AJAY KUMAR GARG ENGINEERING COLLEGE, GHAZIABAD

Subject: Data Analytics (BCS-052)

Complete Unit-1 Notes

Topic: Data Analytics and its Lifecycle

Delivered by:

Mr. Updesh Kumar Jaiswal

Assistant Professor,

Department of CSE,

AKGEC, Ghaziabad.

Unit-1 (Syllabus) according to AKTU

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 Lifecycle: Need, key roles for successful analytic projects, various phases of data analytics lifecycle – discovery, data preparation, model planning, model building, communicating results, operationalization.

Data Analytics

Definition:

"Data Analytics is the scientific method of analyzing raw data to extract meaningful information, identify patterns, and aid decision-making across business, science, and society."

- **Data Analytics** is the process of examining, cleaning, transforming, and interpreting data to discover useful patterns, trends, and insights that support better decision-making.

- In short, we can say that data analytics is the process of manipulating data to extract useful trends and hidden patterns that can help us derive valuable insights to make business predictions.

Data Analytics con…

- **Data analytics** is an important field that involves the process of collecting, processing, and interpreting data to uncover insights and help in making decisions.

- **Data analytics** is the practice of examining raw data to identify trends, draw conclusions, and extract meaningful information. This involves various techniques and tools to process and transform data into valuable insights that can be used for decision-making.

- **Data analytics** encompasses a wide array of techniques for analyzing data to gain valuable insights that can enhance various aspects of operations.

Data Analytics Process

"Data Analytics Process is the systematic sequence of steps used to collect, clean, transform, analyze, and visualize data in order to extract meaningful insights and support informed decision-making."

Data Analytics Process has following:

1- Data Collection – Gather data from databases, sensors, social media, etc.

2- Data Cleaning & Preparation – Remove errors, duplicates, and format data.

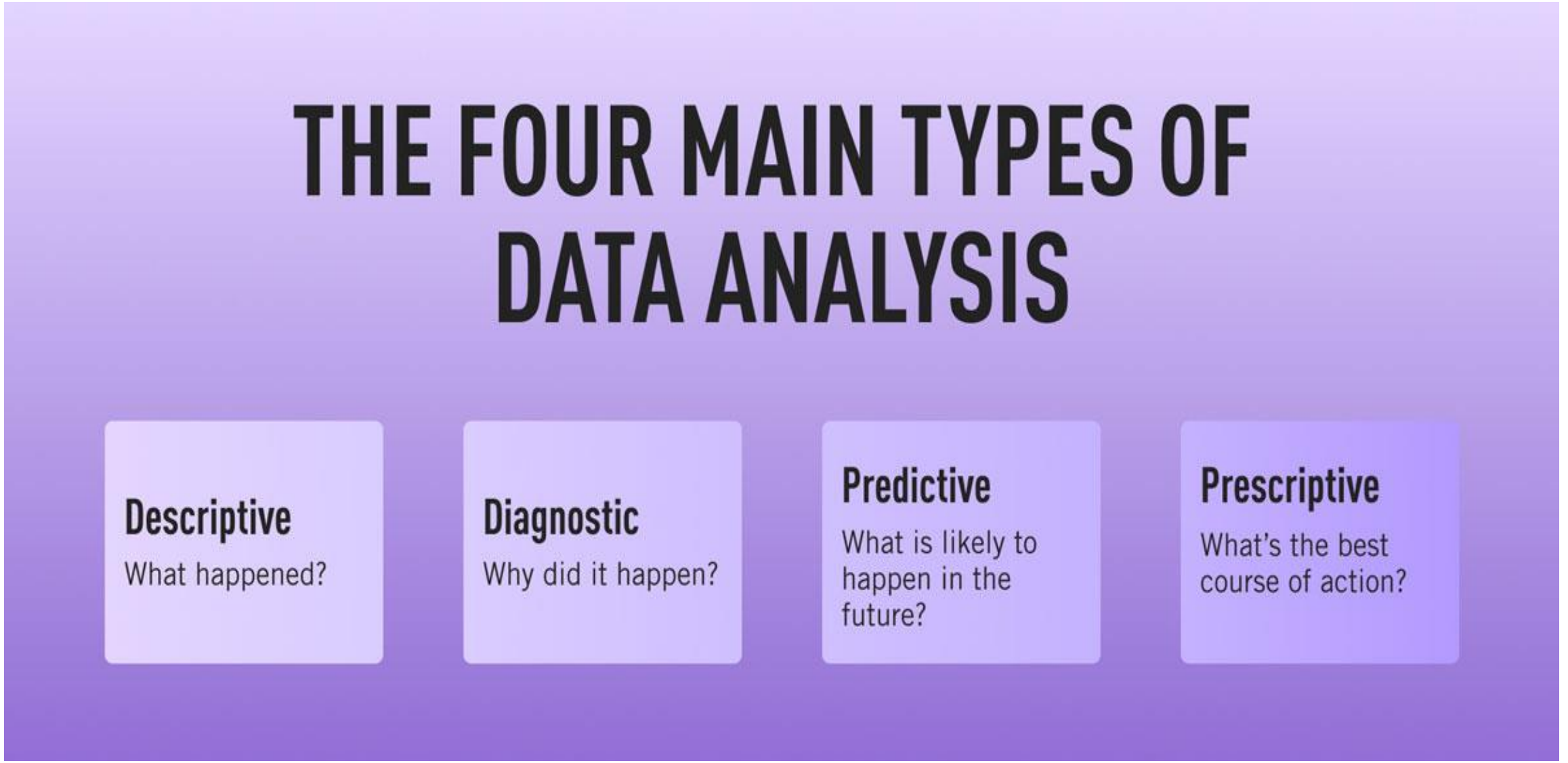
3- Data Transformation – Convert raw data into usable form for analysis.

4- Data Analysis – Apply statistical, ML, or AI techniques to find patterns and trends.

5- Data Visualization – Present insights using charts, graphs, or dashboards.

6- Decision-Making – Use insights to make informed and effective decisions.

Types of Data Analytics



1- Descriptive analytics

- **Descriptive analytics** is a simple, surface-level type of analysis that looks at **what has happened in the past.**
- The two main techniques used in descriptive analytics are **data aggregation** and **data mining**—so, the data analyst first gathers the data and presents it in a summarized format (that's the aggregation part) and then "mines" the data to discover patterns.
- The data is then presented in a way that can be easily understood by a wide audience (not just data experts).
- It's important to note that descriptive analytics doesn't try to explain the historical data or establish cause-and-effect relationships;
- At this stage, it's simply a case of determining and describing the "what".

2- Diagnostic analytics

- While descriptive analytics looks at the “what”, **diagnostic analytics explores the “why”.**
- When running diagnostic analytics, data analysts will first seek to identify anomalies within the data—i.e. anything that cannot be explained by the data in front of them.
- For example: If the data shows that there was a sudden drop in sales for the month of March, the data analyst will need to investigate the cause.
- To do this, they perform discovery phase, and identify any additional data sources that might tell them more about why such anomalies arose.
- Finally, the data analyst will try to uncover causal relationships—for example, looking at any events that may correlate or correspond with the decrease in sales.
- At this stage, data analysts may use probability theory, regression analysis, filtering, and time-series data analytics.

3- Predictive analytics

- Just as the name suggests, **predictive analytics tries to predict what is likely to happen in the future.**
- This is where data analysts start to come up with actionable, data-driven insights that the company can use to inform/perform their next steps.
- Predictive analytics estimates the likelihood of a future outcome based on historical data and probability theory, and while it can never be completely accurate, it does eliminate much of the guesswork from key business decisions.
- Predictive analytics can be used to forecast all sorts of outcomes—from what products will be most popular at a certain time, to how much the company revenue is likely to increase or decrease in a given period.
- Ultimately, predictive analytics is used to increase the business's chances of "hitting the mark" and taking the most appropriate action.

4- Prescriptive Analytics

- Building on predictive analytics, **prescriptive analytics advises on the actions and decisions that should be taken**.
- In other words, prescriptive analytics shows you how you can take advantage of the outcomes that have been predicted.
- When conducting prescriptive analysis, data analysts will consider a range of possible scenarios and assess the different actions the company might take.
- Prescriptive analytics is one of the more complex types of analysis, and may involve working with algorithms, machine learning, and computational modeling procedures.
- However, the effective use of prescriptive analytics can have a huge impact on the company's decision-making process and, ultimately, on the bottom line.
- The type of analysis you carry out will also depend on the kind of data you're working with. If you're not already familiar, it's worth learning about **the four levels of data measurement: nominal, ordinal, interval, ratio**.

Use of Data Analytics

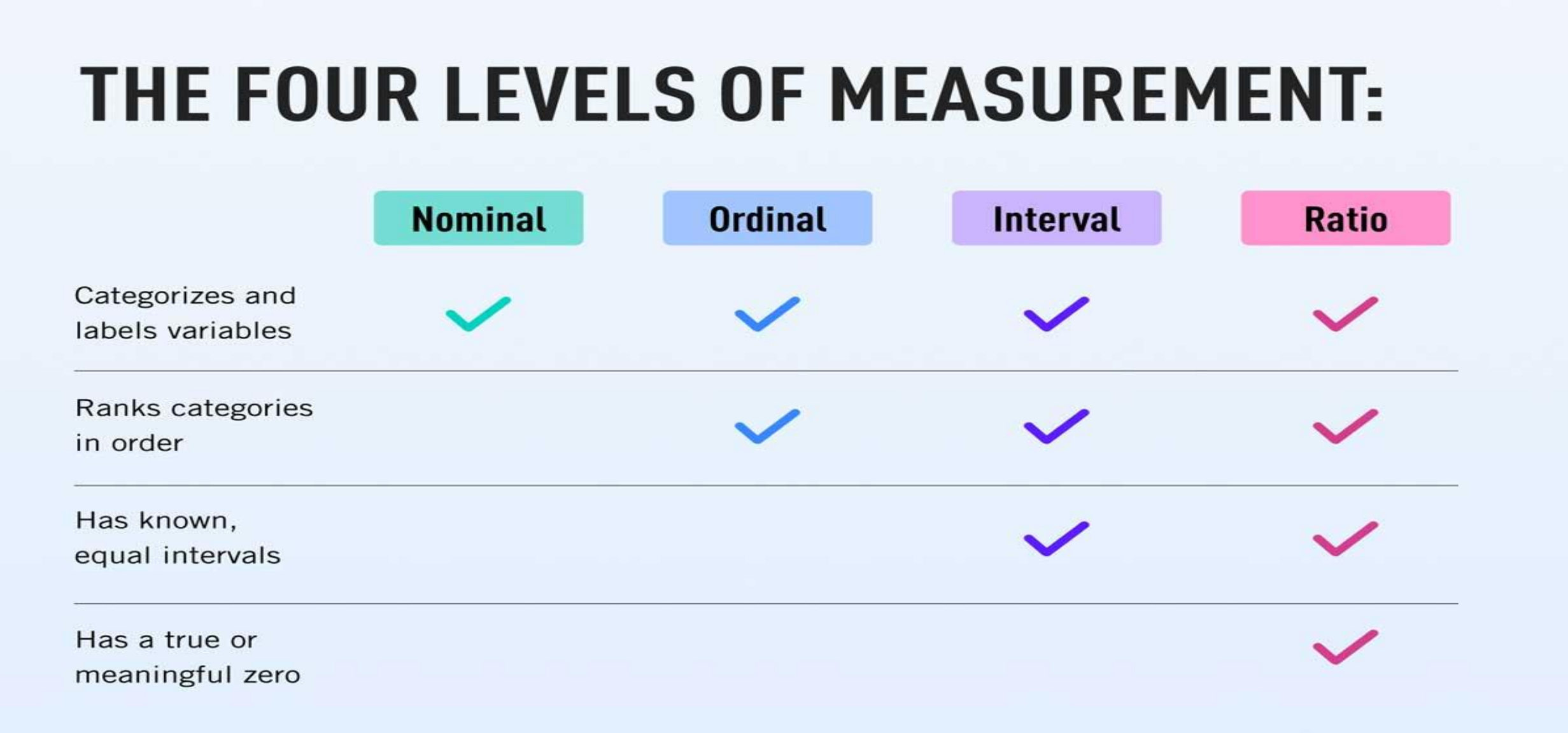
Data analytics is used to **extract meaningful insights from raw data** to support better decisions, improve efficiency, predict trends, and drive innovation.

Key uses include:

1. **Business Decision-Making:** Optimize strategies, improve customer experience, and increase profits.
2. **Predictive Analysis:** Forecast sales, demand, stock prices, or weather.
3. **Risk Management:** Detect fraud, reduce errors, and manage uncertainties.
4. **Operational Efficiency:** Streamline processes, reduce costs, and improve resource allocation.
5. **Scientific & Social Benefits:** Advance healthcare, smart city planning, climate monitoring, and research.

In short: Data analytics helps organizations and individuals make informed, data-driven decisions and gain competitive advantage.

There are four types of measurement (or scales): **nominal, ordinal, interval, and ratio**.



Four Levels/Scales/Types of Measurement con…

(I)- Nominal

- The nominal scale simply categorizes variables according to qualitative labels (or names).

- These labels and groupings don't have any order or hierarchy to them, nor do they convey any numerical value.

NOMINAL DATA

Nominal data divides variables into mutually exclusive, labeled categories.

Examples

Eye color

Smartphone

Transport

How is nominal data analyzed?

Descriptive statistics: Frequency distribution and mode

Non-parametric statistical tests

(II)- Ordinal

- The ordinal scale also categorizes variables into labeled groups, and these categories have an order or hierarchy to them.
- For example, you could measure the variable “income” on an ordinal scale as follows:
  - low income
  - medium income
  - high income.
- Another example could be level of education, classified as follows:
  - high school
  - master’s degree
  - doctorate
- These are still qualitative labels (as with the nominal scale), but you can see that they follow a hierarchical order.

ORDINAL DATA

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades

Education level

Seniority level

How is ordinal data analyzed?

Descriptive statistics: Frequency distribution, mode, median, and range

Non-parametric statistical tests

Four Levels/Scales/Types of Measurement con…

(III)- Interval

- The interval scale is a numerical scale which labels and orders variables, with a known, evenly spaced interval between each of the values.

- A commonly-cited example of interval data is temperature in Fahrenheit, where the difference between 10 and 20 degrees Fahrenheit is exactly the same as the difference between, say, 50 and 60 degrees Fahrenheit.

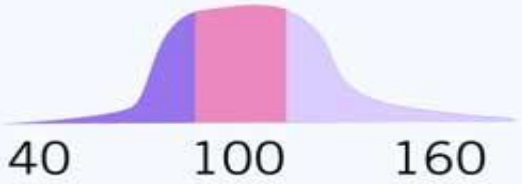
INTERVAL DATA

Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Examples

Temperature

IQ score



Income ranges

How is interval data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, and variance

Parametric statistical tests (e.g. t-test, linear regression)

(IV)- Ratio

- The ratio scale is exactly the same as the interval scale, with one key difference: The ratio scale has what's known as a "true zero."

- A good example of ratio data is weight in kilograms.

- If something weighs zero kilograms, it truly weighs nothing—compared to temperature (interval data), where a value of zero degrees doesn't mean there is "no temperature," it simply means it's extremely cold!

RATIO DATA

Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples

Weight in KG

...50 70 90...

Number of staff

...10 30 50...

Income in USD

...20k 40k 60k...

How is ratio data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, variance, and coefficient of variation

Parametric statistical tests (e.g. ANOVA, linear regression)

Classification of data (Structured, Semi-Structured, and Unstructured Data

- When we talk about data or analytics, the terms structured, unstructured, and semi-structured data often get discussed.
- These are the three forms of data that have now become relevant for all types of business applications.
- Structured data has been around for some time, and traditional systems and reporting still rely on this form of data.
- However, there has been a swift increase in the generation of semi-structured and unstructured data sources in the past few years, due to the rise of Big Data.
- As a result, more and more businesses are now looking to take their business intelligence and analytics to the next level by including all three forms of data.
- Structured, Semi-Structured, and Unstructured data are discussed in detail on coming slides.

1- Structured Data

- Structured data is information that has been formatted and transformed into a well-defined data model.
- The raw data is mapped into predesigned fields that can then be extracted and read through SQL easily.
- SQL relational databases, consisting of tables with rows and columns, are the perfect example of structured data.
- The relational model of this data format utilizes memory since it minimizes data redundancy.
- However, this also means that structured data is more inter-dependent and less flexible.

2- Semi-Structured Data

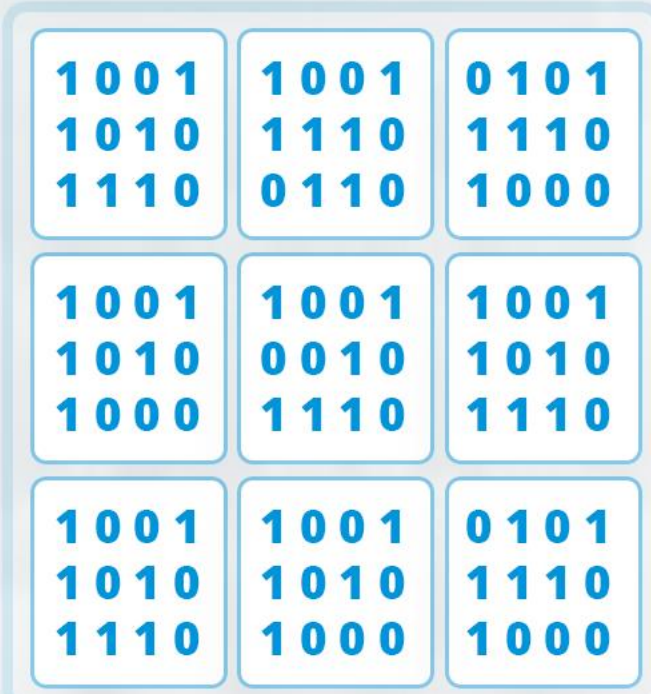
- You may not always find your data sets to be structured or unstructured. Semi-structured data or partially structured data is another category between structured and unstructured data.
- Semi-structured data is a type of data that has some consistent and definite characteristics.
- It does not confine into a rigid structure such as that needed for relational databases.
- Businesses use organizational properties like metadata or semantics tags with semi-structured data to make it more manageable.
- However, it still contains some variability and inconsistency.

3- Unstructured Data

- Unstructured data is defined as data present in absolute raw form.
- This data is difficult to process due to its complex arrangement and formatting.
- Unstructured data includes social media posts, chats, satellite imagery, IoT sensor data, emails, and presentations.
- Unstructured data management takes this data to organize it in a logical, predefined manner in data storage.
- Natural Language Processing (NLP) tools help understand unstructured data that exists in a written format.

Types of Data

Structured

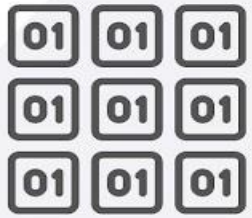


Unstructured

Semi-Structured

Astera
Enabling Data-Driven Innovation

Structured data



Characteristics

Predefined data models
Easy to search
Text-based
Shows what's happening

Resides in

Relational databases
Data warehouses

Stored in

Rows and columns

Examples

Dates, phone numbers, social security numbers, customer names, transaction info

Unstructured data

Characteristics

No predefined data models
Difficult to search
Text, pdf, images, video
Shows the why

Resides in

Applications
Data warehouses and lakes

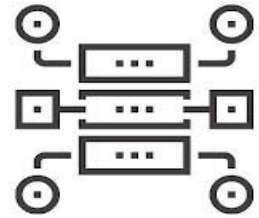
Stored in

Various forms

Examples

Documents, emails and messages, conversation transcripts, image files, open-ended survey answers

Semi-structured data



Characteristics

Loosely organized
Meta-level structure that can contain unstructured data
HTML, XML, JSON

Resides in

Relational databases
Tagged-text format

Stored in

Abstracts & figures

Examples

Server logs, tweets organized by hashtags, emails sorting by folders (inbox; sent; draft)

WHAT'S DATA?

In the subject of Data Analytics, ***data*** refers to the **raw, unprocessed facts and figures** collected from different sources, which can later be analyzed to extract useful insights.

Data act as the **foundation** of analytics because without data, no meaningful patterns, predictions, or decisions can be made.

👉 **Example 1:**

Raw sales data = Jan: 100 units, Feb: 150 units, Mar: 80 units.
After analysis → "Sales increased by 50% in February, but dropped 46% in March due to Holi Holidays."

👉 **Example 2:**

•A list of temperatures recorded every hour (43°C, 42°C, 44°C…) is **data**.
•When we analyze it and say, *"The average temperature this week was 43°C"*, it becomes **information, and** we can take a **decision** that no need to go outside in noon.

“Characteristics of Data” in data analytics

1- Volume – Refers to the large amount of data generated from multiple sources such as sensors, devices, social media, transactions, etc.

2- Variety – Data comes in different formats: structured (tables, databases), semi-structured (XML: eXtensible Markup Language; JSON: JavaScript Object Notation), and unstructured (text, images, videos).

3- Velocity – The speed at which new data is generated, collected, and processed (e.g., Real-time streaming of data means, it is continuously generated and transmitted at high speed, and needs to be processed immediately as it arrives, rather than being stored first and analyzed later, Stock market price updates, or Social media posts).

4- Veracity – Refers to the accuracy, quality, and trustworthiness of data. Noisy or incomplete data may affect analysis.

5- Value – The usefulness of data in generating insights, making decisions, and solving problems.

1- Volume

- The name Big Data itself is related to an enormous size of data.
- Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more.
- **In Facebook** approximately **4.5 billion** times the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day.
- Big data technologies can handle large amounts of data.

2- Variety

- Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources.
- Data will only be collected from **databases** and **sheets** in the past, but these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos,** etc.

3- Velocity

Definition – Velocity means the speed at which data is generated/created, collected, and processed in real-time.

Real-time data – Data often arrives in streams (e.g., stock market updates, IoT sensor readings, online transactions).

High frequency – Modern systems generate data at millisecond or nanosecond intervals (e.g., social media posts every second).

Processing speed – Analytics systems must process data quickly to avoid delays in decision-making.

Challenge – Handling data streams continuously without losing information is difficult.

Example domains – Online fraud detection, real-time traffic monitoring, health monitoring devices, and live recommendation engines.

4- Veracity

- Veracity means how much the data is reliable.
- It has many ways to filter or translate the data.
- Veracity is the process of being able to handle and manage data efficiently.
- Big Data is also essential in business development.
  - For example, **Facebook posts** with hashtags.

5- Value

- Value is an essential characteristic of big data. It is not the data that we process or store.

- It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.

Data → Information → Knowledge → Decision flow

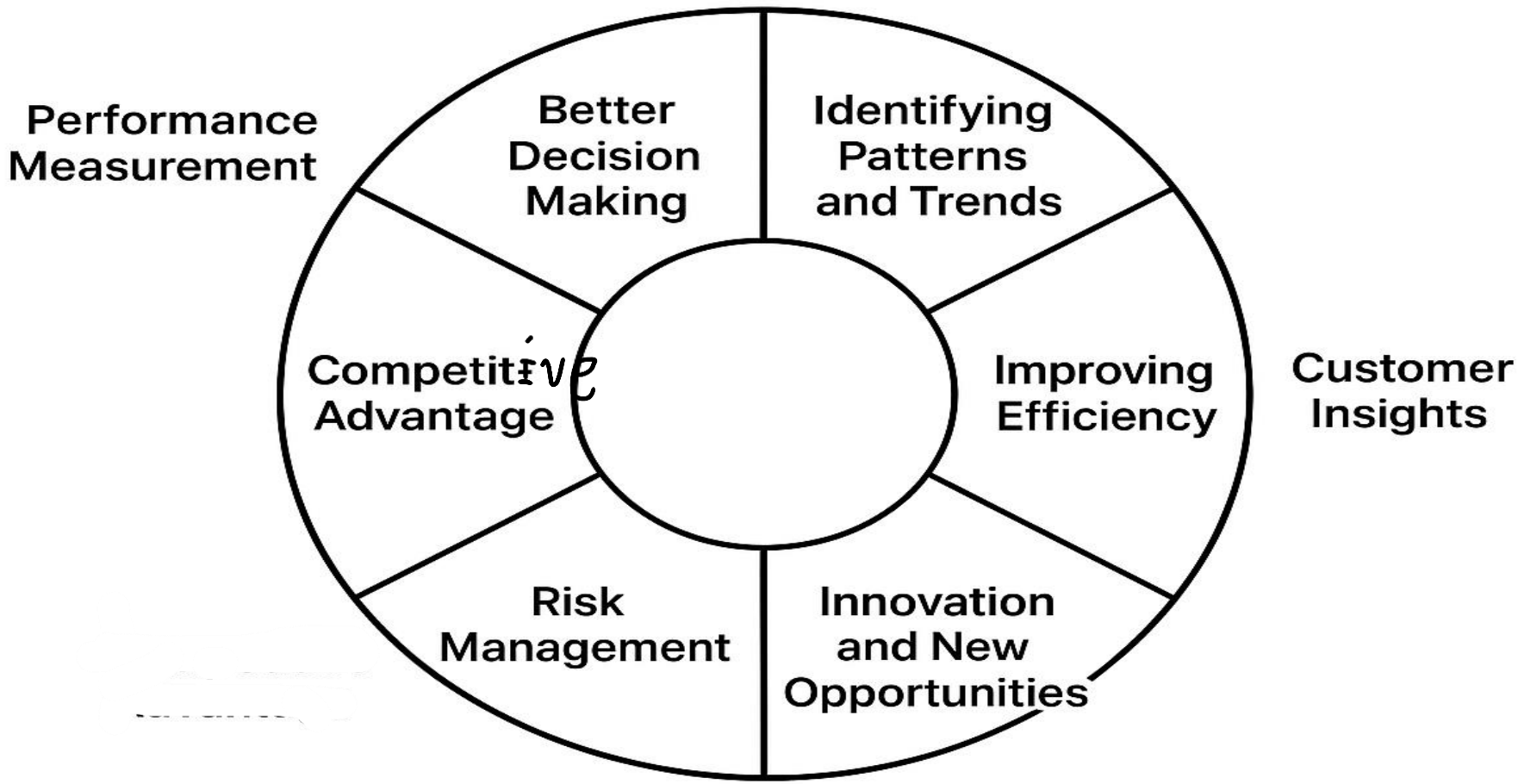
- **Data** = raw input
- **Information** = processed and meaningful
- **Knowledge** = insights & understanding
- **Decision** = action based on knowledge

This flow shows how **analytics transforms meaningless raw data into valuable decision-making power.**

Needs of Data Analytics

1) **Better Decision Making** – Helps organizations make informed, data-driven decisions instead of relying only on intuition. So, explosion of data (Big Data Era) is used for better decision making.
2) **Identifying Patterns and Trends** – Detects hidden patterns, correlations, and trends in large datasets that humans alone cannot easily spot. Data analytics provides scientific and social benefits by improving healthcare, enabling smart governance, and supporting environmental sustainability.
3) **Improving Efficiency** – Optimizes business operations by identifying bottlenecks, reducing waste, and enhancing productivity. Analytics optimizes resources, reduces wastage, and improves efficiency.
4) **Customer Insights** – Provides a deeper understanding of customer behavior, preferences, and feedback for personalized services.
5) **Innovation and New Opportunities** – Enables discovery of new products, services, or markets by analyzing unmet needs and emerging demands.
6) **Risk Management** – Assists in detecting fraud, predicting failures, and managing financial or operational risks.
7) **Competitive Advantage** – Organizations using analytics gain an edge over competitors by acting faster and smarter.
8) **Performance Measurement** – Tracks and evaluates business performance through metrics, KPIs, and dashboards.

NEEDS OF DATA ANALYTICS



Sources of Data Analytics

Data analytics relies on collecting data from multiple sources, which can be broadly classified as:

1. Transactional Data

- Generated from day-to-day business operations.
- Examples: Sales transactions, invoices, ATM withdrawals, e-commerce purchases.

2. Operational Data

- Collected from internal processes of an organization.
- Examples: Manufacturing logs, inventory data, production reports.

3. Social Media Data

- Generated by users on social platforms.
- Examples: Tweets, Facebook posts, Instagram likes, comments, shares.

Sources of Data Analytics con…

4. Machine/IoT Data

- Data from sensors, devices, and Internet of Things (IoT) systems.
- Examples: Smart meters, wearable devices, industrial sensors, connected vehicles.

5. Web & Online Data

- Data from websites, apps, and online interactions.
- Examples: Clickstream data, website traffic, app usage, online reviews.

6. Public & Government Data

- Open datasets available for research or public use.
- Examples: Census data, weather reports, economic indicators, healthcare statistics.

7. Multimedia Data

- Non-textual data used for analysis.
- Examples: Images, videos, audio, satellite imagery.

8. Third-Party & Purchased Data

- Data bought from vendors or third-party aggregators.
- Examples: Market research reports, demographic databases, financial data feeds.

Big Data

- **Big Data** refers to extremely large and complex datasets that are difficult to store, process, and analyze using traditional data management tools.

- **Big Data** means **very large and complex sets of data** that are generated every second from different sources like social media, sensors, transactions, mobile apps, and machines. Traditional databases cannot handle or analyze such massive data efficiently.

- **Big Data** refers to datasets that are **too large, too fast-changing, and too complex** to be captured, stored, managed, and analyzed using traditional database systems or software tools.

- **Big Data** requires **specialized technologies and techniques** (like Hadoop, Spark, NoSQL, cloud platforms, and machine learning) to extract meaningful insights.

Characteristics of Big Data

Key Features of Big Data (5 V's):

1. **Volume** – Huge amount of data (terabytes, petabytes, and beyond).

   Example: Facebook generates petabytes of user data daily.

2. **Velocity** – Data comes in at very high speed.

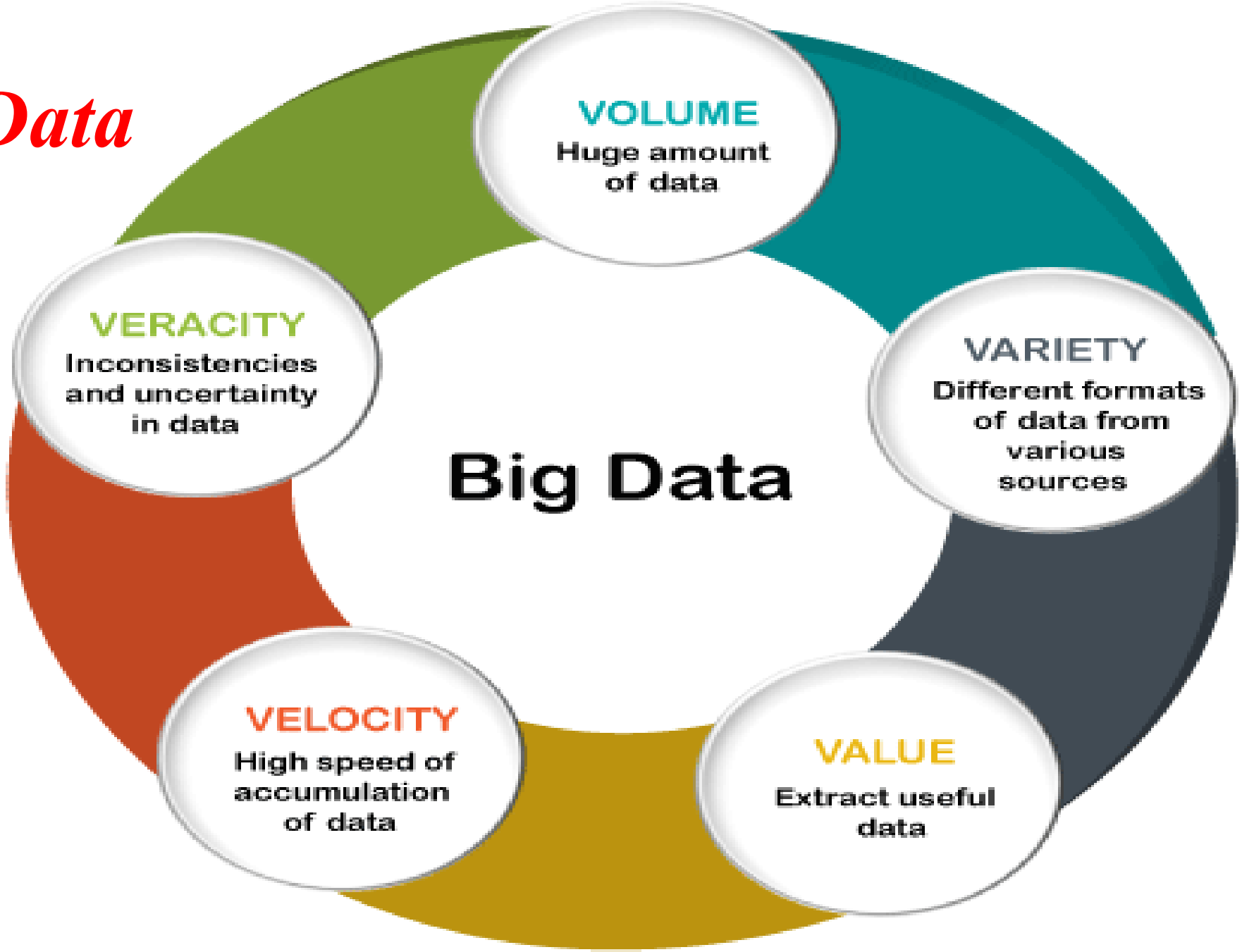
   Example: Stock market tick data, live sensor data from IoT devices.

3. **Variety** – Data comes in different forms.
   - Structured (tables, numbers)
   - Semi-structured (XML, JSON)
   - Unstructured (images, videos, text)

4. **Veracity** – Data quality may vary; some data can be noisy or uncertain.
5. **Value** – The most important aspect; useful insights and decisions can be drawn from Big Data.

5 V's of Big Data

1) Volume
2) Velocity
3) Variety
4) Veracity
5) Value



Big Data Analytics

- **Big Data analytics** is the process of collecting, organizing and analyzing large sets of data (called Big Data) to discover patterns and other useful information.
- **Big Data analytics** can help organizations to better understand the information contained within the data and will also help to identify the data that is most important in business decisions.
- **Big Data analytics** is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forecasting and data optimization.
- Using Big Data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions in the future.

Introduction to Big Data platform

👉 **Definition:** ***A Big Data Platform integrates tools, technologies, and services—such as distributed storage, parallel processing, analytics engines, and visualization tools—to efficiently manage and extract insights from large-scale and diverse data sources.***

- A **Big Data Platform** is a **software framework or ecosystem** that enables the **collection, storage, processing, management, and analysis of massive, complex, and high-velocity datasets** that traditional systems cannot handle.
- Big data platforms are comprehensive frameworks that enable organizations to store, process, and analyze vast amounts of structured and unstructured data.
- At their core, big data platforms are comprehensive ecosystems of tools, technologies, and infrastructure designed to handle the three V's of big data: volume, velocity, and variety.

Big data platform features:

- **Data storage and management**: Data storage and management is a fundamental feature of big data platforms. These platforms provide robust and scalable storage solutions for handling large volumes of structured and unstructured data.
- **Distributed processing**: Distributed processing is a crucial feature of big data platforms that enables processing large volumes of data across multiple nodes or servers in a distributed computing environment.
- **Fault tolerance**: Fault tolerance refers to the ability of a system to continue functioning even in the event of software or hardware failures.
- **Data analytics and visualization**: The big data analysis platforms offer robust tools and algorithms that can process large volumes of data in real-time or near real-time.

Big data platform features con…

- **Data Collection:** This first phase involves gathering data from multiple sources, including databases, social media, and sensors. To have this data at hand, data engineers will employ web scraping, data feeds, APIs, and data extraction tools.
- **Data Storage:** Following data collection is efficient storage for retrieval and processing. Common big data platforms rely on distributed storage systems thanks to their high availability, fault tolerance, and scalability. Some big names include Hadoop Distributed File System (HDFS), Google Cloud Storage, and Amazon S3.
- **Data Processing:** This is a critical phase in the data lifecycle, where raw data is transformed into actionable insights. This stage encompasses a series of sophisticated operations designed to refine and structure the data for analysis. These operations consist of data cleaning, transformation, and aggregation, each serving a unique purpose in preparing data for insightful analysis.

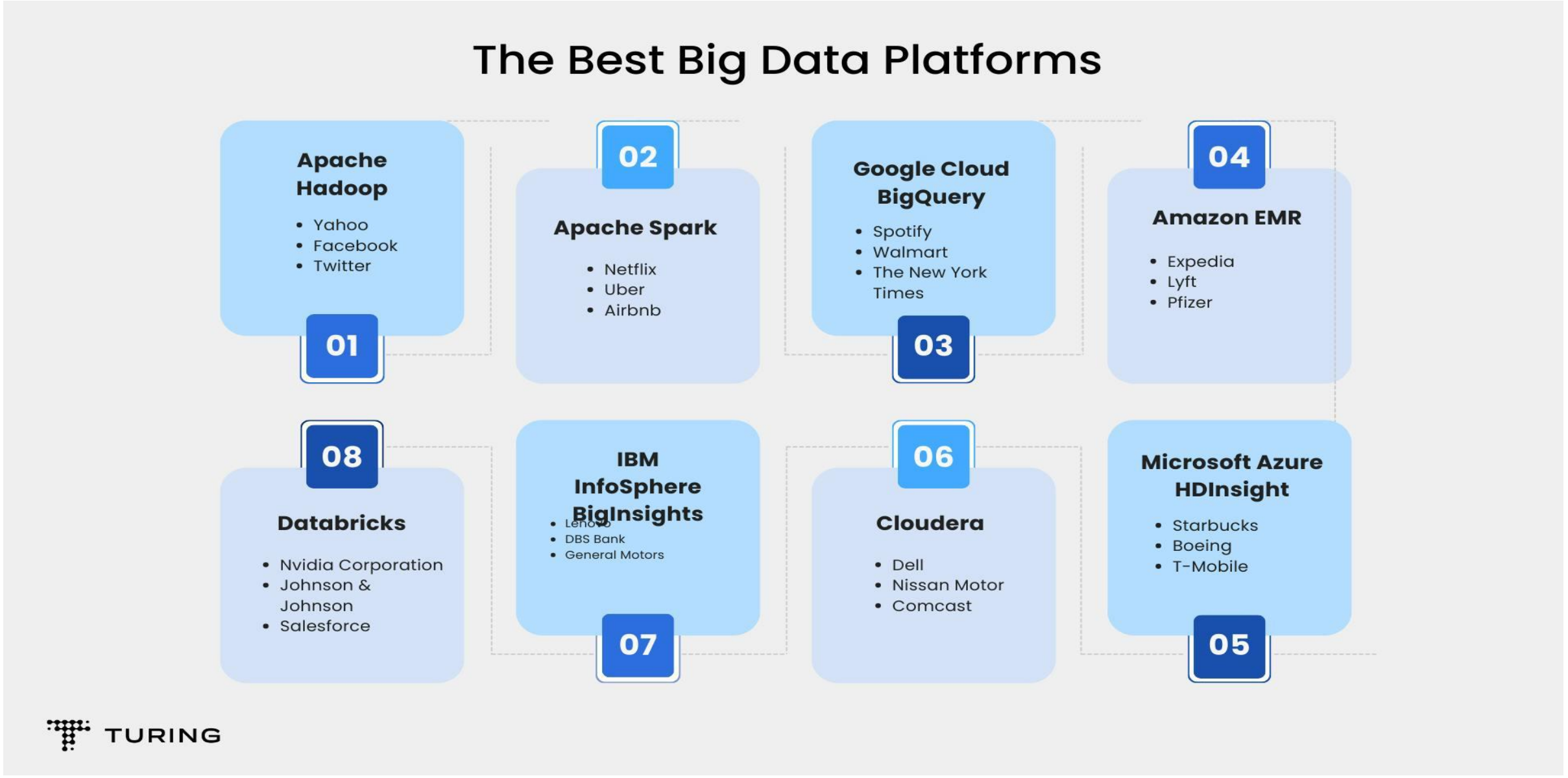
Big data platform features con…

- **Data Analysis:** This step is a pivotal phase in the big data processing pipeline, where the primary goal is to distill vast volumes of complex data into actionable insights and discernible patterns.
- This phase leverages sophisticated methodologies and technologies, including machine learning algorithms, data mining techniques, and advanced visualization tools, to unearth valuable information hidden within the data.
- **Data Quality Assurance:** Data quality assurance (DQA) comes next to ensure the reliability and effectiveness of data used across various business operations and decision-making processes.
- It encompasses a comprehensive approach to maintaining high data governance standards, accuracy, consistency, integrity, relevance, and security.
- By implementing rigorous DQA measures, organizations can significantly enhance the trustworthiness of their data, thereby improving the outcomes of their data-driven initiatives.

Big data platform features con…

- **Data Management:** Data management covers a comprehensive set of disciplines and practices dedicated to the proper handling, maintenance, and utilization of data.
- It's a critical aspect of modern organizations, given the exponential growth of data and its pivotal role in decision-making, strategic planning, and operational efficiency.

Big data platforms



TURING

Factors to Consider When Choosing Big Data Platforms:

- **Scalability:** Scalability stands out as a pivotal aspect when evaluating enterprise data platforms.
- This is because, as your data expands, the platform should seamlessly accommodate the growing volume, velocity, and variety of data without sacrificing performance.
- A scalable platform allows for the smooth expansion of your data infrastructure as your business needs evolve.

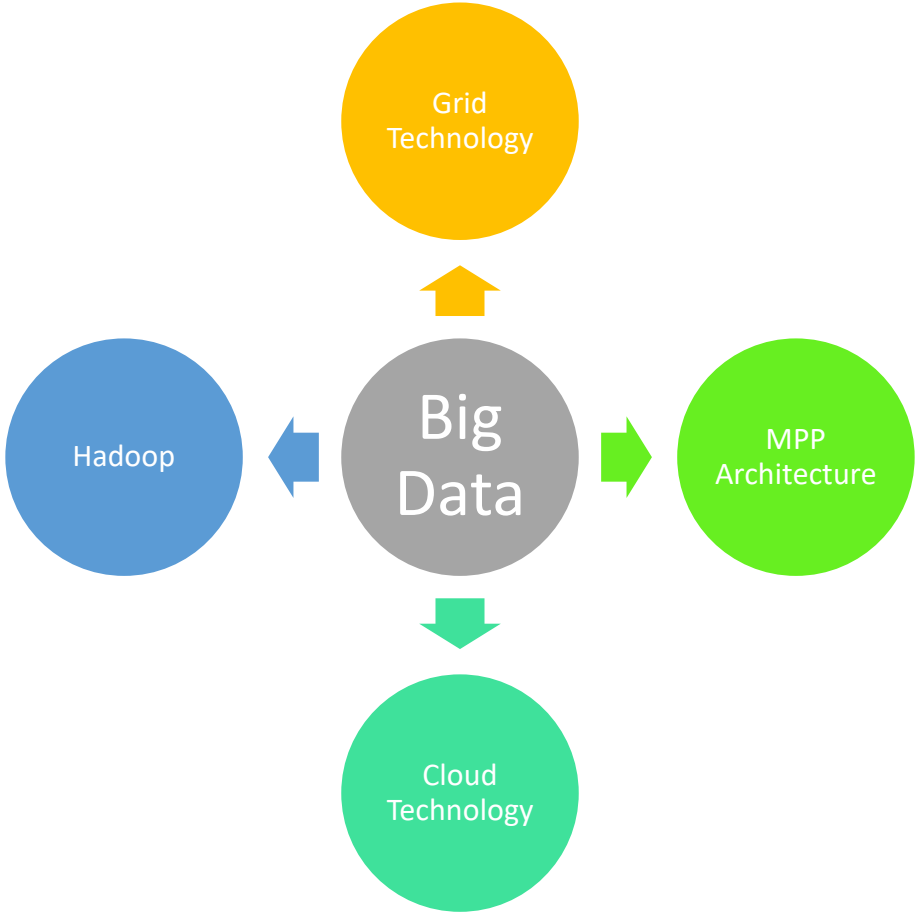
- **Performance:** Performance is another important factor in platform selection.
- Opt for big data platforms that exhibit excellent data processing speeds, efficient scaling, high fault tolerance, and minimal disruptions.

- **Data Security and Compliance:** Today's world is marked by increased risks of breaches and cyber-attacks, making data security and compliance a top priority.
- A robust security framework is thus indispensable for maintaining data integrity, protecting customer privacy, and mitigating legal and regulatory risks.

- **User-friendliness:** Ease of use is the next aspect your business should prioritize. A platform with a steep learning curve can hinder adoption and productivity, resulting in poor performance.
- Seek a platform with a highly intuitive user interface and data tools, so your team can easily navigate functions and execute business-specific tasks without extensive technical expertise.

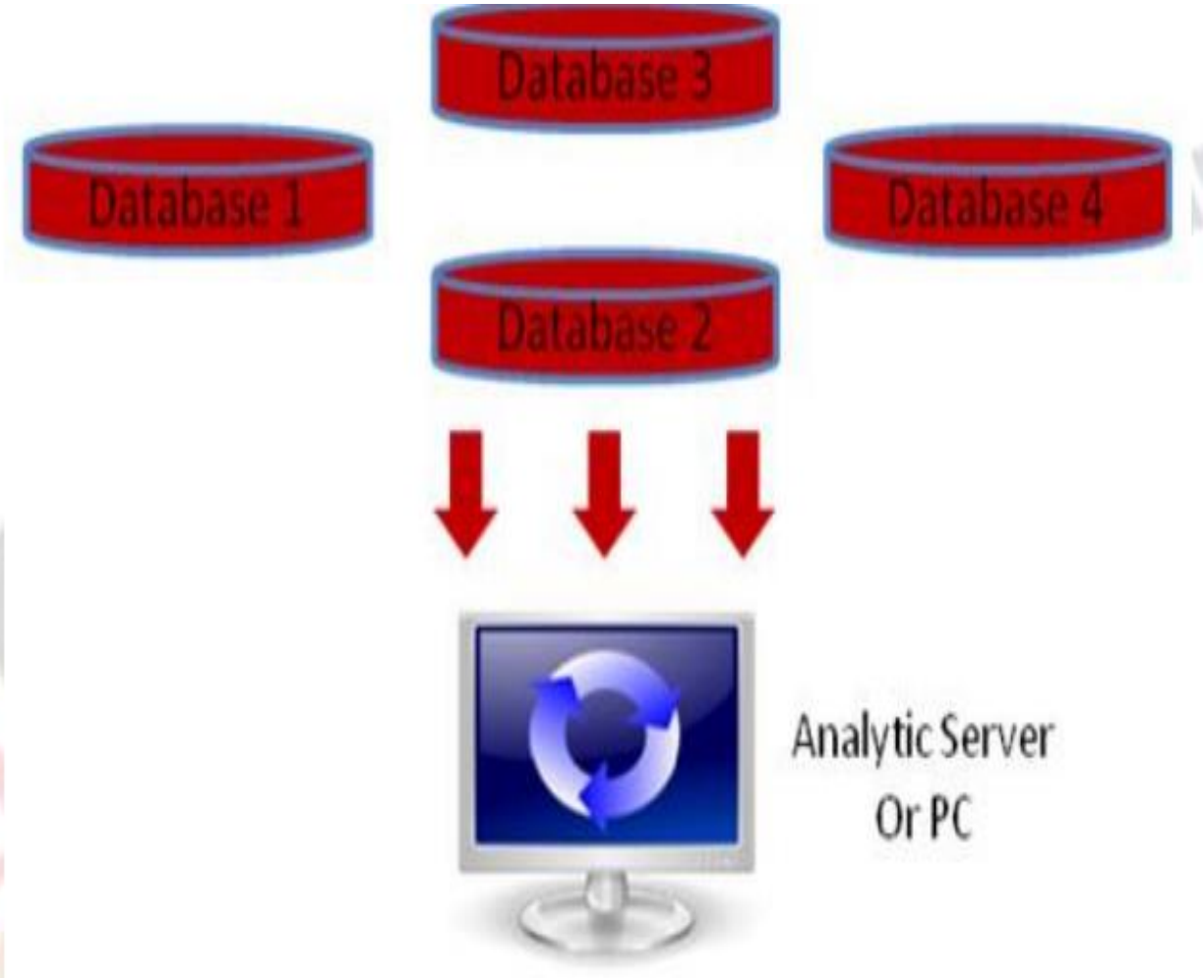
- **Integration Capabilities:** Your chosen big data platform should seamlessly integrate with your existing ecosystem, including your databases and applications.
- This integration streamlines data processing and eliminates the need for complex data migration processes.

Evolution of Analytic Scalability



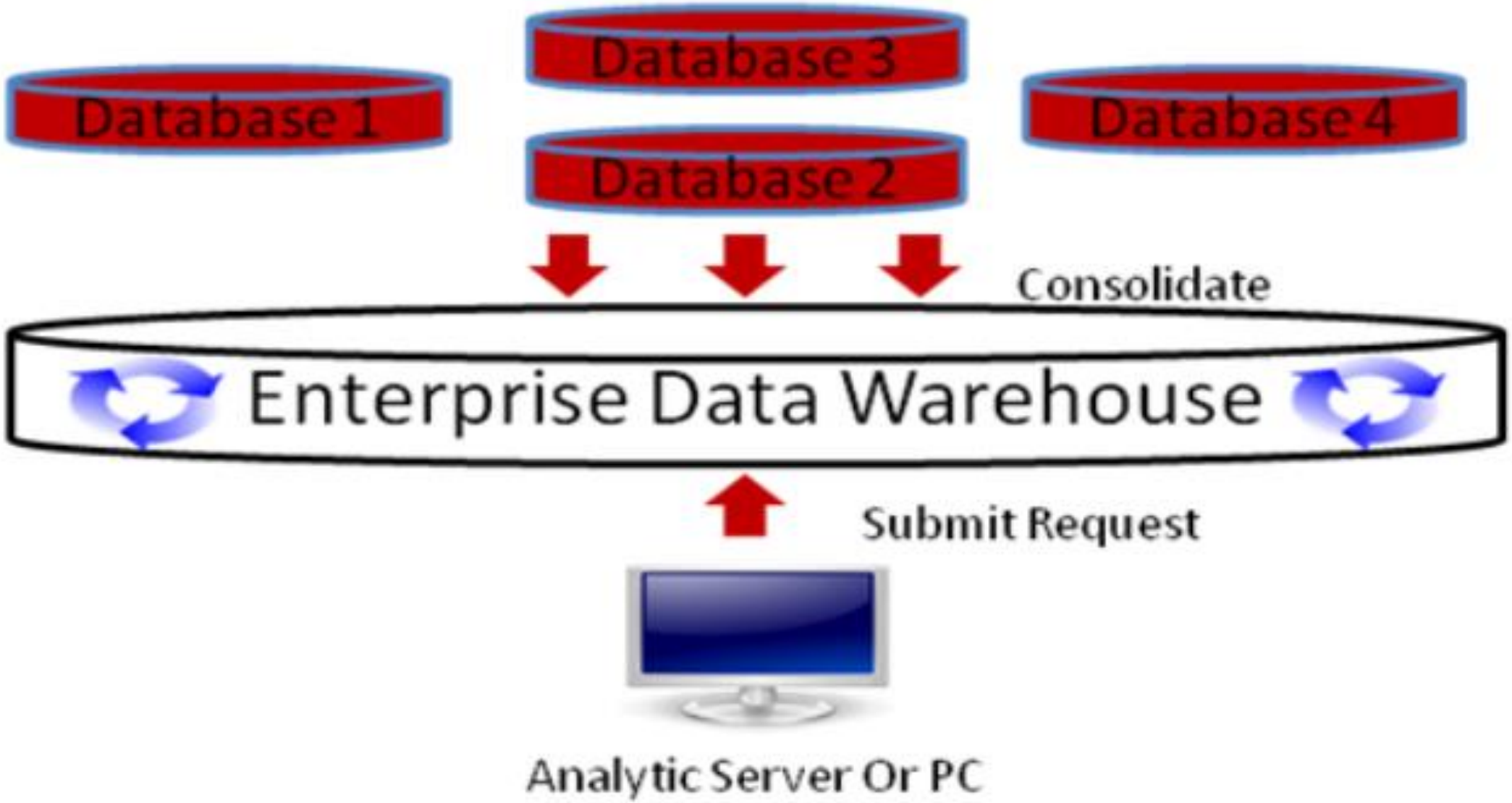
Traditional Analytic Architecture

- Traditional analytics collects data from heterogeneous data sources.
- We had to pull all data together into a separate analytics environment to do analysis which can be an analytical server or a personal computer with more computing capability.
- The heavy processing occurs in the analytic environment.
- In such environments, shipping of data becomes a must, which might result in issues related with security of data and its confidentiality.



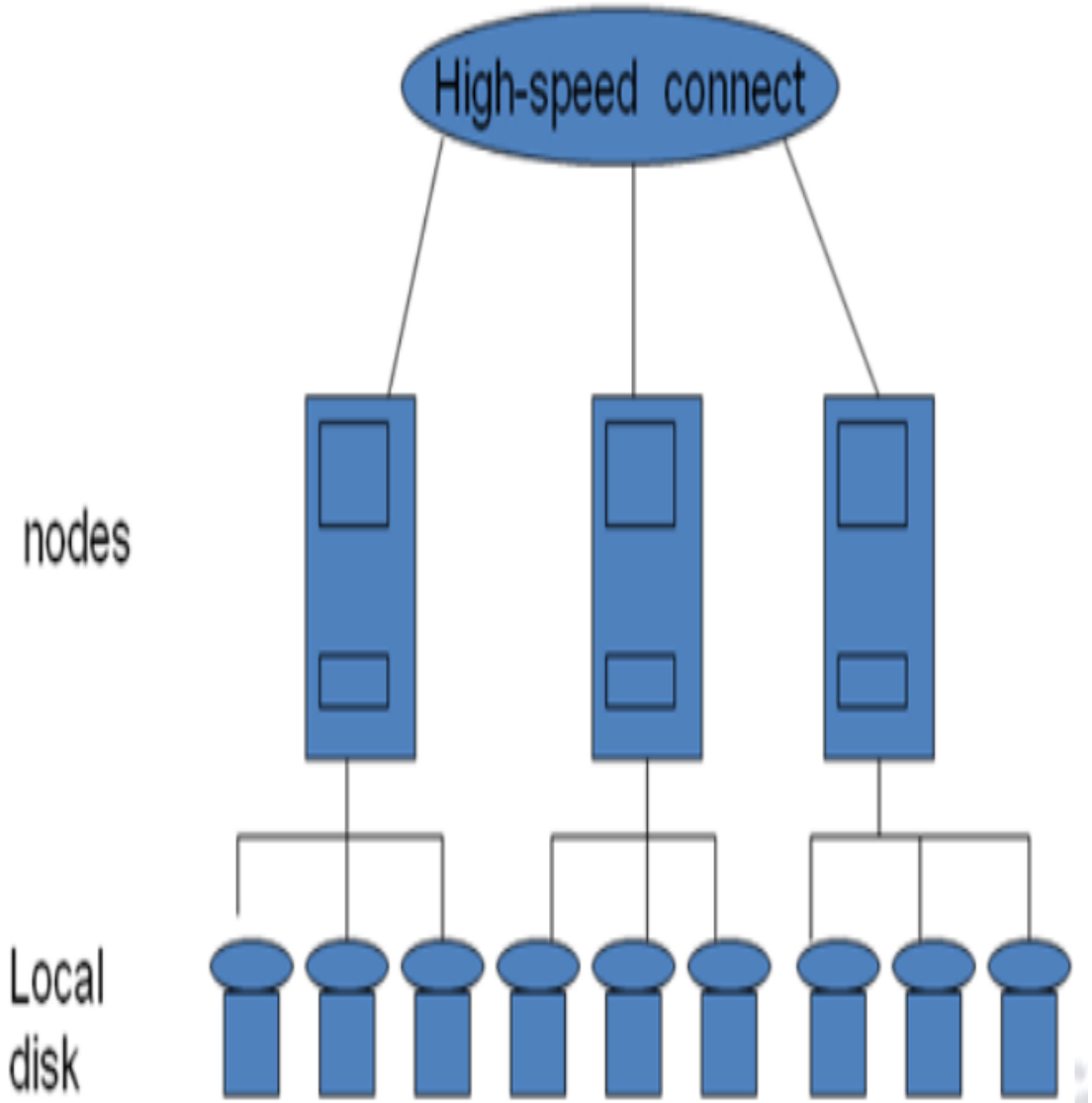
Modern In-Database Architecture

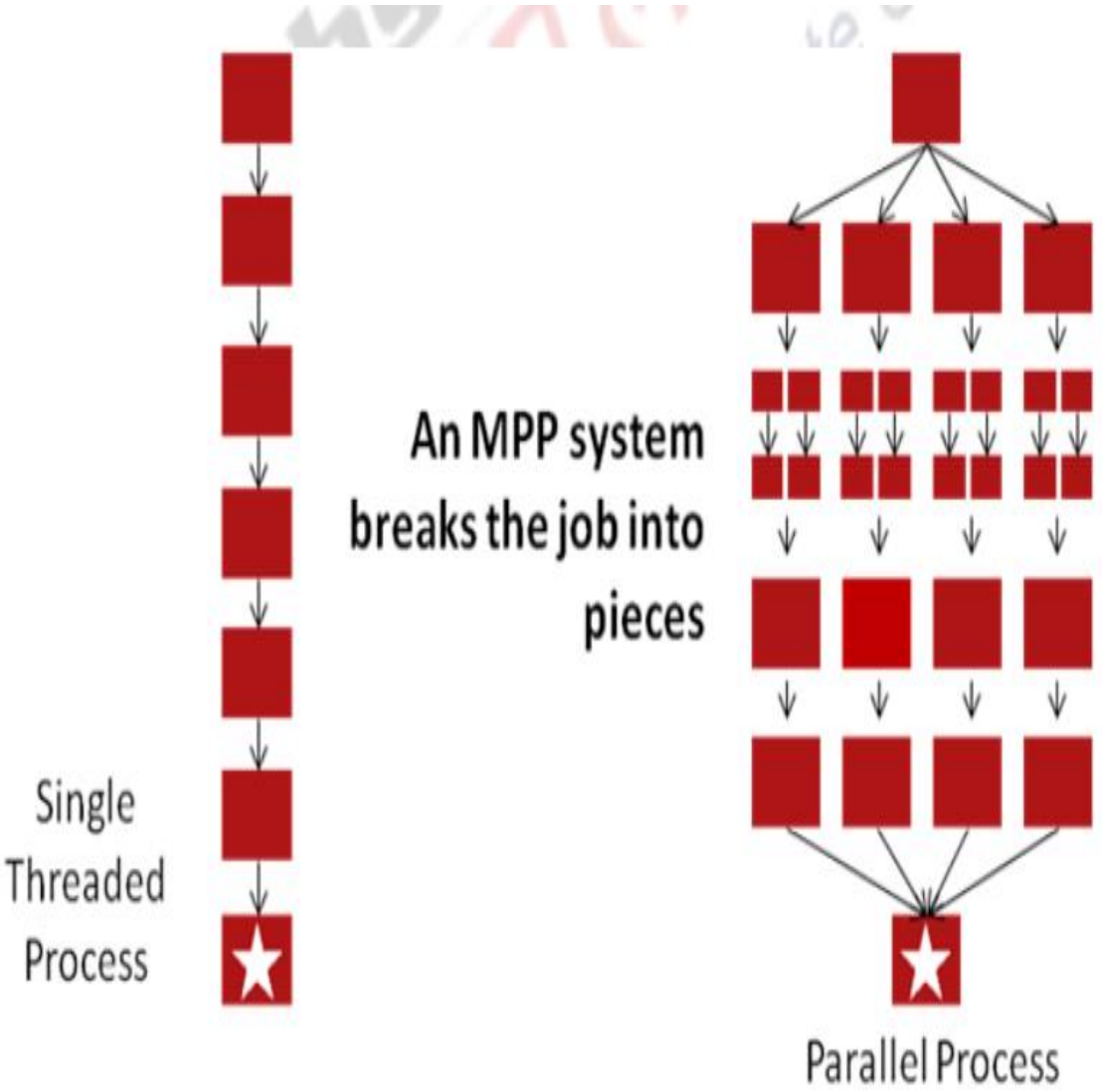
- Data from heterogeneous sources are collected, transformed and loaded into data warehouse for final analysis by decision makers.
- The processing stays in the database where the data has been consolidated.
- The data is presented in aggregated form for querying.
- Queries from users are submitted to OLAP (online analytical processing) engines for execution.
- Such in-database architectures are tested for their query throughput rather than transaction throughput as in traditional database environments.
- More of metadata is required for directing the queries which helps in reducing the time taken for answering queries and hence increase the query throughput.
- Moreover the data in consolidated form are free from anomalies, since they are preprocessed before loading into warehouses which may be used directly for analysis


Database 3
Database 1
Database 4
Database 2
Consolidate
Enterprise Data Warehouse
Submit Request
Analytic Server Or PC

Massively Parallel Processing (MPP)

- Massive Parallel Processing (MPP) is the —shared nothing approach of parallel computing.
- It is a type of computing wherein the process is being done by many CPUs working in parallel to execute a single program.
- One of the most significant differences between a Symmetric Multi-Processing or SMP and Massive Parallel Processing is that with MPP, each of the many CPUs has its own memory to assist it in preventing a possible hold up that the user may experience with using SMP when all of the CPUs attempt to access the memory simultaneously.


High-speed connect
nodes
Local
disk


Single
Threaded
Process
An MPP system
breaks the job into
pieces
Parallel Process

The Cloud Computing

- Cloud computing is the delivery of computing services over the Internet.
- Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations.
- Examples of cloud services include online file storage, social networking sites, webmail, and online business applications.
- The cloud computing model allows access to information and computer resources from anywhere that a network connection is available.
- Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.

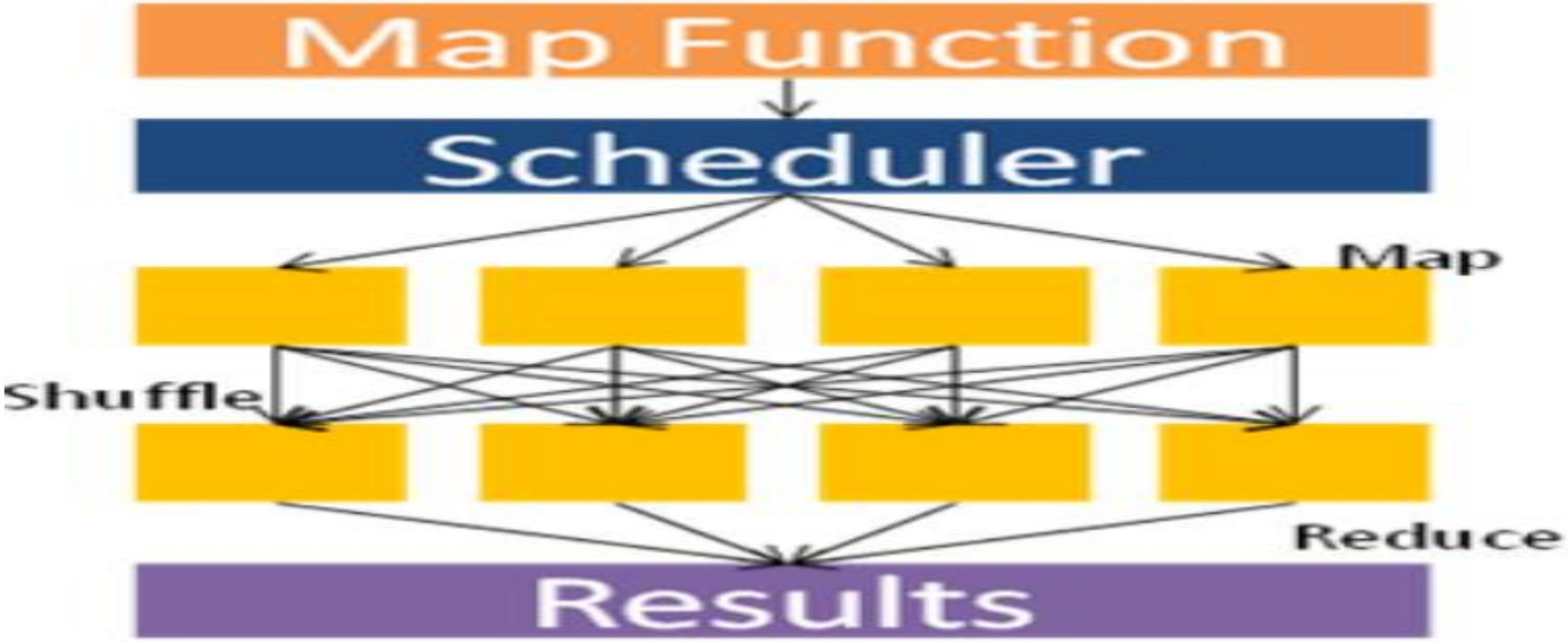
Grid Computing

- Grid computing is a form of distributed computing whereby a "super and virtual computer" is composed of a cluster of networked, loosely coupled computers, acting in concert to perform very large tasks.
- Grid computing (Foster and Kesselman, 1999) is a growing technology that facilitates the executions of large-scale resource intensive applications on geographically distributed computing resources.
- Facilitates flexible, secure, coordinated large scale resource sharing among dynamic collections of individuals, institutions, and resource

Hadoop

- The Apache® Hadoop® project develops open-source software for reliable, scalable, distributed computing.
- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
- Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

- Two main building blocks inside this runtime environment are MapReduce and Hadoop Distributed File System (HDFS).
- Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.



Working of MapReduce

- Let‘s assume there are 20 terabytes of data and 20 MapReduce server nodes for a project. The steps are summarized as follows:
  - Distribute a terabyte to each of the 20 nodes using a simple file copy process
  - Submit two programs(Map, Reduce) to the scheduler
  - The map program finds the data on disk and executes the logic it contains
  - The results of the map step are then passed to the reduce process to summarize and aggregate the final answers

Morden data analytics tools

Here's a list of **modern data analytics tools** widely used today:

---

1. Data Processing & Big Data Platforms

- **Apache Hadoop** – Distributed storage and processing framework.
- **Apache Spark** – Fast in-memory data processing engine.
- **Google BigQuery** – Cloud-based data warehouse for large-scale analytics.
- **Databricks** – Unified analytics platform built on Apache Spark.

---

2. Data Visualization Tools

- **Tableau** – Interactive dashboards and visual analytics.
- **Power BI** – Microsoft's business analytics and visualization tool.
- **QlikView / Qlik Sense** – Data visualization and self-service analytics.

Morden data analytics tools con…

3. Statistical & Predictive Analytics Tools

- **R** – Open-source language for statistical computing and visualization.
- **Python (with Pandas, NumPy, SciPy, scikit-learn)** – Data analysis and machine learning.
- **SAS** – Advanced analytics, predictive modeling, and data management.

---

4. Business Intelligence (BI) Platforms

- **Looker** – Cloud-based BI and data exploration platform.
- **SAP Analytics Cloud** – Business intelligence, planning, and predictive analytics.
- **Oracle Analytics Cloud** – AI-powered analytics platform.

---

5. AI & Machine Learning Platforms

- **TensorFlow** – Open-source machine learning framework.
- **PyTorch** – Deep learning framework for AI models.
- **H2O.ai** – Automated machine learning and AI platform.

Applications of Data Analytics

- Data analytics has become a cornerstone of modern decision-making across various industries. Its ability to extract valuable insights from vast datasets has revolutionized the way businesses operate and solve complex problems. Here are some key applications of data analytics:

- **Healthcare -**
  - **Personalized medicine**: Analyzing patient data to tailor treatments to individual needs.
  - **Disease prediction**: Identifying early signs of diseases for proactive prevention.
  - **Drug discovery**: Accelerating the development of new drugs by analyzing molecular data.
  - **Healthcare operations**: Optimizing resource allocation and improving patient outcomes.

- **Finance –**
  - **Fraud detection**: Identifying suspicious activities and preventing financial losses.
  - **Risk assessment**: Evaluating investment risks and making informed decisions.
  - **Customer segmentation**: Understanding customer behavior to tailor financial products and services.
  - **Market analysis**: Identifying investment opportunities and predicting market trends.

- **Marketing –**
  - **Customer segmentation**: Grouping customers based on demographics, preferences, and behavior.
  - **Targeted advertising**: Delivering personalized ads to the right audience.
  - **Market research**: Understanding customer needs and preferences to improve product development.
  - **Customer prediction**: Identifying customers at risk of leaving to improve retention.

- **Retail –**
  - **Inventory management**: Optimizing stock levels to reduce costs and avoid stockouts.
  - **Personalized recommendations**: Suggesting products based on customer preferences and purchase history.
  - **Price optimization**: Determining the optimal pricing strategy to maximize revenue.
  - **Customer satisfaction analysis**: Measuring customer satisfaction and identifying areas for improvement.

- **Manufacturing –**
  - **Predictive maintenance**: Predicting equipment failures to prevent downtime and reduce costs.
  - **Quality control**: Monitoring product quality and identifying defects.
  - **Supply chain optimization**: Improving efficiency and reducing costs in the supply chain.
  - **Process optimization**: Identifying opportunities to improve manufacturing processes and increase productivity.

- **Transportation –**
  - **Traffic management**: Optimizing traffic flow and reducing congestion.
  - **Route optimization**: Finding the most efficient routes for deliveries and transportation.
  - **Predictive maintenance**: Preventing breakdowns of transportation equipment.
  - **Transportation planning**: Analyzing travel patterns to improve infrastructure and services.

- **Government –**
  - **Public safety**: Analyzing crime data to identify hotspots and improve policing strategies.
  - **Urban planning**: Understanding population trends and optimizing city development.
  - **Economic development**: Identifying growth opportunities and attracting investments.
  - **Environmental monitoring**: Tracking environmental changes and addressing sustainability challenges.

Analysis vs Reporting

The distinction between analysis and reporting in data analytics can be characterized as follows:

- **Analysis** typically involves interpreting data, seeking patterns, and making predictions to aid in decision-making.

- **Reporting** is more about summarizing data in a structured format, primarily used to inform stakeholders about the status or results without deep insights or predictive elements.

Analysis vs Reporting con…

Aspect	Analysis	Reporting
Purpose	To explore data and extract insights.	To inform stakeholders of the results.
Focus	Understanding and prediction.	Summarization and description.
Outcome	Strategic decisions, forecasts, solutions.	Status updates, results documentation.
Tools	Advanced statistical and ML tools.	Business intelligence tools.
Process	Iterative and explorative.	Systematic and structured.
Skills Needed	Statistical, programming, critical thinking	Data presentation, software proficiency

Data Analytics Lifecycle

- The data analytics lifecycle is a structure for doing data analytics that has business objectives at its core.
- It undergoes various stages throughout its life, during its creation, testing, processing, consumption, and reuse.
- Data Analytics Lifecycle maps out these stages for professionals working on data analytics projects.
- These phases are arranged in a circular structure that forms a Data Analytics Lifecycle. Each step has its significance and characteristics.
- The Data Analytics Lifecycle is designed to be used with significant big data projects.
- It is used to portray the actual project correctly; the cycle is iterative.

Importance of Data Analytics Lifecycle
Or Why is Data Analytics Lifecycle Needed?

- Data Analytics Lifecycle defines the roadmap of how data is generated, collected, processed, used, and analyzed to achieve business goals.
- It offers a systematic way to manage data for converting it into information that can be used to fulfill organizational and project goals.
- The process provides the direction and methods to extract information from the data and proceed in the right direction to accomplish business goals.
- Data professionals use the lifecycle's circular form to proceed with data analytics in either a forward or backward direction.

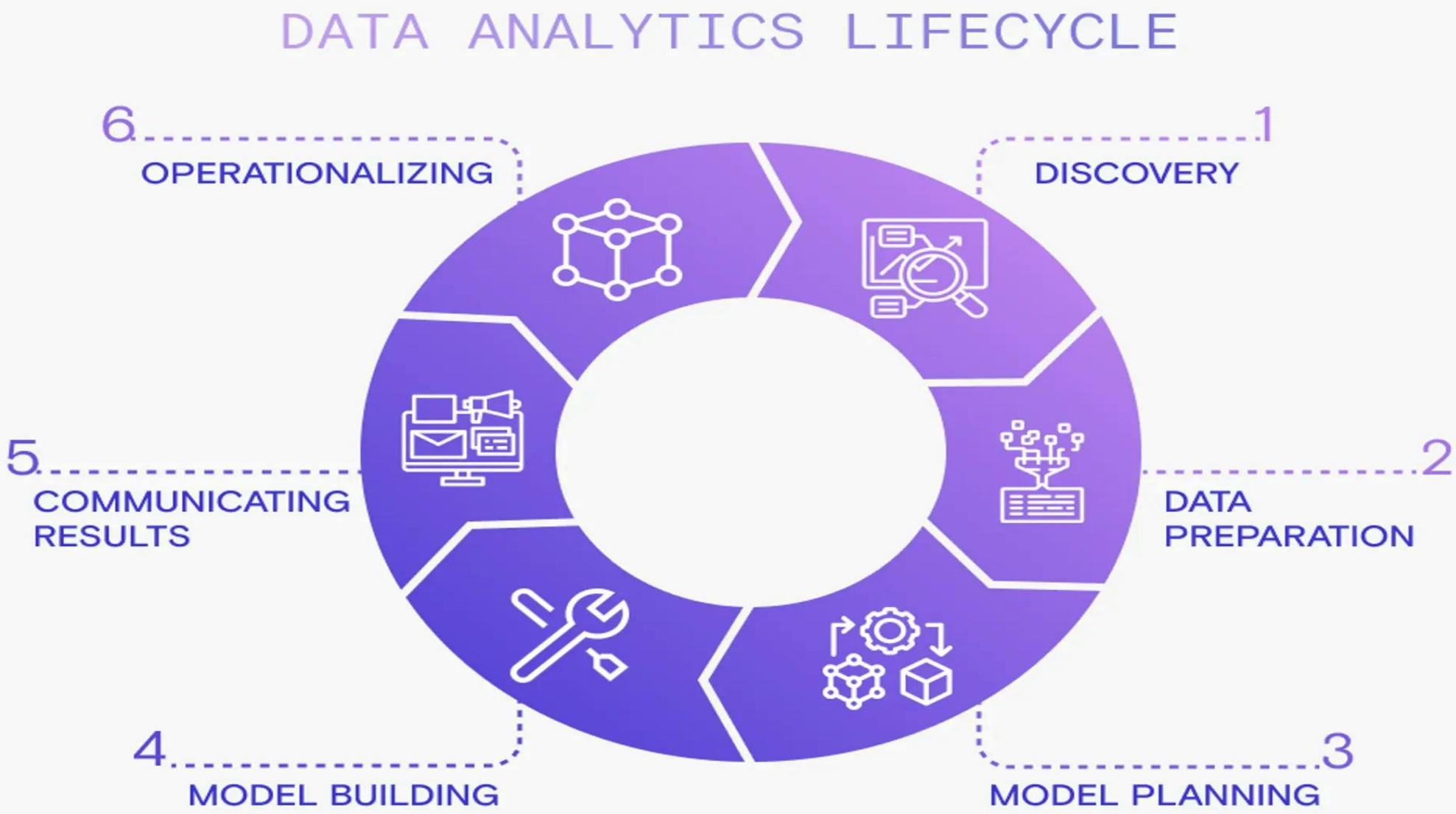


Figure: Various phases of data analytics life cycle

1- Discovery Phase:

- This first phase involves getting the context around your problem: you need to know what problem you are solving and what business outcomes you wish to see.
- You should begin by defining your **business objective and the scope of the work.**
- Work out what data sources will be available and useful to you (for example, Google Analytics, Salesforce, or any marketing campaign information you might have available), and perform a gap analysis of what data is required to solve your business problem analysis compared with what data you have available, working out a plan to get any data you still need.
- Once your objective has been identified, you should formulate an initial hypothesis.
- Design your analysis so that it will determine whether to accept or reject this hypothesis.
- Decide in advance what the criteria for accepting or rejecting the hypothesis will be to ensure that your analysis is rigorous and follows the scientific method.

2- Data preparation:

- In the next stage, you need to decide which data sources will be useful for the analysis, collect the data from all these disparate sources, and load it into a data analytics sandbox so that it can be used for prototyping.
- When loading your data into the sandbox area, you will need to transform it.

- The two main types of transformations are:

  (a)- Preprocessing transformations means cleaning your data to remove things like nulls, defective values, duplicates, and outliers.

  (b)- Analytics transformations can mean a variety of things, such as standardizing or normalizing your data so it can be used more effectively with certain machine learning algorithms, or preparing your datasets for human consumption.

- Depending on whether your transformations take place before or after the loading stage, this whole process is known as either **ETL (extract, transform, load).**

3- Model Planning Phase:

- A model in data analytics is a mathematical or programmatic description of the relationship between two or more variables.
- It allows us to study the effects of different variables on our data and to make statistical assumptions about the probability of an event happening.

- **You may want to think about the following when deciding on a model:**
  - **How large is your dataset?**

    - You may only have a small dataset available, or you may require your dashboards to be fast, which generally requires smaller and pre-aggregated data.

Model Planning Phase con…

- **How will the output be used?**

  - In the Business Intelligence (BI) use case scenario; fast and pre-aggregated data are necessary.

- **Is the data labeled with column headings?**

  - If it is, you could use supervised learning, but if not, unsupervised learning is your only option.

- **Do you want the outcome to be qualitative or quantitative?**

  - If your question expects a **quantitative answer** (for example, "How many sales are forecast for next month?" or "How many customers were satisfied with our product last month?") then you should use a **regression model.**

  - However, if you expect a **qualitative answer** (for example, "Is this email spam?", where the answer can be Yes or No, or "Which of our five products are we likely to have the most success in marketing to customers ?"), then you may want to use a **classification or clustering model.**

Model Planning Phase con…

- **Is accuracy or speed of the model particularly important?**
  - If so, check whether your chosen model will perform well. The size of your dataset will be a factor when evaluating the speed of a particular model.
- **Is your data unstructured?**
  - Unstructured data cannot be easily stored in either relational or graph databases and includes free text data such as emails or files.
  - This type of data is most suited to machine learning.
- **Have you analyzed the contents of your data?**
  - This allows you to work out which variables have the largest effects and to identify new factors (that are a combination of different existing variables) that have a big impact.

4- Model Building and Execution Phase:

- The steps within this phase of the data analytics lifecycle depend on the model you have chosen to use.
- Different Models can be:

(i)- SQL model:

- You will first need to find your source tables and the join keys.
- Next, determine where to build your models.
- Depending on the complexity, building your model can range from saving SQL queries in your warehouse and executing them automatically on a schedule, to building more complex data modeling chains using software tools like dbt or Dataform.

Model Building and Execution Phase con…

(ii)- Statistical model

- Next, you will need to decide which statistical model is appropriate for your use case.
- For example, you could use a correlation test, a linear regression model, or an analysis of variance (ANOVA).
- Finally, you should run your model on your dataset and publish your results.

(iii)- Machine learning (ML) model

- ML models require you to create two samples from this dataset: one for training the model, and another for testing the model.
- If you are using a machine learning model, it will need to be trained.
- This involves executing your model on your training dataset, and tuning various parameters of your model so you get the best predictive results.
- Once this is working well, you can execute your model on your real dataset, which is used for testing your model.
- You can now work out which model gave the most accurate result and use this model for your final results, which you will then need to publish.

5- Communicating results:

- You must communicate your findings clearly, and it can help in use data visualizations.
- Any communication with stakeholders should include a narrative, a list of key findings, and an explanation of the value your analysis adds to the business.
- You should also compare the results of your model with your initial criteria for accepting or rejecting your hypothesis to explain them how confident they can be in your analysis.

6- Operationalizing Phase:

- Once the stakeholders are happy with your analysis, you can execute the same model outside of the analytics sandbox on a production dataset.
- You should monitor the results of this to check if they lead to your business goal being achieved.
- If your business objectives are being met, deliver the final reports to your stakeholders, and communicate these results more widely across the business.

Thank You