Vendor: Databricks

Exam Code:DATABRICKS-CERTIFIED-ASSOCIATE-DEVELOPER-FOR-APACHE-SPARK

Exam Name: Databricks Certified Associate Developer for Apache Spark 3.0

Version: Demo

QUESTION 1

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?

A. itemsDf.persist(StorageLevel.MEMORY_ONLY)

B. itemsDf.cache(StorageLevel.MEMORY_AND_DISK)

C. itemsDf.store()

D. itemsDf.cache()

E. itemsDf.write.option(\\'destination\\', \\'memory\\').save()

Correct Answer: D

QUESTION 2

The code block shown below should return a new 2-column DataFrame that shows one attribute from column attributes per row next to the associated itemName, for all suppliers in column supplier whose name includes Sports. Choose the answer that correctly fills the blanks in the code block to accomplish this.

Sample of DataFrame itemsDf:

1.+------+----------------------------------+---------------------------+-------------------+

2.|itemId|itemName |attributes |supplier |

3.+------+----------------------------------+---------------------------+-------------------+

4.|1 |Thick Coat for Walking in the Snow|[blue, winter, cozy] |Sports Company Inc.| 5.|2 |Elegant Outdoors Summer Dress |[red, summer, fresh, cooling]|YetiX |

6.|3 |Outdoors Backpack |[green, summer, travel] |Sports Company Inc.|

7.+------+----------------------------------+---------------------------+-------------------+

Code block:

itemsDf.__1__(__2__).select(__3__, __4__) A. 1. filter

2.

col("supplier").isin("Sports")

3.

"itemName"

4.

explode(col("attributes"))

B. 1. where

2.

col("supplier").contains("Sports")

3.

"itemName"

4.

"attributes"

C. 1. where

2.

col(supplier).contains("Sports")

3.

explode(attributes)

4.

itemName

D. 1. where

2.

"Sports".isin(col("Supplier"))

3.

"itemName"

4.

array_explode("attributes")

E. 1. filter

2.

col("supplier").contains("Sports")

3.

"itemName"

4.

explode("attributes")

Correct Answer: E

Output of correct code block:

+----------------------------------+------+

|itemName |col |

+----------------------------------+------+

|Thick Coat for Walking in the Snow|blue |

|Thick Coat for Walking in the Snow|winter|

|Thick Coat for Walking in the Snow|cozy |

|Outdoors Backpack |green |

|Outdoors Backpack |summer|

|Outdoors Backpack |travel|

+----------------------------------+------+

The key to solving this is knowing about Spark\\'s explode operator. Using this operator, you can extract

values from arrays into single rows. The following guidance steps through

the

answers systematically from the first to the last gap. Note that there are many ways to solving the gap

---

QUESTION 3

Which of the following describes Spark\\'s way of managing memory?

A. Spark uses a subset of the reserved system memory.

B. Storage memory is used for caching partitions derived from DataFrames.

C. As a general rule for garbage collection, Spark performs better on many small objects than few big objects.

D. Disabling serialization potentially greatly reduces the memory footprint of a Spark application.

E. Spark\\'s memory usage can be divided into three categories: Execution, transaction, and storage.

Correct Answer: B

Spark\\'s memory usage can be divided into three categories: Execution, transaction, and storage.

No, it is either execution or storage.

As a general rule for garbage collection, Spark performs better on many small objects than few big objects.

No, Spark\\'s garbage collection runs faster on fewer big objects than many small objects. Disabling

serialization potentially greatly reduces the memory footprint of a Spark application.

The opposite is true ?serialization reduces the memory footprint, but may impact performance in a

negative way.

Spark uses a subset of the reserved system memory. No, the reserved system memory is separate from

Spark memory. Reserved memory stores Spark\\'s internal objects.

More info: Tuning - Spark 3.1.2 Documentation, Spark Memory Management | Distributed Systems

Architecture, Learning Spark, 2nd Edition, Chapter 7

QUESTION 4

Which of the following describes the role of tasks in the Spark execution hierarchy?

A. Tasks are the smallest element in the execution hierarchy.

B. Within one task, the slots are the unit of work done for each partition of the data.

C. Tasks are the second-smallest element in the execution hierarchy.

D. Stages with narrow dependencies can be grouped into one task.

E. Tasks with wide dependencies can be grouped into one stage.

Correct Answer: A

Stages with narrow dependencies can be grouped into one task. Wrong, tasks with narrow dependencies can be grouped into one stage. Tasks with wide dependencies can be grouped into one stage. Wrong, since a wide transformation causes a shuffle which always marks the boundary of a stage. So, you cannot bundle multiple tasks that have wide dependencies into a stage. Tasks are the second-smallest element in the execution hierarchy. No, they are the smallest element in the execution hierarchy. Within one task, the slots are the unit of work done for each partition of the data. No, tasks are the unit of work done per partition. Slots help Spark parallelize work. An executor can have multiple slots which enable it to process multiple tasks in parallel.

QUESTION 5

Which of the following code blocks reads in the JSON file stored at filePath as a DataFrame?

A. spark.read.json(filePath)

B. spark.read.path(filePath, source="json")

C. spark.read().path(filePath)

D. spark.read().json(filePath)

E. spark.read.path(filePath)

Correct Answer: A

QUESTION 6

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

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

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

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

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

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

Correct Answer: C

QUESTION 7

Which of the following code blocks reorders the values inside the arrays in column attributes of DataFrame

itemsDf from last to first one in the alphabet?

1.+------+-----------------------------+-------------------+

2.|itemId|attributes |supplier |

3.+------+-----------------------------+-------------------+

4.|1 |[blue, winter, cozy] |Sports Company Inc.|

5.|2 |[red, summer, fresh, cooling]|YetiX |

6.|3 |[green, summer, travel] |Sports Company Inc.|

7.+------+-----------------------------+-------------------+

A. itemsDf.withColumn(\\'attributes\\', sort_array(col(\\'attributes\\').desc()))

B. itemsDf.withColumn(\\'attributes\\', sort_array(desc(\\'attributes\\')))

C. itemsDf.withColumn(\\'attributes\\', sort(col(\\'attributes\\'), asc=False))

D. itemsDf.withColumn("attributes", sort_array("attributes", asc=False))

E. itemsDf.select(sort_array("attributes"))

Correct Answer: D

QUESTION 8

The code block displayed below contains multiple errors. The code block should return a DataFrame that

contains only columns transactionId, predError, value and storeId of DataFrame

transactionsDf. Find the errors.

Code block:

transactionsDf.select([col(productId), col(f)])

Sample of transactionsDf:

1.+-------------+---------+-----+-------+---------+----+

2.|transactionId|predError|value|storeId|productId| f| 3.+-------------+---------+-----+-------+---------+----+

4.| 1| 3| 4| 25| 1|null|

5.| 2| 6| 7| 2| 2|null|

6.| 3| 3| null| 25| 3|null|

7.+-------------+---------+-----+-------+---------+----+

A. The column names should be listed directly as arguments to the operator and not as a list.

B. The select operator should be replaced by a drop operator, the column names should be listed directly as arguments to the operator and not as a list, and all column names should be expressed as strings without being wrapped in a col() operator.

C. The select operator should be replaced by a drop operator.

D. The column names should be listed directly as arguments to the operator and not as a list and following the pattern of how column names are expressed in the code block, columns productId and f should be replaced by transactionId, predError, value and storeId.

E. The select operator should be replaced by a drop operator, the column names should be listed directly as arguments to the operator and not as a list, and all col() operators should be removed.

Correct Answer: B

QUESTION 9

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

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

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

C. The default value for spark.sql.autoBroadcastJoinThreshold is 10MB.

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

E. The default number of partitions returned from certain transformations can be controlled by the spark.default.parallelism property.

Correct Answer: D

QUESTION 10

Which of the following code blocks returns a DataFrame that is an inner join of DataFrame itemsDf and DataFrame transactionsDf, on columns itemId and productId, respectively and in which every itemId just appears once?

A. itemsDf.join(transactionsDf, "itemsDf.itemId==transactionsDf.productId").distinct("itemId")

B. itemsDf.join(transactionsDf, itemsDf.itemId==transactionsDf.productId).dropDuplicates(["itemId"])

C. itemsDf.join(transactionsDf, itemsDf.itemId==transactionsDf.productId).dropDuplicates("itemId")

D. itemsDf.join(transactionsDf, itemsDf.itemId==transactionsDf.productId, how="inner").distinct(["itemId"])

E. itemsDf.join(transactionsDf, "itemsDf.itemId==transactionsDf.productId", how="inner").dropDuplicates (["itemId"])

Correct Answer: B

QUESTION 11

The code block displayed below contains an error. The code block should return a copy of DataFrame transactionsDf where the name of column transactionId has been changed to transactionNumber. Find the error.

Code block:

transactionsDf.withColumn("transactionNumber", "transactionId")

A. The arguments to the withColumn method need to be reordered.

B. The arguments to the withColumn method need to be reordered and the copy() operator should be appended to the code block to ensure a copy is returned.

C. The copy() operator should be appended to the code block to ensure a copy is returned.

D. Each column name needs to be wrapped in the col() method and method withColumn should be replaced by method

withColumnRenamed.

E. The method withColumn should be replaced by method withColumnRenamed and the arguments to the method need to be reordered.

Correct Answer: E

QUESTION 12

The code block shown below should show information about the data type that column storeId of DataFrame transactionsDf contains. Choose the answer that correctly fills the blanks in the code block to accomplish this.

Code block:

transactionsDf.__1__(__2__).__3__

A. 1. select

2.

"storeId"

3.

print_schema()

B. 1. limit

2.

1

3.

columns

C. 1. select

2.

"storeId"

3.

printSchema()

D. 1. limit

2.

"storeId"

3.

printSchema()

E. 1. select

2.

storeId

3.

dtypes

Correct Answer: B

Correct code block: transactionsDf.select("storeId").printSchema() The difficulty of this is that it is hard to solve with the stepwise first-to-last- gap approach that has worked well for similar questions, since the answer options are so different from one another. Instead, you might want to eliminate answers by looking for patterns of frequently wrong answers. A first pattern that you may recognize by now is that column names are not expressed in quotes. For this reason, the answer that includes storeId should be eliminated. By now, you may have understood that the DataFrame.limit() is useful for returning a specified amount of rows. It has nothing to do with specific columns. For this reason, the answer that resolves to

limit("storeId") can be eliminated. Given that we are interested in information about the data type, you should