Assignmet -2

1. Which statement best describes the data storage model used by HBase?
    a. Key-value pairs
    b. Document-oriented
    c. Encryption
    d. Relational tables

Ans-

Option a: Key-value pairs - Correct. HBase is a NoSQL database that stores data in a key-value format. Each row is identified by a unique key, and columns within a row are organized into column families.

Option b: Document-oriented - Incorrect. Document-oriented databases like MongoDB store data as self-contained documents, while HBase uses a key-value model.

Option c: Encryption- Incorrect.

Option d: Relational tables - Incorrect. Relational databases use structured tables with rows and columns, while HBase is a NoSQL database with a flexible schema.

2. What is Apache Avro primarily used for in the context of Big Data?
    a. Real-time data streaming
    b. Data serialization
    c. Machine learning
    d. Database management

Ans-

Option a: Real-time data streaming - Incorrect. While Avro can be used in streaming applications, its primary focus is data serialization.

Option b: Data serialization - Correct. Avro is a data serialization format that efficiently encodes data structures for storage and transmission.

Option c: Machine learning - Incorrect. Avro can be used to store data for machine learning models, but its core functionality is data serialization.
Option d: Database management - Incorrect. Avro is not a database management system, but a format for storing data.

Explanation-
Apache Avro is a framework for data serialization. It provides a compact, fast, and efficient way to serialize and deserialize data, making it suitable for communication between different systems or for persisting data in a binary format. Avro is commonly used in Big Data applications for serialization of data in a way that supports schema evolution and provides interoperability across various programming languages.

3.Which component in HDFS is responsible for storing actual data blocks on the DataNodes?

a. NameNode
b. DataNode
c. Secondary NameNode
d. ResourceManager

Ans-option a: NameNode - Incorrect. The NameNode manages metadata about the file system, such as block locations and file permissions.
option b: DataNode - Correct. DataNodes are the physical storage units in HDFS that store data blocks.
option c: Secondary NameNode - Incorrect. The Secondary NameNode is a backup of the NameNode, not for data storage.
option d: ResourceManager - Incorrect. The ResourceManager is part of YARN, responsible for resource management in Hadoop, not data storage.
Explanation-

In the Hadoop Distributed File System (HDFS), DataNodes are responsible for storing the actual data blocks. Each DataNode manages the storage of data blocks and periodically sends heartbeat signals and block reports to the NameNode to confirm its status and the health of the data blocks it stores.

4.Which feature of HDFS ensures fault tolerance by replicating data blocks across multiple DataNodes?

a. Partitioning
b. Compression
c. [✓ Replication]
d. Encryption

Ans- option a: partitioning - Incorrect. Partitioning divides data into smaller chunks for processing, not for fault tolerance.
option b: compression - Incorrect. Compression reduces data size, but doesn't provide redundancy.
option c: replication - Correct. HDFS replicates data blocks across multiple DataNodes to ensure data availability in case of node failures.
option d: encryption - Incorrect. Encryption protects data confidentiality, not availability.

Explanation-
HDFS achieves fault tolerance through the replication of data blocks. Each data block is replicated across multiple DataNodes, which helps in ensuring data availability and reliability even in the event of hardware failures. By default, each block is replicated three times across different nodes to safeguard against data loss.

5.Which component in MapReduce is responsible for sorting and grouping the intermediate key-value pairs before passing them to the Reducer?

a. Mapper
b. Reducer
c. [✓ Partitioner]
d. Combiner

Ans- option a: Mapper - Incorrect. The Mapper generates key-value pairs, but doesn't perform sorting or grouping.
option b: Reducer - Incorrect. The Reducer processes the grouped key-value pairs, but doesn't perform the initial sorting and grouping.
option c: Partitioner - Correct. The Partitioner determines which Reducer will process a specific key-value pair.

option d: Combiner - Incorrect. The Combiner is an optional optimization that can reduce data volume before sending it to the Reducer, but it doesn't perform sorting and grouping.

Explanation-

In the MapReduce framework, the Partitioner is responsible for distributing the intermediate key-value pairs generated by the Mapper to the appropriate Reducer tasks. It handles the sorting and grouping of these pairs to ensure that all values for a given key are sent to the same Reducer. The sorting and grouping happen as part of the shuffle and sort phase of the MapReduce process.

6.What is the default replication factor in Hadoop Distributed File System (HDFS)?

a. 1
b. 2
c. 3
d. 4

Ans- option a: 1 - Incorrect. A replication factor of 1 would offer no fault tolerance.
option b: 2 - Incorrect. A replication factor of 2 provides some fault tolerance, but 3 is the default.
option c: 3 - Correct. The default replication factor in HDFS is 3, providing a balance between fault tolerance and storage efficiency.
option d: 4 - Incorrect. A replication factor of 4 would increase storage overhead without significantly improving fault tolerance.

Explanation-

The default replication factor in HDFS is 3. This means that each data block is replicated three times across different DataNodes. This default replication factor strikes a balance between data redundancy and storage overhead, providing fault tolerance and high availability for the data.

7.In a MapReduce job, what is the role of the Reducer?

a. Sorting input data
b. Transforming intermediate data
c. Aggregating results
d. Splitting input data

Ans- option a: sorting input data - Incorrect. The Mapper and Partitioner handle data sorting and distribution.
option b: transforming intermediate data - Correct. The Reducer can transform intermediate data based on the key-value pairs it receives.
option c: aggregating results - Correct. The Reducer is often used to aggregate values based on the key.
option d: splitting input data - Incorrect. The input data is split into blocks by the InputFormat.
Explanation-

The Reducer in a MapReduce job is responsible for aggregating the intermediate data produced by the Mapper. It takes the sorted and grouped key-value pairs from the shuffle and sort phase and performs a reduction operation, which might involve summing up values, calculating averages, or other forms of aggregation depending on the specific job requirements.

8.Which task can be efficiently parallelized using MapReduce?
a. Real-time sensor data processing
b. Single-row database queries
c. Image rendering
d. Log file analysis

Ans- option a: real-time sensor data processing - Incorrect. MapReduce is better suited for batch processing than real-time processing.
option b: single-row database queries - Incorrect. Single-row database queries are typically handled by relational databases.
option c: image rendering - Incorrect. Image rendering often requires specialized hardware and algorithms.
option d: log file analysis - Correct. Log file analysis involves processing large amounts of data, making it a good candidate for MapReduce.

Explanation-
MapReduce is particularly well-suited for tasks that can be parallelized across a large number of independent data chunks. Log file analysis is an example of such a task, as log files can be split into segments that can be processed in parallel. Each Mapper processes a chunk of log data to extract relevant information, and the Reducer aggregates and processes the results.

9.Which MapReduce application involves counting the occurrence of words in a large corpus of text?

- a. PageRank algorithm
- b. K-means clustering
- c. Word count
- d. Recommender system

Ans- option a: PageRank algorithm - Incorrect. PageRank is used for ranking web pages.
option b: K-means clustering - Incorrect. K-means clustering is used for grouping data points.
option c: word count - Correct. A word count application counts the frequency of words in a text corpus.
option d: recommender system - Incorrect. Recommender systems use collaborative filtering or content-based approaches.

Explanation-
The Word Count application is a classic example of a MapReduce job. It involves counting the frequency of each word in a large corpus of text. The Mapper extracts words from the text and emits them as key-value pairs with a count of 1. The Reducer then sums up these counts for each unique word to produce the final word count results.

10.What does reversing a web link graph typically involve?

- a. Removing dead links from the graph
- b. Inverting the direction of edges
- c. Adding new links to the graph
- d. Sorting links based on page rank

Ans- option a: removing dead links from the graph - Incorrect. Removing dead links is a different task.
option b: inverting the direction of edges - Correct. Reversing a web link graph means changing the direction of links, creating a graph where pages are pointed to instead of pointing to others.
option c: adding new links to the graph - Incorrect. Reversing doesn't involve adding new links.
option d: sorting links based on page rank - Incorrect. Sorting links is a different operation.

Explanation-
Reversing a web link graph involves inverting the direction of the edges between nodes (web pages). In a web link graph, each directed edge represents a hyperlink from one page to another. Reversing the graph means changing the direction of these links, so a link from Page A to Page B becomes a link from Page B to Page A. This is useful for various analyses, such as computing PageRank in a different context or understanding link relationships from a different perspective.