DATABRICKS-CERTIFIED-ASSOCIAT

Q&As

Databricks Certified Associate Developer for Apache Spark 3.0

Pass Databricks DATABRICKS-CERTIFIED-ASSOCIATE-DEVELOPER-FOR-APACHE-SPARK Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.leads4pass.com/databricks-certified-associate-developer-for-apache-spark.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

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

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

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

C. Spark achieves great performance by storing data and performing computation in memory, whereas large jobs in Hadoop require a large amount of relatively slow disk I/O operations.

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

E. Spark achieves performance gains for developers by extending Hadoop\\'s DataFrames with a user-friendly API.

Correct Answer: C

QUESTION 2

Which of the following code blocks returns a 2-column DataFrame that shows the distinct values in column productId and the number of rows with that productId in DataFrame transactionsDf?

A. transactionsDf.count("productId").distinct()

B. transactionsDf.groupBy("productId").agg(col("value").count())

C. transactionsDf.count("productId")

D. transactionsDf.groupBy("productId").count()

E. transactionsDf.groupBy("productId").select(count("value"))

Correct Answer: D

transactionsDf.groupBy("productId").count()

Correct. This code block first groups DataFrame transactionsDf by column productId and then counts the

rows in each group.

transactionsDf.groupBy("productId").select(count("value")) Incorrect. You cannot call select on a

GroupedData object (the output of a groupBy) statement.

transactionsDf.count("productId")

No. DataFrame.count() does not take any arguments.

transactionsDf.count("productId").distinct()

Wrong. Since DataFrame.count() does not take any arguments, this option cannot be right.

transactionsDf.groupBy("productId").agg(col("value").count()) False. A Column object, as returned by col

("value"), does not have a count() method. You can see all available methods for Column object linked in

the Spark documentation below. More info: pyspark.sql.DataFrame.count -- PySpark 3.1.2 documentation,

pyspark.sql.Column -- PySpark 3.1.2 documentation

Static notebook | Dynamic notebook: See test 3, 41 (Databricks import instructions)

QUESTION 3

Which of the following code blocks returns a DataFrame showing the mean value of column "value" of DataFrame transactionsDf, grouped by its column storeId?

A. transactionsDf.groupBy(col(storeId).avg())

B. transactionsDf.groupBy("storeId").avg(col("value"))

C. transactionsDf.groupBy("storeId").agg(avg("value"))

D. transactionsDf.groupBy("storeId").agg(average("value"))

E. transactionsDf.groupBy("value").average()

Correct Answer: C

QUESTION 4

Which of the following code blocks shuffles DataFrame transactionsDf, which has 8 partitions, so that it has 10 partitions?

A. transactionsDf.repartition(transactionsDf.getNumPartitions()+2)

B. transactionsDf.repartition(transactionsDf.rdd.getNumPartitions()+2)

C. transactionsDf.coalesce(10)

D. transactionsDf.coalesce(transactionsDf.getNumPartitions()+2)

E. transactionsDf.repartition(transactionsDf._partitions+2)

Correct Answer: B

QUESTION 5

Which of the following statements about RDDs is incorrect?

A. An RDD consists of a single partition.

B. The high-level DataFrame API is built on top of the low-level RDD API.

C. RDDs are immutable.

D. RDD stands for Resilient Distributed Dataset.

E. RDDs are great for precisely instructing Spark on how to do a query.

Correct Answer: A