**Assessment 1 QUIZZES CSE5BDC**

**Quiz 1.**

**Which of the following is a true benefit of using MapReduce compared to RDBMS?**

**a.**

**MapReduce is more scalable than RDBMS and therefore can handle larger datasets than RDBMS.**

b.

MapReduce can be used to retrieve data faster using an index compared to RDBMS.

c.

MapReduce performs query optimisation and therefore can execute queries faster than RDBMS.

d.

MapReduce always runs faster than RDBMS.

**Which of the following statements about the benefits of MapReduce is false?**

**a.**

**MapReduce shuffles less data as a result of using a scaling-out architecture.**

b.

MapReduce can easily scale out the compute to include more nodes of a cluster.

c.

MapReduce can automatically recover from failed compute nodes.

d.

MapReduce can make use of multiple disks to make data loading faster.

**Which of the following is a true benefit of using local aggregation within MapReduce?**

a. Less data are written out to disk.

b. The mapper will use less memory.

c. Less data will be aggregated at the mapper.

**d. Less data will be shuffled from the mapper to the reducer.**

**Given 1 million employee records (containing each employee's name, phone number, date of birth and company join date), which of the following problems will benefit the least from being solved using MapReduce?**

a.

Select all employees who joined the company after 2018.

b.

Output the name and phone number of each employee.

c.

Count the number of employees that joined the company in the last five years

**d.**

**Perform an operation on the employee records that requires comparing every employee record with every other employee record.**

**Which of the following statements about the way MapReduce works is false?**

**a. Data from the mapper are shuffled and grouped at the reducer according to the output value of the mapper.**

b.

Data in the reducer are sorted according to the output key of the mapper.

c.

A MapReduce job that just requires selecting certain lines of data does not require a reducer.

d.

Data from the mapper are shuffled and grouped at the reducer according to the output key of the mapper.

**Quiz 2.**

**What is the benefit of using subqueries in Hive versus writing multiple separate queries to solve a problem?**

**c.**

**There is a potential for less map reduce jobs to be executed.**

**Which of the following statements is true of Hive?**

**d.**

**SORT BY only sorts the data within each reducer.**

**Suppose you are given a table with the following columns: City Name, Country Name, Population. The table contains every city for every country in the world and their population.  
You are asked to output the city with the highest population among all cities in Australia. Which Hive commands would you use?**

**b.**

**SELECT, WHERE, ORDER BY, LIMIT**

**Why does Hive allow users to define complex data types like structs, whereas traditional relational databases do not allow it?**

**a.**

**Traditional relational databases perform schema on write.**

**Suppose you are asked to output the url, ipAddress and time of the most recent log entry in the mytraffic table of lab 3, where each row of the mytraffic table represents a log entry. Which of the following set of Hive commands best solves this problem?**

**d.**

**SELECT, ORDER BY, LIMIT**

**Which statement about the MapReduce job tracker is false?**

a.

The job tracker is responsible for balancing the workload across the different task trackers.

b.

There is a fixed number of map and reduce slots allocated by the job tracker to each compute node.

c.

The job tracker is responsible for monitoring the health of every task tracker.

**d.**

**In a cluster, there can be multiple job trackers running at the same time.**

**Suppose you are given a table with the following columns: City Name, Country Name, Population. The table contains every city for every country in the world and their populations.  
You are asked to list the top 10 countries in descending order of total population using Hive. Which Hive commands would you use?**

**b.**

**SELECT, GROUP BY, ORDER BY, LIMIT**

**Quiz 3.**

**Given a Scala list called numbers initialised with integer values, what does the following line of Scala code do?  
numbers.map(\_ / 2.0).reduce(\_ \* \_)**

b.

Divides each value in the numbers list by 2.0 and then computes the product of all the values in the numbers list.

**Suppose you are given any list of numbers - for example, List(1, 3, 3, 4, 2, 5, 9, 6, 1) - and you are asked to return the highest even number. Which of the following set of functions best solves the problem?**

**c.**

filter, reduce, max

**How does Apache Spark achieve fault tolerance?**

**d.**

**Uses lineage to record all transformations that have been performed and replay the necessary transformations.**

**Which of the following Yarn managers is responsible for keeping track of all the tasks for big data processing jobs?**

**d.**

**The application master**

**Consider the following Scala code: val cost = 29.19. Which of the following statements is false?**

**b.**

**You can assign cost to a different value but the new value must be of type Double.**

**Which one of the following components of Storm is responsible for performing the computing?**

**d.**

**Bolts**

**Which of the following tasks should NOT be done using Storm?**

**d.**

**Allowing millions of people to update their bank accounts at the same time.**

**Which of the following are NOT contributing reasons for MapReduce to be so bad for processing graph algorithms?**

**d.**

**MapReduce cannot perform the necessary computation required by graph algorithms.**

**Quiz 4.**

**Given the censusSplit RDD from task 3 of lab 5. Suppose you are asked to output the total income of all people in the dataset. Which of the following sets of RDD transforms/actions best solves the problem?**

**b.**

**map, reduce**

**Consider the censusSplit RDD from task 3 of lab 5. Suppose you are asked to output the age and fntwgt of the person with the maximum fntwgt among all people in the dataset. Which of the following sets of RDD transforms/actions best solves the problem?**

**b.**

**map, sortByKey, first**

**You are given an RDD of three-element tuples (<name>, <occupation>, <salary>). You are asked to output the total salary of all people in the ‘teacher' occupation. Which of the following sets of RDD transforms best solves the problem?**

**a.**

**map, filter, reduceByKey**

**Which of following is NOT an advantage of using structured programming with SparkSQL dataframes compared to programming using the Spark RDD API?**

**d.**

**Structured programming allows data to be cached in RAM.**

**Consider the censusSplit RDD from task 3 of lab 5. Suppose you are asked to find the maximum flnwgt wage for each education category. Which of the following set of RDD transforms/actions best solves the problem?**

**c.**

**map, sortByKey, first (this may not be correct)**

**Quiz 5.**

**Which of the following statements regarding data caching in Apache Spark is false?**

**a.**

**Caching only a part of an RDD has no performance benefits.**

b.

Caching data are especially important for the performance of iterative programs.

c.

Caching reduces the amount of disk access and therefore speeds up query execution.

d.

RDDs in Apache Spark are only cached if you explicitly specify that you want the RDD to be cached.

**Which of the following statements about parquet storage format is false?**

a.

Parquet storage format stores the schema with the data.

**b.**

**Given a dataframe with 100 columns, it is faster to query a single column of the dataframe if the data are stored using the CSV storage format compared to the parquet storage format.**

c.

Parquet storage format stores all values of the same column together.

d.

Given a dataframe with 100 columns, it is faster to query a single column of the dataframe if the data are stored using the parquet storage format compared to the data being stored in the CSV storage format.

**Which of the following statements is false?**

a.

Executing queries using SparkSQL DataFrames and DataSets functions are at least as fast as using their RDD counterparts, and often faster.

b.

You can add columns to a dataframe using the withColumn function.

**c.**

**DataSets contain schemas whereas DataFrames do not contain schemas.**

d.

After performing a self-join on a dataframe, the resulting columns will contain duplicate column names.

**What is a benefit of using the partitionBy function in SparkSQL?**

It allows you to quickly retrieve all data associated with a given value on the partitioned column.

**Which of the following statements about query optimisation in Spark is false?**

a.

Query optimisation can change the order by which operations are applied on dataframes to speed up query execution performance.

b.

Query optimisation is automatically used when querying data via dataframes in SparkSQL.

c.

Query optimisation is one of the main benefits of using SparkSQL.

**d.**

**Spark automatically applies query optimisation on a sequence of RDD transformations.**

**Which of the following statements is false?**

a.

Writing a query using SparkSQL dataframes will usually run faster than writing the same query using RDDs instead.

b.

Using reduceByKey instead of groupByKey can reduce the amount of data shuffled.

**c.**

**You need to explicitly invoke a combiner in order to enjoy the benefits of reduced data shuffle when using the reduceByKey function.**

d.

Suppose you want to get the maximum value from an RDD of integers. It is faster to use the reduce function than to sort the values in the RDD in descending order and then return the first value.