INTRODUCTION TO BIG DATA ANALYTICS

UNIT 1

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.

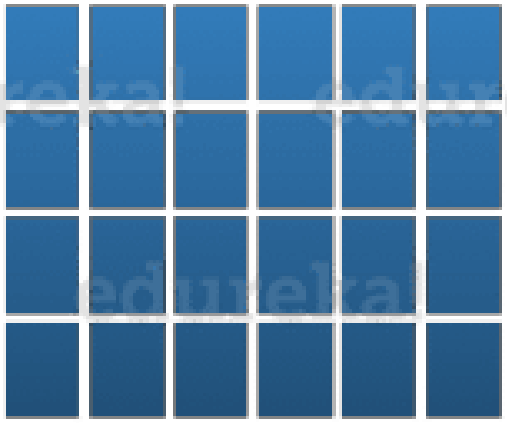
WHAT IS BIG DATA

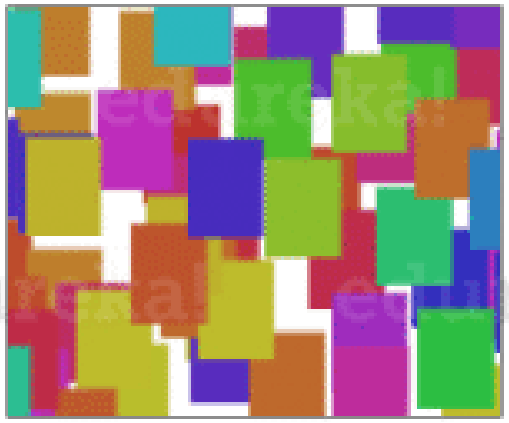
- As Gartner defines it – “Big Data are high volume, high velocity, or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.”
- The term ‘big data’ is self-explanatory – a collection of huge data sets that normal computing techniques cannot process.
- The term not only refers to the data, but also to the various frameworks, tools, and techniques involved.
- Technological advancement and the advent of new channels of communication (like social networking) and new, stronger devices have presented a challenge to industry players in the sense that they have to find other ways to handle the data.
- Big data is an all-inclusive term, representing the enormous volume of complex data sets that companies and governments generate in the present-day digital environment.
- Big data, typically measured in petabytes or terabytes, materializes from three major sources—transactional data, machine data, and social data.

TYPES OF BIG-DATA

Big Data is generally categorized into three different varieties. They are as shown below:

- **Structured Data**
- **Semi-Structured Data**
- **Unstructured Data**





TYPES OF BIG-DATA

•**Structured Data** owns a dedicated data model, It also has a well-defined structure, it follows a consistent order and it is designed in such a way that it can be **easily accessed** and used by a person or a computer. Structured data is usually stored in well-defined columns and also Databases.

Example: Database Management Systems(**DBMS**)

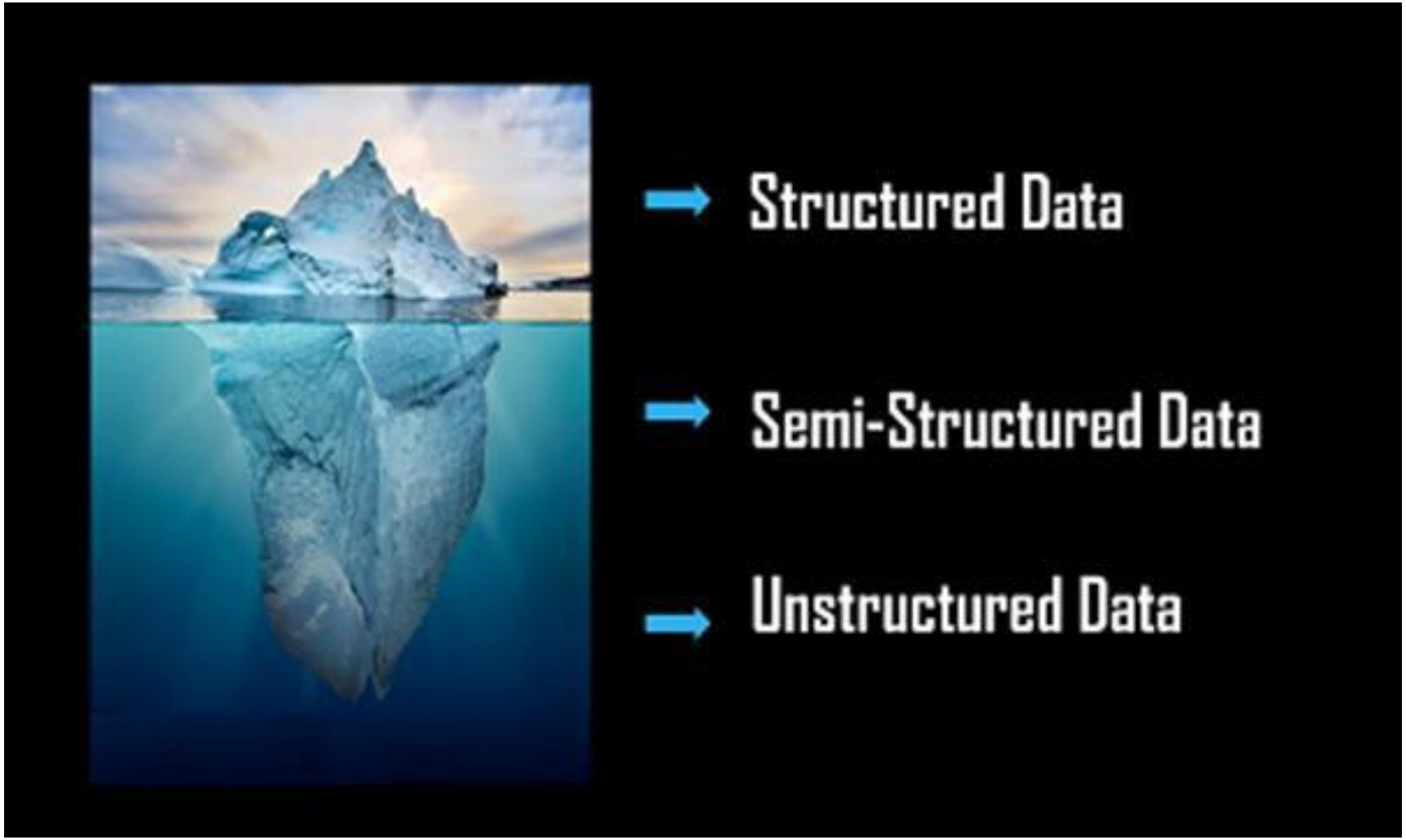
•**Semi-Structured Data** can be considered as another form of Structured Data. It inherits a few properties of Structured Data, but the major part of this kind of data fails to have a definite structure and also, it does not obey the formal structure of data models such as an RDBMS.

Example: Comma Separated Values(**CSV**) File.

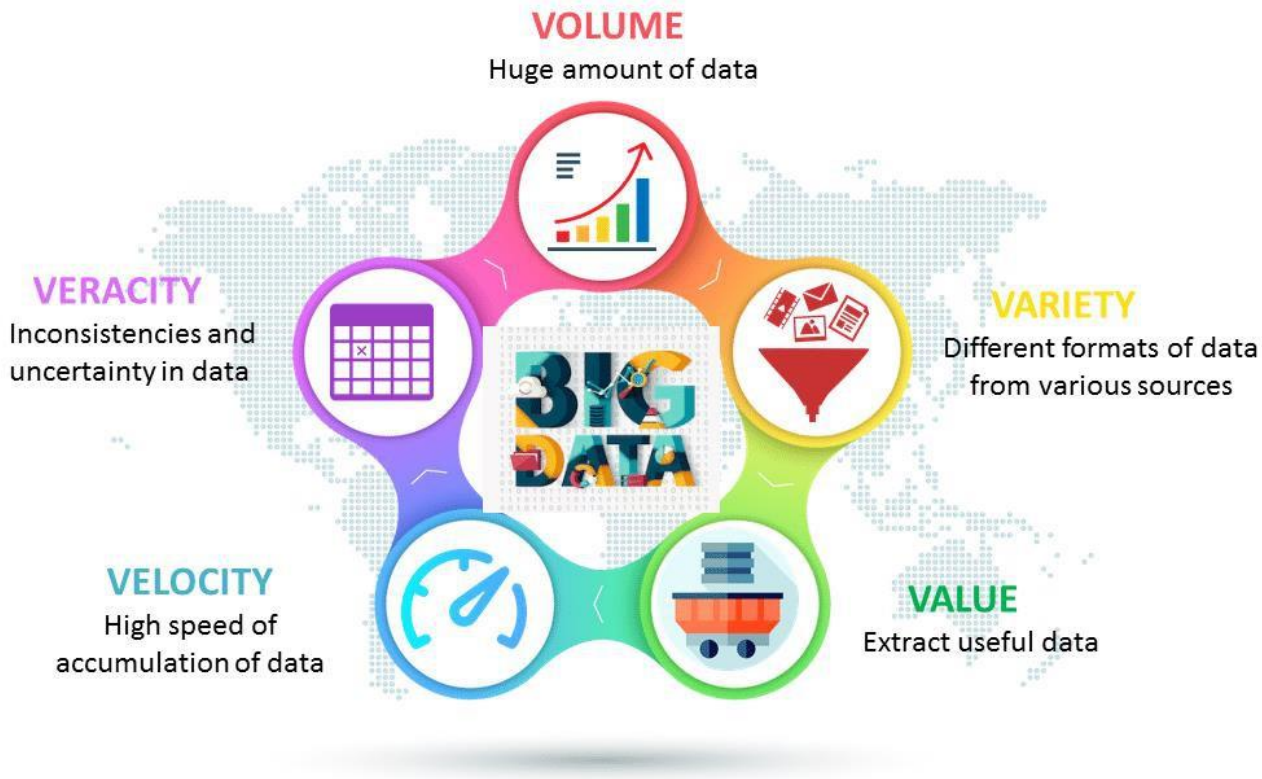
•**Unstructured Data** is completely a different type of which neither has a structure nor obeys to follow the formal structural rules of data models. It does not even have a consistent format and it found to be varying all the time. But, rarely it may have information related to data and time.

Example: Audio Files, Images etc

TYPES OF BIG-DATA



THE CHARACTERISTICS OF BIG DATA

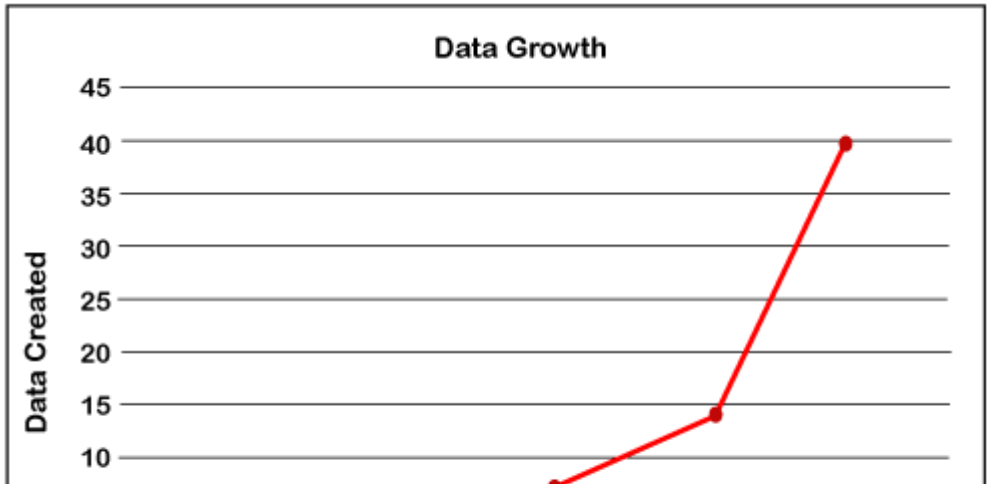

VOLUME
Huge amount of data
VERACITY
Inconsistencies and uncertainty in data
VARIETY
Different formats of data from various sources
BIG DATA
VELOCITY
High speed of accumulation of data
VALUE
Extract useful data

THE CHARACTERISTICS OF BIG DATA

Volume

Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot. We are currently using **distributed systems**, to store data in several locations and brought together by a software Framework like **Hadoop**.

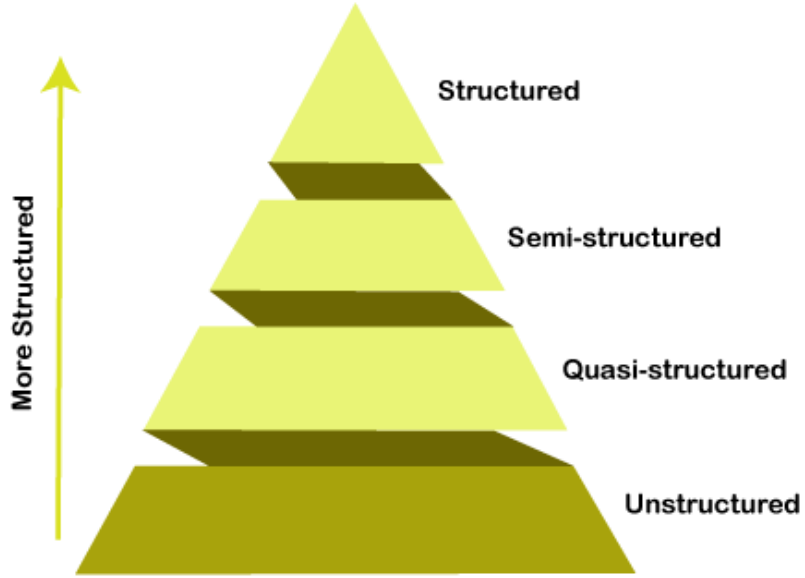
Facebook alone can generate about **billion** messages, **4.5 billion** times that the “like” button is recorded, and over **350 million** new posts are uploaded **each day.** Such a huge amount of data can only be handled by Big Data Technologies



THE CHARACTERISTICS OF BIG DATA

Variety

As Discussed before, **Big Data** is generated in multiple varieties. Compared to the traditional data like phone numbers and addresses, the latest trend of data is in the form of photos, videos, and audios and many more, making about 80% of the data to be completely unstructured



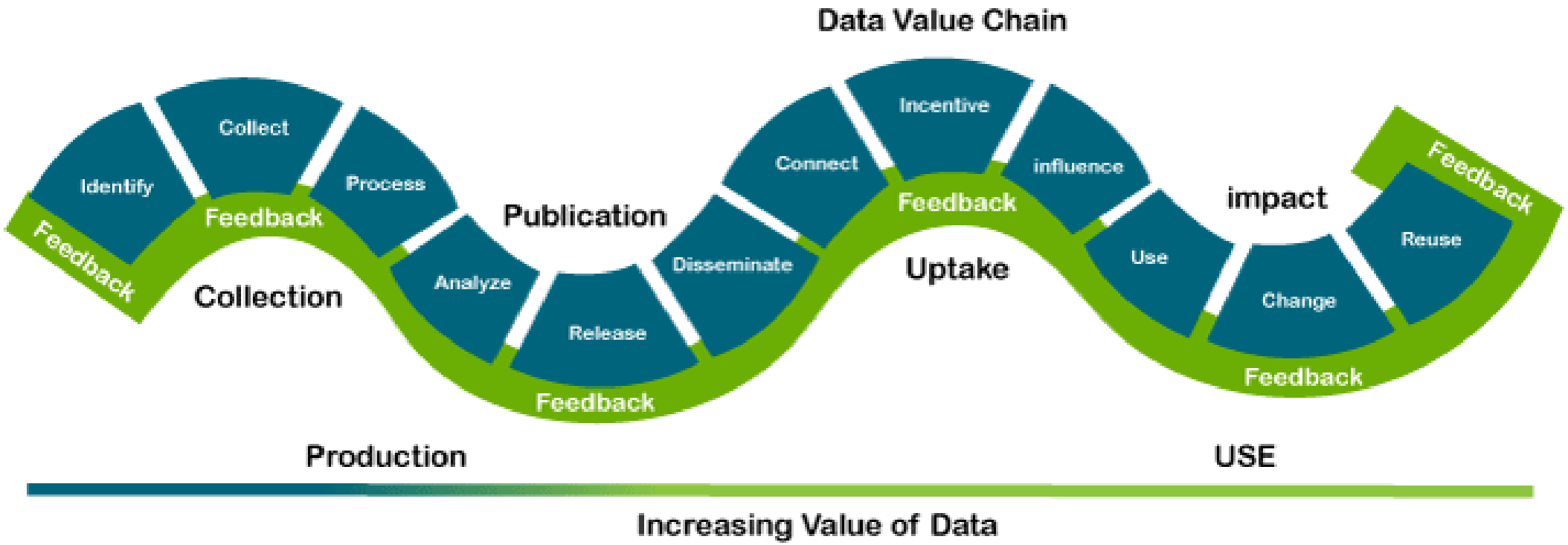
Veracity

Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, Big Data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments

THE CHARACTERISTICS OF BIG DATA

Value

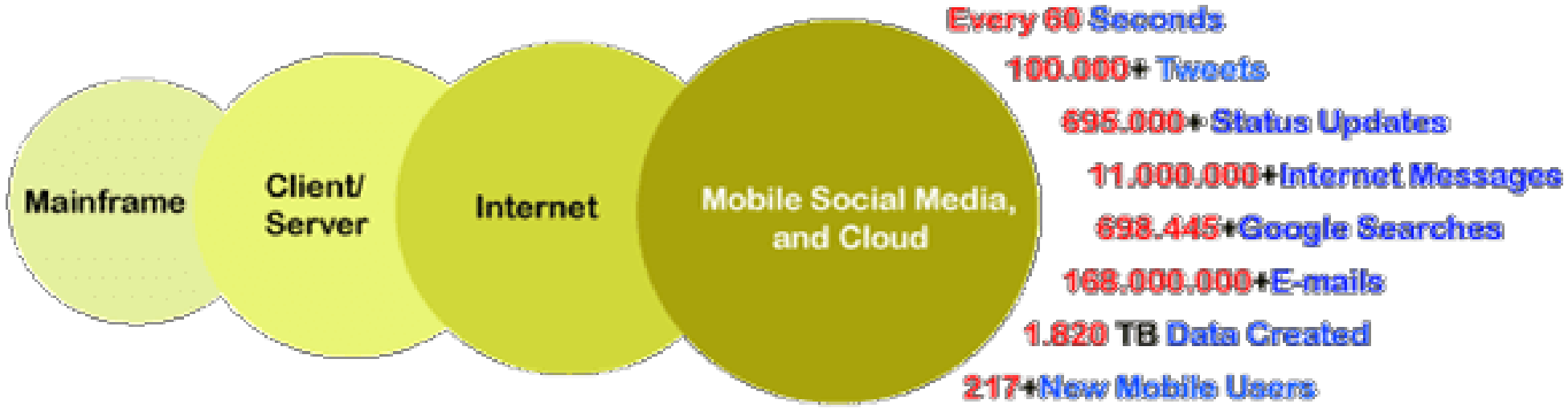
Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights.



THE CHARACTERISTICS OF BIG DATA

Velocity

Last but never least, Velocity plays a major role compared to the others, there is no point in investing so much to end up waiting for the data. So, the major aspect of Big Data is to provide data on demand and at a faster pace.



APPLICATIONS OF BIG DATA

- Retail
  - Leading online retail platforms are wholeheartedly deploying big data throughout a customer's purchase journey, to predict trends, forecast demands, optimize pricing, and identify customer behavioral patterns.
  - Big data is helping retailers implement clear strategies that minimize risk and maximize profit.

- Healthcare
  - Big data is revolutionizing the healthcare industry, especially the way medical professionals in the past diagnosed and treated diseases.
  - In recent times, effective analysis and processing of big data by machine learning algorithms provide significant advantages for the evaluation and assimilation of complex clinical data, which prevent deaths and improve the quality of life by enabling healthcare workers to detect early warning signs and symptoms.

APPLICATIONS OF BIG DATA

- Financial Services and Insurance
  - The increased ability to analyze and process big data is dramatically impacting the financial services, banking, and insurance landscape.
  - In addition to using big data for swift detection of fraudulent transactions, lowering risks, and supercharging marketing efforts, few companies are taking the applications to the next levels.

- Manufacturing
  - Advancements in robotics and automation technologies, modern-day manufacturers are becoming more and more data focused, heavily investing in automated factories that exploit big data to streamline production and lower operational costs.
  - Top global manufacturers are also integrating sensors into their products, capturing big data to provide valuable insights on product performance and its usage.

APPLICATIONS OF BIG DATA

- Energy
  - To combat the rising costs of oil extraction and exploration difficulties because of economic and political turmoil, the energy industry is turning toward data-driven solutions to increase profitability.
  - Big data is optimizing every process while cutting down energy waste from drilling to exploring new reserves, production, and distribution.

- Logistics & Transportation
  - State-of-the-art warehouses use digital cameras to capture stock level data, which, when fed into ML algorithms, facilitates intelligent inventory management with prediction capabilities that indicate when restocking is required.
  - In the transportation industry, leading transport companies now promote the collection and analysis of vehicle telematics data, using big data to optimize routes, driving behavior, and maintenance.

APPLICATIONS OF BIG DATA

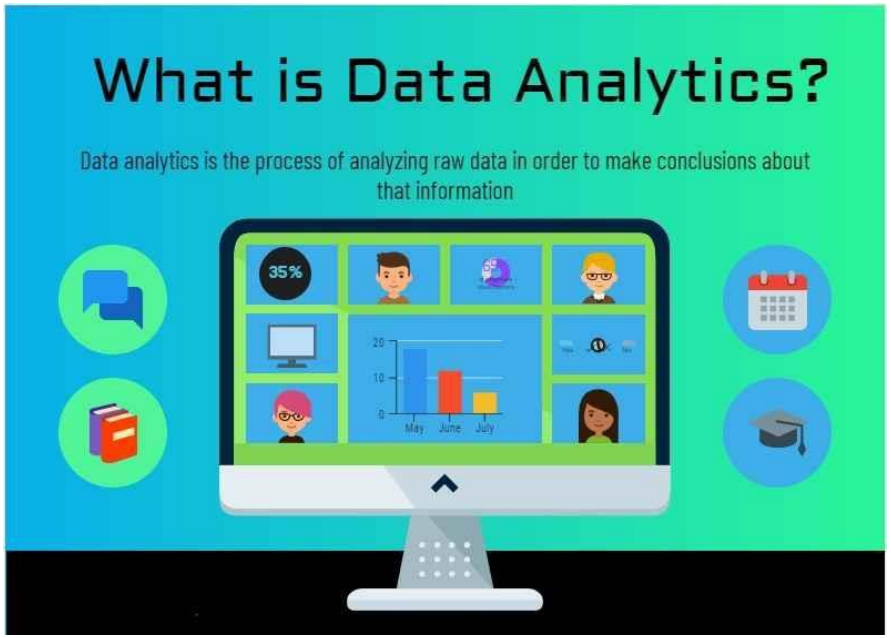
- Government
  - Cities worldwide are undergoing large-scale transformations to become “smart”, through the use of data collected from various Internet of Things (IoT) sensors.
  - Governments are leveraging this big data to ensure good governance via the efficient management of resources and assets, which increases urban mobility, improves solid waste management, and facilitates better delivery of public utility services.

WHAT IS ANALYTICS

Data analytics is a discipline focused on extracting insights from data, including the analysis, collection, organization, and storage of data, as well as the tools and techniques used to do so.

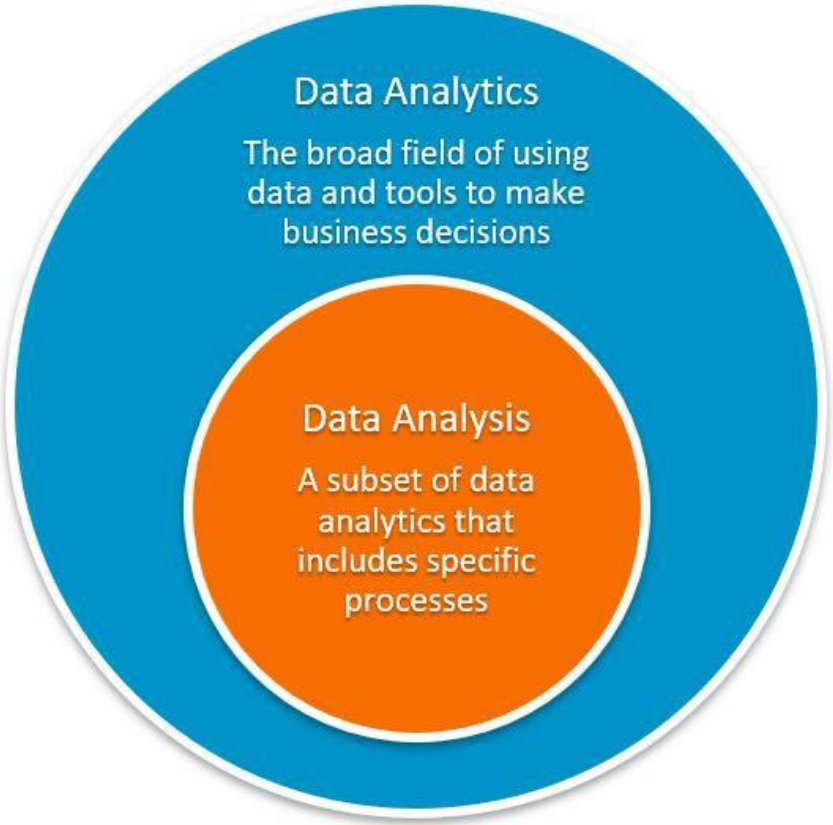
DATA ANALYTICS DEFINITION

- Data analytics is a discipline focused on extracting insights from data.
- It comprises the processes, tools and techniques of data analysis and management, including the collection, organization, and storage of data.
- The chief aim of data analytics is to apply statistical analysis and technologies on data to find trends and solve problems.
- Data analytics has become increasingly important in the enterprise as a means for analyzing and shaping business processes and improving decision-making and business results.
- Data analytics draws from a range of disciplines — including computer programming, mathematics, and statistics — to perform analysis on data in an effort to describe, predict, and improve performance.
- To ensure robust analysis, data analytics teams leverage a range of data management techniques, including data mining, data cleansing, data transformation, data modeling, and more.



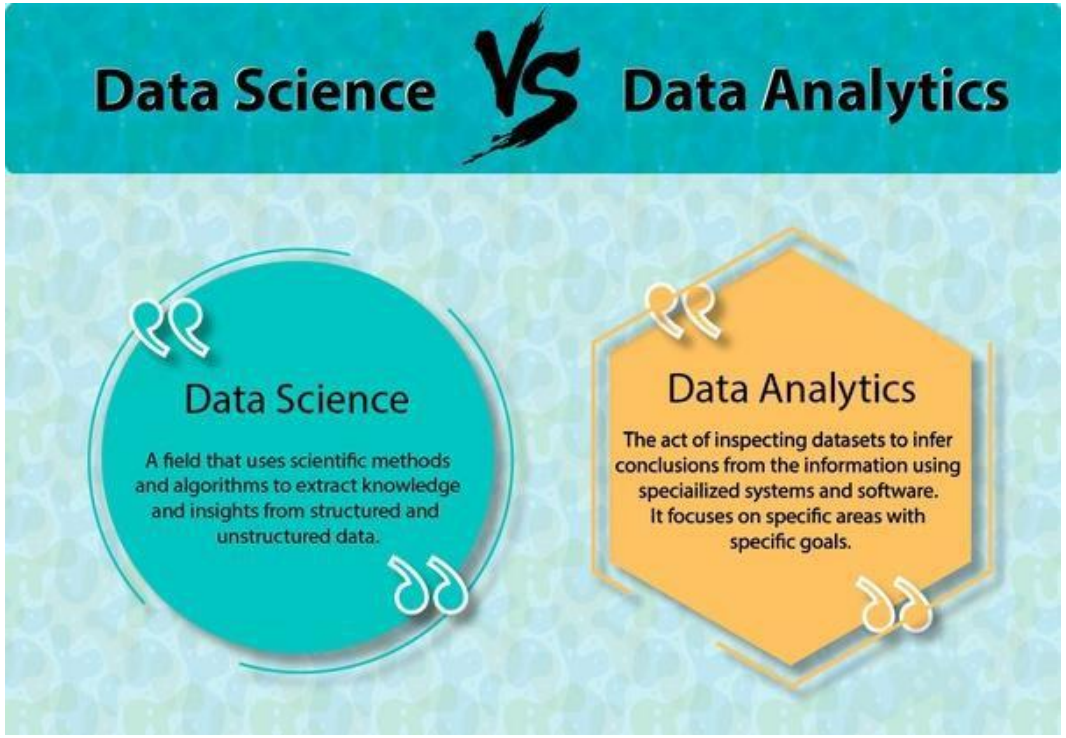
DATA ANALYTICS VS. DATA ANALYSIS

- While the terms data analytics and data analysis are frequently used interchangeably, data analysis is a subset of data analytics concerned with examining, cleansing, transforming, and modeling data to derive conclusions.
- Data analytics includes the tools and techniques used to perform data analysis.



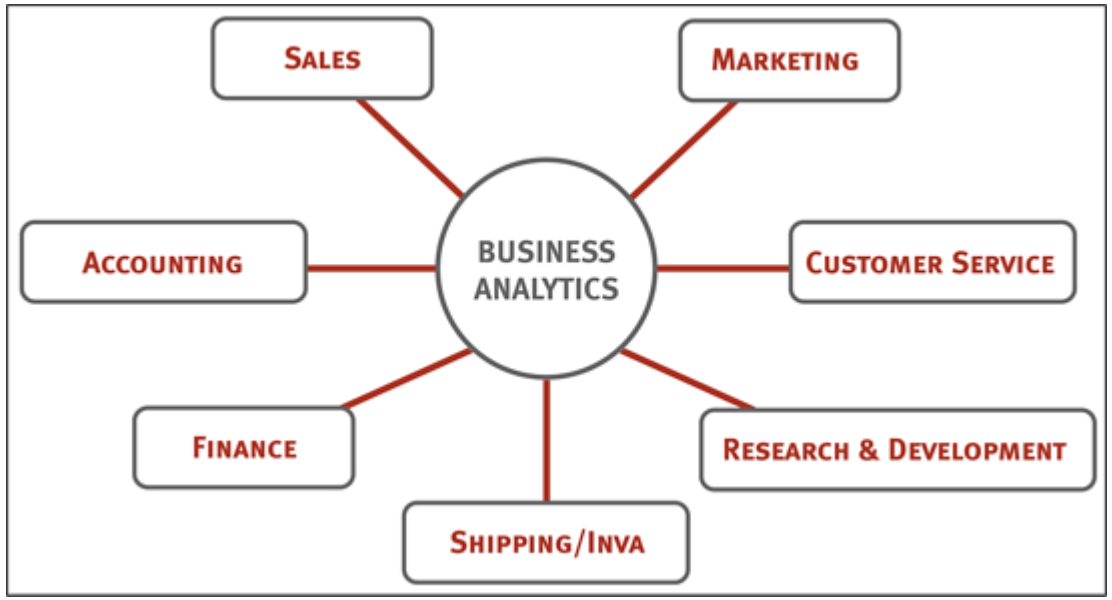
DATA ANALYTICS VS. DATA SCIENCE

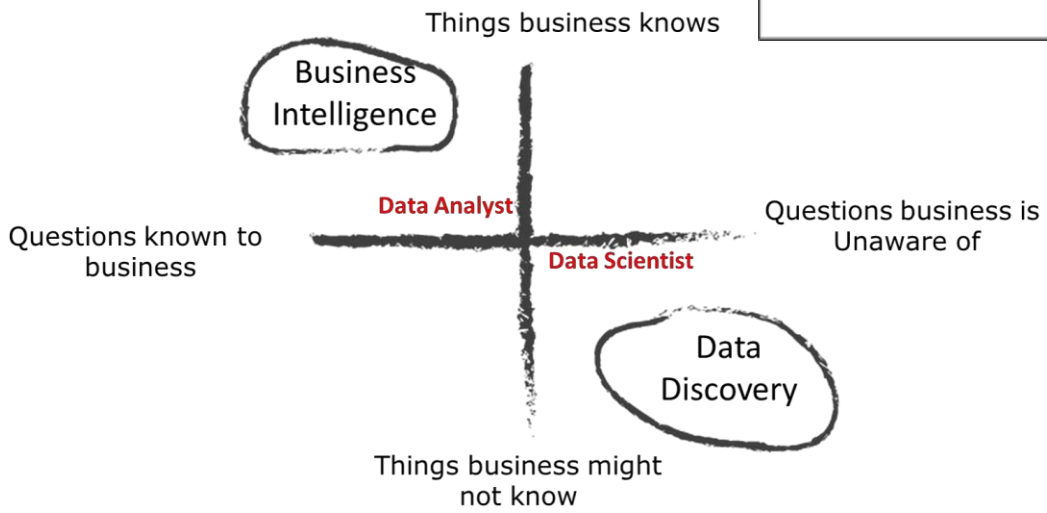
- Data analytics and data science are closely related.
- Data analytics is a component of data science, used to understand what an organization's data looks like.
- Generally, the output of data analytics are reports and visualizations.
- Data science takes the output of analytics to study and solve problems.
- The difference between data analytics and data science is often seen as one of timescale.
- Data analytics describes the current or historical state of reality, whereas data science uses that data to predict and/or understand the future.



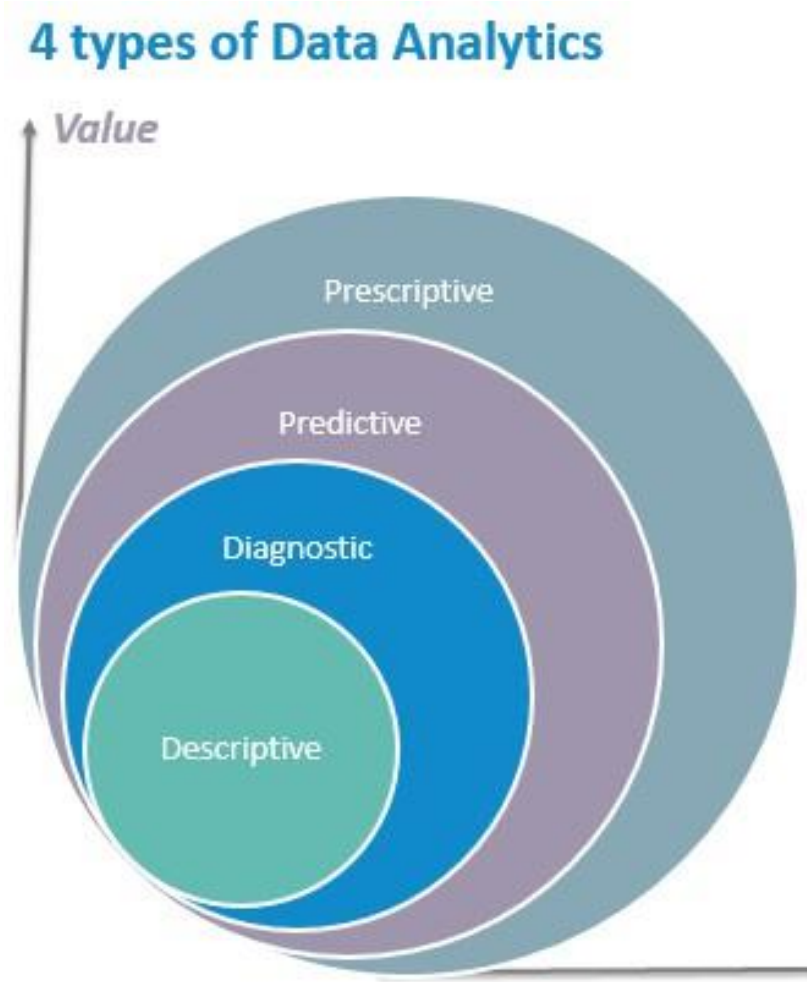
DATA ANALYTICS VS. BUSINESS ANALYTICS

- Business analytics is another subset of data analytics.
- Business analytics uses data analytics techniques, including data mining, statistical analysis, and predictive modeling, to drive better business decisions.
- Gartner defines business analytics as “solutions used to build analysis models and simulations to create scenarios, understand realities, and predict future states.”





TYPES OF DATA ANALYTICS



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

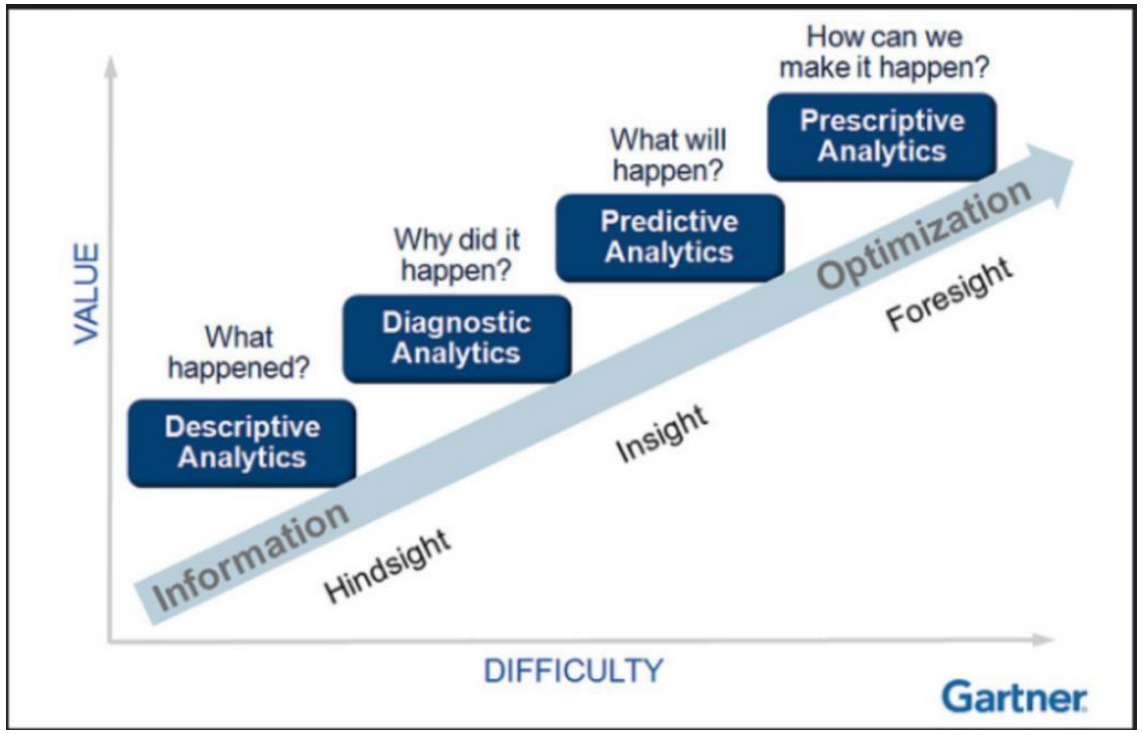
Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

TYPES OF DATA ANALYTICS

1.Descriptive analytics: What has happened and what is happening right now? Descriptive analytics uses historical and current data from multiple sources to describe the present state by identifying trends and patterns. In business analytics, this is the purview of business intelligence (BI).

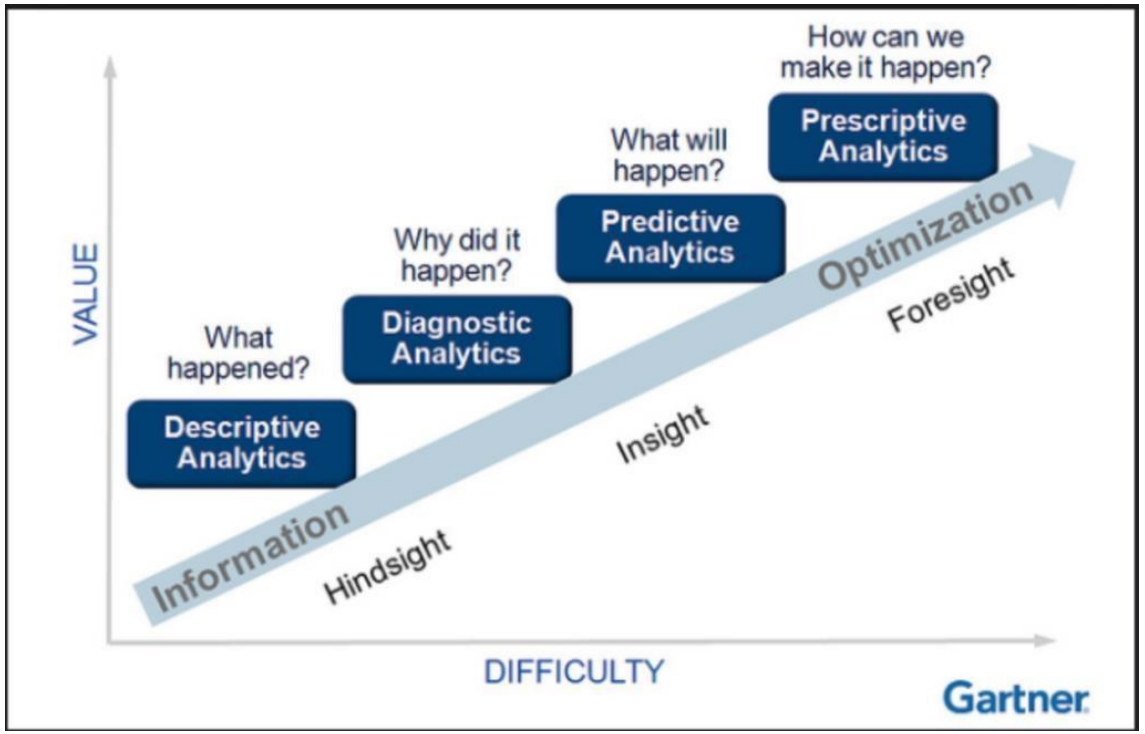
2.Diagnostic analytics: Why is it happening? Diagnostic analytics uses data (often generated via descriptive analytics) to discover the factors or reasons for past performance.



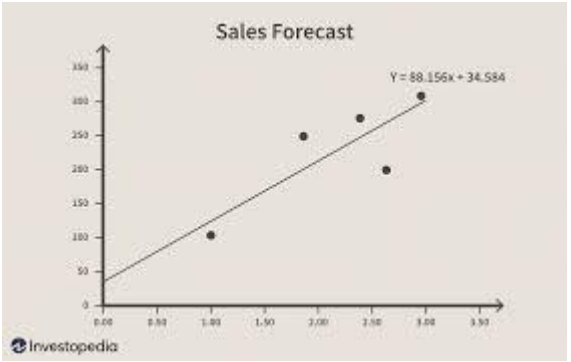
TYPES OF DATA ANALYTICS

3. **Predictive analytics:** What is likely to happen in the future? Predictive analytics applies techniques such as statistical modeling, forecasting, and machine learning to the output of descriptive and diagnostic analytics to make predictions about future outcomes. Predictive analytics is often considered a type of “advanced analytics,” and frequently depends on machine learning and/or deep learning.

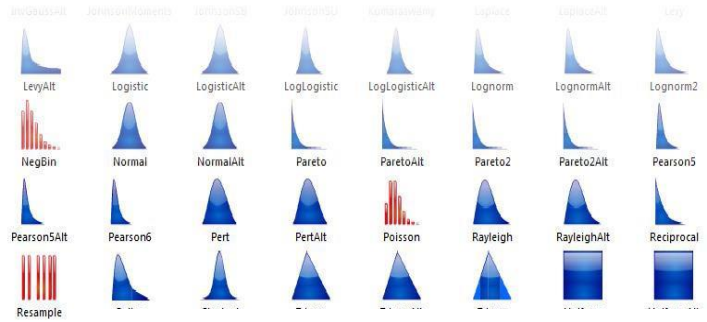
4.Prescriptive analytics: What do we need to do? Prescriptive analytics is a type of advanced analytics that involves the application of testing and other techniques to recommend specific solutions that will deliver desired outcomes. In business, predictive analytics uses machine learning, business rules, and algorithms.



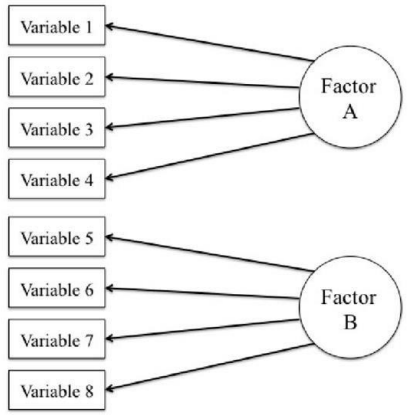
DATA ANALYTICS METHODS AND TECHNIQUES



1.Regression analysis: Regression analysis is a set of statistical processes used to estimate the relationships between variables to determine how changes to one or more variables might affect another. For example, how might social media spending affect sales?



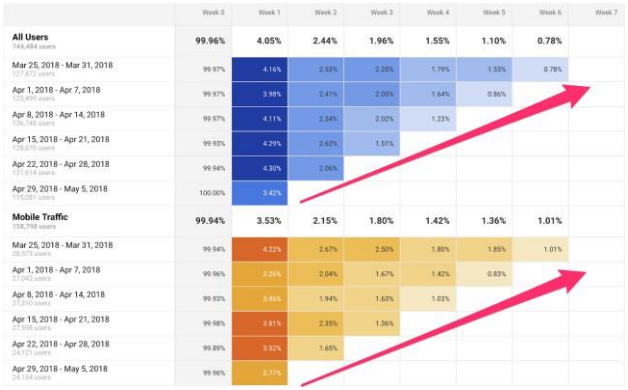
2.Monte Carlo simulation: “Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables.” It is frequently used for risk analysis.



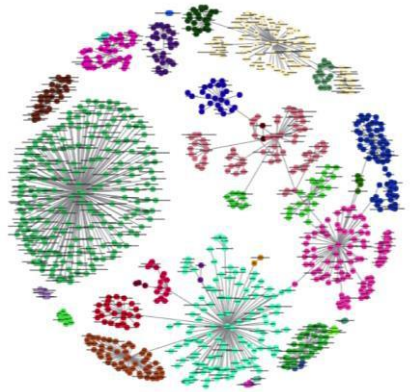
3.Factor analysis: Factor analysis is a statistical method for taking a massive data set and reducing it to a smaller, more manageable one. This has the added benefit of often uncovering hidden patterns. In a

business setting, factor analysis is often used to explore things like customer loyalty.

DATA ANALYTICS METHODS AND TECHNIQUES

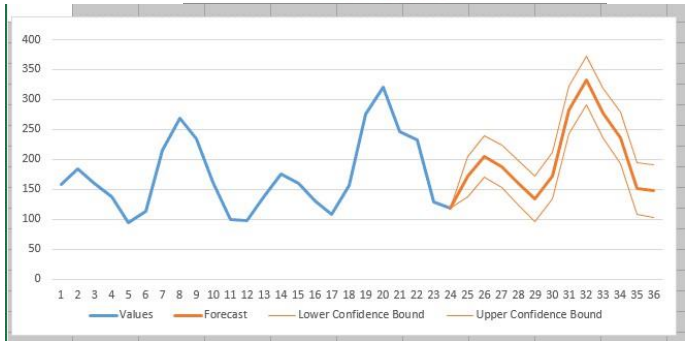


3.Cohort analysis: Cohort analysis is used to break a dataset down into groups that share common characteristics, or cohorts, for analysis. This is often used to understand customer segments.



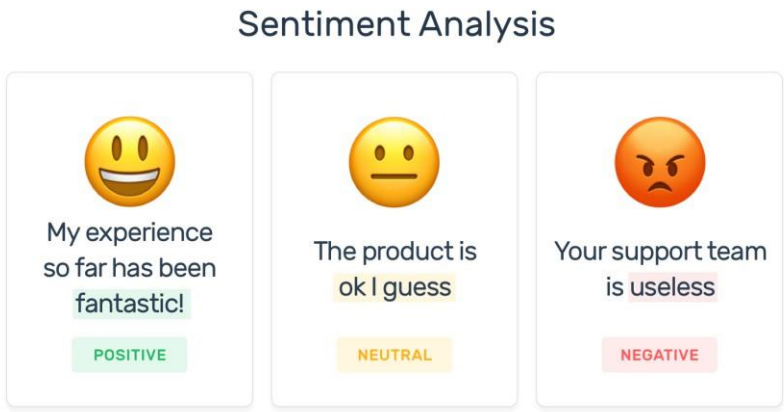
4.Cluster analysis: Cluster analysis as “a class of techniques that are used to classify objects or cases into relative groups called clusters.” It can be used to reveal structures in data — insurance firms might use cluster analysis to investigate why certain locations are associated with particular insurance claims, for instance.

5.Time series analysis: Time series analysis as “a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals. Time series analysis can be used to identify trends and cycles over time, e.g., weekly sales numbers. It is frequently used for economic and sales forecasting.



DATA ANALYTICS METHODS AND TECHNIQUES

6. **Sentiment analysis:** Sentiment analysis uses tools such as natural language processing, text analysis, computational linguistics, and so on, to understand the feelings expressed in the data. While the previous six methods seek to analyze quantitative data (data that can be measured), sentiment analysis seeks to interpret and classify qualitative data by organizing it into themes. It is often used to understand how customers feel about a brand, product, or service.



BIG DATA ANALYTICS

- Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions.
- Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.
- Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable.
- Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing

HOW BIG DATA ANALYTICS WORKS

Big data analytics refers to collecting, processing, cleaning, and analyzing large datasets to help organizations operationalize their big data.

1. Collect Data

- Data collection looks different for every organization. With today's technology, organizations can gather both structured and unstructured data from a variety of sources — from cloud storage to mobile applications to in-store IoT sensors and beyond.
- Some data will be stored in data warehouses where business intelligence tools and solutions can access it easily.
- Raw or unstructured data that is too diverse or complex for a warehouse may be assigned metadata and stored in a data lake.

HOW BIG DATA ANALYTICS WORKS

2. Process Data

- Once data is collected and stored, it must be organized properly to get accurate results on analytical queries, especially when it's large and unstructured.
- Available data is growing exponentially, making data processing a challenge for organizations.
- One processing option is **batch processing**, which looks at large data blocks over time.
- Batch processing is useful when there is a longer turnaround time between collecting and analyzing data.
- **Stream processing** looks at small batches of data at once, shortening the delay time between collection and analysis for quicker decision-making.
- Stream processing is more complex and often more expensive.

3. Clean Data

- Data big or small requires scrubbing to improve data quality and get stronger results; all data must be formatted correctly, and any duplicative or irrelevant data must be eliminated or accounted for.
- Dirty data can obscure and mislead, creating flawed insights.

HOW BIG DATA ANALYTICS WORKS

4. Analyze Data

Getting big data into a usable state takes time. Once it’s ready, advanced analytics processes can turn big data into big insights. Some of these big data analysis methods include:

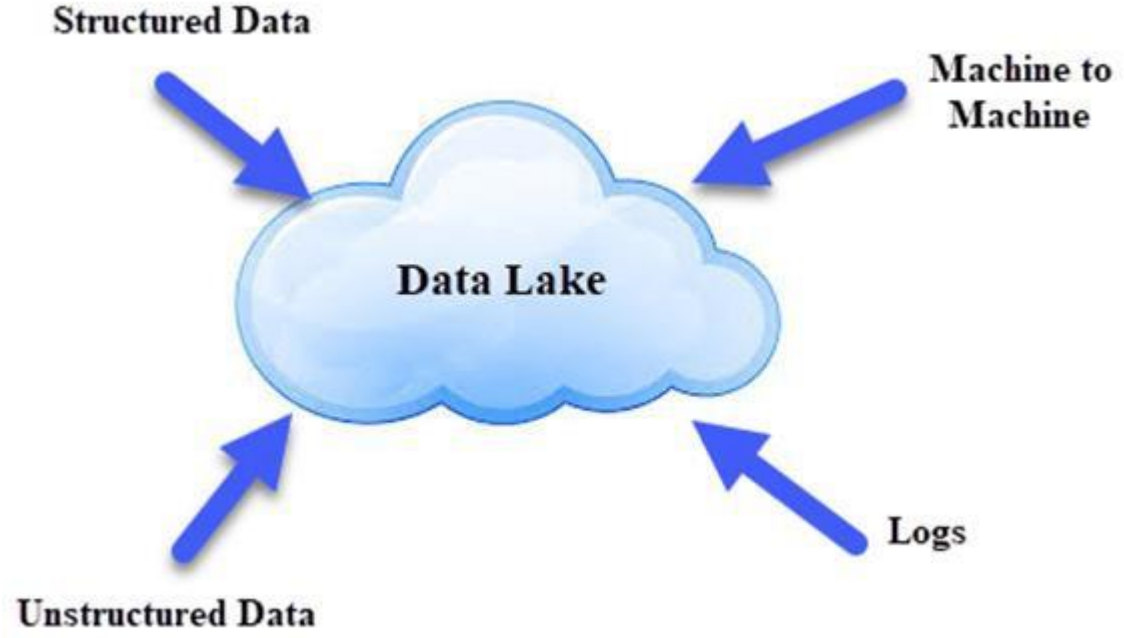
- **Data mining** sorts through large datasets to identify patterns and relationships by identifying anomalies and creating data clusters.
- **Predictive analytics** uses an organization’s historical data to make predictions about the future, identifying upcoming risks and opportunities.
- **Deep learning** imitates human learning patterns by using artificial intelligence and machine learning to layer algorithms and find patterns in the most complex and abstract data.

WHAT IS DATA LAKE?

- A Data Lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data.
- It is a place to store every type of data in its native format with no fixed limits on account size or file.
- It offers high data quantity to increase analytic performance and native integration.
- Data Lake is like a large container which is very similar to real lake and rivers.
- Just like in a lake you have multiple tributaries coming in, a data lake has structured data, unstructured data, machine to machine, logs flowing through in real-time.

WHAT IS DATA LAKE?

- The Data Lake democratizes data and is a cost-effective way to store all data of an organization for later processing.
- Research Analyst can focus on finding meaning patterns in data and not data itself.
- Unlike a hierarchal Dataware house where data is stored in Files and Folder, Data lake has a flat architecture. Every data elements in a Data Lake is given a unique identifier and tagged with a set of metadata information.



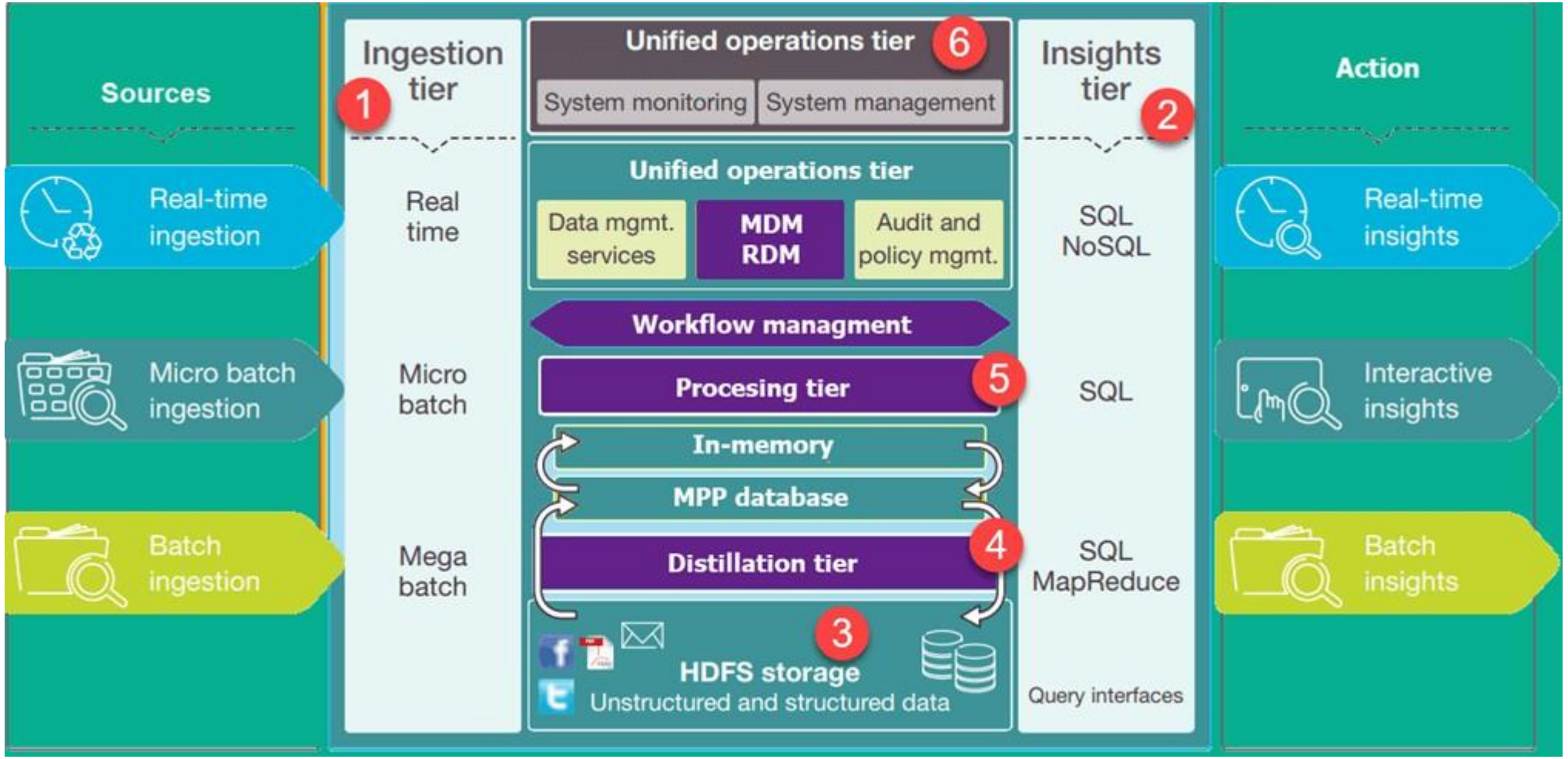
WHY DATA LAKE?

The main objective of building a data lake is to offer an unrefined view of data to data scientists.

Reasons for using Data Lake are:

- With the onset of storage engines like Hadoop storing disparate information has become easy. There is no need to model data into an enterprise-wide schema with a Data Lake.
- With the increase in data volume, data quality, and metadata, the quality of analyses also increases.
- Data Lake offers business Agility
- Machine Learning and Artificial Intelligence can be used to make profitable predictions.
- It offers a competitive advantage to the implementing organization.
- There is no data silo structure. Data Lake gives 360 degrees view of customers and makes analysis more robust.

DATA LAKE ARCHITECTURE



DATA LAKE ARCHITECTURE

The figure shows the architecture of a Business Data Lake. The lower levels represent data that is mostly at rest while the upper levels show real-time transactional data. This data flow through the system with no or little latency.

Following are important tiers in Data Lake Architecture:

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

KEY DATA LAKE CONCEPTS

Following are Key Data Lake concepts that one needs to understand to completely understand the Data Lake Architecture

KEY DATA LAKE CONCEPTS

- Data Ingestion

  Data Ingestion allows connectors to get data from a different data sources and load into the Data lake.

  Data Ingestion supports:

  1. All types of Structured, Semi-Structured, and Unstructured data.
  2. Multiple ingestions like Batch, Real-Time, One-time load.
  3. Many types of data sources like Databases, Webservers, Emails, IoT, and FTP.

- Data Storage

  Data storage should be scalable, offers cost-effective storage and allow fast access to data exploration. It should support various data formats.

- Data Governance

  Data governance is a process of managing availability, usability, security, and integrity of data used in an organization.

KEY DATA LAKE CONCEPTS

- Security

  Security needs to be implemented in every layer of the Data lake. It starts with Storage, Unearthing, and Consumption. The basic need is to stop access for unauthorized users. It should support different tools to access data with easy to navigate GUI and Dashboards.

  Authentication, Accounting, Authorization and Data Protection are some important features of data lake security.

- Data Quality:

  Data quality is an essential component of Data Lake architecture. Data is used to exact business value. Extracting insights from poor quality data will lead to poor quality insights.

- Data Discovery

  Data Discovery is another important stage before you can begin preparing data or analysis. In this stage, tagging technique is used to express the data understanding, by organizing and interpreting the data ingested in the Data lake.

KEY DATA LAKE CONCEPTS

- Data Auditing

  Two major Data auditing tasks are tracking changes to the key dataset.

  1. Tracking changes to important dataset elements
  2. Captures how/ when/ and who changes to these elements.
  3. Data auditing helps to evaluate risk and compliance.

- Data Lineage

  This component deals with data's origins. It mainly deals with where it movers over time and what happens to it. It eases errors corrections in a data analytics process from origin to destination.

- Data Exploration

  It is the beginning stage of data analysis. It helps to identify right dataset is vital before starting Data Exploration.

  All given components need to work together to play an important part in Data lake building easily evolve and explore the environment.

MATURITY STAGES OF DATA LAKE

Stage 1: Handle and ingest data at scale

This first stage of Data Maturity Involves improving the ability to transform and analyze data. Here, business owners need to find the tools according to their skillset for obtaining more data and build analytical applications.

Stage 2: Building the analytical muscle

This is a second stage which involves improving the ability to transform and analyze data. In this stage, companies use the tool which is most appropriate to their skillset. They start acquiring more data and building applications. Here, capabilities of the enterprise data warehouse and data lake are used together.

MATURITY STAGES OF DATA LAKE

Stage 3: EDW and Data Lake work in unison

This step involves getting data and analytics into the hands of as many people as possible. In this stage, the data lake and the enterprise data warehouse start to work in a union. Both playing their part in analytics

Stage 4: Enterprise capability in the lake

In this maturity stage of the data lake, enterprise capabilities are added to the Data Lake. Adoption of information governance, information lifecycle management capabilities, and Metadata management. However, very few organizations can reach this level of maturity, but this tally will increase in the future.

BEST PRACTICES FOR DATA LAKE IMPLEMENTATION

- Architectural components, their interaction and identified products should support native data types
- Design of Data Lake should be driven by what is available instead of what is required. The schema and data requirement is not defined until it is queried
- Design should be guided by disposable components integrated with service API.
- Data discovery, ingestion, storage, administration, quality, transformation, and visualization should be managed independently.
- The Data Lake architecture should be tailored to a specific industry. It should ensure that capabilities necessary for that domain are an inherent part of the design
- Faster on-boarding of newly discovered data sources is important
- Data Lake helps customized management to extract maximum value
- The Data Lake should support existing enterprise data management techniques and methods

END OF UNIT 1