UNIT-01
2 MARKS

1.Summarize the reasons for the domain expertise for any type of data analytics.

Domain expertise is crucial in data analytics because it provides the necessary context to understand data, identify relevant patterns, and interpret results, leading to better data quality, more meaningful insights, and ultimately, more effective decision-making.

Here's a more detailed explanation of why domain expertise is so important:

- **Data Quality Assessment:**

Domain experts can identify anomalies, outliers, and potential biases in the data that might go unnoticed by someone lacking domain knowledge, ensuring the data is reliable and accurate.

- **Feature Engineering:**

Domain expertise helps identify which features are most relevant and meaningful for a particular problem, leading to more effective models and analysis.

- **Model Interpretation:**

Understanding the domain context is crucial for interpreting the output of a data model and its implications in the real world, ensuring that the insights are actionable and relevant.

- **Understanding Data Context:**

Domain knowledge helps analysts understand how data is collected, its sources, and the context in which it exists, which is essential for proper data preprocessing and cleaning.

- **Identifying Actionable Insights:**

Domain expertise allows analysts to translate data insights into practical recommendations and solutions that address real-world business issues.

- **Improved Communication:**

Domain expertise facilitates better communication between data analysts and stakeholders, ensuring that everyone understands the findings and can make informed decisions.

- **Innovation and Opportunity Identification:**

Domain experts can identify opportunities for innovation and new approaches to problem-solving based on their understanding of the industry and data.

- **Tailored Solutions:**

Domain knowledge allows for the development of solutions that are tailored to the specific needs and challenges of a particular industry or business.

1. Summarize the reasons for domain expertise in any type of data analytics.

Domain expertise is essential in data analytics as it helps in **understanding the context** of data, identifying relevant **trends, correlations, and anomalies**, and ensuring **accurate interpretation** of results. Without domain knowledge, analysts may misinterpret insights, leading to incorrect conclusions. Expertise in areas like healthcare, finance, or marketing ensures that data-driven decisions align with business objectives.

Example: In healthcare, a domain expert can help identify meaningful **patterns in patient data**, such as risk factors for diseases, leading to better diagnosis and treatment recommendations.

2.How the accuracy in big data is beneficial?

Data accuracy is vital to the success of all organizations—from sales to accounting and marketing to human resources. Data informs decisions, creates impressions about an organization, and drives revenue. Reasons why data accuracy is important and a priority for the enterprise are that data accuracy:

- Delivers better results to the organization’s users
- Drives more value from artificial intelligence implementations with accurate and consistent data to feed algorithms
- Enables better decision-making
- Enhances efficiency

- Makes it easier to achieve consistent results
- Mitigates risks associated with flawed data
- Provides confidence to users who depend on the data
- Reduces the need to spend time and money finding and fixing errors in the data
- Supports focused audience targeting and marketing efforts

2. How is accuracy in big data beneficial?

Accuracy in big data ensures **better decision-making, predictive analytics, and fraud detection**. High accuracy reduces **errors, biases, and misinformation**, leading to efficient business operations and improved customer experiences.

Example: In **financial services**, accurate big data analysis can detect fraudulent transactions in real-time, preventing losses and enhancing security.

3.What are the top challenges in big data?

The top challenges in big data revolve around managing its volume, velocity, variety, and veracity, along with data quality, security, and integration, as well as finding and retaining skilled talent.

Here's a more detailed breakdown:

1. Volume, Velocity, Variety, and Veracity (The 4 Vs):

- **Volume:**

  The sheer amount of data generated and stored is immense, requiring robust storage and processing infrastructure.

- **Velocity:**

  Data is generated and processed at an increasingly rapid pace, demanding real-time analytics and processing capabilities.

- **Variety:**

  Big data comes in various formats (structured, unstructured, semi-structured) from diverse sources, requiring flexible data management and analysis tools.

- **Veracity:**

  Data quality and accuracy are crucial, as inaccurate or incomplete data can lead to flawed insights and decisions

3. What are the top challenges in big data?

Big data presents challenges such as **data storage, processing speed, data quality, security, and privacy**. Managing the **3Vs (Volume, Velocity, Variety)** is difficult, requiring advanced tools like Hadoop and Spark. Ensuring **data integrity** and protecting sensitive information from cyber threats is also critical.

Example: Social media companies face challenges in analyzing millions of posts per second while filtering out **fake news** and maintaining **user privacy**.

4,12.**Distinguish between Big data and conventional data.**

Traditional Data	Big Data
Traditional data is generated in enterprise level.	**Big data is generated outside the enterprise level.**
Its volume ranges from Gigabytes to Terabytes.	**Its volume ranges from Petabytes to Zettabytes or Exabytes.**
Traditional database system deals with structured data.	**Big data system deals with structured, semi-structured,database, and unstructured data.**
Traditional data is generated per hour or per day or more.	**But big data is generated more frequently mainly per seconds.**
Traditional data source is centralized and it is managed in centralized form.	**Big data source is distributed and it is managed in distributed form.**
Data integration is very easy.	**Data integration is very difficult.**

Traditional Data	Big Data
Normal system configuration is capable to process traditional data.	**High system configuration is required to process big data.**
The size of the data is very small.	**The size is more than the traditional data size.**
Traditional data base tools are required to perform any data base operation.	**Special kind of data base tools are required to perform any database schema based operation.**
Normal functions can manipulate data.	**Special kind of functions can manipulate data.**
Its data model is strict schema based and it is static.	**Its data model is a flat schema based and it is dynamic.**
Traditional data is stable and inter relationship.	**Big data is not stable and unknown relationship.**
Traditional data is in manageable volume.	**Big data is in huge volume which becomes unmanageable.**
It is easy to manage and manipulate the data.	**It is difficult to manage and manipulate the data.**

4. Distinguish between big data and conventional data.

- **Big Data: Large, complex datasets that require specialized tools for analysis. It includes structured, unstructured, and semi-structured data from diverse sources.**
- **Conventional Data: Smaller, structured datasets stored in relational databases (RDBMS) with fixed schemas.**

Example:

- **Big Data: Real-time IoT sensor data from smart cities.**
- **Conventional Data: Employee records stored in an Excel sheet.**

5.Outline the role of big data analytics.

Big data analytics plays a crucial role in modern organizations by enabling them to extract valuable insights from large datasets, leading to better decision-making, improved operations, and increased competitiveness.

Here's a more detailed outline:

1. Uncovering Hidden Patterns and Trends:

- Identifying Opportunities:

Big data analytics helps organizations discover hidden patterns, correlations, and market trends that might otherwise go unnoticed, allowing them to identify new opportunities for growth and innovation.

- Predictive Analytics:

By analyzing historical data, organizations can use big data analytics to predict future outcomes, enabling them to make proactive decisions and mitigate potential risks.

- Understanding Customer Behavior:

Big data analytics provides insights into customer behavior, preferences, and needs, allowing businesses to personalize their offerings and improve customer experiences.

5. Outline the role of big data analytics.
Big data analytics helps in trend identification, predictive modeling, and real-time decision-making. It enhances customer experiences, operational efficiency, and fraud detection.
Example: E-commerce companies like Amazon use big data analytics to recommend products based on user behavior.

6,15.State the four computing resources of Big Data Storage.

The four main computing resources for Big Data Storage are data storage, data mining, data analytics, and data visualization.
Here's a more detailed explanation of each:

- Data Storage:

This involves the methods and technologies used to store large, complex datasets, which can include data lakes, data warehouses, cloud storage, and object storage.

- Data Mining:

This process involves extracting valuable insights and patterns from large datasets using various techniques and algorithms.

- Data Analytics:

This focuses on analyzing and interpreting the extracted data to gain meaningful insights and make informed decisions.

- Data Visualization:

This involves presenting the analyzed data in a visual format, such as charts, graphs, and dashboards, to facilitate understanding and communication of findings.

6. State the four computing resources of Big Data Storage.

1. **Processing Power (CPU & GPU):** Required for computing large datasets.
2. **Storage (HDFS, Cloud Storage):** Stores petabytes of structured and unstructured data.
3. **Memory (RAM):** Supports high-speed processing and real-time analytics.
4. **Networking (High-Speed Data Transfer):** Ensures seamless data movement across systems.

Example: Google Cloud provides **high-performance computing** for real-time big data processing.

7.What role does cloud computing play in Big Data Management?

Cloud Computing: The Big Data Solution

Cloud computing offers an effective solution towards dealing with big size information sets. Organizations can store their big-data efficiently manage them as well analyze them by leveraging scalability provided through clouds on demand resources such as storage capacity

7. What role does cloud computing play in Big Data Management?
Cloud computing provides **scalability, cost-efficiency, remote accessibility, and real-time processing** for big data applications. Platforms like **AWS, Google Cloud, and Azure** offer distributed computing and AI-powered analytics.
Example: Netflix uses **AWS cloud computing** to process and recommend personalized content to millions of users.

8.List the various dimensions of growth of big data.

The growth of big data is often characterized by the "5 V's": Volume, Velocity, Variety, Veracity, and Value.
Here's a breakdown of each dimension:

- Volume:

Refers to the sheer amount of data being generated and collected, often measured in terabytes, petabytes, or even exabytes.

- Velocity:

Describes the speed at which data is generated, collected, and processed, including real-time and batch processing.

- Variety:

Encompasses the different types and formats of data, including structured, unstructured, and semi-structured data.

- Veracity:

Relates to the accuracy, reliability, and trustworthiness of the data.

- Value:

Focuses on the insights and actionable knowledge that can be derived from the data, enabling better decision-making and business outcomes.

8. List the various dimensions of growth of big data.

1. **Volume:** Increase in the amount of data generated daily.
2. **Velocity:** Speed at which data is created and processed.
3. **Variety:** Diverse data types (text, images, audio, video).
4. **Veracity:** Ensuring data accuracy and reliability.
5. **Value:** Extracting meaningful insights for business impact.

Example: Twitter generates terabytes of text data every day, requiring advanced analytics.

9.Identify the risks involved in using big data.

Using big data presents risks including
data privacy breaches
security vulnerabilities
ethical concerns,
potential for misuse
managing large datasets.

9. Identify the risks involved in using big data.

- **Privacy breaches:** Exposure of sensitive user data.
- **Security vulnerabilities:** Risk of cyberattacks.
- **Bias in AI models:** Incorrect conclusions due to skewed data.

- **Data inconsistency:** Errors caused by poor-quality data.

Example: A social media company mishandling user data can lead to **data breaches** and legal consequences.

10.With example explain structured data.

Structured data refers to information organized in a predefined format, making it easily searchable and analyzable by computers, such as data in tables, databases, or spreadsheets. Examples include customer names, addresses, dates, times, and product prices.

- **Spreadsheets (Excel):**

Data organized in rows and columns, similar to relational databases, but with more flexibility in formatting and data types.

10. With an example, explain structured data.

Structured data is **highly organized and stored in relational databases** with predefined schemas. It is easy to **query using SQL**.

Example: A **customer database** containing fields like **Name, Age, Email, and Purchase History** stored in an **SQL table**.

11.State the best practices for Big Data Analytics.

- Implement Data Quality Management Programs.
- Build More Scalable Infrastructures.
- Employ Agile Development Methodologies.
- Safeguard Data With Robust Security Measures.
- Use Data Ethically.
- Monitor and Optimize Continuously.
- Provide Workforce Skill Development.

11. State the best practices for Big Data Analytics.

1. **Ensure data quality and consistency.**
2. **Use scalable and efficient storage solutions (HDFS, Cloud).**
3. **Implement security measures to protect sensitive data.**
4. **Leverage AI and ML for deeper insights.**
5. **Optimize processing speed with parallel computing.**

Example: A **bank** using AI-driven fraud detection models must ensure **data accuracy** for reliable predictions.

13.Compare data analytics and big data analytics.

The purpose of big data is to store huge volume of data and to process it.	The purpose of data analytics is to analyze the raw data and find out insights for the information.
Parallel computing and other complex automation tools are used to handle big data.	Predictive and statistical modelling with relatively simple tools used to handle data analytics.
Big data operations are handled by big data professionals.	Data analytics is performed by skilled data analysts.
Big data analysts need the knowledge of programming, NoSQL databases, distributed systems, and frameworks.	Data Analysts need the knowledge of programming, statistics, and mathematics.
It supports in dealing with huge volumes of data.	It supports in examining raw data and recognizing useful information.

13. Compare data analytics and big data analytics.

Feature	Data Analytics	Big Data Analytics
Data Size	Small to Medium	Very Large (Petabytes)
Processing	SQL, Excel	Hadoop, Spark
Complexity	Simple Queries	AI, ML, Predictive
Example	HR payroll data analysis	Sentiment analysis of social media posts

Identify the four computing resources of Big Data Storage.

11,14. List out the best practices of Big Data Analytics.

- **Define clear business objectives.**
- **Use the right technology stack (Spark, NoSQL).**
- **Ensure compliance with data privacy regulations (GDPR, HIPAA).**
- **Implement real-time monitoring for fraud detection.**

Example: Amazon optimizes inventory using real-time big data analytics.

16. Define big data and under what conditions it is given that name.

Big Data refers to extremely large datasets that traditional databases cannot handle efficiently. It is characterized by Volume, Velocity, Variety, Veracity, and Value.
Example: Facebook stores and processes billions of images, videos, and messages daily.

Why it's called "big data":

The term "big data" is used because these datasets are so large and complex that they require specialized technologies and tools to manage, store, and analyze them effectively.

characterized by the "5 Vs": volume, velocity, variety, veracity, and value.

- Examples:
  - Social media posts, emails, and web traffic data.
  - Financial transaction records, customer databases, and inventory data.
  - Sensor data from IoT devices, medical records, and scientific experiments.