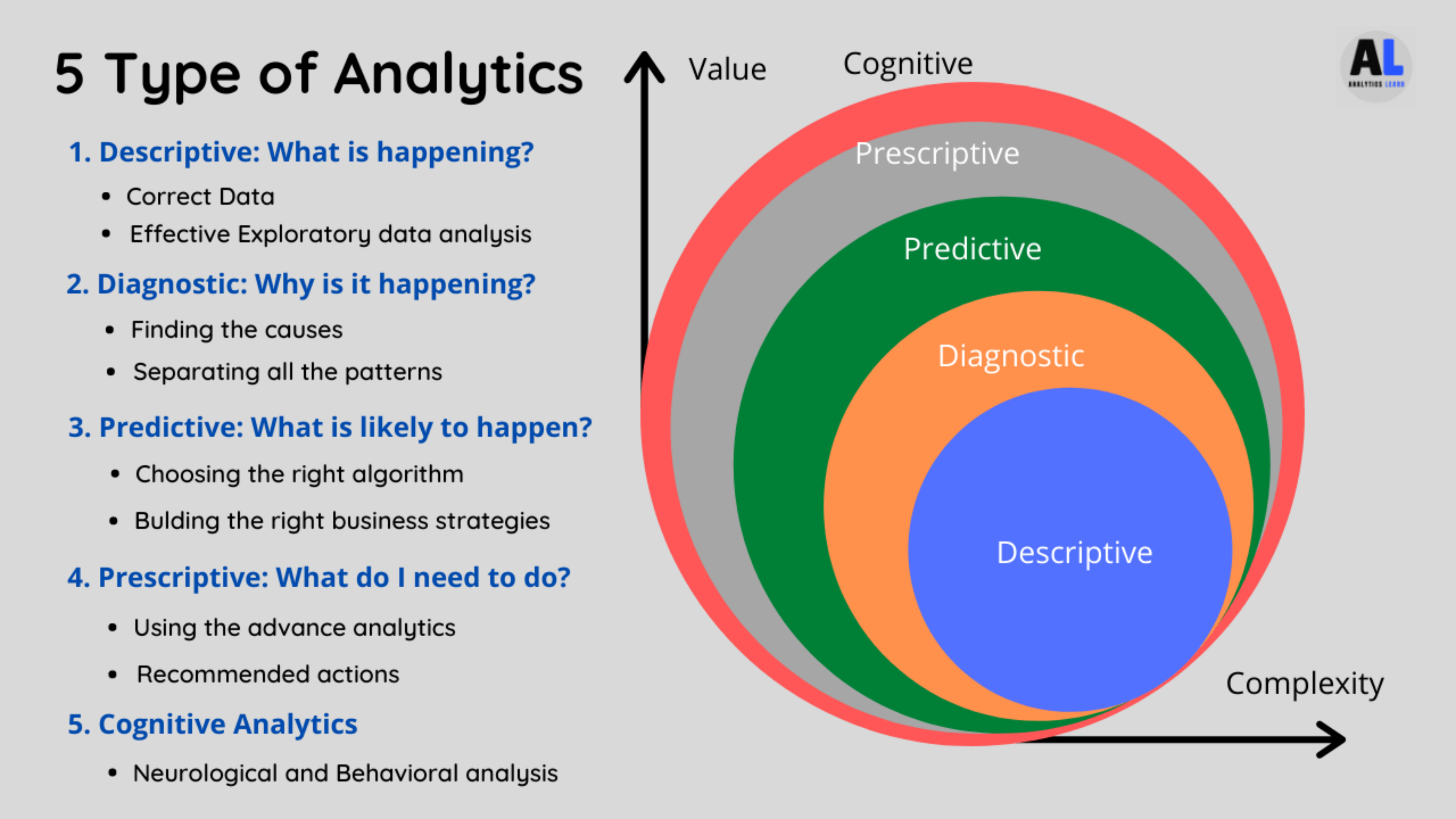
- Data Analytics refers to the techniques used to analyze data to enhance productivity and business gain. Data is extracted from various sources and is cleaned and categorized to analyze various behavioral patterns. The techniques and the tools used vary according to the organization or individual.

- Data analytics refers to the process of inspecting, cleaning, transforming and applying data to extrapolate useful information.
- Data analytics has become a primary tool within business today as it is used for various types of decision-making processes.
- Data analytics can be applied to any size of data sets, but as time progresses, businesses collect big data or a high volume of information.

What is Data Analytics?

- This historical data makes analytics more accurate as techniques can be applied to predict the future. Data tools can pull data from multiple locations and sources.
- We can derive, data analysis is the process of collecting and organizing data in order to draw helpful conclusions from it.
- The process of data analysis uses analytical and logical reasoning to gain information from the data.
- The main purpose of data analysis is to find meaning in data so that the derived knowledge can be used to make informed decisions.

Types of Data Analytics?



5 Type of Analytics

1. Descriptive: What is happening?

- Correct Data
- Effective Exploratory data analysis

2. Diagnostic: Why is it happening?

- Finding the causes
- Separating all the patterns

3. Predictive: What is likely to happen?

- Choosing the right algorithm
- Bulding the right business strategies

4. Prescriptive: What do I need to do?

- Using the advance analytics
- Recommended actions

5. Cognitive Analytics

- Neurological and Behavioral analysis

Types of Data Analytics

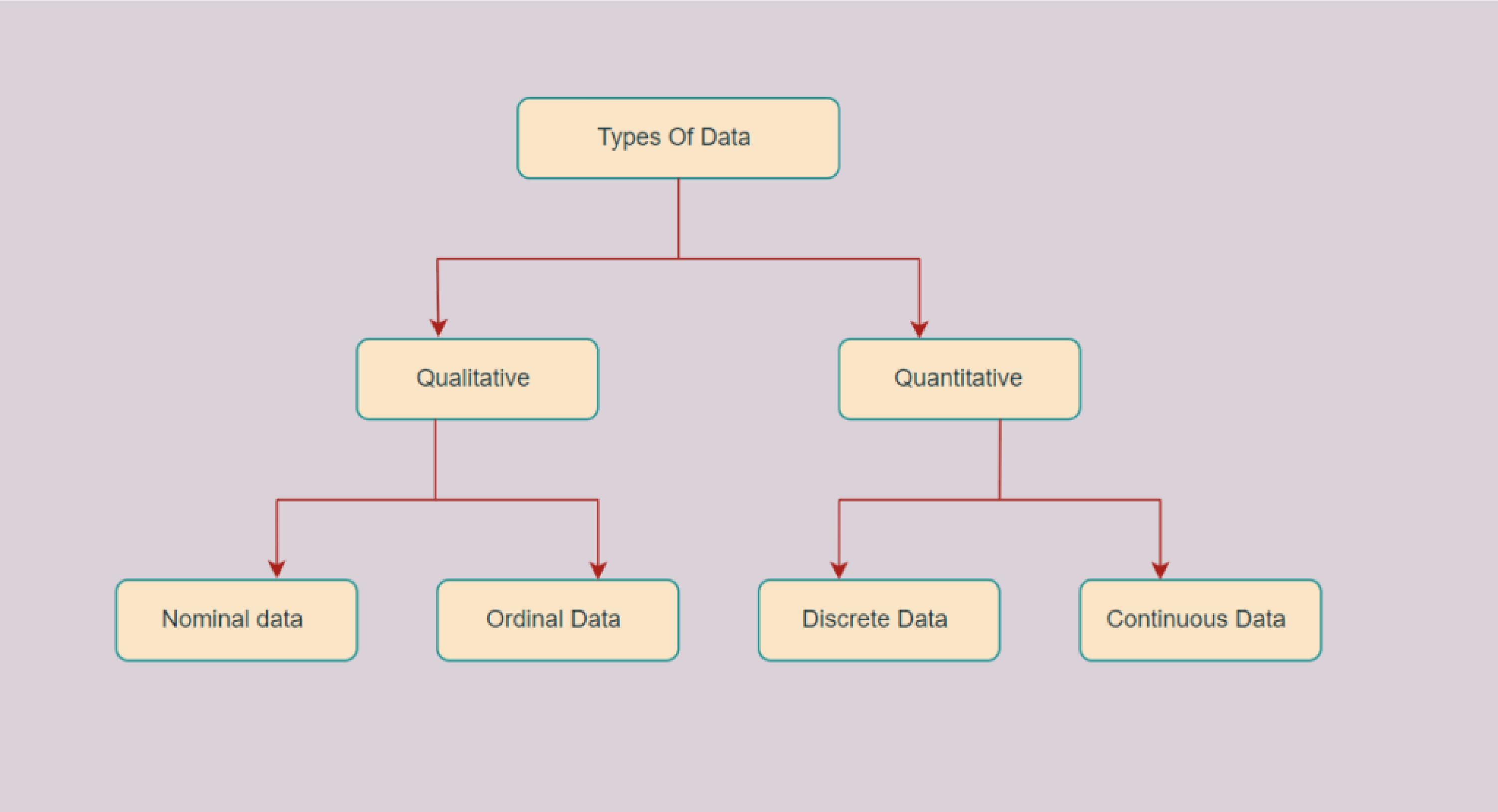
- **1. Descriptive:** If you have a question about something that has already happened, then descriptive analytics can help you answer it. Descriptive analytics is often used as a means to explain something to stakeholders. For example, it can track return on investment (ROI) and other metrics of past performance.
- As useful as descriptive analytics is, it’s best to combine descriptive analytics with another method like diagnostic to go deeper into why something has happened. Descriptive analytics will point out what happened, but you will need to explore the reasoning behind the event still.
- **2. Diagnostic:** Like a diagnosis, diagnostic analytics provides insight as to why something happened. They work hand-in-hand with descriptive analytics to further explain critical findings.
- If you take a look at key performance indicators (KPIs) and want to understand why something improved or got worse, then diagnostic analytics help to:
- Identify anomalies in data
- Collect the data that helps to understand the changes
- Uses statistical techniques to help explain such anomalies

Types of Data Analytics

- **3. Predictive:** Answer questions about what could happen in the future.
- This analytic method leverages past data to evaluate trends and estimate the likelihood of something recurring. Statistical analysis, regression and machine learning is used to make predictive analytics function.
- **4. Prescriptive:** If you find yourself in a critical position to make a decision about the future but feel unsure about what choice to make, prescriptive analytics can be a lifesaver.
- Prescriptive analytics works by finding patterns from large datasets and then estimates the likelihood of different outcomes.
- **5. Cognitive:** Cognitive Analytics applies human-like intelligence to certain tasks, and brings together a number of intelligent technologies, including semantics, artificial intelligence algorithms, deep learning and machine learning.

SOURCES AND NATURE OF DATA

Nature of Data



Qualitative Data

- Qualitative data is defined as the data that approximates and characterizes.
- Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods. Qualitative data in statistics is also known as categorical data – data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon.

Qualitative/Categorical Data

Nominal Data:

- Nominal Data is used to label variables without any order or quantitative value. The colour of hair can be considered nominal data, as one colour can't be compared with another colour.
- The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed to distinct categories.
- **Examples of Nominal Data :**
- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)
- Gender (Male, Female, Others)
- Eye Color (Black, Brown, etc.)

Qualitative/Categorical Data

Ordinal Data

- Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can’t do any arithmetical tasks on them.
- The ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered as “in-between” the qualitative data and quantitative data. Compared to the nominal data, ordinal data have some kind of order that is not present in nominal data.
- **Examples of Ordinal Data :**
- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of peoples in a competition (First, Second, Third, etc.)
- Economic Status (High, Medium, and Low)

Quantitative Data/Numerical Data

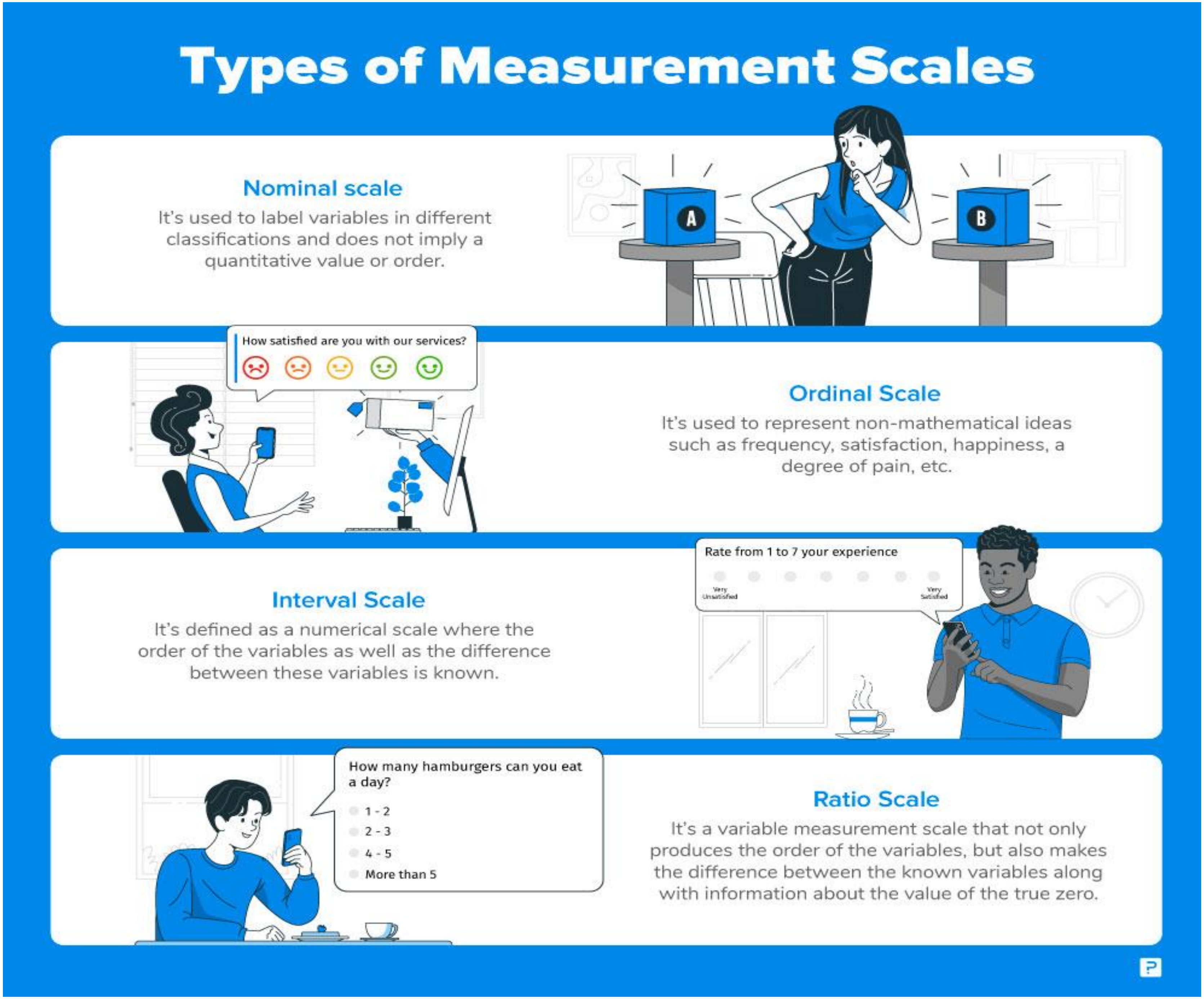
- Quantitative data is the value of data in the form of counts or numbers where each data set has a unique numerical value. This data is any quantifiable information that researchers can use for mathematical calculations and statistical analysis to make real-life decisions based on these mathematical derivations.

Quantitative/Numerical Data

- **Discrete Data**
- The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can’t be broken into decimal or fraction values.
- The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.
- **Examples of Discrete Data :**
- Total numbers of students present in a class
- Cost of a cell phone
- Numbers of employees in a company
- The total number of players who participated in a competition
- Days in a week

Quantitative/Numerical Data

- **Continuous Data**
- Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.
- The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.
- **Examples of Continuous Data :**
- Height of a person
- Speed of a vehicle
- “Time-taken” to finish the work
- Wi-Fi Frequency
- Market share price



Types of Measurement Scales

Nominal scale

It's used to label variables in different classifications and does not imply a quantitative value or order.

Ordinal Scale

It's used to represent non-mathematical ideas such as frequency, satisfaction, happiness, a degree of pain, etc.

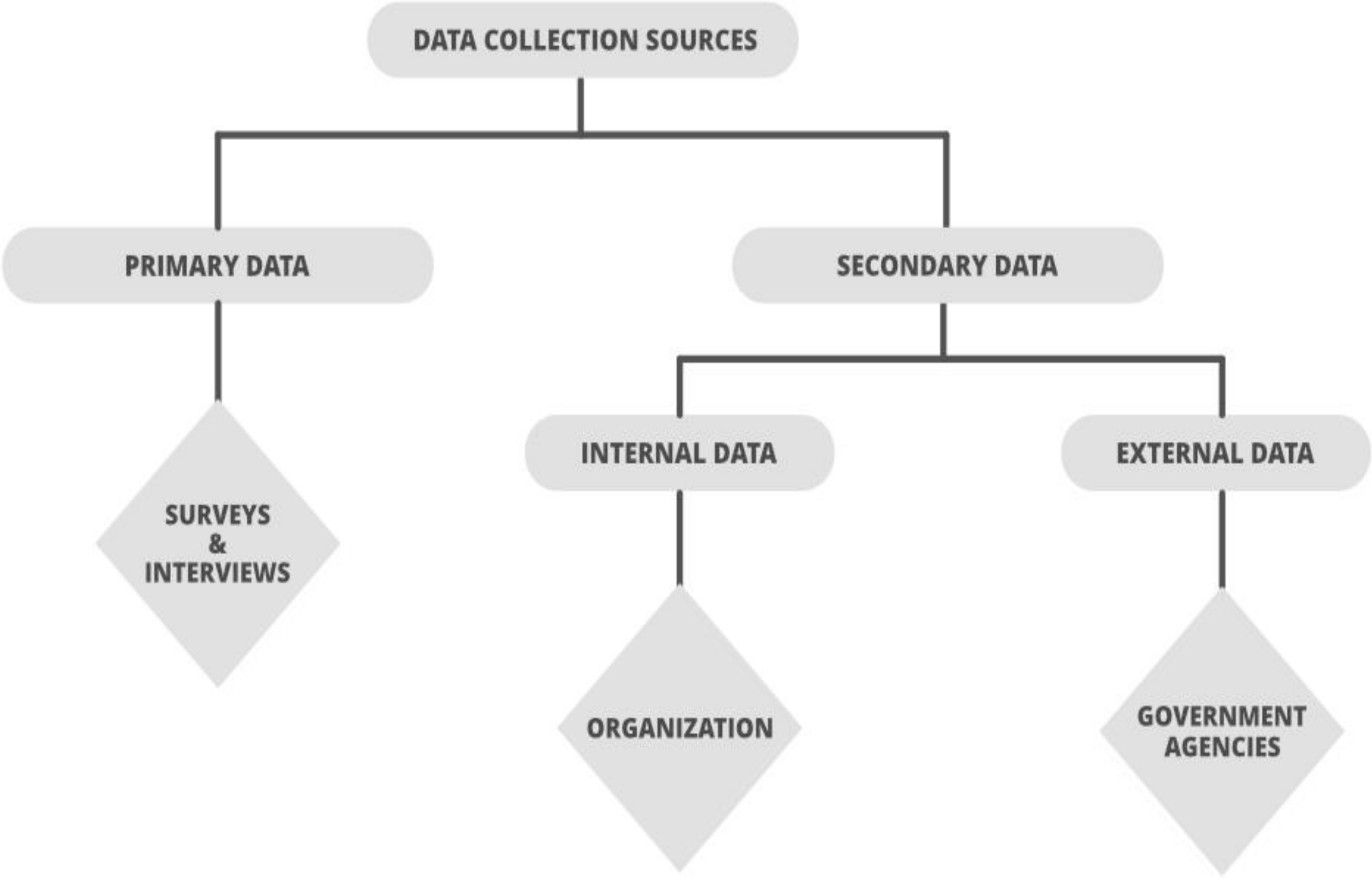
Interval Scale

It's defined as a numerical scale where the order of the variables as well as the difference between these variables is known.

Ratio Scale

It's a variable measurement scale that not only produces the order of the variables, but also makes the difference between the known variables along with information about the value of the true zero.

Data Collection Sources



Sources of Data

- Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis. In the process of big data analysis, “Data collection” is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.
- Data collection starts with asking some questions such as what type of data is to be collected and what the source of collection is. Most of the data collected are of two types known as “qualitative data“ which is a group of non-numerical data such as words, sentences mostly focus on behavior and actions of the group and another one is “quantitative data” which is in numerical forms and can be calculated using different scientific tools and sampling data.

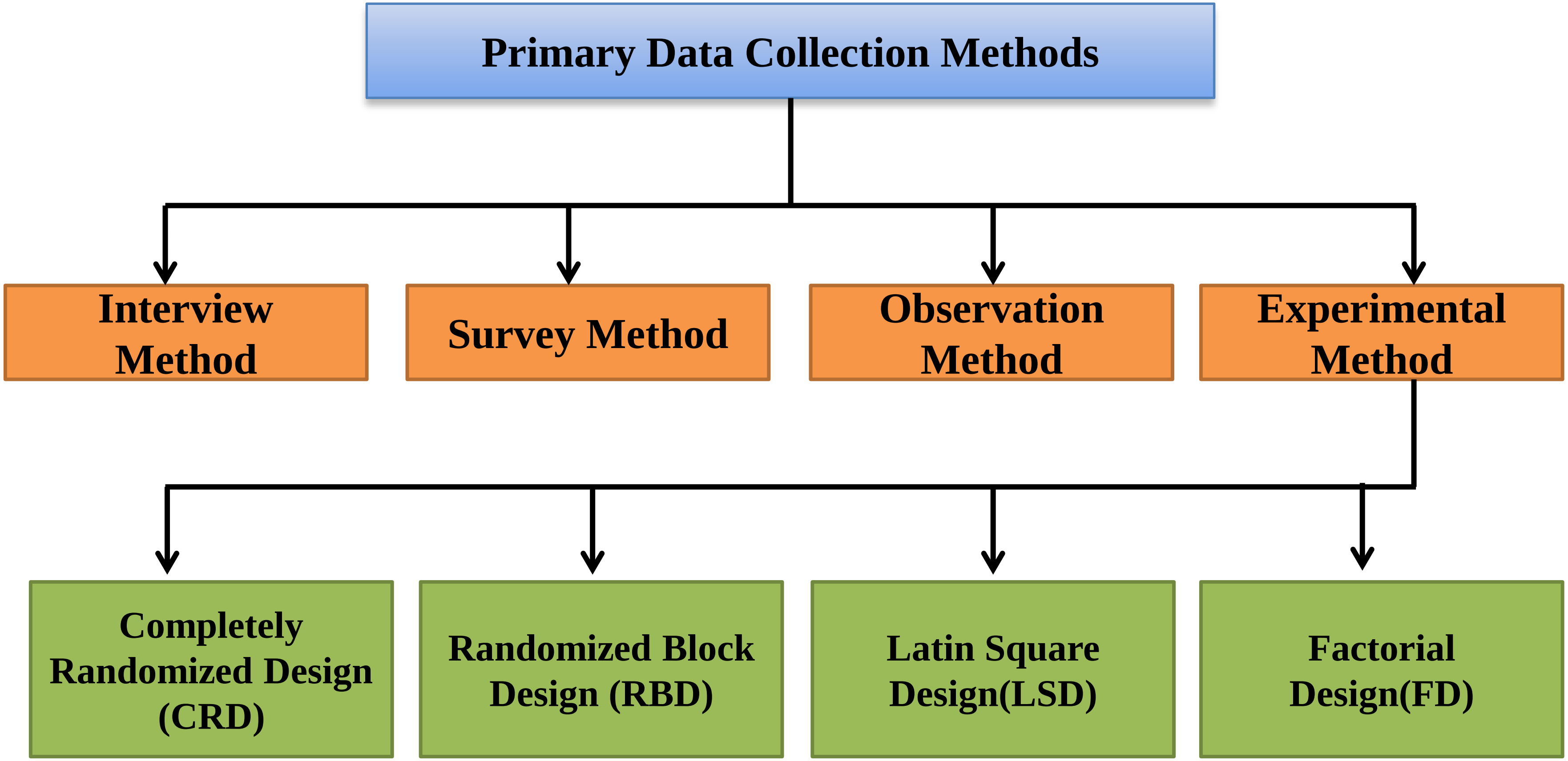
Primary & Secondary Data

- **Primary Data**
  - Primary data is the data that is collected for the first time through personal experiences or evidence, particularly for research. It is also described as raw data or first-hand information. The mode of assembling the information is costly, as the analysis is done by an agency or an external organisation, and needs human resources and investment. The investigator supervises and controls the data collection process directly. The data is mostly collected through observations, physical testing, mailed questionnaires, surveys, personal interviews, telephonic interviews, case studies, and focus groups, etc.
- **Secondary Data**
  - Secondary data is a second-hand data that is already collected and recorded by some researchers for their purpose, and not for the current research problem. It is accessible in the form of data collected from different sources such as government publications, censuses, internal records of the organisation, books, journal articles, websites and reports, etc. This method of gathering data is affordable, readily available, and saves cost and time. However, the one disadvantage is that the information assembled is for some other purpose and may not meet the present research purpose or may not be accurate.

Primary & Secondary Data

Primary Data	Secondary Data
Definition	
Primary data are those that are collected for the first time.	Secondary data refer to those data that have already been collected by some other person.
Originality	
These are original because these are collected by the investigator for the first time.	These are not original because someone else has collected these for his own purpose.
Nature of Data	
These are in the form of raw materials.	These are in the finished form.
Reliability and Suitability	
These are more reliable and suitable for the enquiry because these are collected for a particular purpose.	These are less reliable and less suitable as someone else has collected the data which may not perfectly match our purpose.
Time and Money	
Collecting primary data is quite expensive both in the terms of time and money.	Secondary data requires less time and money; hence it is economical.
Precaution and Editing	
No particular precaution or editing is required while using the primary data as these were collected with a definite purpose.	Both precaution and editing are essential as secondary data were collected by someone else for his own purpose.

Methods of Collecting Primary Data



Methods of Collecting Primary Data

1. Interview method:

- The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

- The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

Methods of Collecting Primary Data

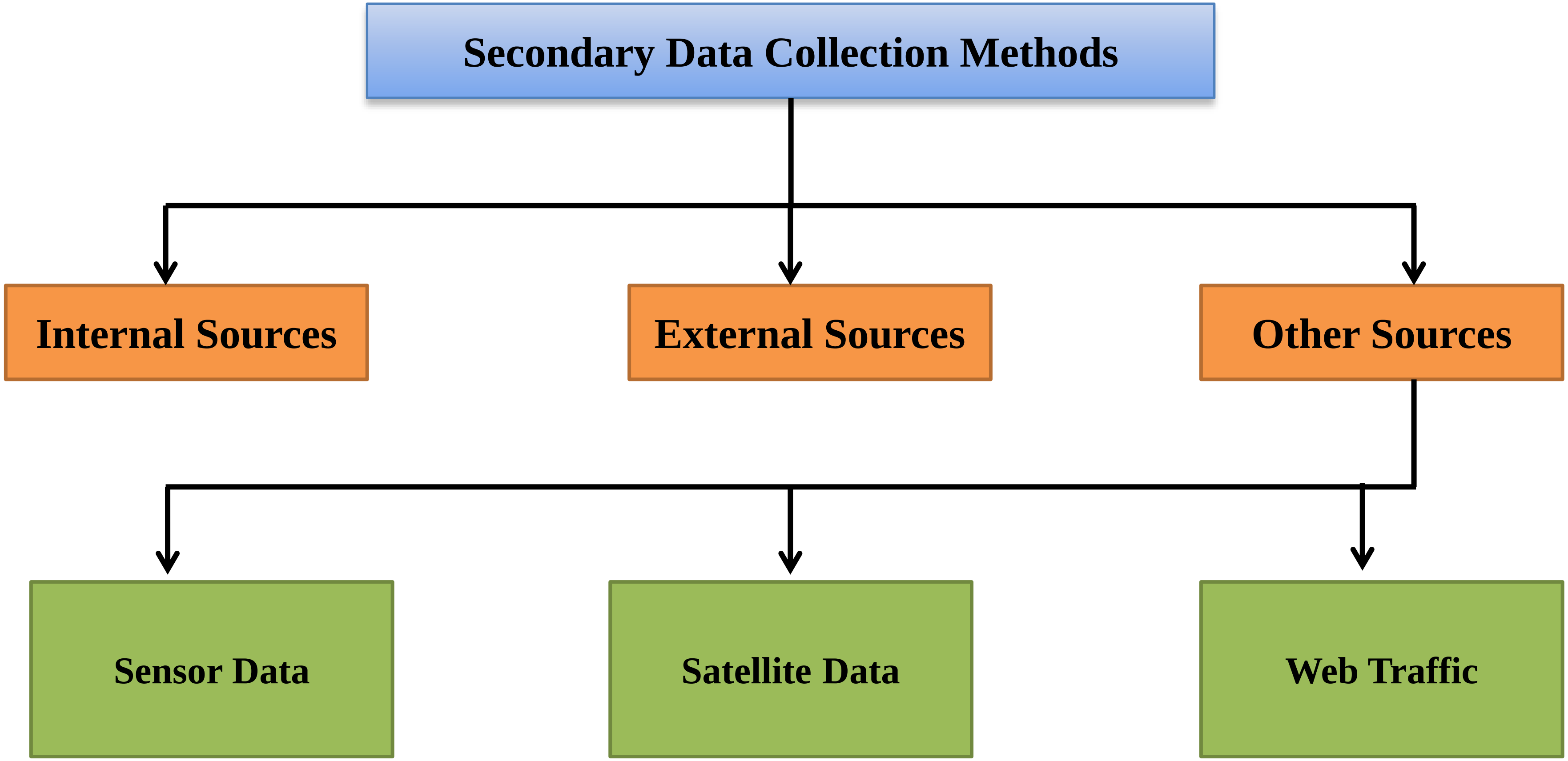
3. Observation method:

- The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behavior towards the products. The data obtained will be sent for processing.
- **4. Experimental method:**
- The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.
- **CRD- Completely Randomized design** is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.

Methods of Collecting Primary Data

- **RBD- Randomized Block Design** is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.
- **LSD – Latin Square Design** is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment.
- **FD- Factorial design** is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

Methods of Collecting Secondary Data



Methods of Collecting Secondary Data

- Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

1. Internal source:

- These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

2. External source:

- The data which can’t be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

Methods of Collecting Secondary Data

3. Other sources:

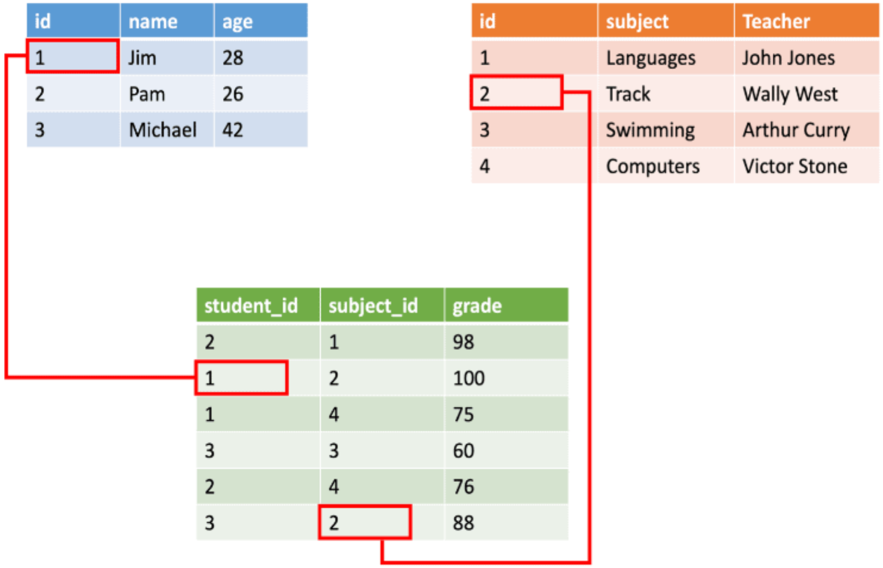
- **Sensors data:** With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data:** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
- **Web traffic:** Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

Classification of Data

- We can classify data as **structured data, semi-structured data, or unstructured data**. **Structured data** resides in predefined formats and models, **Unstructured data** is stored in its natural format until it's extracted for analysis, and **Semi-structured** data basically is a mix of both structured and unstructured data.

Structured Data

- **Structured data** is generally tabular data that is represented by columns and rows in a database.
- Databases that hold tables in this form are called ***relational databases***.
- The mathematical term "*relation*" specify to a formed set of data held as a table.
- In structured data, all row in a table has the same set of columns.
- SQL (Structured Query Language) programming language used for structured data.



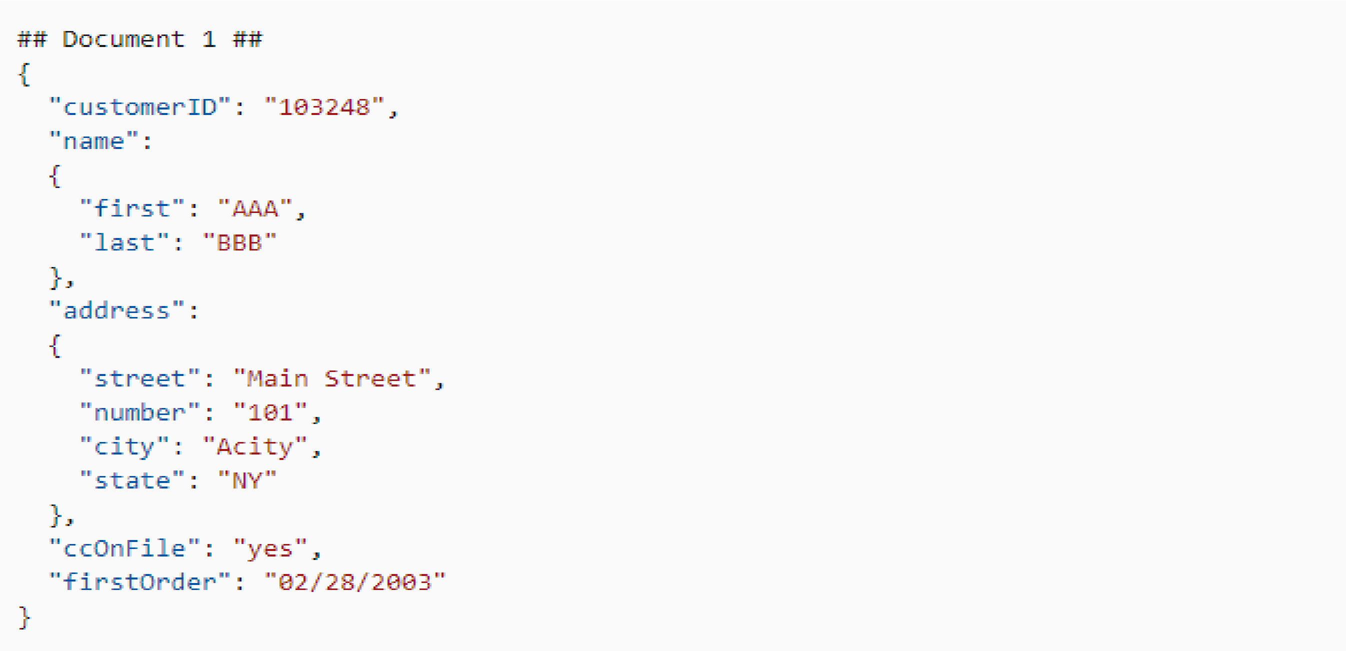
id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88

Semi-Structured Data

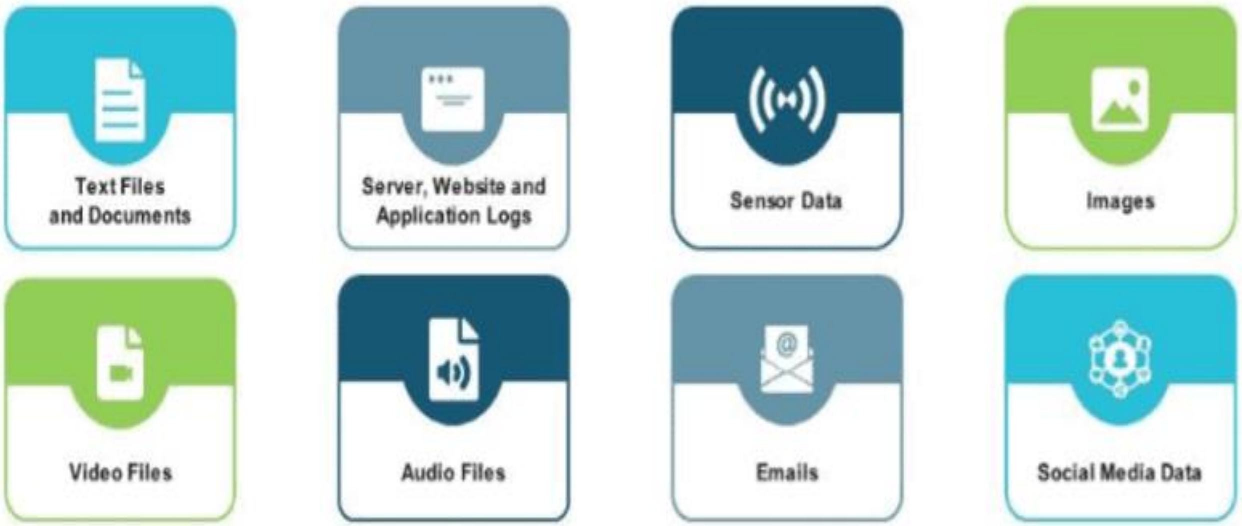
- **Semi-structured** data is information that doesn’t consist of Structured data (relational database) but still has some structure to it.
- Semi-structured data consist of documents held in ***JavaScript Object Notation* (JSON) format**. It also includes ***key-value*** stores and ***graph*** databases.

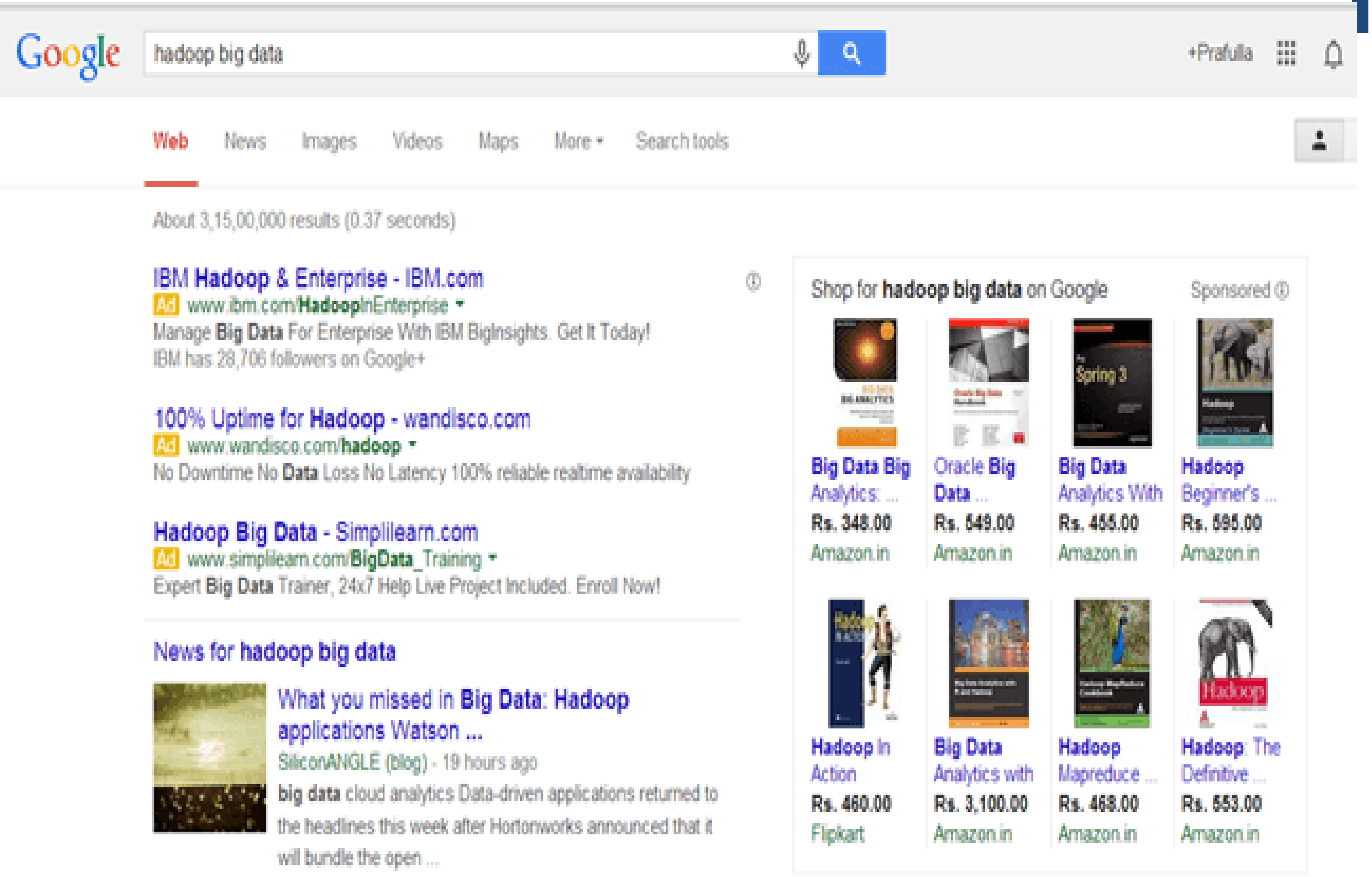


```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Unstructured Data

- **Unstructured data** is information that either does not organize in a pre-defined manner or not have a pre-defined data model.
- Unstructured information is a set of text-heavy but may contain data such as numbers, dates, and facts as well.
- **Videos, audio, and binary** data files might not have a specific structure. They're assigned to as **unstructured** data.





About 3,15,00,000 results (0.37 seconds)

IBM Hadoop & Enterprise - IBM.com
Ad www.ibm.com/**Hadoop**InEnterprise
Manage **Big Data** For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
Ad www.wandisco.com/**hadoop**
No Downtime No **Data** Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/**BigData**_Training
Expert **Big Data** Trainer, 24x7 Help Live Project Included. Enroll Now!

News for **hadoop big data**

What you missed in **Big Data**: **Hadoop** applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for **hadoop big data** on Google — Sponsored

Big Data Big Analytics: ...
Rs. 348.00
Amazon.in

Oracle **Big Data** ...
Rs. 549.00
Amazon.in

Big Data Analytics With
Rs. 455.00
Amazon.in

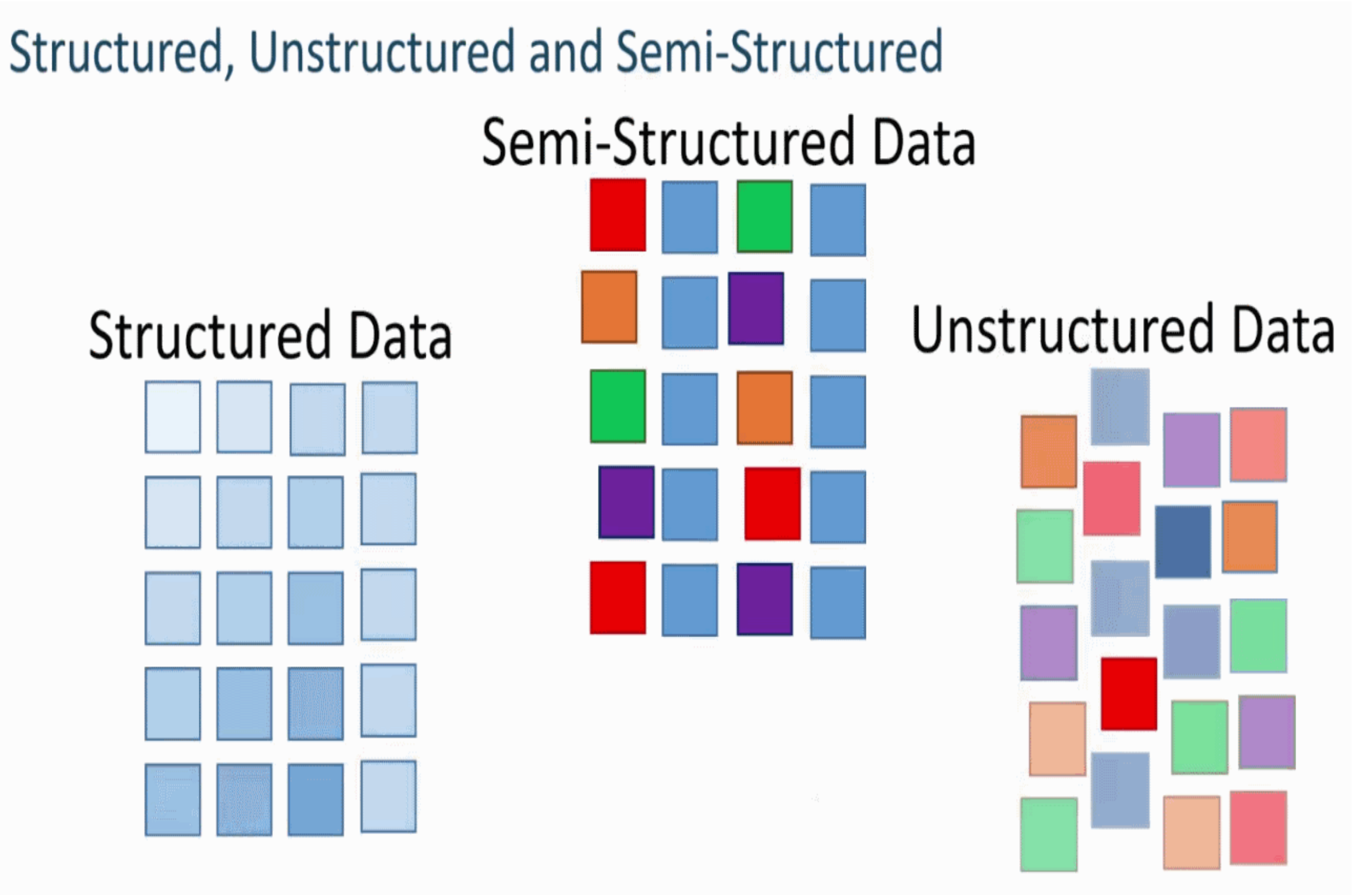
Hadoop Beginner's ...
Rs. 595.00
Amazon.in

Hadoop In Action
Rs. 460.00
Flipkart

Big Data Analytics with
Rs. 3,100.00
Amazon.in

Hadoop Mapreduce ...
Rs. 468.00
Amazon.in

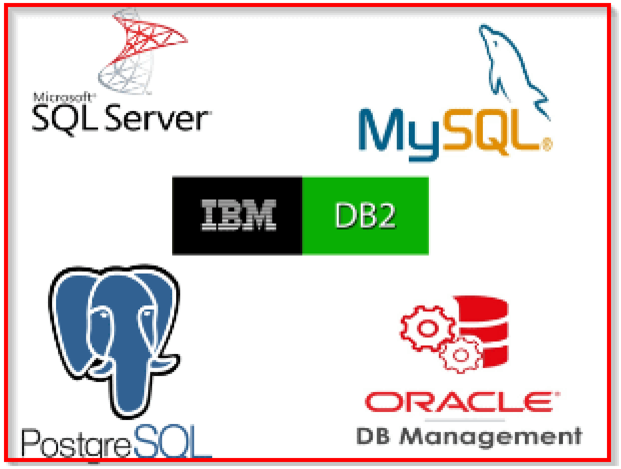
Hadoop: The Definitive ...
Rs. 553.00
Amazon.in



Structured, Unstructured and Semi-Structured

Characteristics Of Structured (Relational) Data

- Relational databases provide undoubtedly the well-understood model for holding data.
- The simplest structure of columns and tables makes them very easy to use initially, but the inflexible structure can cause some problems.
- We can communicate with relational databases using **Structured Query Language (SQL).**
- SQL allows the joining of tables using a few lines of code, with a structure most beginner employees can learn very fast.
- Examples of relational databases:
  - MySQL
  - PostgreSQL
  - Db2



Characteristics Unstructured (Non-Relational) Data

- Non-relational databases permit us to store data in a format that more closely meets the original structure.
- A ***non-relational database*** is a database that does not use the tabular schema of columns and rows found in most traditional database systems.
- It uses a storage model that is enhanced for the specific requirements of the type of data being stored.
- In a non-relational database the data may be stored as **JSON documents**, as simple **key/value pairs**, or as a graph consisting of edges and vertices.
- Examples of non-relational databases:
  - Redis
  - JanusGraph
  - MongoDB
  - RabbitMQ

Amazon DocumentDB
APACHE HBASE
mongoDB
cassandra
redis

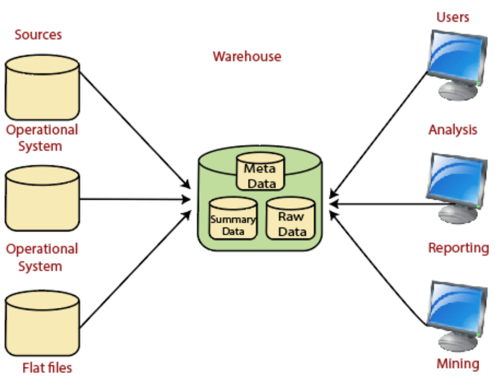
Structured Vs Unstructured Data

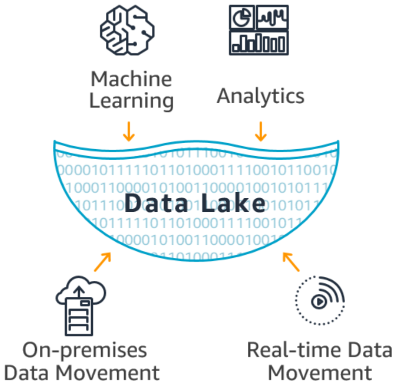
- **1) <u>Defined Vs Undefined Data</u>**
- Structured data is undoubtedly a defined type of data in a structure.
- Structured data lives in columns and rows and it can be mapped into pre-defined fields.
- Unstructured data does not have a predefined data format.
- **2) <u>Quantitative Vs Qualitative Data</u>**
- Structured data is generally quantitative data, it usually consists of hard numbers or things that can be counted.
- Methods for analysis include **classification**, **regression**, and **clustering** of data.
- Unstructured data is generally categorized as qualitative data, and cannot be analyzed and processed using conventional tools and methods.
- Understanding qualitative data requires advanced analytics techniques like **data stacking** and **data mining**.

Structured Vs Unstructured Data

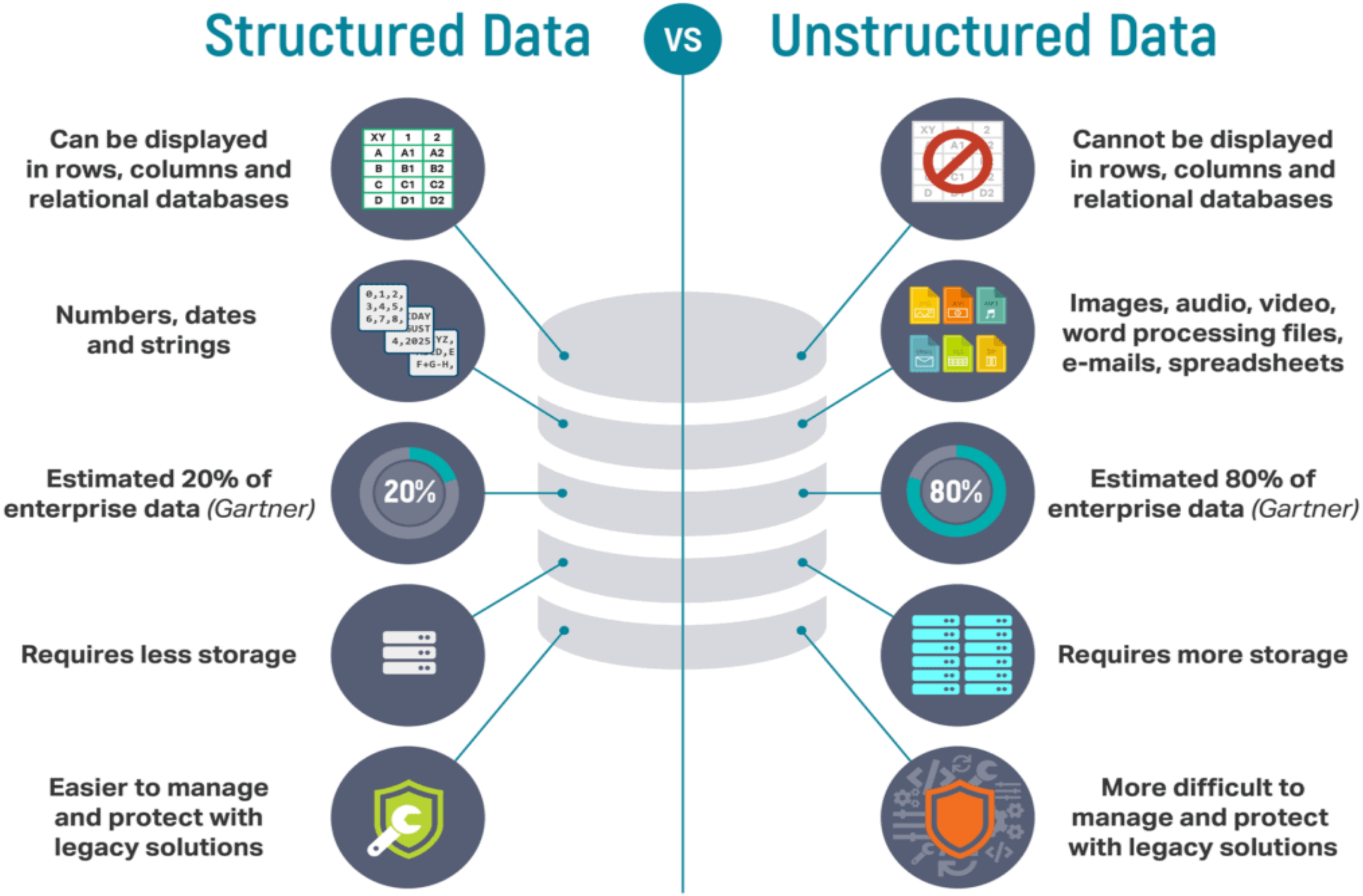
- **3) <u>Storage in Data Lakes Vs Data Houses</u>**
- Structured data is generally stored in **data warehouses**.
- Unstructured data is stored in data lakes. A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semi-structured, and unstructured data. It can store data in its native format and process any variety of it, ignoring size limits.
- Unstructured data requires **more storage** space, while structured data requires **less storage** space.

Architecture of a Data Warehouse





- **4) <u>Ease of Analysis</u>**
- Structured data is **easy** to search, both for algorithms and for humans.
- Unstructured data is more **difficult** to search and requires processing to become understandable.



Structured Data vs Unstructured Data

Structured Data

Can be displayed in rows, columns and relational databases

Numbers, dates and strings

Estimated 20% of enterprise data *(Gartner)*

Requires less storage

Easier to manage and protect with legacy solutions

Unstructured Data

Cannot be displayed in rows, columns and relational databases

Images, audio, video, word processing files, e-mails, spreadsheets

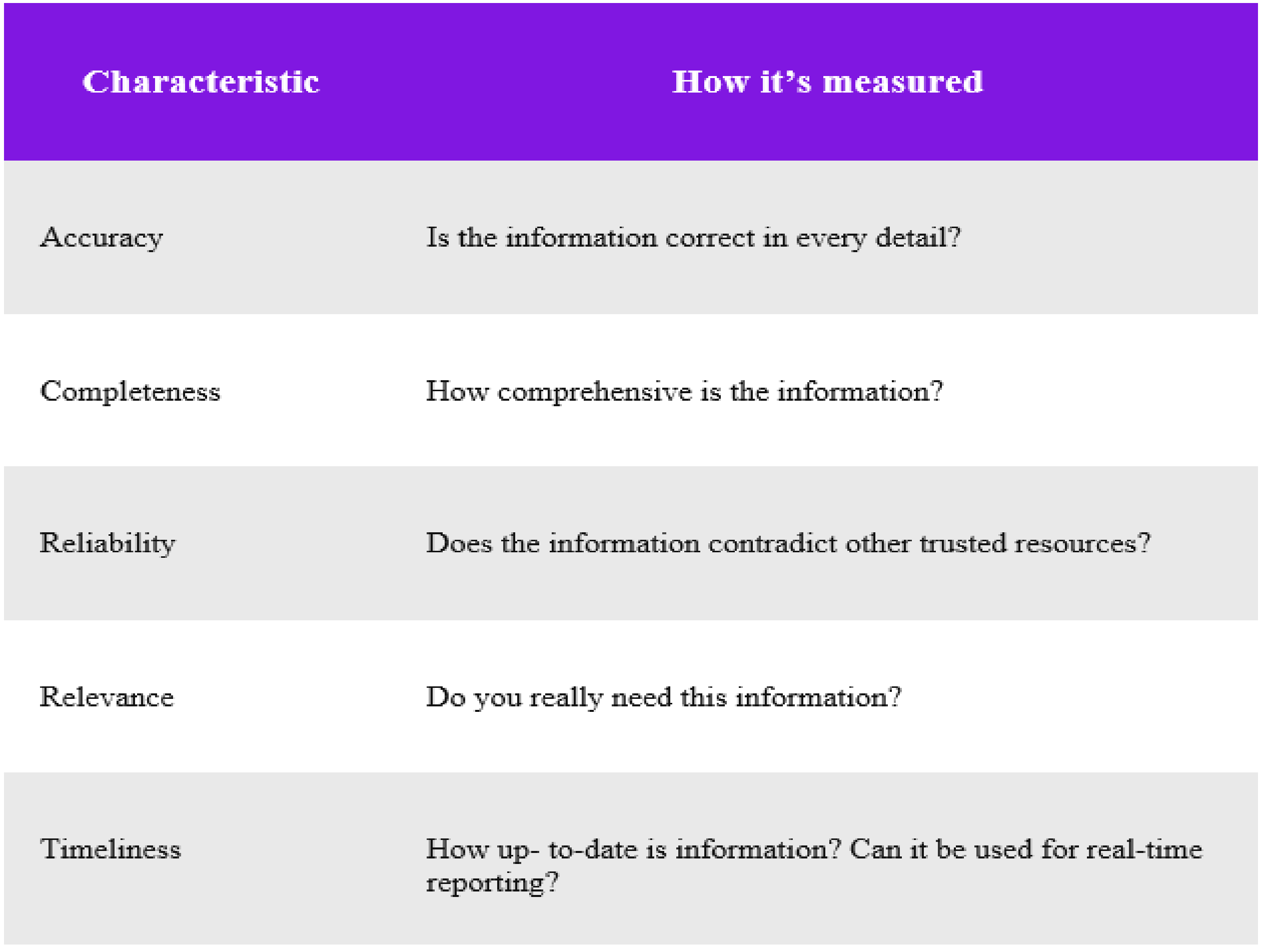
Estimated 80% of enterprise data *(Gartner)*

Requires more storage

More difficult to manage and protect with legacy solutions

Characteristics of Data

- **Data quality is crucial** – It assesses whether information can serve its purpose in a particular context (such as data analysis, for example). So, how do you determine the quality of a given set of information? There are data quality characteristics of which you should be aware.
- **There are five traits that you’ll find within data quality:**
- Accuracy
- Completeness
- Reliability
- Relevance
- Timeliness



Characteristic	How it’s measured
Accuracy	Is the information correct in every detail?
Completeness	How comprehensive is the information?
Reliability	Does the information contradict other trusted resources?
Relevance	Do you really need this information?
Timeliness	How up- to-date is information? Can it be used for real-time reporting?

Characteristics of Data

Accuracy

- As the name implies, this data quality characteristic means that information is correct. To determine whether data is accurate or not, ask yourself if the information reflects a real-world situation. For example, in the realm of financial services, does a customer really have $1 million in his bank account?
- Accuracy is a crucial data quality characteristic because inaccurate information can cause significant problems with severe consequences. We'll use the example above – if there's an error in a customer's bank account, it could be because someone accessed it without his knowledge.

Completeness

- "Completeness" refers to how comprehensive the information is. When looking at data completeness, think about whether all of the data you need is available; you might need a customer's first and last name, but the middle initial may be optional.
- Why does completeness matter as a data quality characteristic? If information is incomplete, it might be unusable. Let's say you're sending a mailing out. You need a customer's last name to ensure the mail goes to the right address – without it, the data is incomplete.

Characteristics of Data

Reliability

- In the realm of data quality characteristics, reliability means that a piece of information doesn't contradict another piece of information in a different source or system. We'll use an example from the healthcare field; if a patient's birthday is January 1, 1970 in one system, yet it's June 13, 1973 in another, the information is unreliable.
- Reliability is a vital data quality characteristic. When pieces of information contradict themselves, you can't trust the data. You could make a mistake that could cost your firm money and reputational damage.

Relevance

- When you're looking at data quality characteristics, relevance comes into play because there has to be a good reason as to why you're collecting this information in the first place. You must consider whether you really need this information, or whether you're collecting it just for the sake of it.
- Why does relevance matter as a data quality characteristic? If you're gathering irrelevant information, you're wasting time as well as money. Your analyses won't be as valuable.

Characteristics of Data

Timeliness

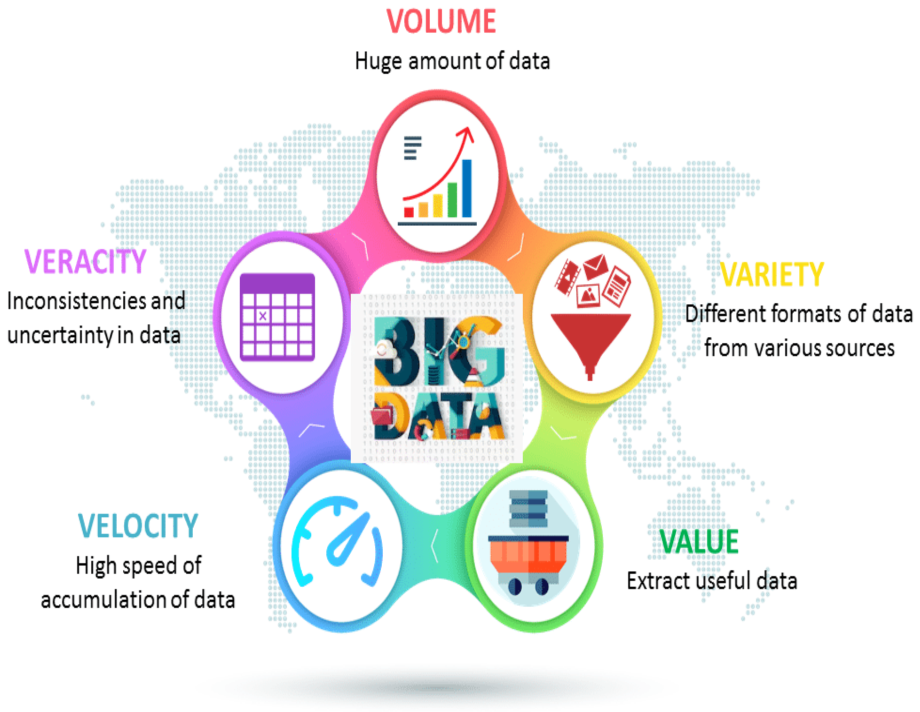
- Timeliness, as the name implies, refers to how up to date information is. If it was gathered in the past hour, then it’s timely – unless new information has come in that renders previous information useless.
- The timeliness of information is an important data quality characteristic, because information that isn’t timely can lead to people making the wrong decisions. In turn, that costs organizations time, money, and reputational damage.
- “Timeliness is an important data quality characteristic – out-of-date information costs companies time and money”

What is Big Data?

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.
- **Examples of Big Data**
- Following are some of the Big Data examples-
- **The New York Stock Exchange** is an example of Big Data that generates about one terabyte of new trade data per day.
- **Social Media:** The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.
- **Types of Big Data**
- Following are the types of Big Data:
- Structured
- Unstructured
- Semi-structured

5 V's of Big Data

- In recent years, Big Data was defined by the "*3Vs*" but now there is "*5Vs*" of Big Data which are also termed as the characteristics of Big Data as follows:



VOLUME
Huge amount of data

VARIETY
Different formats of data from various sources

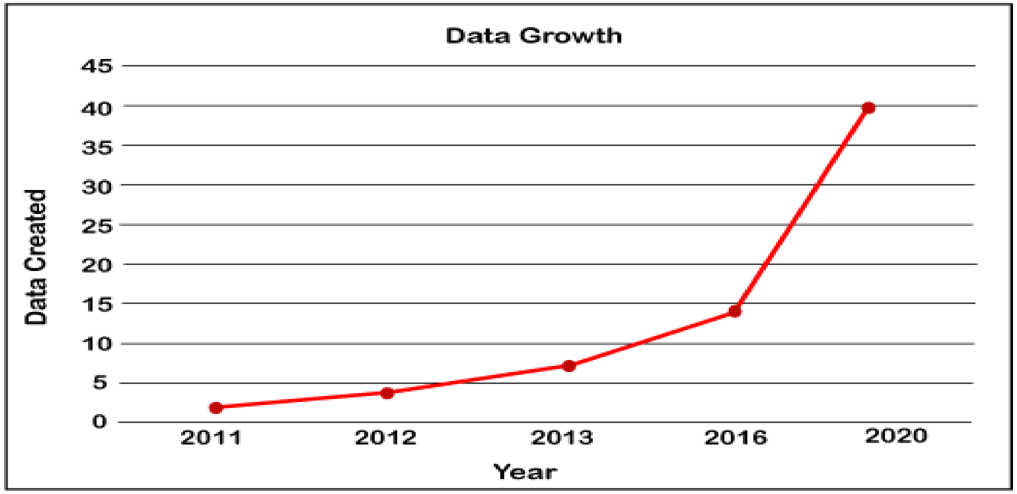
VALUE
Extract useful data

VELOCITY
High speed of accumulation of data

VERACITY
Inconsistencies and uncertainty in data

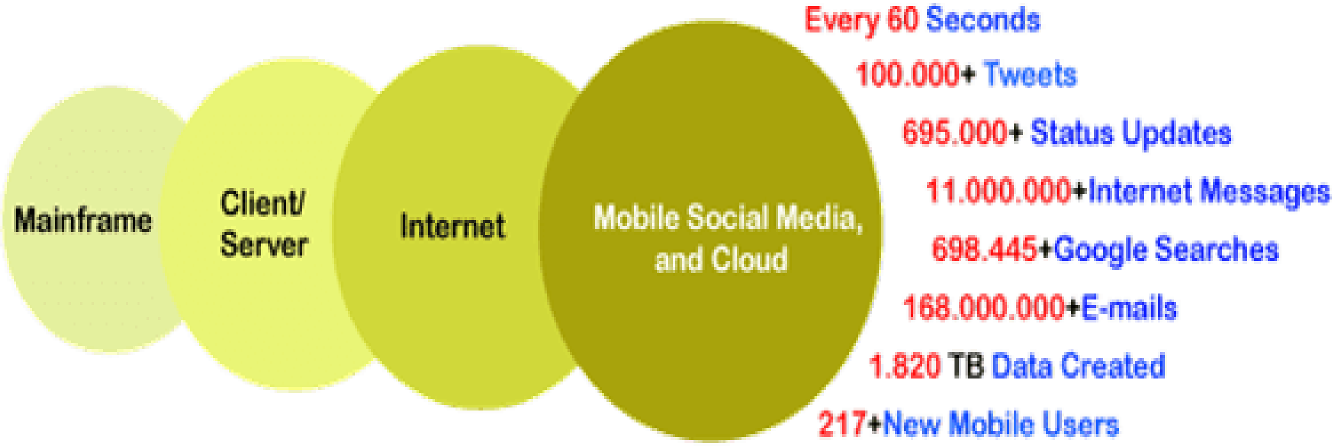
5 V's of Big Data

- **1. <u>Volume:</u>**
- 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'.
- *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabyte (6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 Exabyte of data.



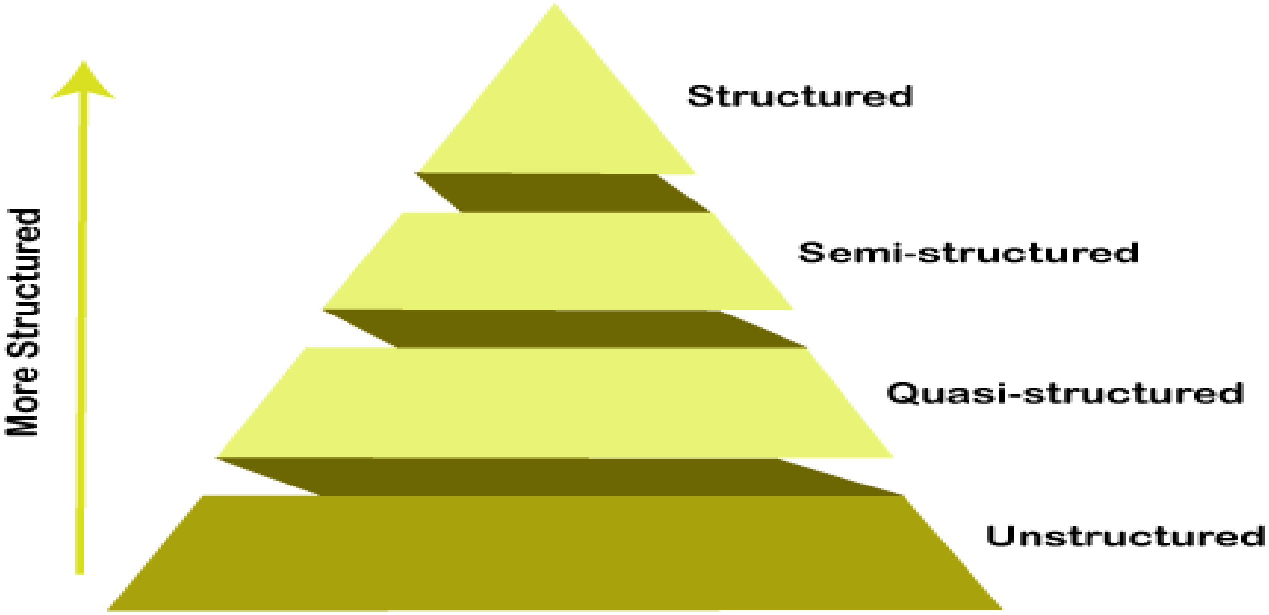
5 V's of Big Data

- **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.



5 V’s of Big Data

- **3. <u>Variety:</u>**
- It refers to 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: In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

Semi-structured: In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.

Unstructured Data: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.

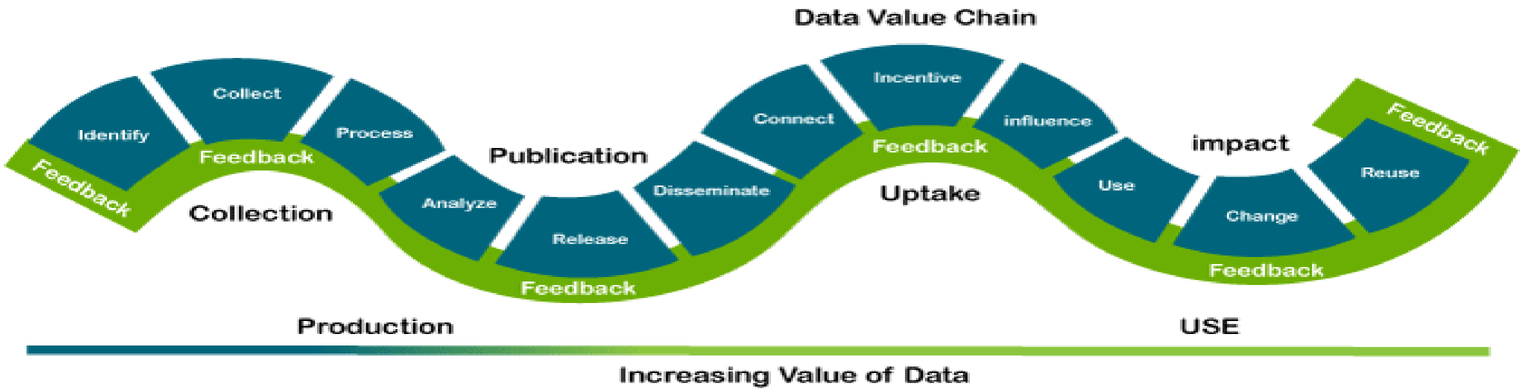
Quasi-structured Data: The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

5 V’s of Big Data

- **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.

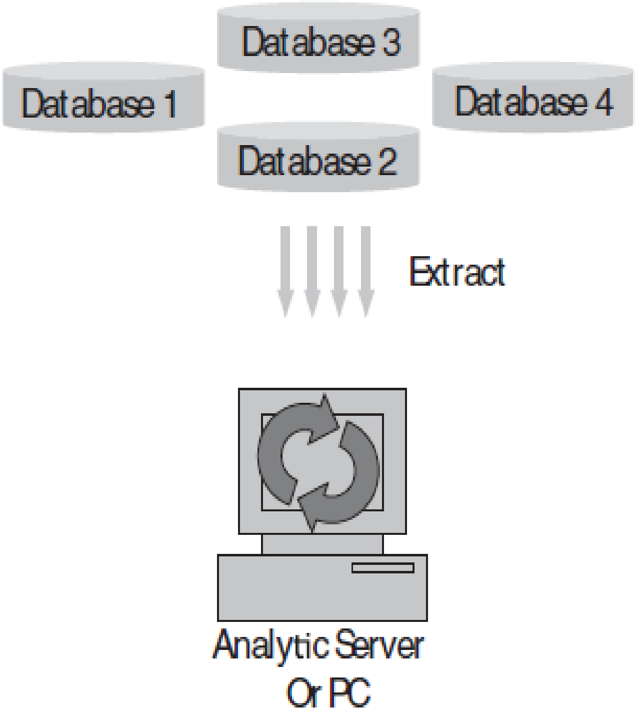
5 V's of Big Data

- 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 5V's.



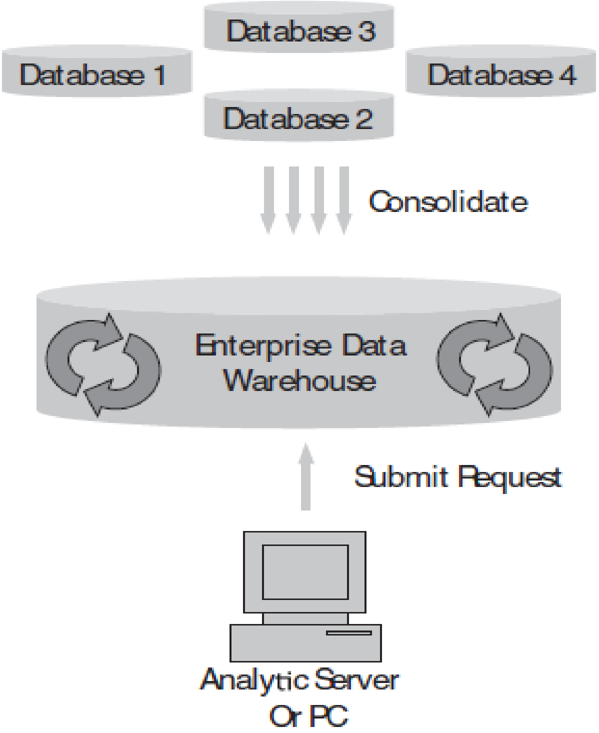
Evolution of Analytic Scalability

- Scalability: The ability of system, to handle increasing amount of work required to perform its task.
- It goes without saying that the world of big data requires new levels of scalability. As the amount of data organizations process continues to increase, the same old methods for handling data just won’t work anymore. Organizations that don’t update their technologies to provide a higher level of scalability will quite simply choke on big data. Luckily, there are multiple technologies available that address different aspects of the process of taming big data and making use of it in analytic processes. Some of these advances are quite new, and organizations need to keep up with the times.



In traditional architectures, the heavy processing occurs in the analytic environment. This may even be a PC!

Traditional Analytical Architecture



In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

Modern Analytical Architecture

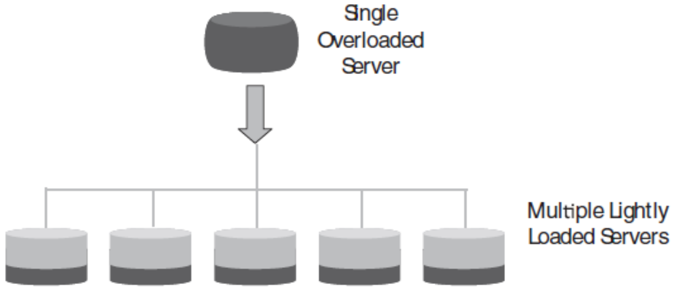
Technologies to address Big Data

- The world of big data requires new level of scalability. There has to be a convergence of the analytics and data environment.
- Massive Parallel Processing (MPP)
- The Cloud
- Grid Computing
- Map Reduce Processing

- Initially, databases were built for each specific purpose or team, and relational databases were spread all over an organization. Such single - purpose databases are often called “data marts”. While many organizations still leverage data marts heavily, leading organizations now see value in combining the various database systems into one big system called an Enterprise Data Warehouse (EDW).
- With an EDW, the goal is to get all corporate data that has importance together in one central database that has a single version of the truth.

MASSIVELY PARALLEL PROCESSING SYSTEMS

- Massively parallel processing (MPP) database systems have been around for decades. While individual vendor architectures may vary, MPP is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data.
- An MPP database spreads data out into independent pieces managed by independent storage and central processing unit (CPU) resources. Conceptually, it is like having pieces of data loaded onto multiple network connected personal computers around a house. It removes the constraints of having one central server with only a single set CPU and disk to manage it. The data in an MPP system gets split across a variety of disks managed by a variety of CPUs spread across a number of servers.



Instead of a single overloaded database, an MPP database breaks the data into independent chunks with independent disk and CPU.

Figure 4.3 Massively Parallel Processing System Data Storage

CLOUD COMPUTING

The concept of cloud computing is getting a lot of attention these days. As with many technologies, cloud computing is going through a hype cycle. Let’s start by defining what cloud computing is all about and how it can help with advanced analytics and big data. As with any new and emerging technology, there are conflicting definitions of what cloud computing is. We’ll consider two that serve as a good foundation for our discussion.

- They list five essential characteristics of a cloud environment.
- 1. On - demand self - service
- 2. Broad network access
- 3. Resource pooling
- 4. Rapid elasticity
- 5. Measured service

Characteristics of a Cloud Computing

- **On-demand self-services:**

The Cloud computing services does not require any human administrators, user themselves are able to provision, monitor and manage computing resources as needed.

- **Broad network access:**

The Computing services are generally provided over standard networks and heterogeneous devices.

- **Rapid elasticity:**

The Computing services should have IT resources that are able to scale out and in quickly and on as needed basis. Whenever the user require services it is provided to him and it is scale out as soon as its requirement gets over.

- **Resource pooling:**

The IT resource (e.g., networks, servers, storage, applications, and services) present are shared across multiple applications and occupant in an uncommitted manner. Multiple clients are provided service from a same physical resource.

- **Measured service:**

The resource utilization is tracked for each application and occupant, it will provide both the user and the resource provider with an account of what has been used. This is done for various reasons like monitoring billing and effective use of resource.

Types of Cloud Computing

- The two primary types of cloud environments: (1) public clouds and (2) private clouds.
- **Public Clouds**
- Public clouds have gotten the most hype and attention. With a public cloud users are basically loading their data onto a host system and they are then allocated resources as they need them to use that data.
- They will get charged according to their usage. There are definitely some advantages to such a setup:
- The bandwidth is as - needed and users only pay for what they use.
- It isn’t necessary to buy a system sized to handle the maximum capacity ever required and then risk having half of the capacity sitting idle much of the time.
- If there are short bursts where a lot of processing is needed then it is possible to get it with no hassle. Simply pay for the extra resources.
- There’s typically very fast ramp - up. Once granted access to the cloud environment, users load their data and start analyzing.
- It is easy to share data with others regardless of their location since a public cloud by definition is outside of a corporate firewall. Anyone can be given permission to log on to the environment created.

CLOUD COMPUTING

- **Private Clouds**
- A private cloud has the same features of a public cloud, but it’s owned exclusively by one organization and typically housed behind a corporate firewall. A private cloud is going to serve the exact same function as a public cloud, but just for the people or teams within a given organization.
- One huge advantage of an onsite private cloud is that the organization will have complete control over the data and system security.
- Data is never leaving the corporate firewall so there’s absolutely no concern about where it ’ s going.

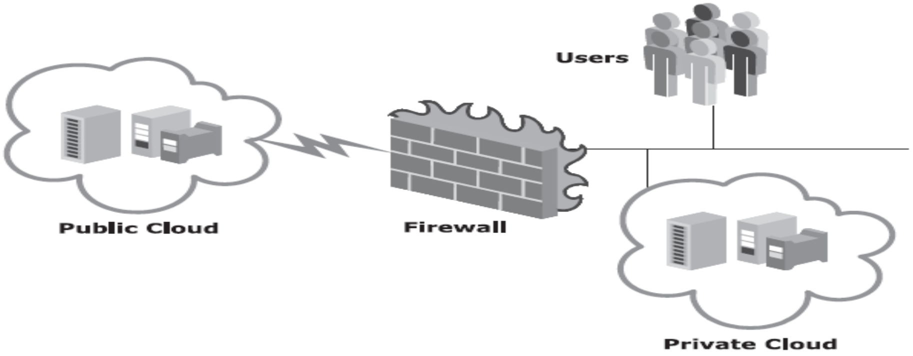
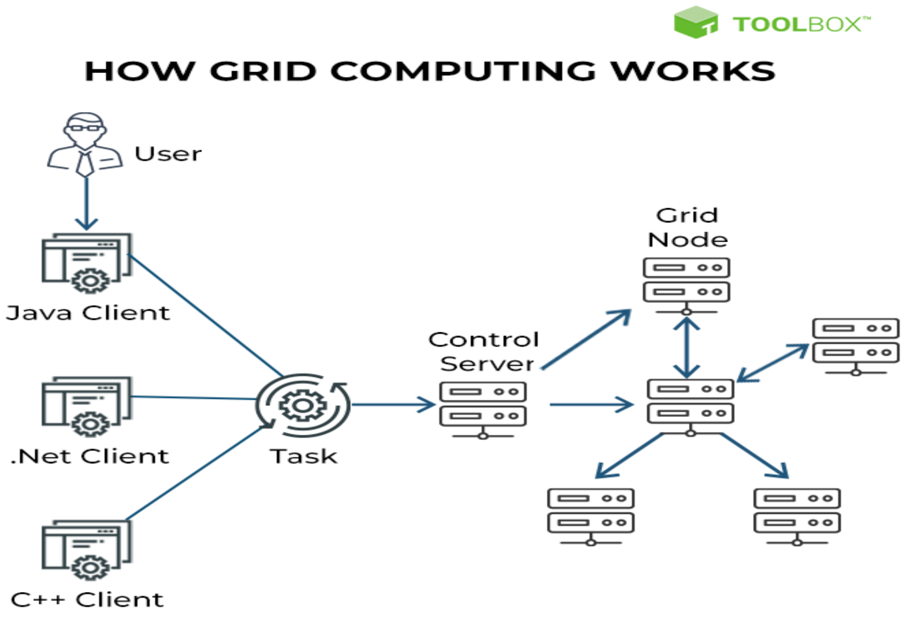


Figure 4.5 Public Clouds versus Private Clouds

GRID COMPUTING

- Grid computing is a distributed architecture of multiple computers connected by networks to accomplish a joint task. These tasks are compute-intensive and difficult for a single machine to handle. Several machines on a network collaborate under a common protocol and work as a single virtual supercomputer to get complex tasks done. This offers powerful virtualization by creating a single system image that grants users and applications seamless access to IT capabilities.



Need of Data Analytics

1. Product Development

- Data analytics offers both estimation and exploration capability for information. It allows one to understand the market or process's current state and offers a solid base for forecasting future results. Data analysis helps companies to comprehend the current business situation and change the processor cause the need for a new product creation that meets market requirements.

2. Target Content

- Learning what consumers wish in advance improves consumer orientation in marketing campaigns. It encourages advertisers to tailor their advertising to a subset of the entire consumer base. It also allows you to figure out which client base group can better respond to the initiative. It also saves money to convince a buyer to buy and increases the overall performance of the marketing activities.

3. Efficiency in Operations

- The importance of data analytics in marketing finds more viable ways to streamline operations or increase benefit levels. It helps to recognize possible issues, avoids the waiting period, and takes action on them.

Why is Data Analytics Growing

- **1. Analysis of the business value chain**
- Any business will allow you to get insights into the value chains that are currently in your business and are achieved through data mining. The analytics would thus clarify how the current knowledge will allow the company to locate the gold mine that is the path to a company's growth.
- **2. Industry Knowledge**
- Industry awareness is another thing you can understand once you have evaluated results, it can reveal your company in the near future and what is the economy's strength now. This is because before anyone else you can profit.
- **3. Opportunities**
- Since the economy tends to change to keep track of dynamic developments, while benefit generation is one that is most commonly pursued by the enterprise, the Data Analytics provides evaluated data, which enable them to look before the time, for more alternatives.

Analytic Process & Tool

- **Ensemble Methods**
  - Multiple models are built using multiple techniques. Once the results from all of the models are known, all of the results are combined together to come up with a final answer.
  - The process of combining the various results can be anything from a simple average of each model ’ s predictions to a much more complex formula.
  - It is important to note that ensemble models go beyond picking the best individual performer from a set of models.
  - They actually combine the results of multiple models in order to get to a single, final answer
  - The power of ensemble models stems from the fact that different techniques have different strengths and weaknesses

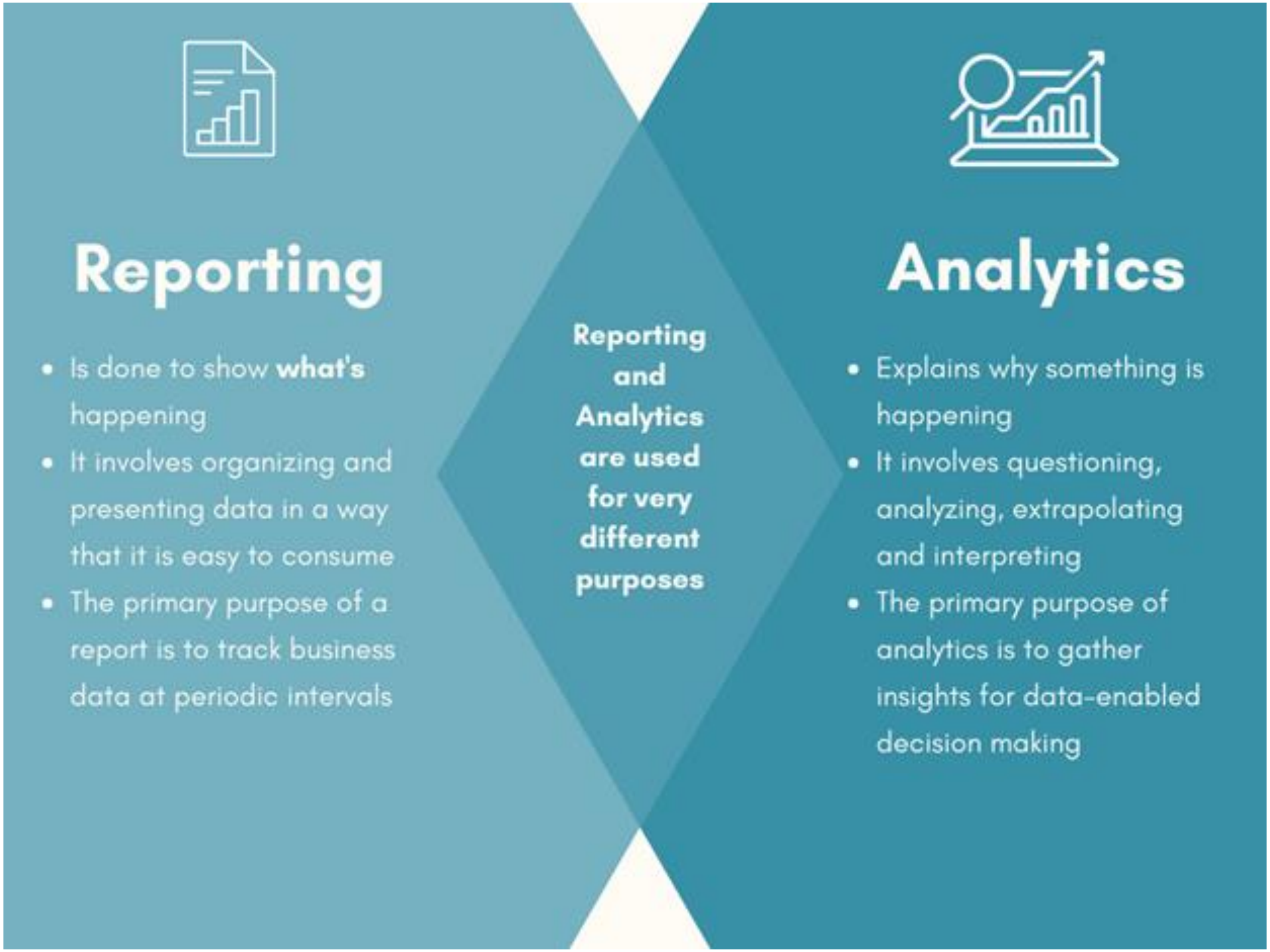
Analytic Process & Tool

- **Commodity Models**
- The goal of a commodity model is not to get the best model, but to quickly get a model that will lead to a better result than if there had been no model at all
- In evaluating a commodity model, the primary concern is that there ’ s a benefit being achieved by using it.
- There may be much room for improvement if more effort was put in.
- But, if a quick model can help in a situation that otherwise wouldn ’ t have a model, it is utilized.

Analytic Process & Tools

- **Text Analysis**
- One of the most rapidly growing methods utilized by organizations today is the analysis of text and other unstructured data sources.
- A lot of big data falls into these classifications.
- Text analysis, as the name implies, takes some sort of text as input.
- This text can be written material like an e - mail, transcribed material such as a medical dictation, or even text that has been scanned from a hard copy and converted to electric form like old courthouse records.
- The reason text analysis has grown in prominence is because of the wealth of new sources of text data
- Text is a very common type of big data and text analysis tools and methods have come a long way.
- Today there are tools that will help you parse text into its component words and phrases and then assist in determining the meaning of those words and phrases.

Analytic vs Reporting



Reporting

- Is done to show **what's** happening
- It involves organizing and presenting data in a way that it is easy to consume
- The primary purpose of a report is to track business data at periodic intervals

Reporting and Analytics are used for very different purposes

Analytics

- Explains why something is happening
- It involves questioning, analyzing, extrapolating and interpreting
- The primary purpose of analytics is to gather insights for data-enabled decision making

Analytic vs Reporting

- **1. Purpose:** Reporting involves extracting data from different sources within an organization and monitoring it to gain an understanding of the performance of the various functions. By linking data from across functions, it helps create a cross-channel view that facilitates comparison to understand data easily. An analysis is being able to interpret data at a deeper level, interpreting it and providing recommendations on actions.
- **2. The Specifics:** Reporting involves activities such as building, consolidating, organizing, configuring, formatting, and summarizing. It requires clean, raw data and reports that may be generated periodically, such as daily, weekly, monthly, quarterly, and yearly. Analytics includes asking questions, examining, comparing, interpreting, and confirming. Enriching data with big data can help predict future trends as well.

Analytic vs Reporting

- **3. The Final Output:** In the case of reporting, outputs such as canned reports, dashboards, and alerts push information to users. Through analysis, analysts try to extract answers using business queries and present them in the form of ad hoc responses, insights, recommended actions, or a forecast. Understanding this key difference can help businesses leverage analytics better.
- **4. People:** Reporting requires repetitive tasks that can be automated. It is often used by functional business heads who monitor specific business metrics. Analytics requires customization and therefore depends on data analysts and scientists. Also, it is used by business leaders to make data-driven decisions.
- **5. Value Proposition:** This is like comparing apples to oranges. Both reporting and analytics serve a different purpose. By understanding the purpose and using them correctly, businesses can derive immense value from both.

Popular Data Analytic Tools

- **1. TIBCO Spotfire**
- is a data analytics platform that provides natural language search and AI-powered data insights. It’s a comprehensive visualization tool that can publish reports to both mobile and desktop applications. Spotfire also provides point-and-click tools for building predictive analytics models.
- **2. Thoughtspot**
- is an analytics platform that allows users to explore data from various types of sources through reports and natural language searches. Its AI system, SpotIQ, finds insights automatically to help users uncover patterns they didn’t know to look for. The platform also allows users to automatically join tables from different data sources to help break down.
- **3. Qlik**
- provides a self-service data analytics and business intelligence platform that supports both cloud and on-premises deployment. The tool boasts strong support for data exploration and discovery by technical and nontechnical users alike. Qlik supports many types of charts that users can customize with both embedded SQL and drag-and-drop modules.

Popular Data Analytic Tools

- **4. SAS Business Intelligence**
- provides a suite of applications for self-service analytics. It has many built-in collaboration features, such as the ability to push reports to mobile applications. While SAS Business Intelligence is a comprehensive and flexible platform, it can be more expensive than some of its competitors. Larger enterprises may find it worth the price due to its versatility.
- **5. Tableau**
- is a data visualization and analytics platform that allows users to create reports and share them across desktop and mobile platforms, within a browser, or embedded in an application. It can run on the cloud or on-premises. Much of the Tableau platform runs on top of its core query language, VizQL. This translates drag-and-drop dashboard and visualization components into efficient back-end queries and minimizes the need for end-user performance optimizations. However, Tableau lacks support for advanced SQL queries.
- **6. Google Data Studio**
- is a free dash boarding and data visualization tool that automatically integrates with most other Google applications, such as **Google Analytics**, Google Ads, and **Google BigQuery**. Thanks to its integration with other Google services, Data Studio is great for those who need to analyze their Google data. For instance, marketers can build dashboards for their Google Ads and Analytics data to better understand customer conversion and retention. Data Studio can work with data from a variety of other sources as well, provided that the data is first replicated to BigQuery using a **data pipeline** like Stitch.

Popular Data Analytic Tools

- **7. Redash**
- is a lightweight and cost-effective tool for querying data sources and building visualizations. The code is open source, and an affordable hosted version is available for organizations that want to get started fast. The core of Redash is the query editor, which provides a simple interface for writing queries, exploring schemas, and managing integrations. Query results are cached within Redash and users can schedule updates to run automatically.
- **8. Periscope Data**
- now owned by Sisense — is a business intelligence platform that supports integrations for a variety of popular data warehouses and databases. Technical analysts can transform data using SQL, Python, or R, and less technical users can easily create and share dashboards. Periscope Data also boasts a number of security certifications, such as HIPAA-HITECH.
- **9. Metabase**
- is a free, open source analytics and business intelligence tool. Metabase allows users to "ask questions" about data, which is a way for nontechnical users to use a point-and-click interface for query construction. This works well for simple filtering and aggregations; more technical users can go straight to raw SQL for more complex analysis. Metabase also has the ability to push analytics results to external systems like Slack.

Popular Data Analytic Tools

- **10. Jupyter Notebook**
- is a free, open source web application that can be run in a browser or on desktop platforms after installation using the Anaconda platform or Python’s package manager, pip. It allows developers to create reports with data and visualizations from live code. The system supports more than 40 programming languages. Jupyter Notebook — formerly IPython Notebook — was originally programmed using Python, and allows developers to make use of the wide range of Python packages for analytics and visualizations. The tool has a wide developer community using other languages as well.
- **11. IBM Cognos**
- is a business intelligence platform that features built-in AI tools to reveal insights hidden in data and explain them in plain English. Cognos also has automated data preparation tools to automatically cleanse and aggregate data sources, which allows for quickly integrating and experimenting with data sources for analysis.
- **12. Chartio**
- is a self-service business intelligence system that integrates with various data warehouses and allows for easy import of files such as spreadsheets. Chartio has a unique visual representation of SQL that allows for point-and-click construction of queries, which lets business analysts who aren’t familiar with SQL syntax modify and experiment with queries without having to dig into the language.

Popular Data Analytic Tools

- **13 Mode**
- is an analytics platform focused on giving data scientists an easy and iterative environment. It provides an interactive SQL editor and notebook environment for analysis, along with visualization and collaboration tools for less technical users. Mode has a unique data engine called Helix that streams data from external databases and stores it in memory to allow for fast and interactive analysis. It supports in-memory analysis of up to 10GB of data.
- **14. KNIME**
- short for the Konstanz Information Miner — is a free, open source data analytics platform that supports data integration, processing, visualization, and reporting. It plugs in machine learning and data mining libraries with minimal or no programming requirements. KNIME is great for data scientists who need to integrate and process data for machine learning and other statistical models but don't necessarily have strong programming skills. The graphical interface allows for point-and-click analysis and modeling.
- **15. Looker**
- is a cloud-based business intelligence and data analytics platform. It features automatic data model generation that scans data schemas and infers relationships between tables and data sources. Data engineers can modify the generated models through a built-in code editor.
- **16. Excel**
- is the most common tool used for manipulating spreadsheets and building analyses. With decades of development behind it, Excel can support almost any standard analytics workflow and is extendable through its native programming language, Visual Basic. Excel is suitable for simple analysis, but it is not suited for **analyzing big data** — it has a limit of around 1 million rows — and it does not have good support for collaboration or versioning. Enterprises should consider more modern cloud-based analytics platforms for large and collaborative analyses.

Applications of Data Analytics

- **1. Transportation**
- Data analytics can be applied to help in improving Transportation Systems and the intelligence around them. The predictive method of the analysis helps find transport problems like Traffic or network congestion. It helps synchronize the vast amount of data and uses them to build and design plans and strategies to plan alternative routes and reduce congestion and traffic, which in turn reduces the number of accidents and mishappenings.
- **2. Web Search or Internet Web Results**
- The web search engines like Yahoo, Bing, Duckduckgo, and Google use a set of data to give you when you search a data. Whenever you hit on the search button, the search engines use algorithms of data analytics to deliver the best-searched results within a limited time frame. The set of data that appears whenever we search for any information is obtained through data analytics.
- **3. Manufacturing**
- Data analytics helps the manufacturing industries maintain their overall work through certain tools like prediction analysis, regression analysis, budgeting, etc. The unit can figure out the number of products needed to be manufactured according to the data collected and analyzed from the demand samples and likewise in many other operations increasing the operating capacity as well as the profitability.

Applications of Data Analytics

- **4. Security**
- Data analyst provides utmost security to the organization, Security Analytics is a way to deal with online protection zeroed in on the examination of information to deliver proactive safety efforts. No business can foresee the future, particularly where security dangers are concerned, yet by sending security investigation apparatuses that can dissect security occasions it is conceivable to identify danger before it gets an opportunity to affect your framework and main concern.
- **6. Education**
- Data analytics applications in education are the most needed data analyst in the current scenario. It is mostly used in adaptive learning, new innovations, adaptive content, etc. Is the estimation, assortment, investigation, and detailing of information about students and their specific circumstances, for reasons for comprehension and streamlining learning and conditions in which it happens.
- **7. Healthcare**
- Applications of data analytics in healthcare can be utilized to channel enormous measures of information in seconds to discover treatment choices or answers for various illnesses. This won't just give precise arrangements dependent on recorded data yet may likewise give accurate answers for exceptional worries for specific patients.

Applications of Data Analytics

- **9. Insurance**
- There is a lot of data analysis taking place during the insurance process. Several data, such as actuarial data and claims data, help insurance companies realize the risk involved in insuring the person. Analytical software can be used to identify risky claims and bring them before the authorities for further investigation.
- **10. Digital Advertisement**
- Digital advertising has also been transformed as a result of the **application of data science**. Data analytics and data algorithms are used in a wide range of advertising mediums, including digital billboards in cities and banners on websites.
- **11. Fraud and Risk Detection**
- Detecting fraud may have been the first **application of data analytics**. They applied data analytics because they already had a large amount of customer data at their disposal. Data analysis was used to examine recent spending patterns and customer profiles to determine the likelihood of default. It eventually resulted in a reduction in fraud and risk.
- **12. Travel**
- **Data analysis applications** can be used to improve the traveller’s purchasing experience by analyzing social media and mobile/weblog data. Companies can use data on recent browse-to-buy conversion rates to create customized offers and packages that take into account the preferences and desires of their customers.