SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A+' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

COURSE NAME : 19CSO302 & Introduction to Data Analytics

III YEAR/ VI SEMESTER

UNIT – I INTRODUCTION

Topic: Big Data

P.Poonkodi

Assistant Professor

Department of Computer Science and Engineering

Big Data

- Types
- Characteristics
- Tools
- Applications

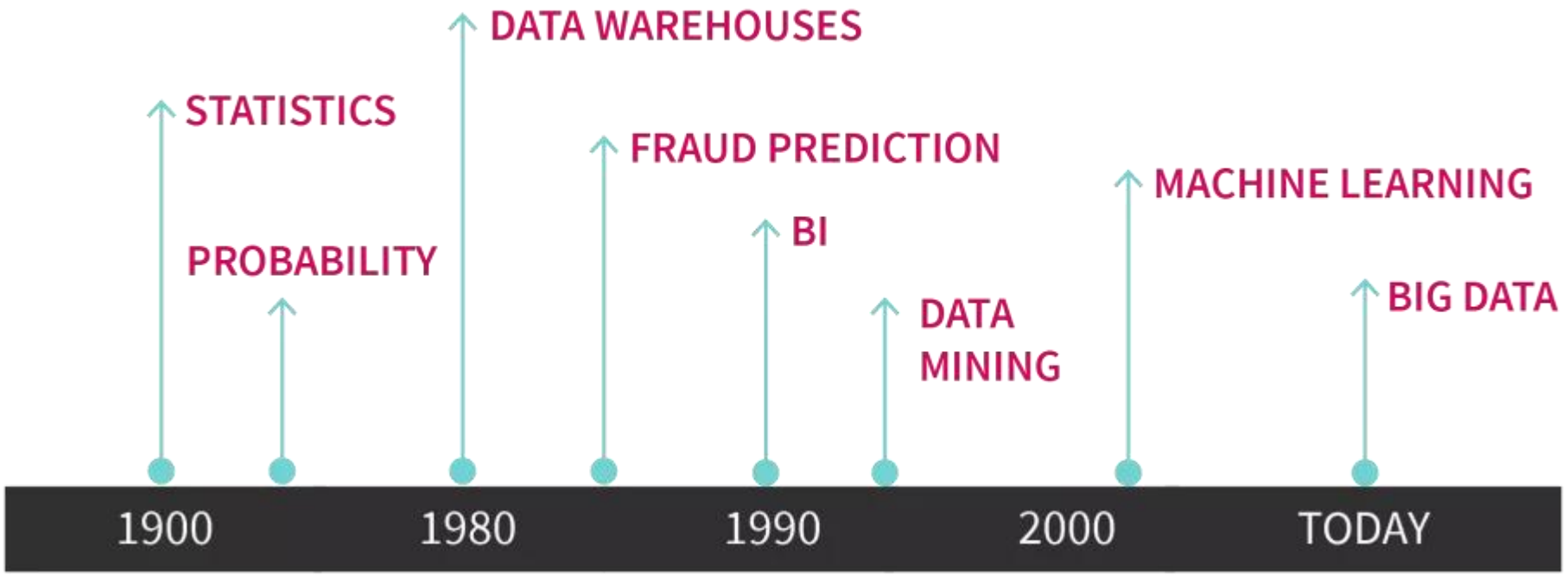
Introduction

- **Big data** is a combination of structured, semi structured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications
- Systems that process and store big data have become a common component of data management architectures in organizations, combined with tools that support big data analytics uses

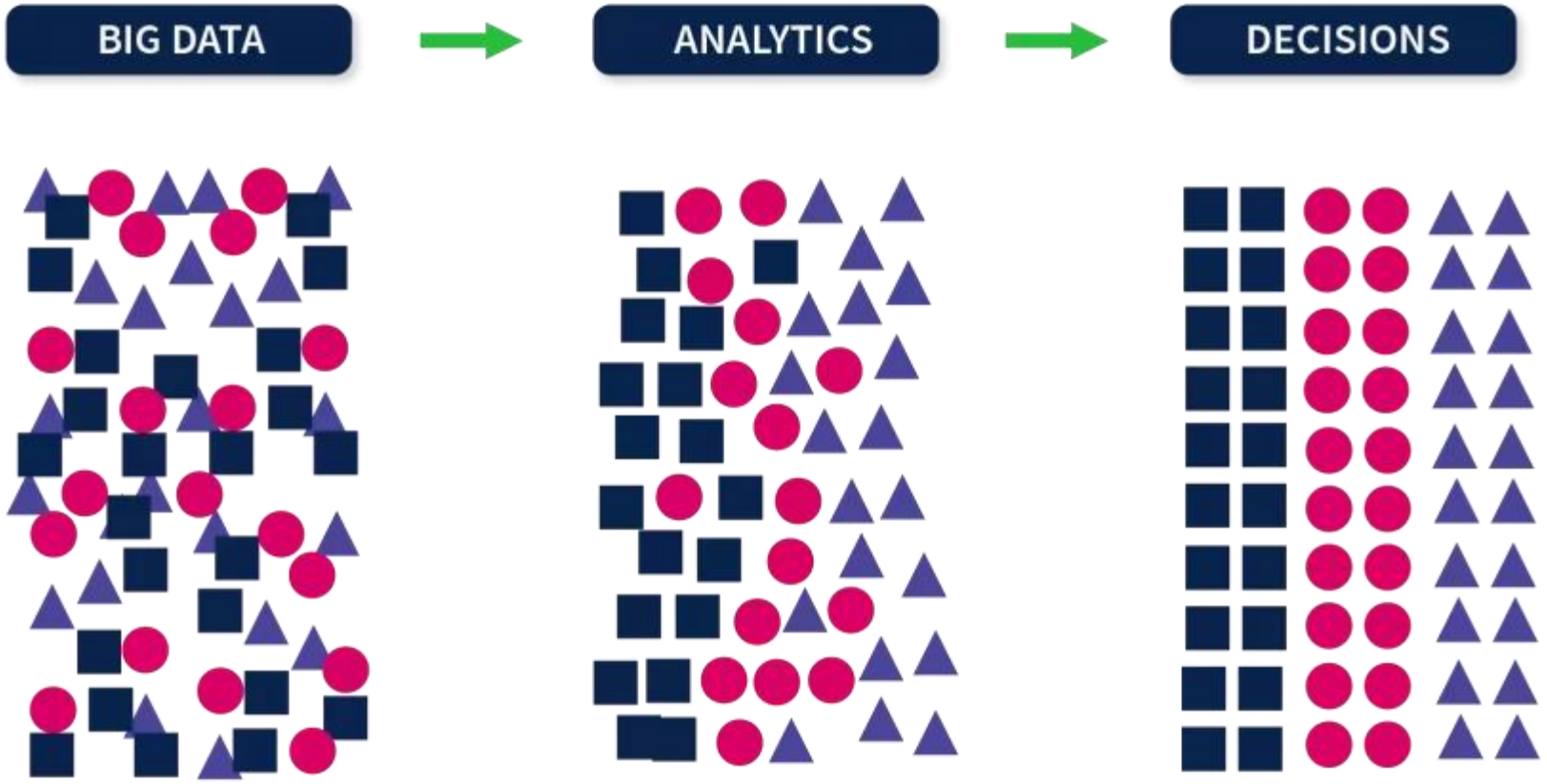
Why is Big Data important?

- Big data is instrumental for organizations, enabling data-driven decision-making and fostering innovation by uncovering patterns and opportunities in extensive datasets
- Its role in personalization enhances customer satisfaction, while in healthcare, it advances research and personalized medicine
- Predictive analytics aids in forecasting, fraud detection, and security
- Overall, big data's impact spans business efficiency, scientific research, and societal initiatives, making it a cornerstone for progress in various fields

History of Big Data

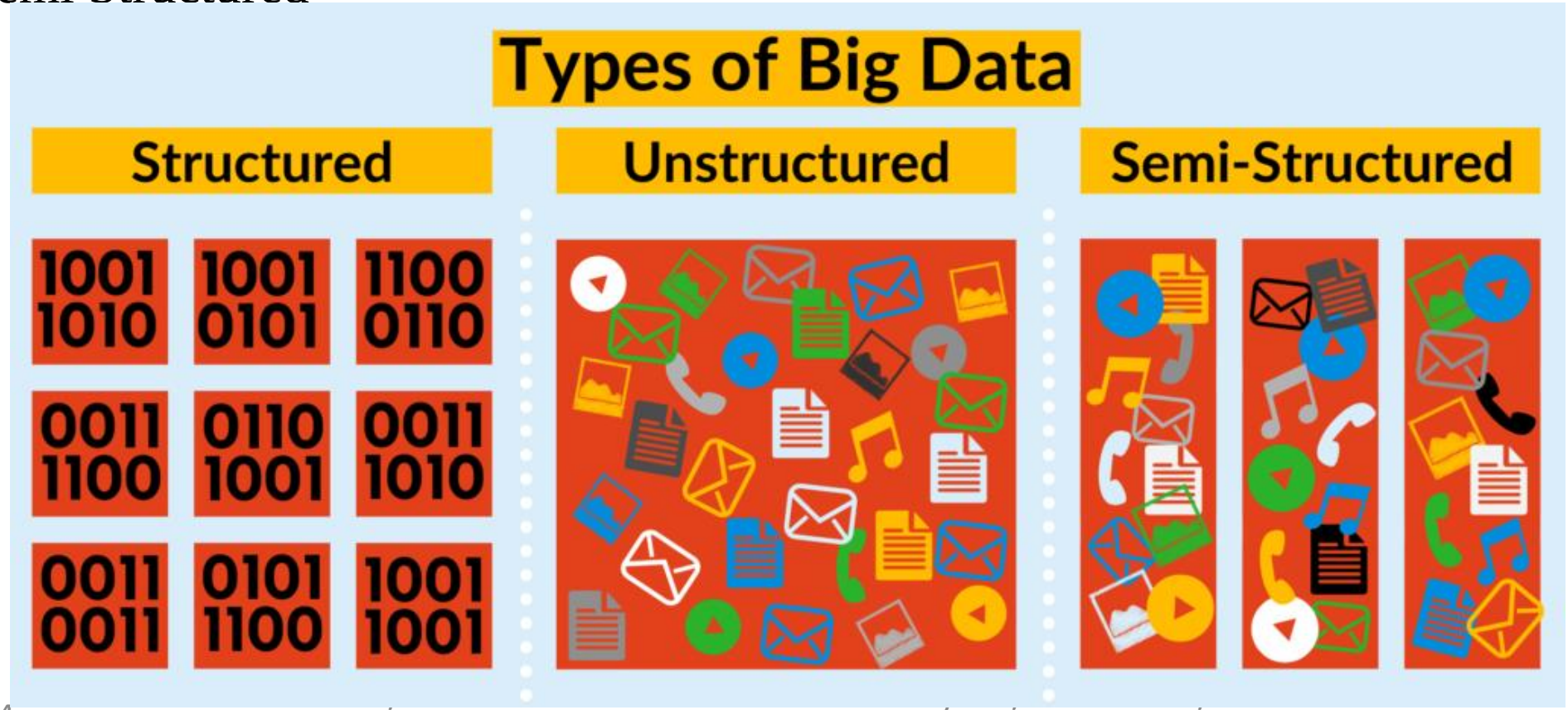


How Big Data Works?



Types

- Big data can be classified into three main types based on its source, structure, and characteristics:
  - Structured
  - Unstructured
  - Semi Structured



Structured

- When the data is processed, stored, and retrieved in the desired format, and can be stored in relational databases, which makes it easier to analyze
- Names, dates, addresses, credit card numbers, and other structured data are examples
- Examples include data stored in relational databases, spreadsheets, or CSV files
- This type of data is easily searchable, query able, and analyzable using traditional database management systems (DBMS) like SQL databases

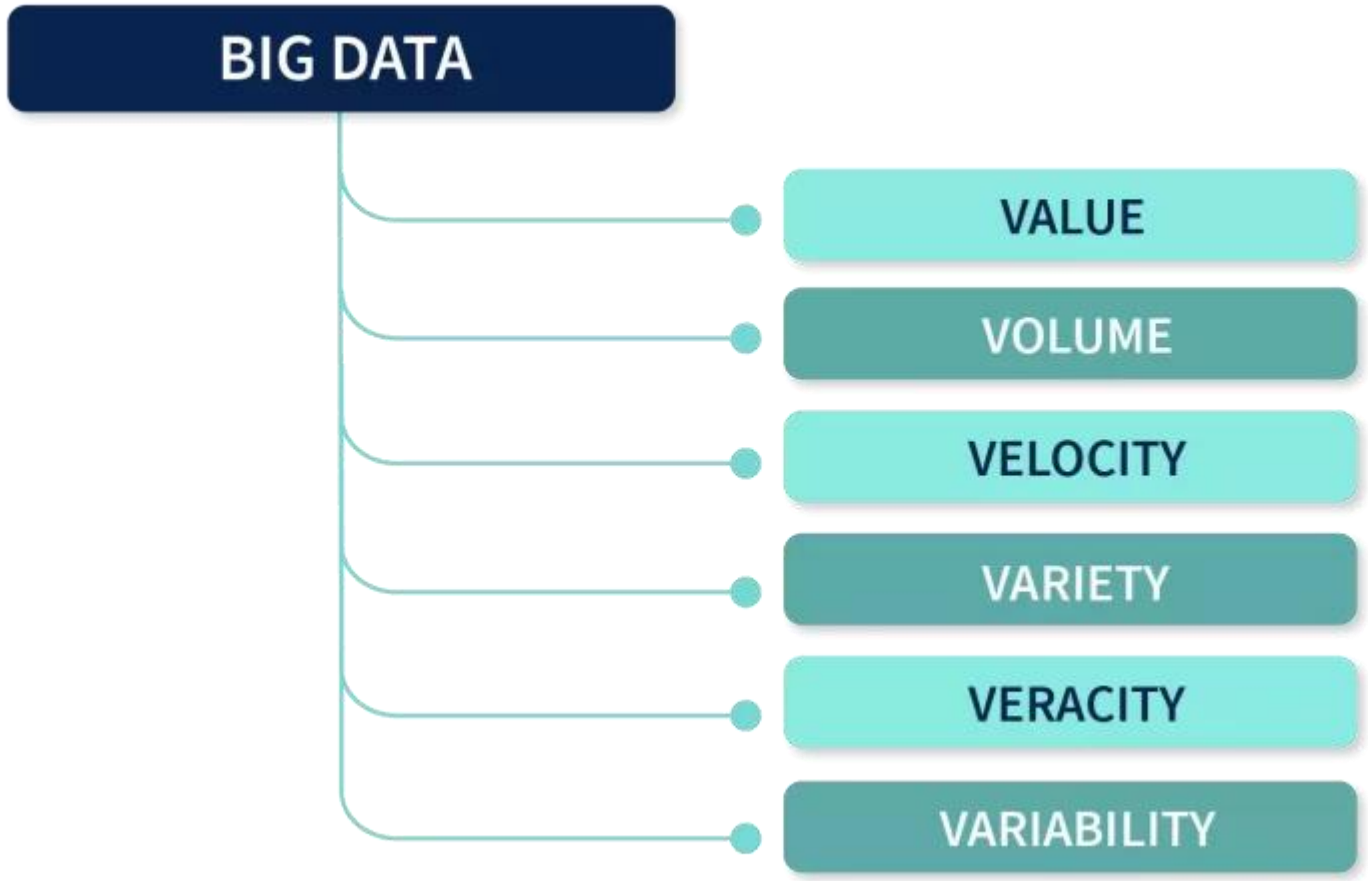
Unstructured

- Unstructured data is information that lacks a specific format and is stored in original form without any presets
- It’s demanding and time-consuming to process and analyze unstructured data
- Rich media like Data from the media, entertainment industries, and surveillance data are examples of unstructured data
- It includes text data, images, videos, audio recordings, social media posts, emails, and sensor data
- Unstructured data poses significant challenges for analysis due to its complexity, but advancements in natural language processing (NLP), computer vision, and other techniques have made it possible to extract valuable insights from this type of data

Semi-structured

- When the data is not segregated properly but yet contains some vital information
- Email, NoSQL databases, and other semi-structured data are examples
- It may be self-describing and includes metadata, tags, or markers to provide context or structure
- Examples include XML (eXtensible Markup Language) files, JSON (JavaScript Object Notation), and log files
- While semi-structured data is more flexible than structured data, it still requires some level of organization for analysis

Characteristics



Characteristics

- The key characteristics of big data are commonly summarized as the Six Vs:
  - **Volume-** Volume is a huge set of data that may or may not be in our desired format and may need further processing to extract valuable information. An example of a high volume of data can be daily transactions through Online payment services like UPI, credit cards, etc.
  - **Velocity -** Velocity is the rate of growth of data. Since big data is rapidly changing or keeps getting updated in data storage, we need to process structured and unstructured data and keep updating our data stores to take advantage of the information. Social media posts are an example of data generated at a high Velocity
  - **Variety -** It refers to whether the data is structured semi-structured or unstructured. Which we talked about earlier. A variety of data can be anything from email, phone number, XML file, audio, videos, etc.

Characteristics

- **Veracity-** It is the quality, validity, and trustworthiness of the data. Suppose you are willing to go to a restaurant, you search on Google to find out which one will be best suited for you. However, due to excess unstructured data from all the sources, you end up getting confused. Here comes the use of hashtags which helps you to sleek down your interest, and along with suggestions, you also get the images of that place that have been shared by other customers thereby, building a sense of trust in you that other people liked it too. This is how veracity can be justified through this example and helps us understand its importance in explaining big data.
- **Value -** It is the most important V of big data, which can be defined as the ability to transform the data into business. So, to better grasp it, let's look at an example. If data lacks Value, we will not be able to use it in other Vs to solve problems associated with a specific hypothesis and thus won't transform it for successful business activities.

Characteristics

- **Variability -** It refers to data whose meaning changes over time. Organizations frequently need to create complex systems to comprehend context and decode the exact meaning of raw data. Let us understand this change through a real life example. Let's understand this change through a real-life example. Every hostel student might relate to it. Suppose every week Mess serves paneer. Although that paneer does not taste good every week there is a change in taste. This change is variable, why does the taste change? Because on someday the chef pours extra spices and on other days he does not. The same is true for data, which might have an impact on the quality of your data if it is constantly changing.

Big Data Use Cases

- **Recommendation Engines**
- **Financial Fraud Detection**
- **Health Care Sector**
- **Agriculture and Food Industries**

Big Data Use Cases

- **Recommendation Engines**
  - While watching the usual Netflix and Prime. These streaming services tend to show us a few shows under “recommended for you”
  - So, this recommendation system works with the help of big data where large amounts of structured and unstructured data can be gathered just from a few clicks and this information helps companies predict the recommendations for the user

Big Data Use Cases

- **Financial Fraud Detection**
  - With the use of Big Data and machine learning algorithms, Even if there are any subtle changes in consumer purchase or credit card activity, they can be automatically analyzed and can be marked for fraud because of this each year these financial institutes millions of dollars in fraud
  - Even in the Insurance sector because of big data billions of dollars are saved each year as claims can be analyzed and a similar machine learning model then detects the potential fraud

Big Data Use Cases

- **Health Care Sector**
  - Doctors, Researchers, and health care companies are rapidly adopting big data solutions to solve different problems
  - Researchers and doctors can conduct more effective research for some of the noncurable diseases like HIV, Cancer, and Alzheimer's and can develop more effective drugs by analyzing the pattern
  - Even hospitals are adapting to big data solutions to do customized ways of testing as opposed to trivial ways of testing and in some countries, big data helps in faster and more efficient analysis of healthcare information

Big Data Use Cases

- **Agriculture and Food Industries**
  - Big data is used to improve the quality and quantity of crops by using geospatial data, graphical data, seismic activities, etc
  - Due to the growing population, improving the quantity of food has been a priority in the last thirty years
  - Due to big data, the fight against global hunger has been made possible

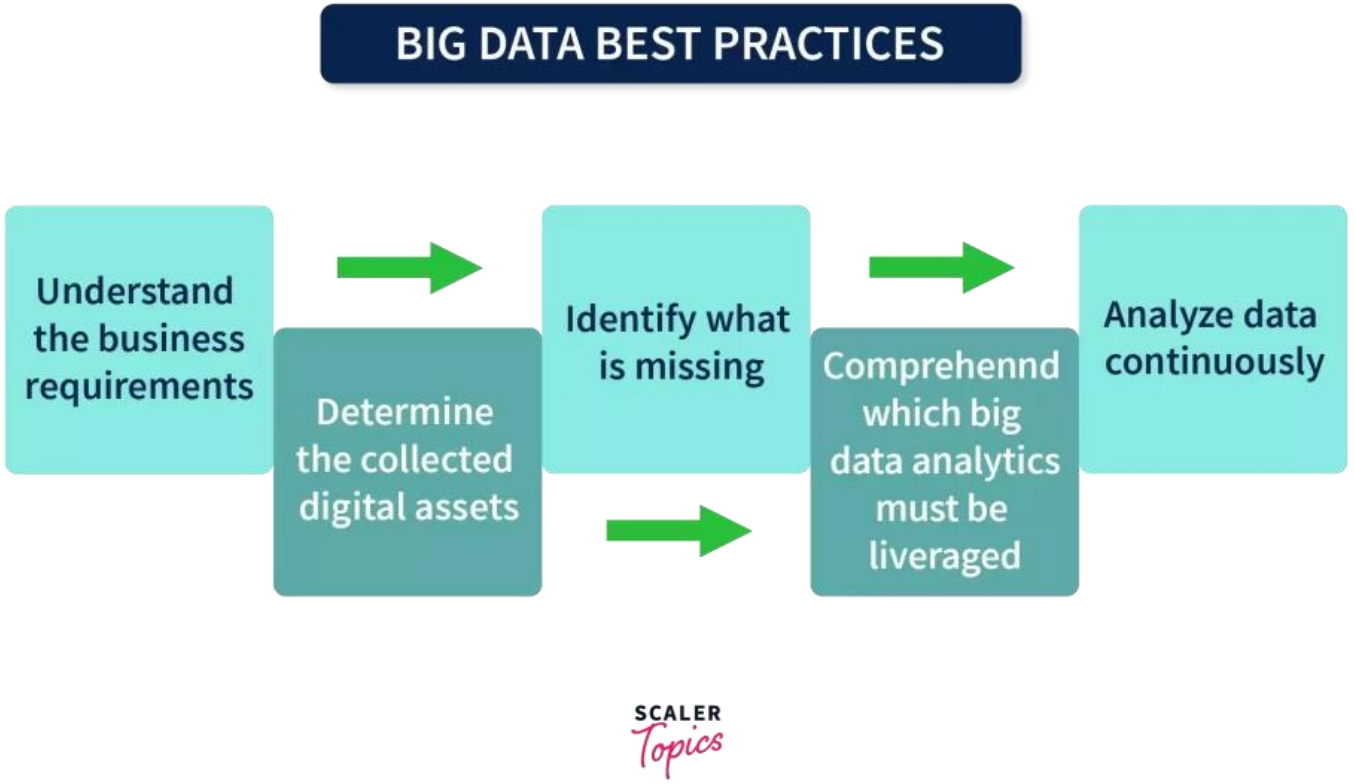
Big Data Challenges

- **Incomplete Understanding of Big Data**
- **Exponential Data Growth**
- **Security of Data (**cyber security professionals to protect their data)
- **Data Integration(**solve these problems by purchasing the right technologies like IBM Info Sphere, Microsoft SQL, etc)

Conclusion

- Big data is a large and complex collection of data that is difficult to process and analyze using traditional methods.
- It is characterized by the six Vs: volume, velocity, variety, veracity, value and variability.
- Big data can be used to make better decisions, improve products and services, and make new discoveries.
- Big data is significantly reducing the cost of storing, analyzing, and processing large amounts of data

Big Data Best Practices



Tools

- **Big data tools** have become vital in today's data-driven economy for businesses and organizations to leverage the potential of big data
- These tools, including Hadoop, Spark, and NoSQL databases, allow for efficient capture, storage, and analysis of huge datasets
- Big data tools enable decision-makers to make educated choices, improve consumer experiences, and discover emerging trends by translating raw data into actionable insights
- Big data tools are crucial for modern organizations due to their role in improving operational efficiency, personalizing offerings, and attaining competitive advantage

Need for Big Data Tools

- **Data Variety**
- **Speed and Real-Time Analytics**
- **Enhanced Customer Experiences**
- **Predictive Analytics**

Big Data Analytics Tools

- Apache Hadoop
- Apache Spark
- Apache Kafka
- Apache Storm
- Apache Cassandra
- Apache Hive
- Qubole
- Xplenty
- MongoDB
- SAS
- Data Pine
- Hevo Data
- Zoho Analytics
- Cloudera
- RapidMiner
- OpenRefine
- Kylin
- Apache Samza
- Lumify
- Trino

Factors to Consider While Selecting the Big Data Tools

- **Project Requirements**
- **Scalability**
- **Usability**
- **Integration**
- **Cost**
- **Security**

Advantages

- **Data Insights:**
  Big Data tools unravel hidden patterns and insights from vast datasets, enabling businesses to make informed strategic choices.
- **Scalability:**
  These tools provide scalable infrastructure to meet expanding data demands without sacrificing speed.
- **Real-time Analysis:**
  Big Data tools enable real-time data processing, which is essential for quick reactions and preemptive measures.
- **Competitive Advantage:**
  Using Big Data tools helps businesses to remain ahead by anticipating trends and customer preferences.

Use Cases

- **Customer Analytics:**
  Understanding customer behavior and preferences assists in the development of personalized marketing tactics.
- **Risk Management:**
  Big Data tools evaluate risks by analyzing historical and real-time data, which is critical in the financial and insurance industries.
- **Healthcare Insights:**
  Big Data tools aid in illness diagnosis, epidemic prediction, and patient care.
- **Supply Chain Optimization:**
  Big Data tools improve inventory management, logistics, and demand forecasting.
- **Fraud Detection:**
  Financial businesses use Big Data analytics to detect fraudulent operations using pattern recognition.

Conclusion

- **Big data tools** have transformed how we handle, analyze, and interpret large datasets.
- Apache Hadoop's distributed processing capabilities have opened the road for efficiently addressing big data volumes.
- In-memory computing of Spark has introduced lightning speed to data processing.
- **Apache Cassandra** and **MongoDB** provide scalable NoSQL database solutions.
- A crucial factor when choosing the right big data tool is its scalability and performance; the tool should be able to handle large datasets and processing demands efficiently.
- High performance of big data tools ensures that data processing and analysis can be completed within reasonable time frames, enabling timely and informed decision-making.

Applications of Big Data

1. Banking
2. Education
3. Media
4. Healthcare
5. Agriculture
6. Travel
7. Manufacturing
8. Government
9. Retail

References

- Runkler TA, “Data Analytics: Models and algorithms for intelligent data analysis”, Springer, Third Edition 2020


Thank You!