BIG DATA ANALYTICS

DR.GOKULAKRISHNAN.S

WHAT IS BIG DATA?

- The term "Big Data" usually refers to datasets that are too large, complex and unable to be processed by ordinary data processing systems to manage efficiently. These datasets can be derived from a variety of sources, including social media, sensors, internet activity, and mobile devices. The data can be structured, semi-structured and unstructured type of data.

WHAT IS BIG DATA?

- Gartner defines Big Data as Big data is high-volume, high-velocity and/or high-variety information that demands cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

BIG DATA ANALYTICS

- A process of analysing large and diverse data sets is known as "Big Data," It discovers hidden patterns, unknown relationships, market trends, user preferences, and other important information. It uses advanced analytics techniques such as statistical analysis, machine learning, data mining, and predictive modelling to extract insights from enormous datasets.

BIG DATA ANALYTICS

- Organisations across the world capture terabytes of data about their users' interactions, business, social media, and also sensors from devices such as mobile phones and automobiles. The challenge of this era is to make sense of this sea of data. This is where big data analytics comes into the picture.

WHERE BIG DATA ANALYTICS USED?

- Big Data Analytics strives to assist organisations in making more informed business decisions, increasing operational efficiency, improving customer experiences and services, and making sure to sustain industries in a competitive world with their respective industries.

WHERE BIG DATA ANALYTICS USED?

- The Big Data Analytics process involves data gathering, storage, processing, analysis, and visualisation of outcomes to make strategic business decisions. The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Big Data Analytics.

STEPS OF BIG DATA ANALYTICS

Data Collection

- This is the initial step, in which data is collected from different sources like social media, sensors, online channels, commercial transactions, website logs etc. Collected data might be structured (predefined organisation, such as databases), semi-structured (like log files) or unstructured (text documents, photos, and videos).

STEPS OF BIG DATA ANALYTICS

Data Cleaning (Data Pre-processing)

- The next step is to process collected data by removing errors and making it suitable and proper for analysis. Collected raw data generally contains errors, missing values, inconsistencies, and noisy data. Data cleaning entails identifying and correcting errors to ensure that the data is accurate and consistent. Pre-processing operations may also involve data transformation, normalisation, and feature extraction to prepare the data for further analysis.

STEPS OF BIG DATA ANALYTICS

Data Analysis

- This is a key phase of big data analytics. Different techniques and algorithms are used to analyse data and derive useful insights. This can include descriptive analytics (summarising data to better understand its characteristics), diagnostic analytics (identifying patterns and relationships), predictive analytics (predicting future trends or outcomes), and prescriptive analytics (making recommendations or decisions based on the analysis).

STEPS OF BIG DATA ANALYTICS

Data Visualization

- Its a step to present data in a visual form using charts, graphs and interactive dashboards. Hence, data visualisation techniques are used to visually portray the data using charts, graphs, dashboards, and other graphical formats to make data analysis insights more clear and actionable.

STEPS OF BIG DATA ANALYTICS

Interpretation and Decision Making

- Once data analytics and visualisation are done and insights gained, stakeholders analyse the findings to make informed decisions. This decision-making includes optimising corporate operations, increasing consumer experiences, creating new products or services, and directing strategic planning.

STEPS OF BIG DATA ANALYTICS

Data Storage and Management

- Once collected, the data must be stored in a way that enables easy retrieval and analysis. Traditional databases may not be sufficient for handling large amounts of data, hence many organisations use distributed storage systems such as Hadoop Distributed File System (HDFS) or cloud-based storage solutions like Amazon S3.

STEPS OF BIG DATA ANALYTICS

Continuous Learning and Improvement

- Big data analytics is a continuous process of collecting, cleaning, and analyzing data to uncover hidden insights. It helps businesses make better decisions and gain a competitive edge.

TYPES OF BIG-DATA

Big Data is generally categorized into three different varieties.

- Structured Data
- Semi-Structured Data
- Unstructured Data

TYPES OF BIG-DATA

Structured Data

- Structured data has a dedicated data model, a well-defined structure, and a consistent order, and is designed in such a way that it can be easily accessed and used by humans or computers. Structured data is usually stored in well-defined tabular form means in the form of rows and columns. Example: MS Excel, Database Management Systems (DBMS)

TYPES OF BIG-DATA

Semi-Structured Data

- Semi-structured data can be described as another type of structured data. It inherits some qualities from Structured Data; however, the majority of this type of data lacks a specific structure and does not follow the formal structure of data models such as an RDBMS. Example: Comma Separated Values (CSV) File.

TYPES OF BIG-DATA

Unstructured Data

- Unstructured data is a type of data that doesnt follow any structure. It lacks a uniform format and is constantly changing. However, it may occasionally include data and time-related information. Example: Audio Files, Images etc.

TYPES OF BIG DATA ANALYTICS

Descriptive Analytics

- Descriptive analytics gives a result like **What is happening in my business?"** if the dataset is business-related. Overall, this summarises prior facts and aids in the creation of reports such as a company's income, profit, and sales figures. It also aids the tabulation of social media metrics. It can do comprehensive, accurate, live data and effective visualisation.

TYPES OF BIG DATA ANALYTICS

Diagnostic Analytics

- Diagnostic analytics determines root causes from data. It answers like **Why is it happening?** Some common examples are drill-down, data mining, and data recovery. Organisations use diagnostic analytics because they provide an in-depth insight into a particular problem. Overall, it can drill down the root causes and ability to isolate all confounding information.

TYPES OF BIG DATA ANALYTICS

Diagnostic Analytics

- **For example** – A report from an online store says that sales have decreased, even though people are still adding items to their shopping carts. Several things could have caused this, such as the form not loading properly, the shipping cost being too high, or not enough payment choices being offered. You can use diagnostic data to figure out why this is happening.

TYPES OF BIG DATA ANALYTICS

Predictive Analytics

- This kind of analytics looks at data from the past and the present to guess what will happen in the future. Hence, it answers like **What will be happening in future?** Data mining, AI, and machine learning are all used in predictive analytics to look at current data and guess what will happen in the future. It can figure out things like market trends, customer trends, and so on.

TYPES OF BIG DATA ANALYTICS

For example –

The rules that Bajaj Finance has to follow to keep their customers safe from fake transactions are set by PayPal. The business uses predictive analytics to look at all of its past payment and user behaviour data and come up with a program that can spot fraud.

TYPES OF BIG DATA ANALYTICS

Prescriptive Analytics

- Perspective analytics gives the ability to frame a strategic decision, the analytical results answer **What do I need to do?** Perspective analytics works with both descriptive and predictive analytics. Most of the time, it relies on AI and machine learning.

TYPES OF BIG DATA ANALYTICS

For example – Prescriptive analytics can help a company to maximise its business and profit. For example in the airline industry, Perspective analytics applies some set of algorithms that will change flight prices automatically based on demand from customers, and reduce ticket prices due to bad weather conditions, location, holiday seasons etc.

TOOLS AND TECHNOLOGIES OF BIG DATA ANALYTICS

Hadoop

A tool to store and analyze large amounts of data. Hadoop makes it possible to deal with big data, It's a tool which made big data analytics possible.

MongoDB

- A tool for managing unstructured data. It's a database which specially designed to store, access and process large quantities of unstructured data.

TOOLS AND TECHNOLOGIES OF BIG DATA ANALYTICS

Cassandra

- A distributed database used to handle chunks of data. Cassandra is an open-source distributed NoSQL database management system that handles massive amounts of data over several commodity servers, ensuring high availability and scalability without sacrificing performance.

TOOLS AND TECHNOLOGIES OF BIG DATA ANALYTICS

Talend

- A tool to use for data integration and management. Talend's solution package includes complete capabilities for data integration, data quality, master data management, and data governance. Talend integrates with big data management tools like Hadoop, Spark, and NoSQL databases allowing organisations to process and analyse enormous amounts of data efficiently.

TOOLS AND TECHNOLOGIES OF BIG DATA ANALYTICS

Spark

- Used for real-time processing and analyzing large amounts of data. Apache Spark is a robust and versatile distributed computing framework that provides a single platform for big data processing, analytics, and machine learning, making it popular in industries such as e-commerce, finance, healthcare, and telecommunications.

TOOLS AND TECHNOLOGIES OF BIG DATA ANALYTICS

Storm

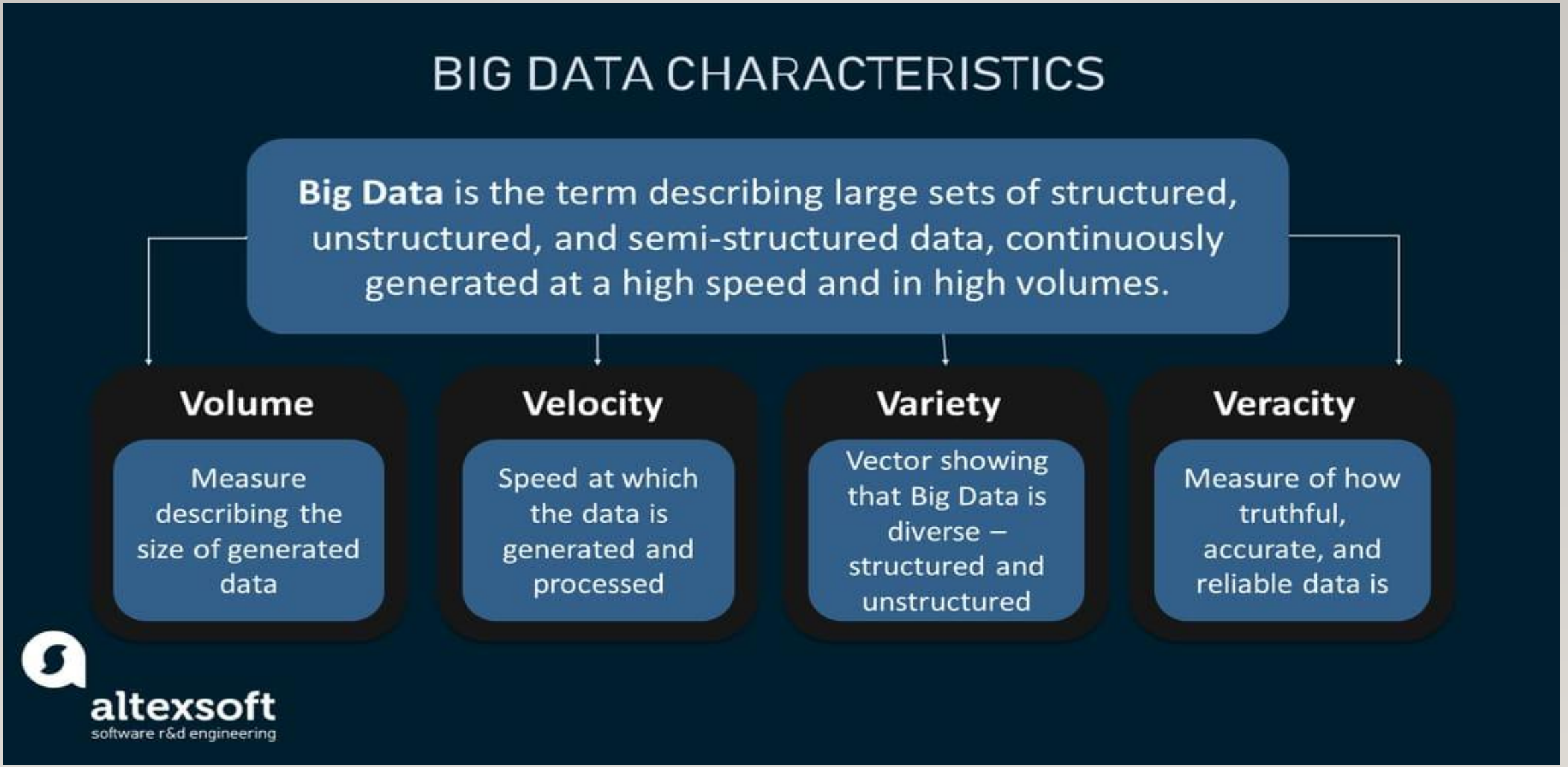
- It is an open-source real-time computational system. Apache Storm is a robust and versatile stream processing framework that allows organisations to process and analyse real-time data streams on a large scale, making it suited for a wide range of use cases in industries such as banking, telecommunications, e-commerce, and IoT.

TOOLS AND TECHNOLOGIES OF BIG DATA ANALYTICS

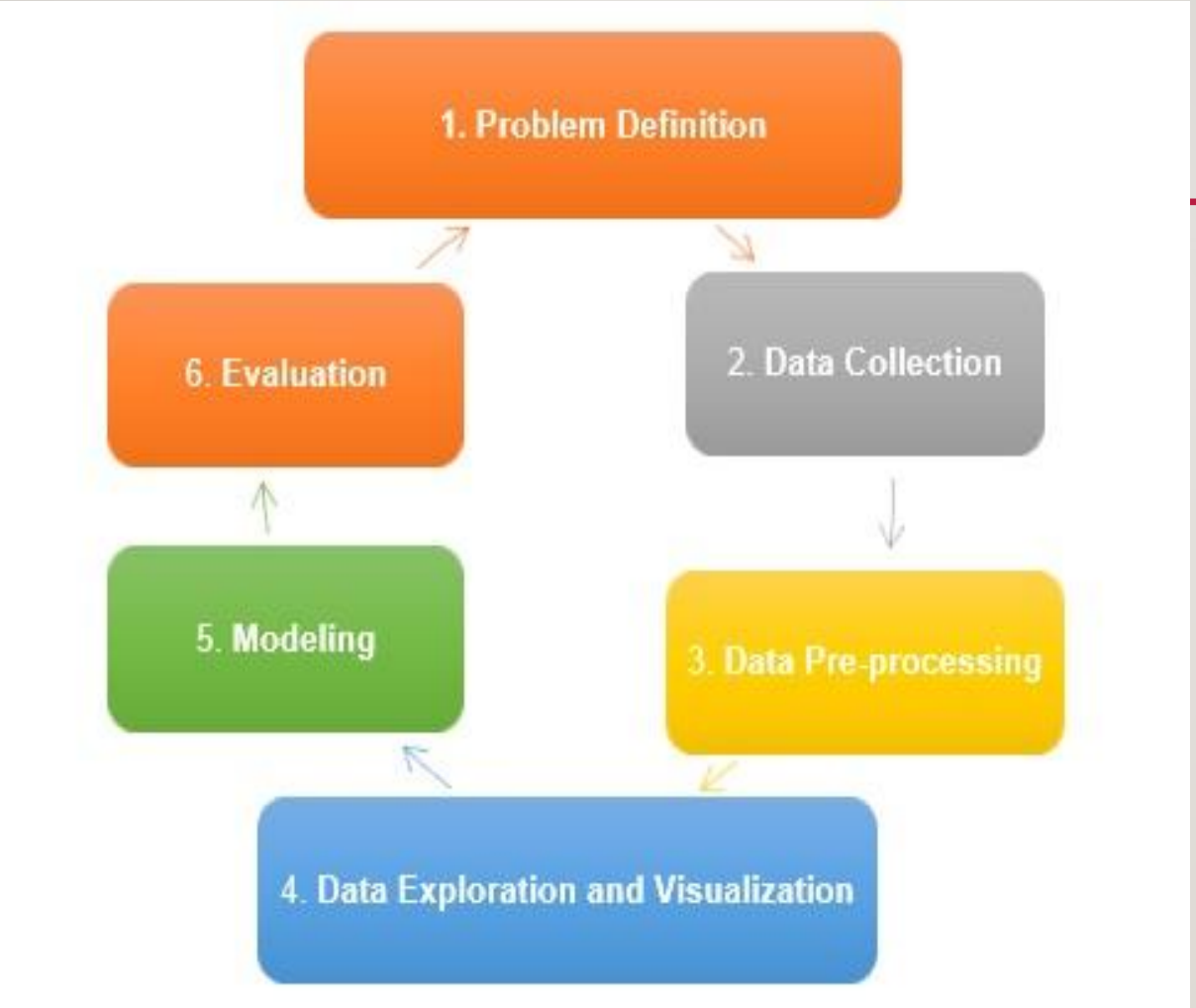
Kafka

- It is a distributed streaming platform that is used for fault-tolerant storage. Apache Kafka is a versatile and powerful event streaming platform that allows organisations to create scalable, fault-tolerant, and real-time data pipelines and streaming applications to efficiently meet their data processing requirements.

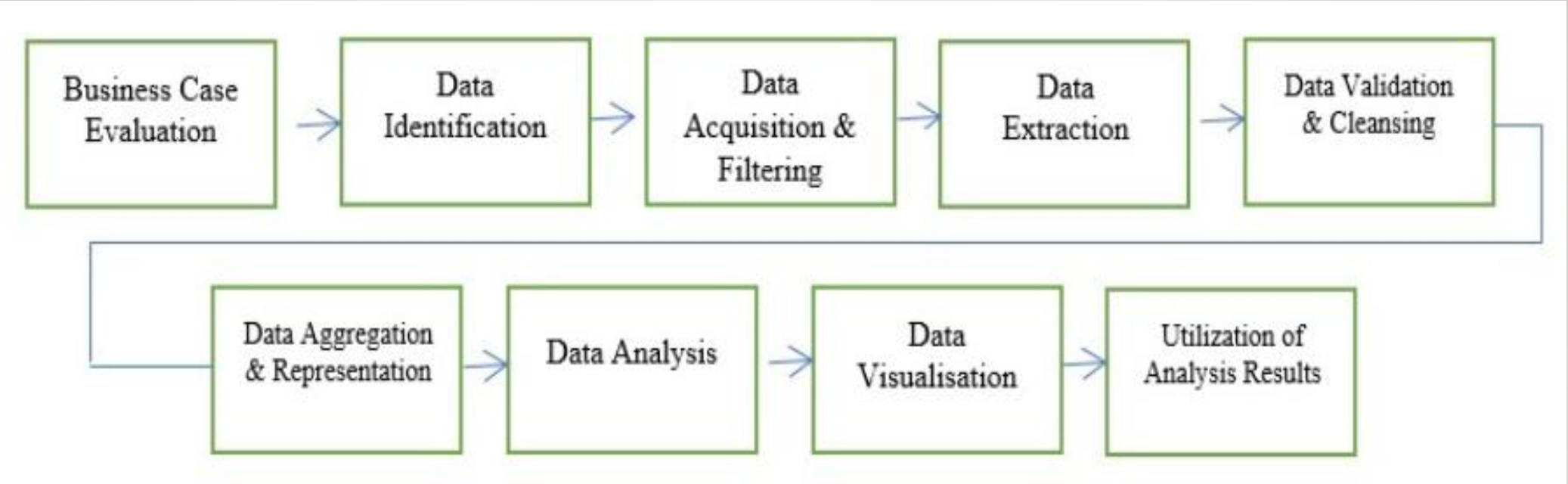
BIG DATA ANALYTICS - CHARACTERISTICS



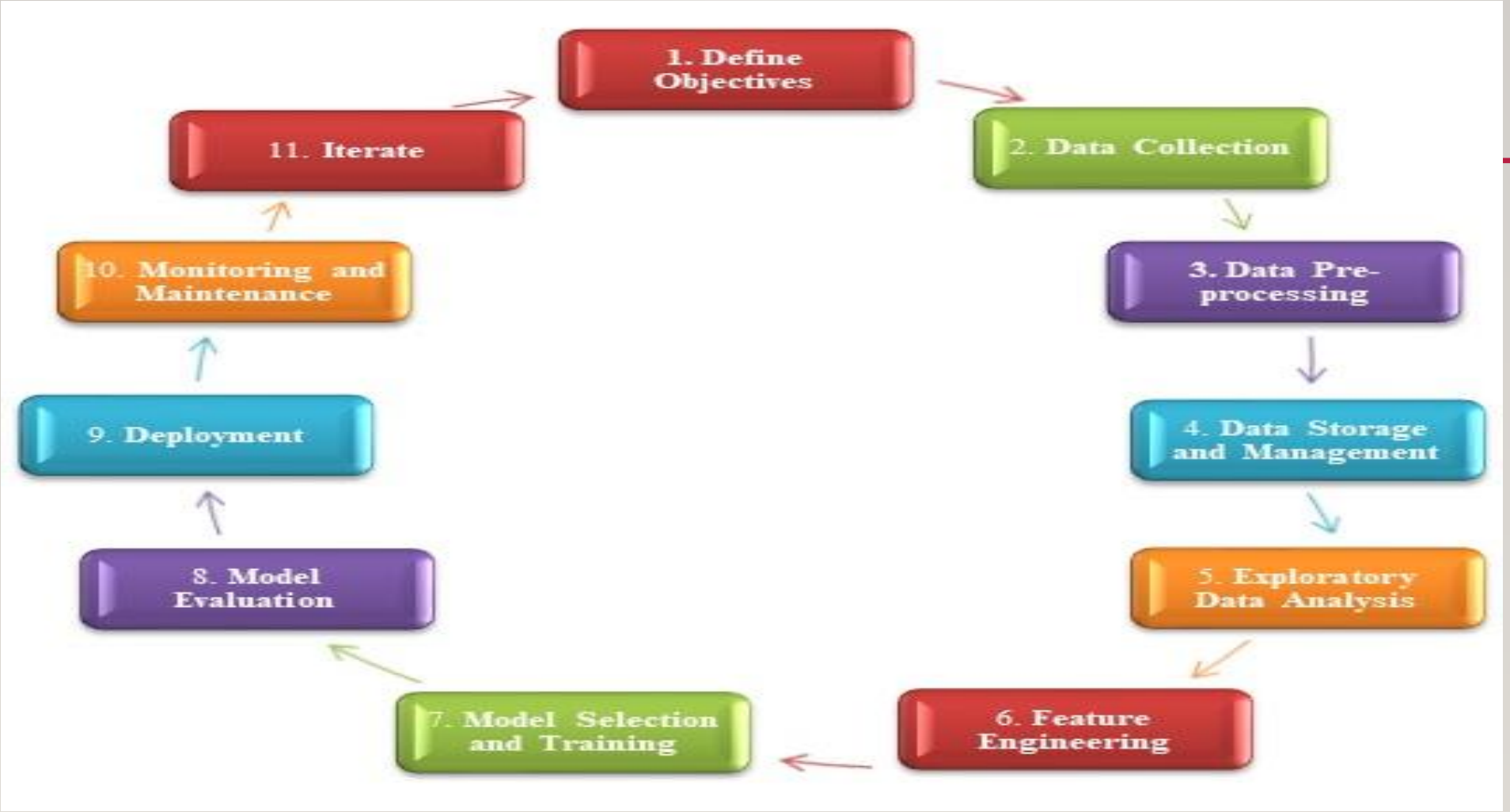
TRADITIONAL DATA MINING LIFE CYCLE

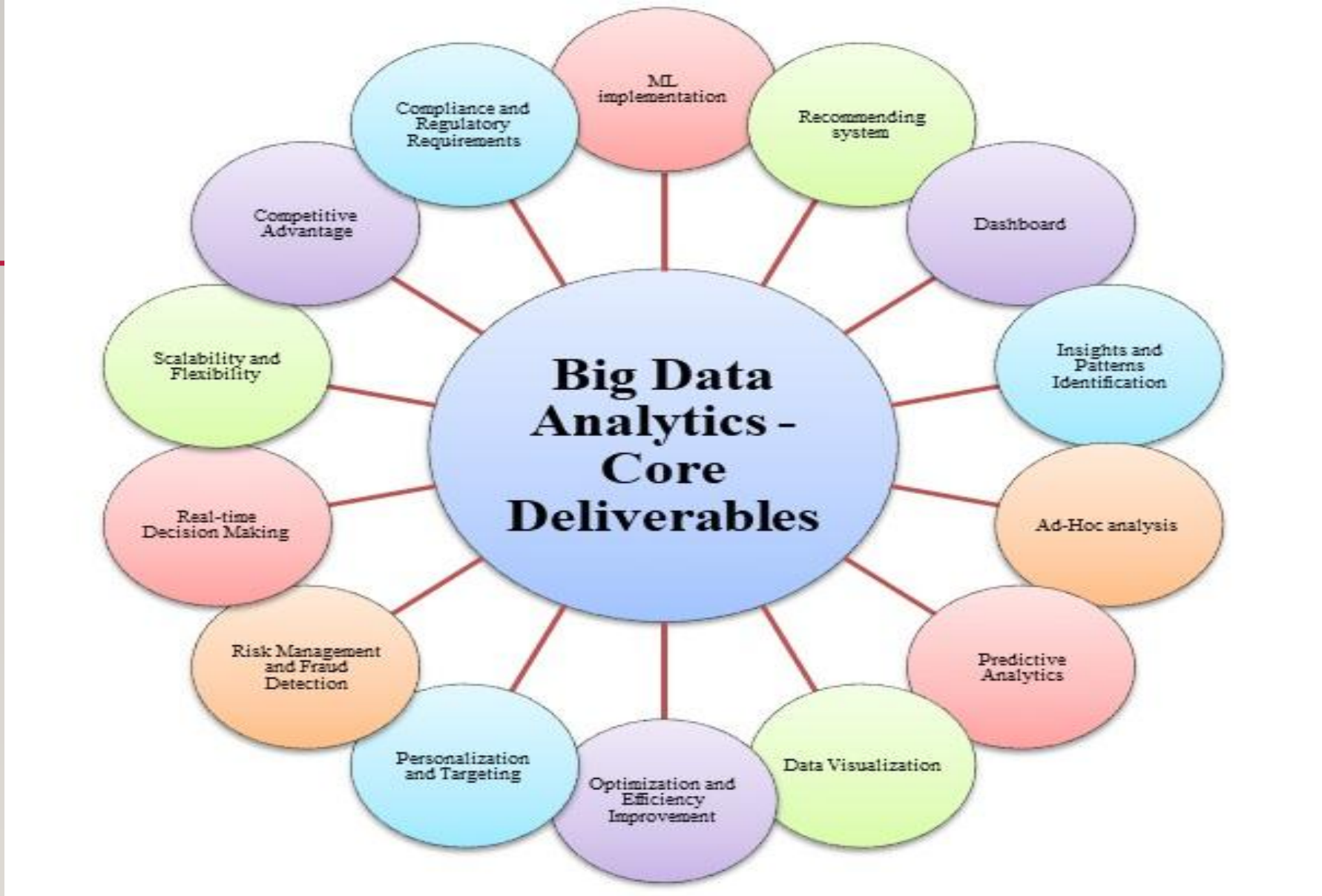


BIG DATA LIFE CYCLE

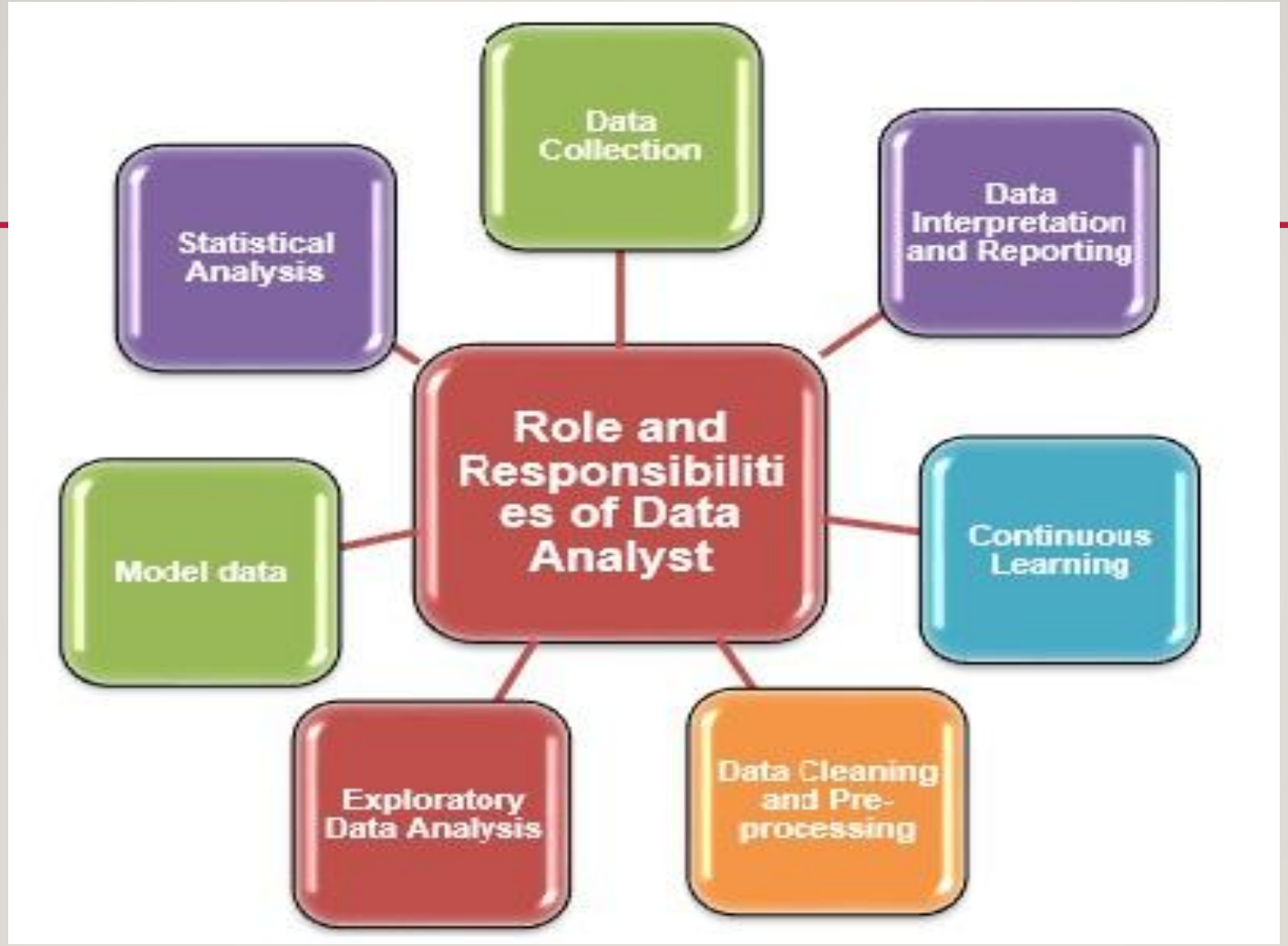


BIG DATA ANALYTICS METHODOLOGY




Big Data Analytics - Core Deliverables
ML implementation
Recommending system
Dashboard
Insights and Patterns Identification
Ad-Hoc analysis
Predictive Analytics
Data Visualization
Optimization and Efficiency Improvement
Personalization and Targeting
Risk Management and Fraud Detection
Real-time Decision Making
Scalability and Flexibility
Competitive Advantage
Compliance and Regulatory Requirements

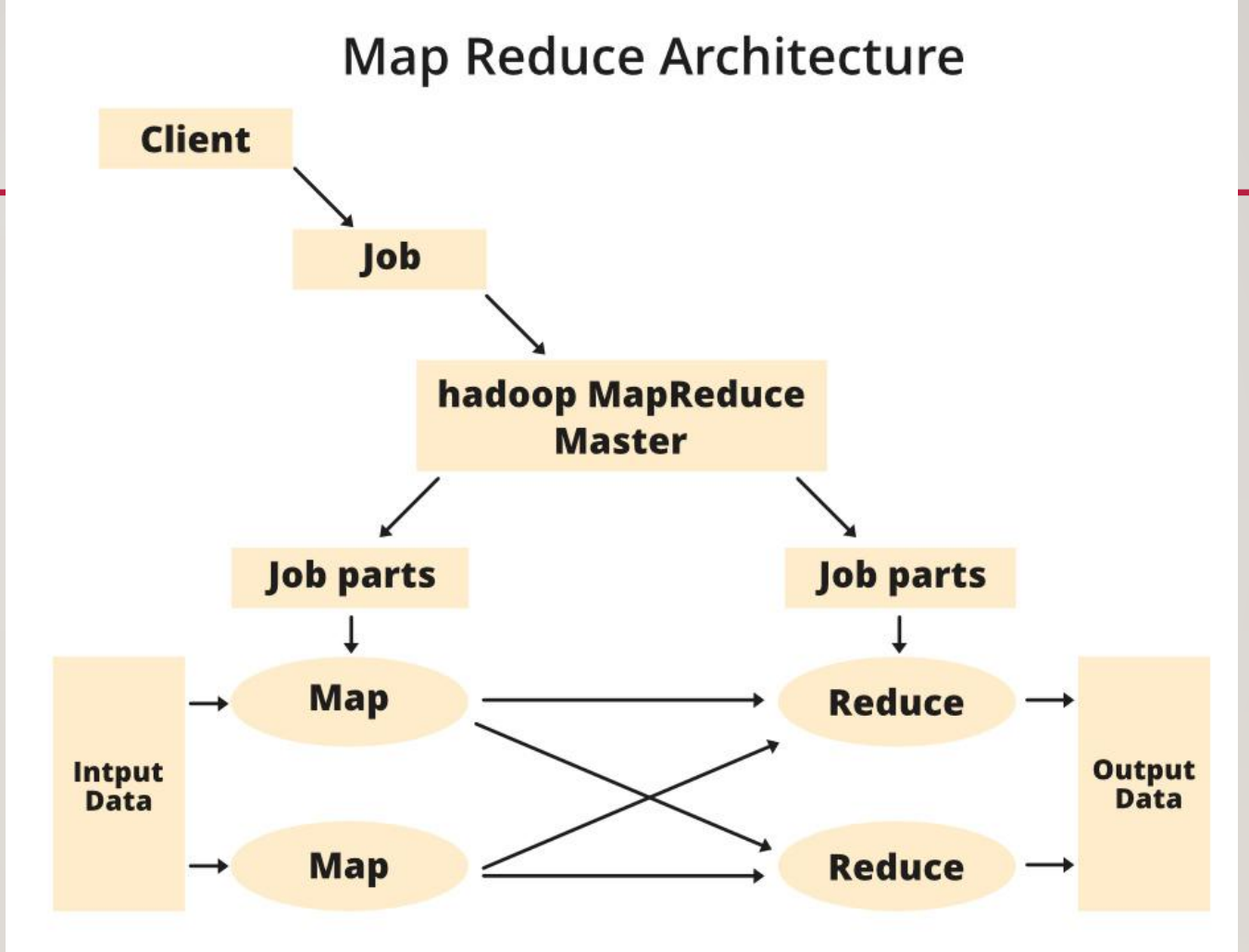
Role and Responsibilities of Data Analyst



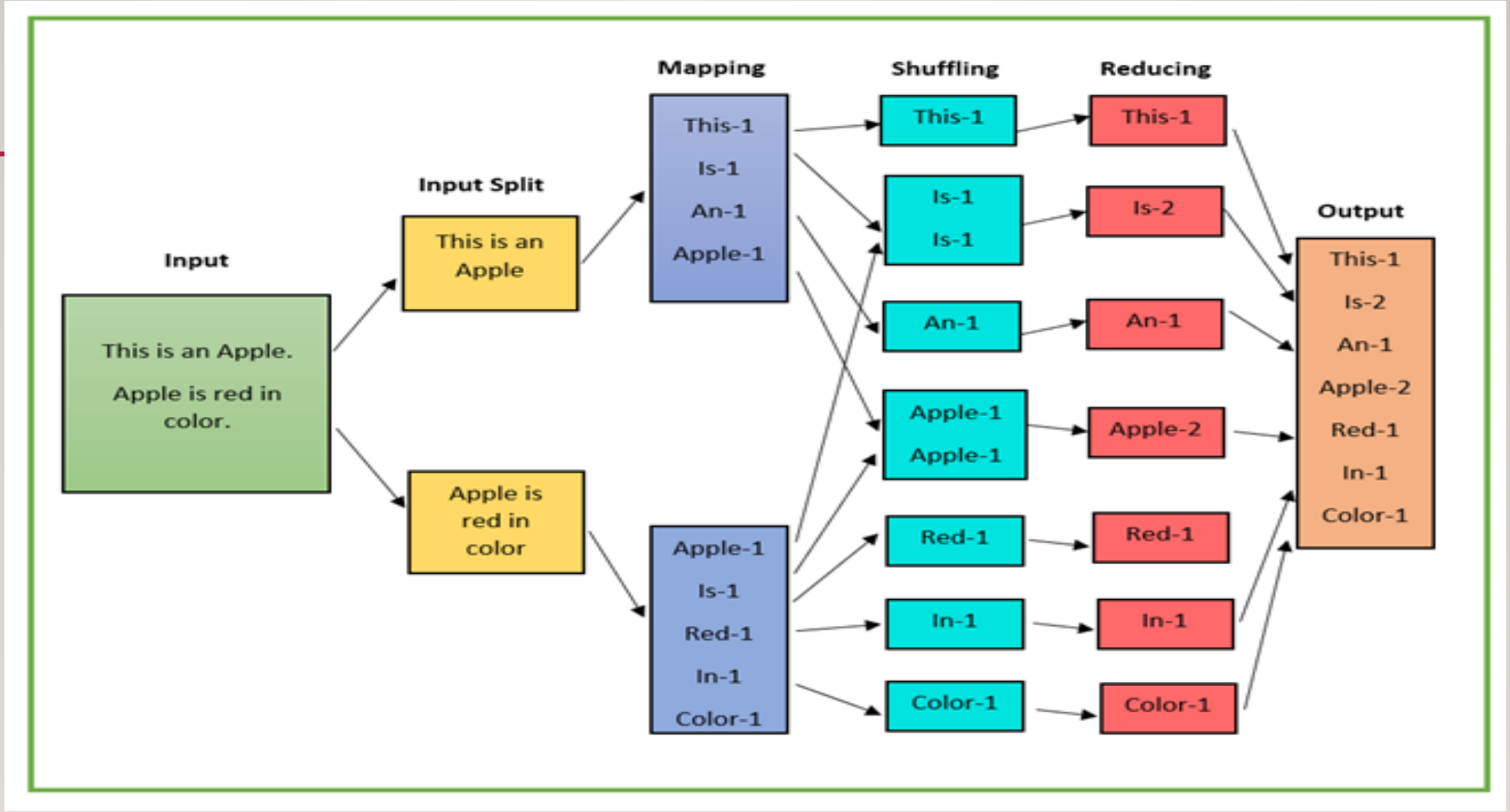
MAPREDUCE

- MapReduce is a programming model used for efficient processing in parallel over large data-sets in a distributed manner. The data is first split and then combined to produce the final result. The libraries for MapReduce is written in so many programming languages with various different-different optimizations. The purpose of MapReduce in Hadoop is to Map each of the jobs and then it will reduce it to equivalent tasks for providing less overhead over the cluster network and to reduce the processing power. The MapReduce task is mainly divided into two phases Map Phase and Reduce Phase.

MAPREDUCE


Map Reduce Architecture
Client
Job
hadoop MapReduce Master
Job parts
Job parts
Intput Data
Map
Map
Reduce
Reduce
Output Data

MAPREDUCE


Input
This is an Apple.
Apple is red in color.
Input Split
This is an Apple
Apple is red in color
Mapping
This-1
Is-1
An-1
Apple-1
Apple-1
Is-1
Red-1
In-1
Color-1
Shuffling
This-1
Is-1
Is-1
An-1
Apple-1
Apple-1
Red-1
In-1
Color-1
Reducing
This-1
Is-2
An-1
Apple-2
Red-1
In-1
Color-1
Output
This-1
Is-2
An-1
Apple-2
Red-1
In-1
Color-1