`

UNIT-I

1. What is big data analytics? Describe the main Characteristics of Big Data with suitable example?

Big Data analytics is the process of examining and interpreting large and complex datasets to uncover hidden patterns, correlations, trends, and valuable insights. It involves using advanced analytics techniques, tools, and technologies to analyze vast amounts of structured, semi-structured, and unstructured data. The primary goal of Big Data analytics is to extract meaningful information that can be used to make informed business decisions, optimize processes, and gain a competitive advantage.

Key aspects of Big Data analytics include:

Data Collection:

Gathering data from diverse sources, such as social media, sensors, logs, transaction records, and other data repositories.

Data Storage:

Storing and managing large datasets efficiently, often utilizing distributed storage systems like Hadoop Distributed File System (HDFS) or NoSQL databases.

Data Processing:

Performing complex data processing tasks using parallel and distributed computing frameworks, such as Apache Spark or Hadoop MapReduce, to handle the volume and variety of Big Data.

Data Analysis:

Applying various analytical techniques, including statistical analysis, machine learning algorithms, data mining, and predictive modeling, to discover patterns and insights within the data.

Data Visualization:

Presenting the results of the analysis in a visual format, such as charts, graphs, and dashboards, to make complex information more accessible and understandable for decision-makers.

Real-time Analytics:

Analyzing data in real-time or near real-time to gain immediate insights and respond quickly to changing conditions or events.

Big Data is characterized by three main dimensions, often referred to as the three Vs: Volume, Velocity, and Variety.

Volume:

Definition: Refers to the vast amounts of data generated or collected.

Example: Social media platforms generate enormous volumes of data every second, including user posts, comments, likes, and shares. The massive amount of data generated by these platforms is a classic example of dealing with high volume in Big Data.

Velocity:

Definition: Denotes the speed at which data is generated, processed, and analyzed.

Example: Financial transactions in real-time trading systems require rapid data processing. Stock exchanges and online trading platforms need to process and analyze market data at extremely high speeds to make split-second decisions. This exemplifies the velocity aspect of Big Data.

Variety:

Definition: Describes the diverse types of data, including structured, semi-structured, and unstructured data.

Example: A retail business might deal with structured data like transaction records, semi-structured data like customer reviews, and unstructured data like images or videos. The ability to handle and analyze this variety of data types is a characteristic of Big Data.

Refer Text book : Page -22-25

2. Classify various types of digital data dealt in Big Data systems? Explain each in detail.

Big data refers to large and complex datasets that cannot be easily managed, processed, or analyzed using traditional data processing tools. Digital data can be classified into various types based on different characteristics. Here are some common classifications of digital data in the context of big data:

Structured Data:

Definition: Data that is organized in a structured format with a well-defined schema.

Examples: Relational databases, CSV files, Excel spreadsheets.

Unstructured Data:

Definition: Data that lacks a predefined data model or structure.

Examples: Text documents, images, videos, audio files.

Semi-Structured Data:

Definition: Data that is not fully structured but has some level of organization.

Examples: JSON, XML, log files.

Definition: Data that lacks a predefined data model or structure.

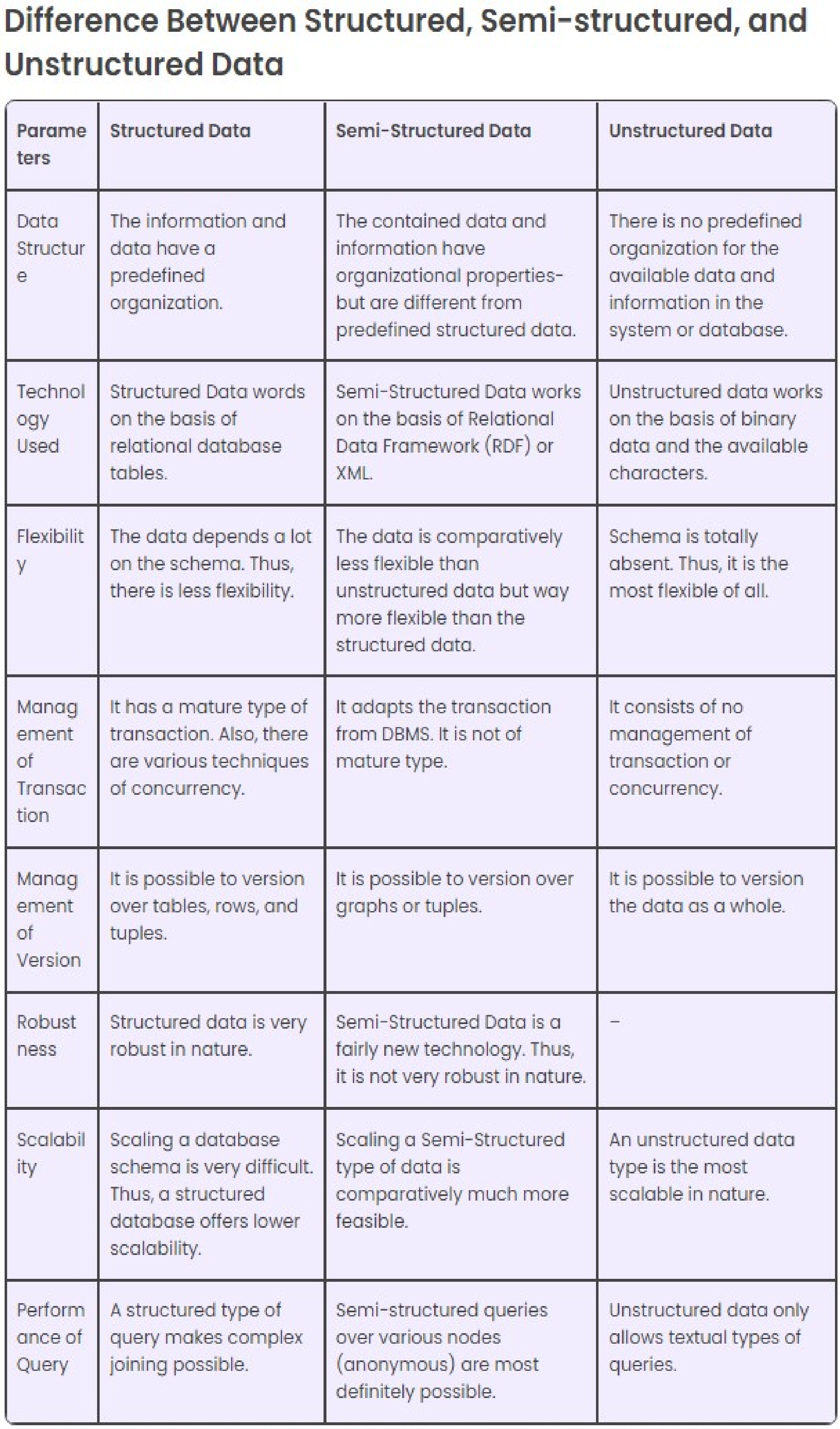
Examples: Text documents, images, videos, audio files.

Semi-Structured Data:

Definition: Data that is not fully structured but has some level of organization.

Examples: JSON, XML, log files.

`



Difference Between Structured, Semi-structured, and Unstructured Data

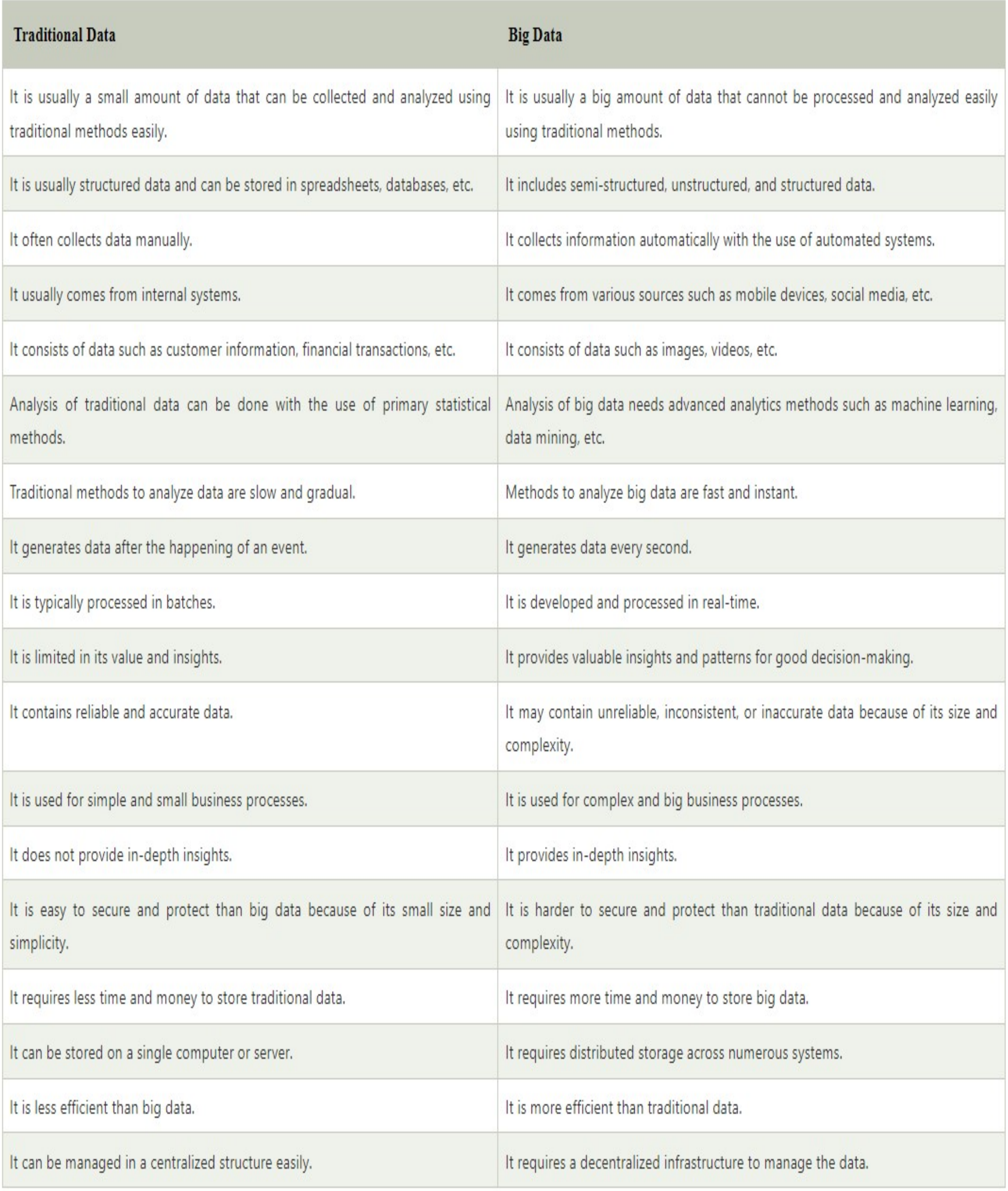
Parameters	Structured Data	Semi-Structured Data	Unstructured Data
Data Structure	The information and data have a predefined organization.	The contained data and information have organizational properties- but are different from predefined structured data.	There is no predefined organization for the available data and information in the system or database.
Technology Used	Structured Data words on the basis of relational database tables.	Semi-Structured Data works on the basis of Relational Data Framework (RDF) or XML.	Unstructured data works on the basis of binary data and the available characters.
Flexibility	The data depends a lot on the schema. Thus, there is less flexibility.	The data is comparatively less flexible than unstructured data but way more flexible than the structured data.	Schema is totally absent. Thus, it is the most flexible of all.
Management of Transaction	It has a mature type of transaction. Also, there are various techniques of concurrency.	It adapts the transaction from DBMS. It is not of mature type.	It consists of no management of transaction or concurrency.
Management of Version	It is possible to version over tables, rows, and tuples.	It is possible to version over graphs or tuples.	It is possible to version the data as a whole.
Robustness	Structured data is very robust in nature.	Semi-Structured Data is a fairly new technology. Thus, it is not very robust in nature.	–
Scalability	Scaling a database schema is very difficult. Thus, a structured database offers lower scalability.	Scaling a Semi-Structured type of data is comparatively much more feasible.	An unstructured data type is the most scalable in nature.
Performance of Query	A structured type of query makes complex joining possible.	Semi-structured queries over various nodes (anonymous) are most definitely possible.	Unstructured data only allows textual types of queries.

Refer Text book : Page 2-10

`

3. Compare and contrast Traditional Business Intelligence and Big Data.

Traditional Business Intelligence (BI) and Big Data are both approaches to handling and analyzing data, but they differ in terms of data sources, processing methods, scalability, and the types of insights they provide. Here's a comparison and contrast between Traditional Business Intelligence and Big Data:



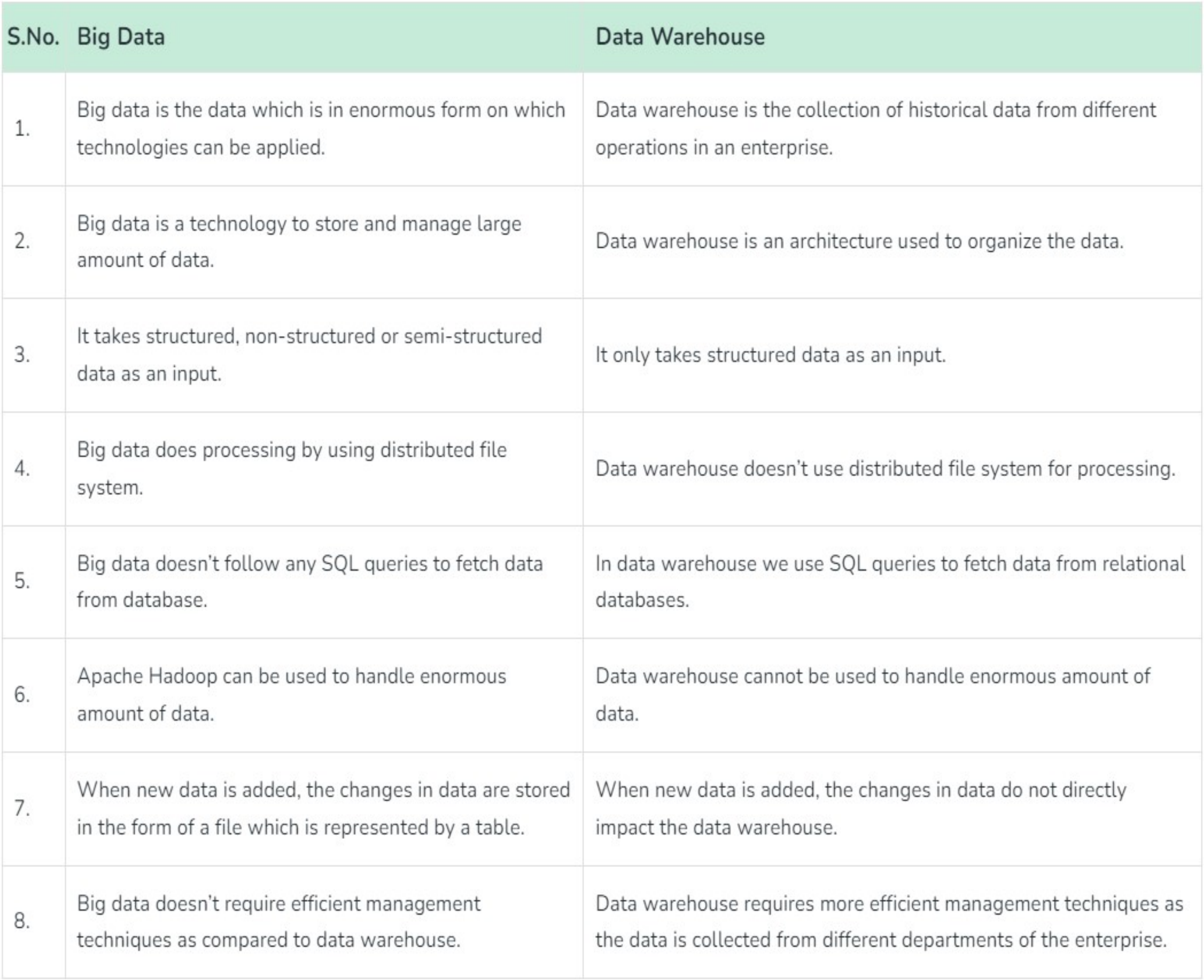
Traditional Data	Big Data
It is usually a small amount of data that can be collected and analyzed using traditional methods easily.	It is usually a big amount of data that cannot be processed and analyzed easily using traditional methods.
It is usually structured data and can be stored in spreadsheets, databases, etc.	It includes semi-structured, unstructured, and structured data.
It often collects data manually.	It collects information automatically with the use of automated systems.
It usually comes from internal systems.	It comes from various sources such as mobile devices, social media, etc.
It consists of data such as customer information, financial transactions, etc.	It consists of data such as images, videos, etc.
Analysis of traditional data can be done with the use of primary statistical methods.	Analysis of big data needs advanced analytics methods such as machine learning, data mining, etc.
Traditional methods to analyze data are slow and gradual.	Methods to analyze big data are fast and instant.
It generates data after the happening of an event.	It generates data every second.
It is typically processed in batches.	It is developed and processed in real-time.
It is limited in its value and insights.	It provides valuable insights and patterns for good decision-making.
It contains reliable and accurate data.	It may contain unreliable, inconsistent, or inaccurate data because of its size and complexity.
It is used for simple and small business processes.	It is used for complex and big business processes.
It does not provide in-depth insights.	It provides in-depth insights.
It is easy to secure and protect than big data because of its small size and simplicity.	It is harder to secure and protect than traditional data because of its size and complexity.
It requires less time and money to store traditional data.	It requires more time and money to store big data.
It can be stored on a single computer or server.	It requires distributed storage across numerous systems.
It is less efficient than big data.	It is more efficient than traditional data.
It can be managed in a centralized structure easily.	It requires a decentralized infrastructure to manage the data.

Refer Text book : Page 26

`

4. Compare Big Data and Data Warehouse.

Big Data and Data Warehousing are both concepts related to managing and analyzing large volumes of data, but they have distinct characteristics and purposes. Here's a comparison between Big Data and Data Warehousing



S.No.	Big Data	Data Warehouse
1.	Big data is the data which is in enormous form on which technologies can be applied.	Data warehouse is the collection of historical data from different operations in an enterprise.
2.	Big data is a technology to store and manage large amount of data.	Data warehouse is an architecture used to organize the data.
3.	It takes structured, non-structured or semi-structured data as an input.	It only takes structured data as an input.
4.	Big data does processing by using distributed file system.	Data warehouse doesn't use distributed file system for processing.
5.	Big data doesn't follow any SQL queries to fetch data from database.	In data warehouse we use SQL queries to fetch data from relational databases.
6.	Apache Hadoop can be used to handle enormous amount of data.	Data warehouse cannot be used to handle enormous amount of data.
7.	When new data is added, the changes in data are stored in the form of a file which is represented by a table.	When new data is added, the changes in data do not directly impact the data warehouse.
8.	Big data doesn't require efficient management techniques as compared to data warehouse.	Data warehouse requires more efficient management techniques as the data is collected from different departments of the enterprise.

Organizations often use both Big Data and Data Warehousing solutions complementarily to address different aspects of their data management and analysis needs. Big Data technologies are employed for handling the challenges posed by large volumes, variety, and velocity of data, while Data Warehousing is used for structured data analysis and reporting.

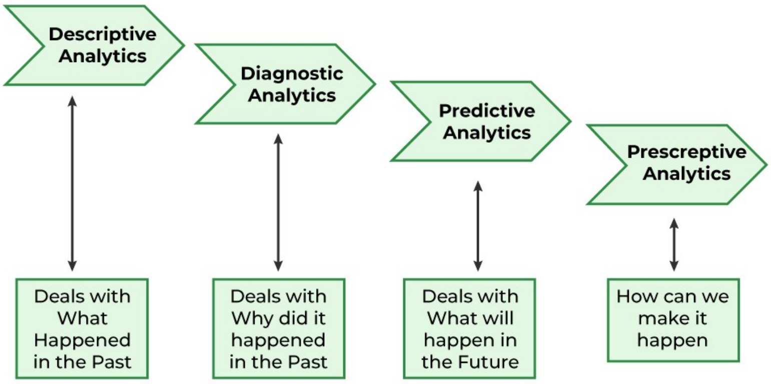
Refer Text book : Page 27

5.Classify different types of classification of analytics.

Classify different types of classification of analytics.

Types of Data Analytics
There are four major types of data analytics:

1. Predictive (forecasting)
2. Descriptive (business intelligence and data mining)
3. Prescriptive (optimization and simulation)
4. Diagnostic analytics


Descriptive Analytics
Diagnostic Analytics
Predictive Analytics
Prescreptive Analytics
Deals with What Happened in the Past
Deals with Why did it happened in the Past
Deals with What will happen in the Future
How can we make it happen

Predictive Analytics

Predictive analytics turn the data into valuable, actionable information. predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring. Predictive analytics holds a variety of statistical techniques from modeling, machine learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event. Techniques that are used for predictive analytics are:

Linear Regression
Time Series Analysis and Forecasting
Data Mining
Basic Corner Stones of Predictive Analytics
Predictive modeling
Decision Analysis and optimization
Transaction profiling

Descriptive Analytics

Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.

The descriptive model quantifies relationships in data in a way that is often used to classify customers or prospects into groups. Unlike a predictive model that focuses on predicting the behavior of a single customer, Descriptive analytics identifies many different relationships between customer and product.

Common examples of Descriptive analytics are company reports that provide historic reviews like:

Data Queries
Reports
Descriptive Statistics
Data dashboard
Prescriptive Analytics

`

Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction.

Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefits from the predictions and showing the decision maker the implication of each decision option. Prescriptive Analytics not only anticipates what will happen and when to happen but also why it will happen. Further, Prescriptive Analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option.

For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.

Diagnostic Analytics

In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.

For example, companies go for this analysis because it gives a great insight into a problem, and they also keep detailed information about their disposal otherwise data collection may turn out individual for every problem and it will be very time-consuming. Common techniques used for Diagnostic Analytics are:

Data discovery
Data mining
Correlations
These four types of data analytics form a continuum, with descriptive analytics providing the foundation by summarizing historical data, diagnostic analytics explaining why things happened, predictive analytics forecasting future events, and prescriptive analytics recommending the best actions to achieve desired outcomes. Organizations often use a combination of these analytics types to gain comprehensive insights into their data and support informed decision-making.

6.Compare and contrast descriptive analytics, predictive analytics and prescriptive analytics

The four major types of data analytics are often categorized based on their objectives and the questions they seek to answer. Here's a brief overview of each type:
Descriptive Analytics:
Objective: To understand what has happened in the past.
Description: Descriptive analytics involves summarizing historical data to provide insights into patterns, trends, and key performance indicators (KPIs). It often includes reporting, dashboards, and data visualization techniques. This type of analytics helps organizations gain a retrospective view of their operations.

`

Diagnostic Analytics:

Objective: To understand why something has happened.

Description: Diagnostic analytics involves digging deeper into historical data to identify the root causes of events or trends. It aims to uncover relationships and patterns in data to explain specific outcomes. Diagnostic analytics helps in understanding the factors that contributed to particular successes or failures.

Predictive Analytics:

Objective: To forecast future trends and outcomes.

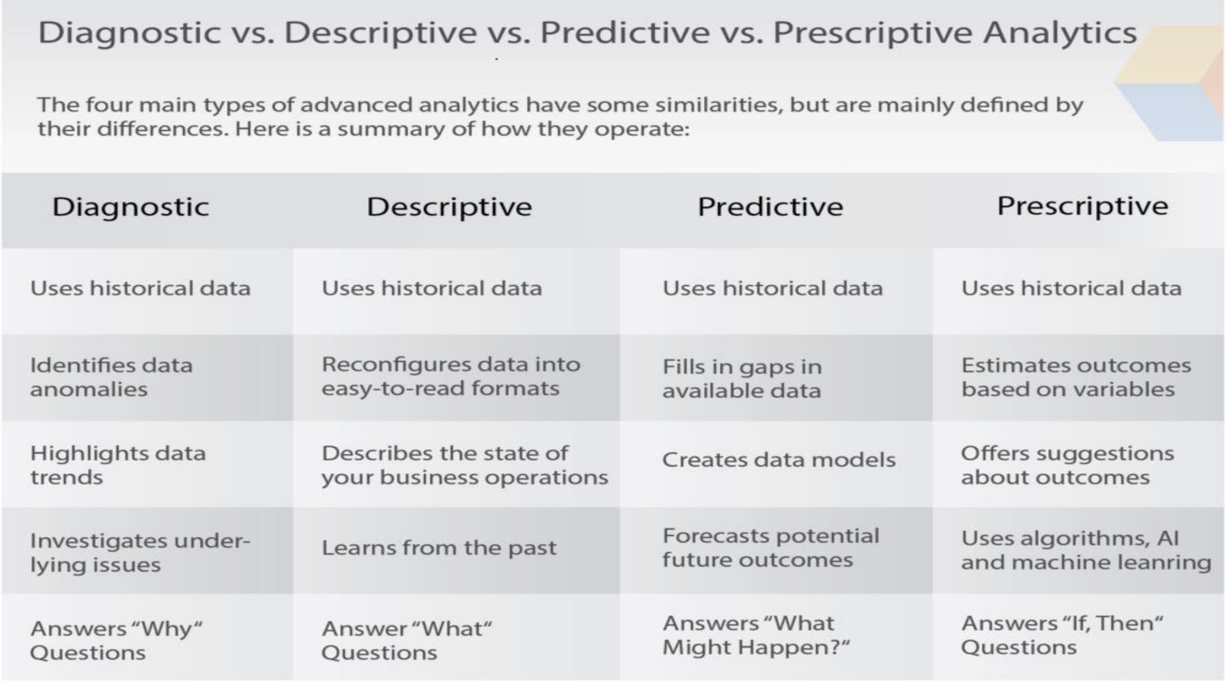
Description: Predictive analytics uses statistical algorithms and machine learning techniques to analyze historical data and make predictions about future events. It helps organizations anticipate trends, behavior, and outcomes, enabling proactive decision-making. Common applications include sales forecasting, demand planning, and risk management.

Prescriptive Analytics:

Objective: To recommend actions to optimize outcomes.

Description: Prescriptive analytics goes beyond predicting future scenarios to provide recommendations on the best course of action. It leverages optimization and simulation techniques to suggest decision strategies that will lead to the most favorable outcomes. Prescriptive analytics is valuable for decision-makers who seek to maximize efficiency and effectiveness.

These four types of data analytics form a continuum, with descriptive analytics providing the foundation by summarizing historical data, diagnostic analytics explaining why things happened, predictive analytics forecasting future events, and prescriptive analytics recommending the best actions to achieve desired outcomes. Organizations often use a combination of these analytics types to gain comprehensive insights into their data and support informed decision-making.



Diagnostic vs. Descriptive vs. Predictive vs. Prescriptive Analytics

The four main types of advanced analytics have some similarities, but are mainly defined by their differences. Here is a summary of how they operate:

Diagnostic	Descriptive	Predictive	Prescriptive
Uses historical data	Uses historical data	Uses historical data	Uses historical data
Identifies data anomalies	Reconfigures data into easy-to-read formats	Fills in gaps in available data	Estimates outcomes based on variables
Highlights data trends	Describes the state of your business operations	Creates data models	Offers suggestions about outcomes
Investigates under-lying issues	Learns from the past	Forecasts potential future outcomes	Uses algorithms, AI and machine leanring
Answers "Why" Questions	Answer "What" Questions	Answers "What Might Happen?"	Answers "If, Then" Questions

7.Explain latest analytics tools in detail.

Refer Question 10:

8.Explain various applications of big data in detail.
Big data has diverse applications across various industries, and its use continues to grow as organizations recognize its potential for extracting valuable insights and driving informed decision-making. Here are some common applications of big data:

Business Intelligence and Analytics:

Big data analytics helps businesses analyze large datasets to gain insights into customer behavior, market trends, and operational efficiency. It supports data-driven decision-making and strategic planning.

Healthcare Analytics:

Big data is used in healthcare for patient data management, predictive analytics for disease prevention, personalized medicine, clinical research, and optimizing healthcare operations.

Financial Services:

Banks and financial institutions use big data for fraud detection, risk management, customer segmentation, algorithmic trading, and analyzing market trends.

Retail and E-commerce:

Big data analytics helps retailers understand customer preferences, optimize pricing strategies, manage inventory, and enhance the overall customer experience through personalized recommendations.

Manufacturing and Supply Chain:

Big data is applied in manufacturing for predictive maintenance, quality control, supply chain optimization, and demand forecasting. It helps improve production efficiency and reduce operational costs.

Telecommunications:

Telecommunication companies use big data for network optimization, predictive maintenance of infrastructure, customer churn analysis, and to enhance the quality of services.

Smart Cities:

Big data is utilized in urban planning to manage traffic, optimize public transportation, monitor air quality, and improve overall city infrastructure for efficiency and sustainability.

Energy Sector:

Big data is applied in the energy industry for optimizing resource exploration, predictive maintenance of equipment, smart grid management, and improving energy efficiency.

Social Media and Marketing:

Social media platforms leverage big data for user behavior analysis, sentiment analysis, targeted advertising, and improving user engagement.

Education Analytics:

Educational institutions use big data for student performance analysis, personalized learning experiences, predictive analytics for identifying at-risk students, and optimizing educational resources.

Human Resources:

Big data supports HR analytics for talent acquisition, employee retention, workforce planning, and performance management.

Security and Fraud Detection:

Big data is crucial for cybersecurity, helping organizations detect and prevent cyber threats. It's also used for fraud detection in various industries, such as finance and healthcare.

Environmental Monitoring:

Big data aids in monitoring and analyzing environmental data, such as weather patterns, climate change, and natural disasters, enabling better preparedness and response.

Genomics and Personalized Medicine:

Big data analytics plays a significant role in genomics research, enabling the analysis of large-scale genomic data for personalized medicine, drug discovery, and understanding genetic factors in diseases.

9.Outline various terminologies used in big data environments and briefly discuss.

In-Memory Analytics:

Definition: In-memory analytics refers to the use of computer memory (RAM) to store and process data rather than traditional disk storage. By keeping data in memory, analytical queries and operations can be performed much faster, leading to quicker insights.

In-Database Processing:

Definition: In-database processing involves performing data analysis and computations directly within a database management system (DBMS). Instead of extracting and moving data to a separate analytics platform, computations are executed within the database, reducing data movement and enhancing performance.

Symmetric Multiprocessor System (SMP):

Definition: SMP is a type of multiprocessing architecture where multiple identical processors are connected to a single shared main memory. All processors have equal access to the memory, and tasks can be executed on any processor. SMP systems are commonly used in servers and high-performance computing.

Massively Parallel Processing (MPP):

Definition: MPP is an architecture where processing tasks are divided among many processors that work in parallel to solve a problem or process data. It is often used in large-scale data warehousing and analytics systems to distribute processing across multiple nodes for improved performance.

Difference Between Parallel and Distributed Systems:

Parallel Systems: Parallel systems involve multiple processors or cores working together on a single task, typically within the same machine. The primary goal is to improve performance by dividing the workload among processors.

`

Distributed Systems: Distributed systems involve multiple independent machines or nodes working together to achieve a common goal. Each node may have its own memory and resources. Communication between nodes is essential for collaborative processing.

Shared Nothing Architecture:

Definition: Shared Nothing Architecture is a distributed computing architecture in which each node in the system has its own dedicated resources, including memory and storage. Nodes operate independently, and communication between nodes involves passing messages. This architecture is often used in distributed databases and parallel processing systems.

Refer Text Book Page 45-47

10.List and explain a few top analytics tools.

Top Analytics Tools

* R is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

Features:

□ Effective data handling and storage facility,
□ It provides a suite of operators for calculations on arrays, in particular, matrices,
□ It provides coherent, integrated collection of big data tools for data analysis
□ It provides graphical facilities for data analysis which display either on-screen or on hardcopy

* **Apache Spark** is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets.

Features:

□ It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk

□ It offers lighting Fast Processing

□ Support for Sophisticated Analytics

□ Ability to Integrate with Hadoop and Existing Hadoop Data

*** Plotly is** an analytics tool that lets users create charts and dashboards to share online.

Features:

□ Easily turn any data into eye-catching and informative graphics

□ It provides audited industries with fine-grained information on data provenance

□ Plotly offers unlimited public file hosting through its free community plan

*** Lumify** is a big data fusion, analysis, and visualization platform. It helps users to discover connections and explore relationships in their data via a suite of analytic options.

Features:

□ It provides both 2D and 3D graph visualizations with a variety of automatic layouts

□ It provides a variety of options for analyzing the links between entities on the graph

□ It comes with specific ingest processing and interface elements for textual content, images, and videos

□ It spaces feature allows you to organize work into a set of projects, or workspaces

□ It is built on proven, scalable big data technologies

`

* **IBM SPSS** Modeler is a predictive big data analytics platform. It offers predictive models and
delivers to individuals, groups, systems and the enterprise. It has a range of advanced algorithms
and analysis techniques.

Features:

□ Discover insights and solve problems faster by analyzing structured and unstructured data
□ Use an intuitive interface for everyone to learn
□ You can select from on-premises, cloud and hybrid deployment options
□ Quickly choose the best performing algorithm based on model performance

* **MongoDB** is a NoSQL, document-oriented database written in C, C++, and JavaScript. It is free
to use and is an open source tool that supports multiple operating systems including Windows Vista (and later versions), OS X (10.7 and later versions), Linux, Solaris, and FreeBSD.
Its main features include Aggregation, Adhoc-queries, Uses BSON format, Sharding, Indexing, Replication, Server-side execution of javascript, Schemaless, Capped collection, MongoDB management service (MMS), load balancing and file storage.

Features:

□ Easy to learn.
□ Provides support for multiple technologies and platforms.
□ No hiccups in installation and maintenance.
□ Reliable and low cost.

Hive:

Description: Built on top of Hadoop, Apache Hive is a data warehousing and SQL-like query language for large-scale data processing. It enables users to query and analyze data using HiveQL, similar to SQL.

Apache Cassandra:

Description: A highly scalable, NoSQL database designed for handling large amounts of data across multiple commodity servers without a single point of failure. Cassandra is suitable for real-time analytics and can handle high write and read throughput.

Tableau:

Description: A data visualization tool that allows users to connect, visualize, and share insights from big data. Tableau supports various data sources, including Hadoop, and provides interactive dashboards and reports.