assignment 3

1. Which abstraction in Apache Spark allows for parallel execution and distributed data processing?
    a. DataFrame
    b. RDD (Resilient Distributed Dataset)
    c. Dataset
    d. Spark SQL

Option a: DataFrame - Incorrect. DataFrames provide a higher-level API for structured data, but they are not the fundamental abstraction for parallel execution.
Option b: RDD (Resilient Distributed Dataset) - **Correct**. RDDs are the fundamental abstraction in Spark for distributed, fault-tolerant, and parallel processing of large datasets.
Option c: Dataset - Incorrect. Datasets are a more recent addition to Spark, combining the benefits of RDDs and DataFrames. However, RDDs are the core concept for parallel execution.
Option d: Spark SQL - Incorrect. Spark SQL is a SQL engine built on top of Spark, providing SQL-like capabilities. While it uses RDDs internally, it's not the direct abstraction for parallel execution.

2. What component resides on top of Spark Core?

A) Spark Streaming

B) Spark SQL

C) RDDs

D) None of the above

Ans - B) Spark SQL

Option A: Spark Streaming - Incorrect. Spark Streaming is a component that provides stream processing capabilities and builds on top of Spark Core, but it is not the answer to this specific question, as Spark SQL is more directly related to structured data processing.

Option B: Spark SQL - Correct. Spark SQL is a component that resides on top of Spark Core. It provides a higher-level API for querying structured data using SQL syntax and integrates with Spark's core functionalities through DataFrames and Datasets.

Option C: RDDs - Incorrect. RDDs are the fundamental abstraction in Spark Core for parallel execution and distributed processing. They are not a component that resides on top of Spark Core but rather part of the core abstraction.

Option D: None of the above - Incorrect. Spark SQL is indeed a component that resides on top of Spark Core, making this option incorrect.

3. Which statements about Cassandra and its Snitches are correct?

Statement 1: In Cassandra, during a write operation, when a hinted handoff is enabled and if any replica is down, the coordinator writes to all other replicas and keeps the write locally until the down replica comes back up.

Statement 2: In Cassandra, Ec2Snitch is an important snitch for deployments, and it is a simple snitch for Amazon EC2 deployments where all nodes are in a single region. In Ec2Snitch, the region name refers to the data center, and the availability zone refers to the rack in a cluster.

A) Only Statement 1 is correct.

B) Only Statement 2 is correct.

C) Both Statement 1 and Statement 2 are correct.

D) Neither Statement 1 nor Statement 2 is correct.

Ans- C) Both Statement 1 and Statement 2 are correct.

Statement 1: Correct. In Cassandra, when a hinted handoff is enabled, if any replica is down during a write operation, the coordinator writes to all other available replicas and keeps a hint for the down replica. Once the down replica comes back online, the coordinator will hand off the hinted write to that replica.

Statement 2: Correct. Ec2Snitch is a snitch used in Cassandra for Amazon EC2 deployments. It assumes that all nodes are within a single region. In Ec2Snitch, the term "region" corresponds to the data center, and "availability zone" corresponds to the rack within the cluster, helping to optimize data placement and replication.

4.Which of the following is a module for Structured data processing?

a. GraphX
b. MLlib
c. Spark SQL
d. Spark R

Option a: GraphX- Incorrect. This module is for graph processing and analytics, allowing for the manipulation and analysis of graph data structures.
Option b: MLlib- Incorrect. This is Spark’s machine learning library, providing algorithms and utilities for machine learning tasks, not specifically for structured data processing.
Option c: Spark SQL- Correct. Spark SQL is the module designed specifically for structured data processing. It provides a programming interface for working with structured and semi-structured data. It allows querying of data via SQL, integrates with DataFrames and Datasets, and provides optimizations through the Catalyst optimizer and Tungsten execution engine.

Option d: Spark R - Incorrect. This module provides support for using R with Spark, primarily aimed at statistical computing and data analysis rather than structured data processing specifically.

5. A healthcare provider wants to store and query patient records in a NoSQL database with high write throughput and low-latency access. Which Hadoop ecosystem technology is most suitable for this requirement?

A) Apache Hadoop

B) Apache Spark

C) Apache HBase

D) Apache Pig

Ans- C) Apache HBase

Option A: Apache Hadoop - Incorrect. Apache Hadoop is a framework for distributed storage and processing of large data sets using the Hadoop Distributed File System (HDFS) and MapReduce. It is not specifically optimized for low-latency access or high write throughput.

Option B: Apache Spark - Incorrect. Apache Spark is a fast, in-memory data processing engine that can handle large-scale data analytics and processing. While it offers low-latency data processing, it is not a NoSQL database and is not designed primarily for high write throughput.

Option C: Apache HBase - Correct. Apache HBase is a distributed, scalable, NoSQL database that runs on top of HDFS. It is designed for high write throughput and low-latency access to large volumes of data, making it suitable for storing and querying patient records efficiently.

Option D: Apache Pig - Incorrect. Apache Pig is a high-level platform for creating MapReduce programs used with Hadoop. It is primarily used for data transformation and analysis, not for high write throughput or low-latency NoSQL data storage.

6.The primary Machine Learning API for Spark is now the _____ based API

a. DataFrame
b. Dataset
c. RDD
d. All of the above

Option A: DataFrame - Correct. The primary Machine Learning API for Spark is now based on DataFrames. Spark's MLlib, the machine learning library, has adopted DataFrames as the primary API for building and training machine learning models. This approach provides a higher-level API and better integration with Spark SQL, offering optimized performance and ease of use.

Option B: Dataset - Incorrect. While Datasets are a powerful API in Spark that provides type safety and functional programming constructs, the primary Machine Learning API is not based on Datasets. Instead, DataFrames are used.

Option C: RDD - Incorrect. RDDs (Resilient Distributed Datasets) were the original abstraction in Spark and were used in earlier versions of MLlib. However, the primary Machine Learning API has shifted to DataFrames for better integration and performance.

Option D: All of the above - Incorrect. While RDDs, DataFrames, and Datasets are all important abstractions in Spark, the primary Machine Learning API is now specifically based on DataFrames

7. How does Apache Spark's performance compare to Hadoop MapReduce?

a) Apache Spark is up to 10 times faster in memory and up to 100 times faster on disk.

b) Apache Spark is up to 100 times faster in memory and up to 10 times faster on disk.

c) Apache Spark is up to 10 times faster both in memory and on disk compared to Hadoop MapReduce.

d) Apache Spark is up to 100 times faster both in memory and on disk compared to Hadoop MapReduce.

Ans- b) Apache Spark is up to 100 times faster in memory and up to 10 times faster on disk.

Option a: Incorrect. While Spark is indeed faster than Hadoop MapReduce, the comparison numbers are not accurate. Spark can be up to 100 times faster in memory, but the figure of 10 times faster on disk is not the correct characterization.

Option b: Correct. Apache Spark is known for its significant performance improvements over Hadoop MapReduce. It can be up to 100 times faster when processing data in memory, due to its in-memory computation capabilities. On disk, Spark is up to 10 times faster compared to MapReduce because of its efficient data processing and reduced disk I/O.

Option c: Incorrect. Spark is not just 10 times faster both in memory and on disk. It achieves up to 100 times faster performance in memory and up to 10 times faster on disk.

Option d: Incorrect. While Spark can be up to 100 times faster in memory, it is not typically characterized as being up to 100 times faster on disk. The correct comparison indicates up to 10 times faster on disk.

8.Which DAG action in Apache Spark triggers the execution of all previously defined transformations in the DAG and returns the count of elements in the resulting RDD or DataFrame?

a. collect()
b. count()
c. take()
d. first()

Option a: collect() - **Incorrect.** The `collect()` action triggers the execution of all previously defined transformations and retrieves all elements of the RDD or DataFrame to the driver program. It does not return the count of elements but rather returns the complete dataset.

Option b: count() - **Correct.** The `count()` action triggers the execution of all previously defined transformations in the DAG and returns the number of elements in the resulting RDD or DataFrame. It is specifically designed to return the count of elements.

Option c: take() - **Incorrect.** The `take()` action triggers execution and retrieves a specified number of elements from the RDD or DataFrame, but it does not return the count of all elements.

Option d: first() - **Incorrect.** The `first()` action triggers execution and retrieves the first element of the RDD or DataFrame, but it does not return the count of elements.

9. What is Apache Spark Streaming primarily used for?

  a. Real-time processing of streaming data
  b. Batch processing of static datasets
  c. Machine learning model training
  d. Graph processing

Option a: Real-time processing of streaming data - **Correct**. Spark Streaming is designed for processing continuous streams of data in real-time.
Option b: Batch processing of static datasets - **Incorrect.** Batch processing is better suited for static datasets.
Option c: Machine learning model training - **Incorrect.** While Spark can be used for machine learning, Spark Streaming is specifically for streaming data.
Option d: Graph processing - **Incorrect.** Graph processing is another area where Spark can be used, but Spark Streaming is focused on streaming data.

10.Which of the following represents the smallest unit of data processed by Apache Spark Streaming?

  a. Batch
  b. Window
  c. Micro-batch
  d. Record

Option a: Batch - **Incorrect.** A batch is a collection of micro-batches.
Option b: Window - **Incorrect.** A window is a time interval used for processing data.
Option c: Micro-batch - **Correct**. Micro-batches are the smallest unit of data processed in Spark Streaming.
Option d: Record - **Incorrect.** A record is a single unit of data within a micro-batch.