1

We have a 500MB file that we attempt to upload to HDFS with replication factor 3 and block size 128 MB as default settings. Blocks are produced. Which HDFS component is in charge of this?

**SELECT THE CORRECT ANSWER**

Top of Form

Namenode

Datanode

Zookeeper

Client

Bottom of Form

Correct Option:D

**EXPLANATION**

When we upload files to HDFS, the HDFS client breaks the original file into multiple blocks of the specified block size (128 MB in this case)

2

What procedure is utilised in distributed systems to store redundant data for fault recovery?

**SELECT THE CORRECT ANSWER**

Top of Form

Partition

Erasure Coding

Bucketing

Replication

Bottom of Form

Correct Option:B

**EXPLANATION**

Erasure coding is a process to keep HA in distributed systems without need of replicating the data as we used to have in earlier HDFS.

3

What is the usage for edit logs in HDFS?

**SELECT THE CORRECT ANSWER**

Top of Form

Edit logs are kept to store daily transaction logs and assist in updating fsimage

Edit logs are logs that record the time a client first began sending data to HDFS

Edit logs are just for fault recovery

All the above

Bottom of Form

Correct Option:C

**EXPLANATION**

The edit log is a logical structure that functions similarly to a transaction log.

4

What is the need for a secondary name node?

**SELECT THE CORRECT ANSWER**

Top of Form

Used to store real-time sync information about primary namenode

Used to store file blocks created and all metadata required

Keeps modifications to the fs image as a log added to native file system edit logs, which is often kept in a different location from the primary namenode

None of the above

Bottom of Form

Correct Option:C

**EXPLANATION**

Secondary namenode's main objective is to merge fsimage with and edit logs periodically.

5

Which type of scheduler guarantees that each job receives a fair proportion of resources on time in the average?

**SELECT THE CORRECT ANSWER**

Top of Form

Capacity Scheduling

FIFO

Equal&Faireness Scheduling

Fairness Scheduling

Bottom of Form

Correct Option:D

**EXPLANATION**

Fairness scheduling ensures every job should receive an equal or roughly equal share of resources whenever they are required

6

Given that HDFS has rack awareness, does this imply that one block replica on the same rack will ensure fault tolerance when putting HDFS blocks?

**SELECT THE CORRECT ANSWER**

Top of Form

TRUE

FALSE

Bottom of Form

Correct Option:B

**EXPLANATION**

In an effort to improve fault tolerance, HDFS tries to move your file block to a new rack rather than the same rack.

7

What file will have the dfs.replication property?

**SELECT THE CORRECT ANSWER**

Top of Form

core-site.xml

yarn-default.xml

hdfs-default.xml

mapred-default.xml

Bottom of Form

Correct Option:C

**EXPLANATION**

dfs.replication defines the default replication factor in HDFS file, and this belongs to hdfs-default or hdfs-site.xml

8

Which of the following phases does not belong in the Map Reduce process?

**SELECT THE CORRECT ANSWER**

Top of Form

Partition

Shuffling

Sorting

InputFormat

Bottom of Form

Correct Option:D

**EXPLANATION**

Inputformat provides specification from the very beginning. Even the map-reduce phase did not count this.

9

Users read and write Avro data as Hive tables using \_\_\_\_\_\_

**SELECT THE CORRECT ANSWER**

Top of Form

HiveSerde

SqlSerde

AvroSerde

HiveQLSerde

Bottom of Form

Correct Option:C

**EXPLANATION**

Hive table data can be read and written in Avro format using the AvroSerde.

10

What exactly is the usage of OutputCollector while writing a MapReduce program?

**SELECT THE CORRECT ANSWER**

Top of Form

To dump output of reducer to filesystem

Read output to the reducer which is collected by the mapper

Redirect collected metrics to the partitoner

All the above

Bottom of Form

Correct Option:A

**EXPLANATION**

OutputCollector's job is to output the reduced task to the disk.

11

On which model HDFS works?

**SELECT THE CORRECT ANSWER**

Top of Form

Master,Executor

Master,Slave

Master,Worker

Admin,Slave

Bottom of Form

Correct Option:B

**EXPLANATION**

HDFS works on master slave model, master being Name node and slave being datanode.

12

DUMP command is used for showing evnets on screen, what is the shorcut to invoke this?.

**SELECT THE CORRECT ANSWER**

Top of Form

\dump

\e

\de

\d

Bottom of Form

Correct Option:D

**EXPLANATION**

\d is alias for dump operator in pig.

13

Which is the popular DB which supports horizontal scaling in Hadoop?

**SELECT THE CORRECT ANSWER**

Top of Form

Cassandra

MongoDB

Hbase

Ozie

Bottom of Form

Correct Option:C

**EXPLANATION**

Hbase stands for Hadoop database and this come natively with hadoop and supports horizontal scaling.

14

By default in hive , Dynamic partition is disabled . In case we want to enable this what will be the command for this?

**SELECT THE CORRECT ANSWER**

Top of Form

SET hive.exec.dynamic.partition = true;

SET hive.set.dynamic.partition = true;

SET hive.exec.dynamic.partition = enable;

SET hive.exec.dynamic = true;

Bottom of Form

Correct Option:A

**EXPLANATION**

SET hive.exec.dynamic.partition = true; command with SET hive.exec.dynamic.partition.mode = nonstrict; helps to enable dynamic partition in hive.

15

What does the following command do? hdfs dfs -ls -R /user/thomas/

**SELECT THE CORRECT ANSWER**

Top of Form

Recursively lists the contents of /user/thomas in HDFS.

Recursively lists the contents of /user/thomas in HDFS and all of its subfolders

Recursively lists the contents of /user/hive/warehouse/thomas in HDFS and all of its subfolders

None of the above

Bottom of Form

Correct Option:B

**EXPLANATION**

Recursively lists the contents of /user/thomas in HDFS and all of its subfolders

16

What if you add combiner in Map reducer?

**SELECT THE CORRECT ANSWER**

Top of Form

Combiner will act as mini reducer at reducer level and sends data to disk in compressed form.

Combiner will act as mini reducer at mapper level and sends data to disk in compressed form.

Combiner will act as mini reducer at mapper level and sends data to reducer in compressed form.

Combiner sends data to reducer in compressed form to be stored in HDFS.

Bottom of Form

Correct Option:C

**EXPLANATION**

Combiner works at mapper level and main objective is to get do local reducing before sending to actual reducer.

17

You are working in hadoop startup and figured out that in 1 project you need to implement custom partitioner and override getPartiton method, what will the arguments you will take care of?

**SELECT THE CORRECT ANSWER**

Top of Form

key,Value,mapper

no of mapper,no of reducer and no of partition

No of reducer,key,mapper

key,value and no of reducers

Bottom of Form

Correct Option:D

**EXPLANATION**

getPartition method requires you to provide key,value and no of reducers.

18

You are required to call custom partition,you override getPartition method , what will the expected return type from same?

**SELECT THE CORRECT ANSWER**

Top of Form

random string with number of reducers

int between 0 number of reducers.

BigInt between 0 number of reducers.

None of the above

Bottom of Form

Correct Option:B

**EXPLANATION**

getPartition method gives back integer and no of redcuers as return type.

19

In map reduce, sorting happens by ?

**SELECT THE CORRECT ANSWER**

Top of Form

Using Bubble Sort

Their Natural Order

By Quicksort

by Merge Sort

Bottom of Form

Correct Option:D

**EXPLANATION**

The partitioner will also sort the data based on the intermediate keys before sending it to downstream reducers. The reducers then combine the sorted data from multiple mappers, producing the final output.

20

What will the benefit of using sequence files in Hadoop?

**SELECT THE CORRECT ANSWER**

Top of Form

Sequence files are very much performant due to their capabilities of serialization and deserialization objects.

Sequence file are not human readable so give more security and are less performant.

Sequence file are just another alternative and does not support serde framework

All of the above

Bottom of Form

Correct Option:A

**EXPLANATION**

Sequence files are binary files in Hadoop that can store serialized key-value pairs. These files are specifically designed for storing large amounts of binary data, and they can be compressed, splittable, and performant to read and write. Because sequence files are not human-readable, they also provide an added layer of security.

21

In case in MR, you figured out that there are lots of disk spliis happening while executing job, what step you will take to control it?

**SELECT THE CORRECT ANSWER**

Top of Form

This is not in developer control, its framework thing so we can just monitor and hope for best.

mapreduce.task.io.sort.mb value can be increased.

mapreduce.task.io.shuffle.mb value can be increased.

cut down the number of mappers

Bottom of Form

Correct Option:B

**EXPLANATION**

it better to increase the value of mapreduce.task.io.sort.mb value which is the total amount of buffer memory to use while sorting files.

22

Which join technique is generally recommened and more efficient in MR?

**SELECT THE CORRECT ANSWER**

Top of Form

Reducer Side

Reducer – Full Outer Join

Reducer – AntiJoin

Mapper Side

Bottom of Form

Correct Option:D

**EXPLANATION**

mapper side are more faster as it does not have to wait for all mappers to complete as in case of reducer.

23

Hbase Is standalone DB comes with hadoop?

**SELECT THE CORRECT ANSWER**

Top of Form

FALSE

TRUE

Bottom of Form

Correct Option:B

**EXPLANATION**

HBase is a NoSQL database that runs on top of Hadoop

24

A hive table stores schema in -------- and actual data in -------- ?

**SELECT THE CORRECT ANSWER**

Top of Form

Hbase,Metastore

Metatable,Hbase

Metastore,HDFS

Oracle, Hbase

Bottom of Form

Correct Option:C

**EXPLANATION**

Data will always be stored in HDFS and metadata(schema) will be stored in metastore by default Derby DB.

25

While creating UDF, is there any limitation on arguments tha can be passed to evaluate method of a Hive UDF?

**SELECT THE CORRECT ANSWER**

Top of Form

yes,Max 4 is allowed

Yes,Min 3 is allowed

Not sure

No there is no such limit.

Bottom of Form

Correct Option:D

**EXPLANATION**

As many as you want but with valid datatype.

26

Where should SORT BY to be used instead of ORDER BY?

**SELECT THE CORRECT ANSWER**

Top of Form

SORT by and ORDER BY does same thing except one works at reducer level and other one at mapper level

SORT by and ORDER BY does same thing

SORT BY is generally used in case of dealing with huge datasets and result outcome needs to be faster as it uses multiple reducers wheras ORDER BY use only 1 reducer

All of the Above

Bottom of Form

Correct Option:C

**EXPLANATION**

SORT BY is generally used in case of dealing with huge datasets and result outcome needs to be faster as it uses multiple reducers wheras ORDER BY use only 1 reducer

27

Why .hiverc is kept in Hive?

**SELECT THE CORRECT ANSWER**

Top of Form

used to store metastore data

Used to keep username,pswd to securely login in HIVE GUI.

Set any hive configuration/customization values

None of the above

Bottom of Form

Correct Option:C

**EXPLANATION**

It is a file that is executed when you launch the hive shell - making it an ideal place for adding any hive configuration/customization you want set, on start of the hive shell.

28

Which of the following options describes the responsibility of the executors in Spark?

**SELECT THE CORRECT ANSWER**

Top of Form

Executors accepts jobs from the driver, analyze those jobs and return results to the driver.

Executors accepts tasks from the driver, analyze those jobs and return results to the cluster manager.

Executors accepts tasks from the driver, executes those tasks and return results to driver.

Executors accepts tasks from the cluster manager, execute those tasks and return results to the cluster manager.

Bottom of Form

Correct Option:C

**EXPLANATION**

Executors accepts tasks from the driver, executes those tasks and return results to driver.

29

Which of the following describes the role of the cluster manager?

**SELECT THE CORRECT ANSWER**

Top of Form

The cluster manager allocates resources to Spark applications and maintains the executor processes in remote mode.

The cluster manager allocates resources to Spark applications and maintains the executor processes in client mode.

The cluster manager allocates resources to the DataFrame manager

The cluster manager schedules tasks on the cluster in client mode.

Bottom of Form

Correct Option:A

**EXPLANATION**

The cluster manager divides resources across applications. It works as an external service for acquiring resources on the cluster.

30

Which of the following is the idea behind dynamic partition pruning in Spark?

**SELECT THE CORRECT ANSWER**

Top of Form

Dynamic partition pruning reoptimizes physical plans based on data types and broadcast variables.

Dynamic partition pruning reoptimizes query plans based on runtime statistics collected during query execution.

Dynamic partition pruning concatenates columns of similar data types to optimize join performance.

Move the data into S3 and use Amazon Redshift from multiple regions to query it.

Bottom of Form

Correct Option:A

**EXPLANATION**

Dynamic partition pruning is an optimization technique in Spark that dynamically prunes the unnecessary partitions of data to avoid reading all data.

31

Which of the following statements about Spark's execution hierarchy is correct?

**SELECT THE CORRECT ANSWER**

Top of Form

In Spark's execution hierarchy, tasks are one layer above slots.

In Spark's execution hierarchy, a job may reach over multiple stage boundaries.

In Spark's execution hierarchy, a stage comprises multiple jobs.

In Spark's execution hierarchy, executors are the smallest unit.

In Spark's execution hierarchy, manifests are one layer above jobs.

Bottom of Form

Correct Option:B

**EXPLANATION**

A job is a sequence of stages, and thus may reach over multiple stage boundaries.

32

Which of the following describes the conversion of a computational query into an execution plan in Spark?

**SELECT THE CORRECT ANSWER**

Top of Form

Spark uses the catalog to resolve the optimized logical plan.

The executed physical plan depends on a cost optimization from a previous stage.

The catalog assigns specific resources to the physical plan.

Depending on whether DataFrame API or SQL API are used, the physical plan may differ.

Bottom of Form

Correct Option:B

**EXPLANATION**

In Spark, the conversion of a computational query into an execution plan is a multi-step process.

33

Which of the following describes characteristics of the Spark driver?

**SELECT THE CORRECT ANSWER**

Top of Form

The Spark driver requests the transformation of operations into DAG computations from the worker nodes.

If it sets in the Spark configuration, Spark scales the Spark driver horizontally to improve parallel processing performance.

The Spark driver processes partitions in an optimized,  distributed fashion.

In a non-interactive Spark application, the Spark driver automatically creates the SparkSession object.

The Spark driver's main responsibility includes scheduling queries for execution on worker nodes.

Bottom of Form

Correct Option:E

**EXPLANATION**

The Spark driver main responsibility includes scheduling queries for execution on worker nodes.

34

Which of the following describes how Spark achieves fault tolerance?

**SELECT THE CORRECT ANSWER**

Top of Form

Due to the mutability of DataFrames after transformations, Spark reproduces them using observed lineage in case of worker node failure.

Spark builds a fault-tolerant layer on top of the legacy RDD data system, which by itself is not fault tolerant.

Spark helps fast recovery of data in case of a worker fault by providing the MEMORY\_AND\_DISK storage level option.

If an executor on a worker node fails while calculating an RDD, that RDD can be recomputed by another executor using the lineage.

Bottom of Form

Correct Option:D

**EXPLANATION**

If an executor on a worker node fails while calculating an RDD, that RDD can be recomputed by another executor using the lineage.

35

Which of the following statements about Spark's DataFrames is incorrect?

**SELECT THE CORRECT ANSWER**

Top of Form

Spark's DataFrames are equal to Python's or R's DataFrames.

Spark's DataFrames are immutable.

RDDs are at the core of DataFrames.

Data in DataFrames is organized into named columns.

The data in DataFrames may be split into multiple chunks.

Bottom of Form

Correct Option:C

**EXPLANATION**

RDDs (Resilient Distributed Datasets) are not at the core of DataFrames in Spark.

36

Which of the following statements about Spark's configuration properties is incorrect?

**SELECT THE CORRECT ANSWER**

Top of Form

The maximum number of tasks that an executor can process at the same time is controlled by the spark.executor.cores property.

The default number of partitions to use when shuffling data for joins or aggregations is 300.

The maximum number of tasks that an executor can process at the same time is controlled by the spark.task.cpus property.

None of the above

Bottom of Form

Correct Option:A

**EXPLANATION**

The maximum number of tasks that an executor can process at the same time is controlled by the spark.executor.cores property.

37

Which of the following describes a way for resizing a DataFrame from 16 to 8 partitions in the most efficient way?

**SELECT THE CORRECT ANSWER**

Top of Form

Use operation DataFrame.coalesce(8) to fully shuffle the DataFrame and reduce the number of partitions.

Use a narrow transformation to reduce the number of partitions.

Use operation DataFrame.coalesce(0.5) to halve the number of partitions in the DataFrame.

Use operation DataFrame.repartition(8) to shuffle the DataFrame and reduce the number of partitions.

Bottom of Form

Correct Option:A

**EXPLANATION**

The most efficient way to reduce the number of partitions in a DataFrame from 16 to 8 is to use the `repartition()` method with the argument as the number of partitions you want to have

38

Which of the following describes the characteristics of accumulators?

**SELECT THE CORRECT ANSWER**

Top of Form

If an action including an accumulator fails during execution and Spark manages to restart the action and complete it successfully, only the successful attempt will be counted in the accumulator.

Accumulators are immutable.

All accumulators used in a Spark application are listed in the Spark UI.

Accumulators are used to pass around lookup tables across the cluster.

Accumulators can be instantiated directly via the accumulator(n) method of the pyspark.RDD module.

Bottom of Form

Correct Option:B

**EXPLANATION**

Accumulators are immutable which means the accumulator variable data cannot be changed by slave nodes.

39

Which of the following statements about data skew is incorrect?

**SELECT THE CORRECT ANSWER**

Top of Form

To mitigate skew, Spark automatically disregards null values in keys when joining.

In skewed DataFrames, the largest and the smallest partition consume very different amounts of memory.

Salting can resolve data skew.

Spark does not automatically optimize skew joins by default.

Bottom of Form

Correct Option:A

**EXPLANATION**

Spark does not automatically disregard null values in skew keys while defining Partitioner. Instead, Spark provides the option to use a custom Partitioner (such as RangePartitioner) that can be used to specify how to divide the data into partitions.

40

Which of the following code blocks, stores DataFrame itemsDf in executor memory and, if insufficient memory is available, serializes it and saves it to disk?

**SELECT THE CORRECT ANSWER**

Top of Form

itemsDf.persist(storagelevel.MEMORY\_ONLY)

itemsDf.cache(storagelevel.MEMORY\_AND\_DISK\_ONLY)

itemsDf.store()

itemsDf.cache()

Bottom of Form

Correct Option:B

**EXPLANATION**

MEMORY\_AND\_DISK\_ONLY spill to disk if there is too much data to fit in memory.

41

The code block displayed below contains an error. The code block should configure Spark so that DataFrames up to a size of 20 MB will be broadcast to all worker nodes when performing a join. Find the error. Code block: spark.conf.set("spark.sql.autoBroadcastJoinThreshold", 20)

**SELECT THE CORRECT ANSWER(S)**

Top of Form

Spark will only broadcast DataFrames that are much smaller than the default value.

The command is evaluated lazily and needs to be followed by an action.

Spark will only apply the limit to threshold joins and not to other joins.

The passed limit has the wrong variable type.

The correct option to write configurations is through spark.config and not spark.conf.

Bottom of Form

Correct Option:A, B, D

**EXPLANATION**

The default value is 10 MB (10485760 bytes). Since the configuration for spark.sql.autoBroadcastJoinThreshold expects a number in bytes (and not megabytes), the code block sets the limits to merely 20 bytes, instead of the requested 20 \* 1024 \* 1024 (= 20971520) bytes.

42

Which of the following describes a narrow transformation?

**SELECT THE CORRECT ANSWER**

Top of Form

A narrow transformation is an operation in which data is exchanged across partitions.

A narrow transformation is an operation in which no data is exchanged across the cluster.

A narrow transformation is an operation in which data is exchanged across the cluster.

A narrow transformation is a process in which 32-bit float variables are cast to smaller float variables, like 16-bit or 8-bit float variables.

A narrow transformation is a process in which data from multiple RDDs is used.

Bottom of Form

Correct Option:B

**EXPLANATION**

narrow transformations, no data is exchanged across the cluster, since these transformations do not require any data from outside of the partition they are applied on. Typical narrow transformations include filter, drop, and coalesce.

43

Which of the following statements about stages is correct?

**SELECT THE CORRECT ANSWER**

Top of Form

Different stages in a job may be executed in parallel.

Stages may contain multiple actions, narrow, and wide transformations.

Tasks in a stage may be executed by multiple machines at the same time.

Stages ephemerally store transactions, before they are committed through actions.

Bottom of Form

Correct Option:C

**EXPLANATION**

Stages are a fundamental unit of execution in Apache Spark that consist of sets of tasks that can be executed in parallel. The tasks within a stage can be executed by multiple machines at the same time, in order to increase the speed of the computation

44

Which of the following describes a difference between Spark's cluster and client execution modes?

**SELECT THE CORRECT ANSWER**

Top of Form

In cluster mode, the driver resides on a worker node, while it resides on an edge node in client mode.

In cluster mode, the cluster manager resides on a worker node, while it resides on an edge node in client execution mode.

In cluster mode, executor processes run on worker nodes, while they run on gateway nodes in client mode.

In cluster mode, the Spark driver is not co-located with the cluster manager, while it is co-located in client mode.

In cluster mode, a gateway machine hosts the driver, while it is co-located with the executor in client mode.

Bottom of Form

Correct Option:D

**EXPLANATION**

The main difference between cluster mode and client mode in Spark is the location of the Spark driver.

45

Which of the following describes Spark's standalone deployment mode?

**SELECT THE CORRECT ANSWER**

Top of Form

Standalone mode is a viable solution for clusters that run multiple frameworks.

Standalone mode uses a single JVM to run Spark driver and executor processes.

Standalone mode uses only a single executor per worker per application.

Standalone mode is how Spark runs on YARN and Mesos clusters.

Bottom of Form

Correct Option:D

**EXPLANATION**

Spark can run on different cluster managers like YARN, Mesos, Kubernetes, and Spark Standalone mode.

46

Which of the following describes properties of a shuffle?

**SELECT THE CORRECT ANSWER**

Top of Form

A shuffle is one of many actions in Spark.

In a shuffle, Spark writes data to disk.

Shuffles involve only single partitions.

Operations involving shuffles are never evaluated lazily.

Bottom of Form

Correct Option:B

**EXPLANATION**

During a shuffle operation, Spark writes data to disk. This happens when the data being shuffled is too large to fit in memory and needs to be spilled to disk.

47

Which of the following statements about the differences between actions and transformations is correct?

**SELECT THE CORRECT ANSWER**

Top of Form

Actions can trigger Adaptive Query Execution, while transformation cannot.

Actions can be queued for delayed execution, while transformations can only be processed immediately.

Actions are evaluated lazily, while transformations are not evaluated lazily.

Actions generate RDDs, while transformations do not.

Bottom of Form

Correct Option:D

**EXPLANATION**

Adaptive Query Execution (AQE) is a feature in Apache Spark that allows the optimizer to dynamically adjust the execution plan of a query based on the characteristics of the data and the cluster. AQE is triggered by certain actions, such as `count` and `collect`, but not by transformations.

48

Which of the following is a characteristic of the cluster manager?

**SELECT THE CORRECT ANSWER**

Top of Form

In client mode, the cluster manager runs on the edge node.

The cluster manager receives input from the driver through the SparkContext.

The cluster manager does not exist in standalone mode.

The cluster manager transforms jobs into DAGs.

Bottom of Form

Correct Option:B

**EXPLANATION**

In order for the driver to contact the cluster manager, the driver launches a SparkContext. The driver then asks the cluster manager for resources to launch executors.

49

Which of the following describes Spark actions?

**SELECT THE CORRECT ANSWER**

Top of Form

Actions are Spark's way of exchanging data between executors.

Writing data to disk is the primary purpose of actions.

Actions are Spark's way of modifying RDDs.

The driver receives data upon request by actions.

Bottom of Form

Correct Option:D

**EXPLANATION**

Actions in Spark are operations that trigger computation and return results to the driver program or write data to an external storage system.

50

Which of the following are valid execution modes?

**SELECT THE CORRECT ANSWER**

Top of Form

Standalone, Client, Cluster

Kubernetes, Local, Client

Cluster, Server, Local

Client, cluster, and local execution

Bottom of Form

Correct Option:A

**EXPLANATION**

Standalone mode is the default mode where Spark runs on a single machine. Client mode is used when the driver program runs on a machine separate from the executor machines. Cluster mode is used when the driver program runs on one of the executor machines in the cluster.

51

Which of the following describes a valid concern about partitioning?

**SELECT THE CORRECT ANSWER**

Top of Form

A shuffle operation returns 200 partitions if not explicitly set.

The coalesce() method should be used to increase the number of partitions.

Decreasing the number of partitions reduces the overall runtime of narrow transformations if there are more executors available than partitions.

No data is exchanged between executors when coalesce() is run.

Bottom of Form

Correct Option:C

**EXPLANATION**

202 is the default value for the Spark property spark.sql.shuffle.partitions. This property determines how many partitions Spark uses when shuffling data for joins or aggregations.

52

Which of the following statements about reducing out-of-memory errors is incorrect?

**SELECT THE CORRECT ANSWER**

Top of Form

Concatenating multiple string columns into a single column may guard against out-of-memory errors.

Reducing partition size can help against out-of-memory errors.

Decreasing the number of cores available to each executor can help against out-of-memory errors.

Setting a limit on the maximum size of serialized data returned to the driver may help prevent out-of-memory errors.

Bottom of Form

Correct Option:C

**EXPLANATION**

Decreasing the number of cores available to each executor can increase the likelihood of out-of-memory errors.

53

Which of the following statements about storage levels is incorrect?

**SELECT THE CORRECT ANSWER**

Top of Form

DISK\_ONLY will not use the worker node's memory.

MEMORY\_AND\_DISK replicates cached DataFrames both on memory and disk.

In client mode, DataFrames cached with the MEMORY\_ONLY\_2 level will not be stored in the edge node's memory.

Caching can be undone using the DataFrame.unpersist() operator.

Bottom of Form

Correct Option:C

**EXPLANATION**

The MEMORY\_ONLY\_2 storage level is not part of Spark's storage levels.

54

Which of the following is not a feature of Adaptive Query Execution?

**SELECT THE CORRECT ANSWER**

Top of Form

Replace a sort merge join with a broadcast join, where appropriate.

Coalesce partitions to accelerate data processing.

Reroute a query in case of an executor failure.

Split skewed partitions into smaller partitions to avoid differences in partition processing time.

Bottom of Form

Correct Option:B

**EXPLANATION**

Coalesce partitions to accelerate data processing is not a feature of Adaptive Query Execution.

55

Which of the following code blocks creates a new 6-column DataFrame by appending the rows of the 6-column DataFrame yesterdayTransactionsDf to the rows of the 6-column DataFrame todayTransactionsDf, ignoring that both DataFrames have different column names?

**SELECT THE CORRECT ANSWER**

Top of Form

todayTransactionsDf.union(yesterdayTransactionsDf)

todayTransactionsDf.unionByName(yesterdayTransactionsDf)

todayTransactionsDf.unionByName(yesterdayTransactionsDf, allowMissingColumns=True)

Bottom of Form

Correct Option:A

**EXPLANATION**

The union command appends rows of yesterdayTransactionsDf to the rows of todayTransactionsDf, ignoring that both DataFrames have different column names. The resulting DataFrame will have the column names of DataFrame todayTransactionsDf.

56

Code block: transactionsDf.persist(StorageLevel.MEMORY\_AND\_DISK)

**SELECT THE CORRECT ANSWER**

Top of Form

The storage level is inappropriate for fault-tolerant storage.

The code block uses the wrong command for caching.

Caching is not supported in Spark, data are always recomputed.

Data caching capabilities can be accessed through the spark object, but not through the DataFrame API.

Bottom of Form

Correct Option:A

**EXPLANATION**

Typically, when thinking about fault tolerance and storage levels, you would want to store redundant copies of the dataset. This can be achieved by using a storage level such as StorageLevel.MEMORY\_AND\_DISK\_2.

57

Which of the following is one of the big performance advantages that Spark has over Hadoop?

**SELECT THE CORRECT ANSWER**

Top of Form

Spark achieves great performance by storing data in the DAG format, whereas Hadoop can only use parquet files.

Spark achieves great performance by storing data in the HDFS format, whereas Hadoop can only use parquet files.

Spark achieves higher resiliency for queries since, different from Hadoop, it can be deployed on Kubernetes.

Spark achieves performance gains for storing and computation in memory

Bottom of Form

Correct Option:D

**EXPLANATION**

Spark achieves performance gains for storing and computation in memory

58

Which of the following describes characteristics of the Dataset API?

**SELECT THE CORRECT ANSWER**

Top of Form

The Dataset API does not provide compile-time type safety.

In Python, the Dataset API's schema is constructed via type hints.

The Dataset API is available in Scala, but it is not available in Python.

In Python, the Dataset API mainly resembles Pandas' DataFrame API.

Bottom of Form

Correct Option:B

**EXPLANATION**

The Dataset API in Spark is available in both Java/Scala and Python, and it provides compile-time type safety.

59

Which of the following describes characteristics of the Spark UI?

**SELECT THE CORRECT ANSWER**

Top of Form

Some of the tabs in the Spark UI are named Jobs, Stages, Storage, DAGs, Executors, and SQL.

There is a place in the Spark UI that shows the property spark.executor.memory.

Via the Spark UI, workloads can be manually distributed across distributors.

Via the Spark UI, stage execution speed can be modified.

Bottom of Form

Correct Option:A

**EXPLANATION**

The Spark UI is a web-based interface that provides insights into the internals of Spark applications.

60

The code block displayed below contains an error. The code block is intended to join DataFrame itemsDf with the larger DataFrame transactionsDf on column itemId. Find the error. Code block: transactionsDf.join(itemsDf, "itemId", how="broadcast")

**SELECT THE CORRECT ANSWER**

Top of Form

broadcast is not a valid join type.

The larger DataFrame transactionsDf is being broadcasted, rather than the smaller DataFrame itemsDf.

Spark will only perform the broadcast operation if this behavior has been enabled on the Spark cluster.

The join method should be replaced by the broadcast method.

Bottom of Form

Correct Option:A

**EXPLANATION**

The broadcasting should be done on the smaller DataFrame, which is itemsDf in this case so that its contents can broadcast

61

Which of the following code blocks efficiently converts DataFrame transactionsDf from 12 into 24 partitions?

**SELECT THE CORRECT ANSWER**

Top of Form

transactionsDf.coalesce(24)

transactionsDf.repartition()

transactionsDf.repartition(24)

Bottom of Form

Correct Option:C

**EXPLANATION**

transactionsDf.repartition(24), no need to pass any argument.

62

Which of the following code blocks removes all rows in the 6-column DataFrame transactionsDf that have missing data in at least 3 columns?

**SELECT THE CORRECT ANSWER**

Top of Form

transactionsDf.dropna(thresh=4)

transactionsDf.dropna(thresh=2)

transactionsDf.dropna("any")

transactionsDf.drop.na("",2)

Bottom of Form

Correct Option:A

**EXPLANATION**

In the given options, `dropna(thresh=4)` will remove all rows in the 6-column DataFrame `transactionsDf` that have missing data in at least 3 columns because `thresh` parameter specifies the minimum number of non-missing values required to keep a row.