Assignment 1

1. Which of the following best describes the concept of 'Big Data'?
    a. Data that is physically large in size
    b. Data that is collected from multiple sources and is of high variety, volume, and velocity
    c. Data that requires specialized hardware for storage
    d. Data that is highly structured and easily analyzable

Ans- Big Data is characterized by the "Three Vs": variety (different types of data), volume (large amounts of data), and velocity (speed at which data is generated and processed). This definition captures the essence of Big Data, distinguishing it from merely large or structured datasets.

2. Which technology is commonly used for processing and analyzing Big Data in distributed computing environments?
    a. MySQL
    b. Hadoop
    c. Excel
    d. SQLite

Ans- Hadoop is a widely-used framework designed for processing and analyzing large datasets in distributed computing environments. It provides a scalable and fault-tolerant way to handle Big Data, unlike MySQL, Excel, or SQLite, which are not typically used for large-scale distributed processing.

3. What is a primary limitation of traditional RDBMS when dealing with Big Data?
    a. They cannot handle structured data
    b. They are too expensive to implement
    c. They struggle with scaling to manage very large datasets
    d. They are not capable of performing complex queries

Ans- Traditional Relational Database Management Systems (RDBMS) often face challenges with scalability when handling Big Data, primarily due to their limited ability to distribute data across multiple nodes. They are not inherently designed for the scale required by Big Data.

4. Which component of Hadoop is responsible for distributed storage?

a. YARN
b. HDFS
c. MapReduce
d. Pig

Ans- The Hadoop Distributed File System (HDFS) is the component responsible for storing data across a distributed cluster, providing redundancy and fault tolerance. YARN is for resource management, MapReduce is a processing framework, and Pig is a high-level data flow language.

5. Which Hadoop ecosystem tool is primarily used for querying and analyzing large datasets stored in Hadoop's distributed storage?

a. HBase
b. Hive
c. Kafka
d. Sqoop

Ans- Hive is a data warehousing and SQL-like query language tool used to query and analyze large datasets in Hadoop. HBase is a NoSQL database, Kafka is a messaging system, and Sqoop is used for data transfer between Hadoop and relational databases.

6. Which YARN component is responsible for coordinating the execution of tasks within containers on individual nodes in a Hadoop cluster?

a. NodeManager
b. ResourceManager
c. ApplicationMaster
d. DataNode

Ans- NodeManager is the YARN component responsible for managing resources and monitoring the execution of tasks on individual nodes. ResourceManager manages overall cluster resources, ApplicationMaster handles application-specific resource requests, and DataNode is part of HDFS.

7. What is the primary advantage of using Apache Spark over traditional MapReduce for data processing?

a. Better fault tolerance
b. Lower hardware requirements
c. Real-time data processing
d. Faster data processing

Ans- Apache Spark provides faster data processing compared to traditional MapReduce due to its in-memory processing capabilities, which reduce the need for disk I/O operations. This leads to significant performance improvements for iterative algorithms and complex data processing tasks.

8. What is Apache Spark Streaming primarily used for?

a. Real-time data visualization
b. Batch processing of large datasets
c. Real-time stream processing
d. Data storage and retrieval

Ans- Apache Spark Streaming is designed for real-time stream processing, enabling the analysis of live data streams. It is not used for batch processing, real-time visualization, or data storage and retrieval.

9. Which operation in Apache Spark GraphX is used to perform triangle counting on a graph?

a. `connectedComponents`
b. `triangleCount`
c. `shortestPaths`
d. `pageRank`

And-The `triangleCount` operation in Apache Spark GraphX is used to count the number of triangles in a graph, which helps in analyzing the structure and connectivity of the graph.

10. Which component in Hadoop is responsible for executing tasks on individual nodes and reporting back to the JobTracker?

a. HDFS Namenode
b. TaskTracker
c. YARN ResourceManager
d. DataNode

Ans- The TaskTracker is responsible for executing MapReduce tasks on individual nodes and reporting the progress and status back to the JobTracker. The HDFS Namenode manages the file system namespace, the YARN ResourceManager allocates resources, and DataNode stores the actual data.