70-775 Q&As

Perform Data Engineering on Microsoft Azure HDInsight

Pass Microsoft 70-775 Exam with 100% Guarantee

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

https://www.lead4pass.com/70-775.html

100% Passing Guarantee
100% Money Back Assurance

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

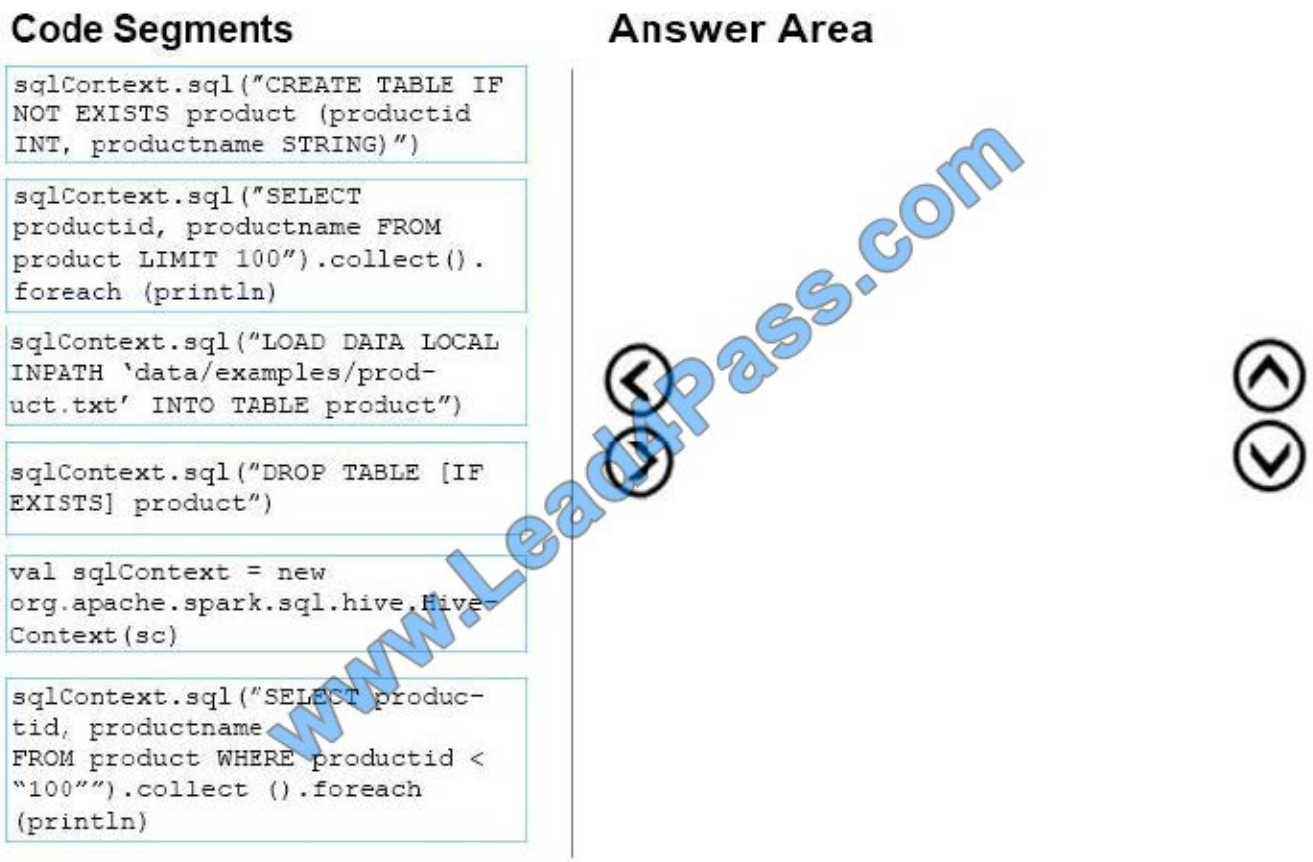
QUESTION 1

You have a text file named Data/examples/product.txt that contains product information.

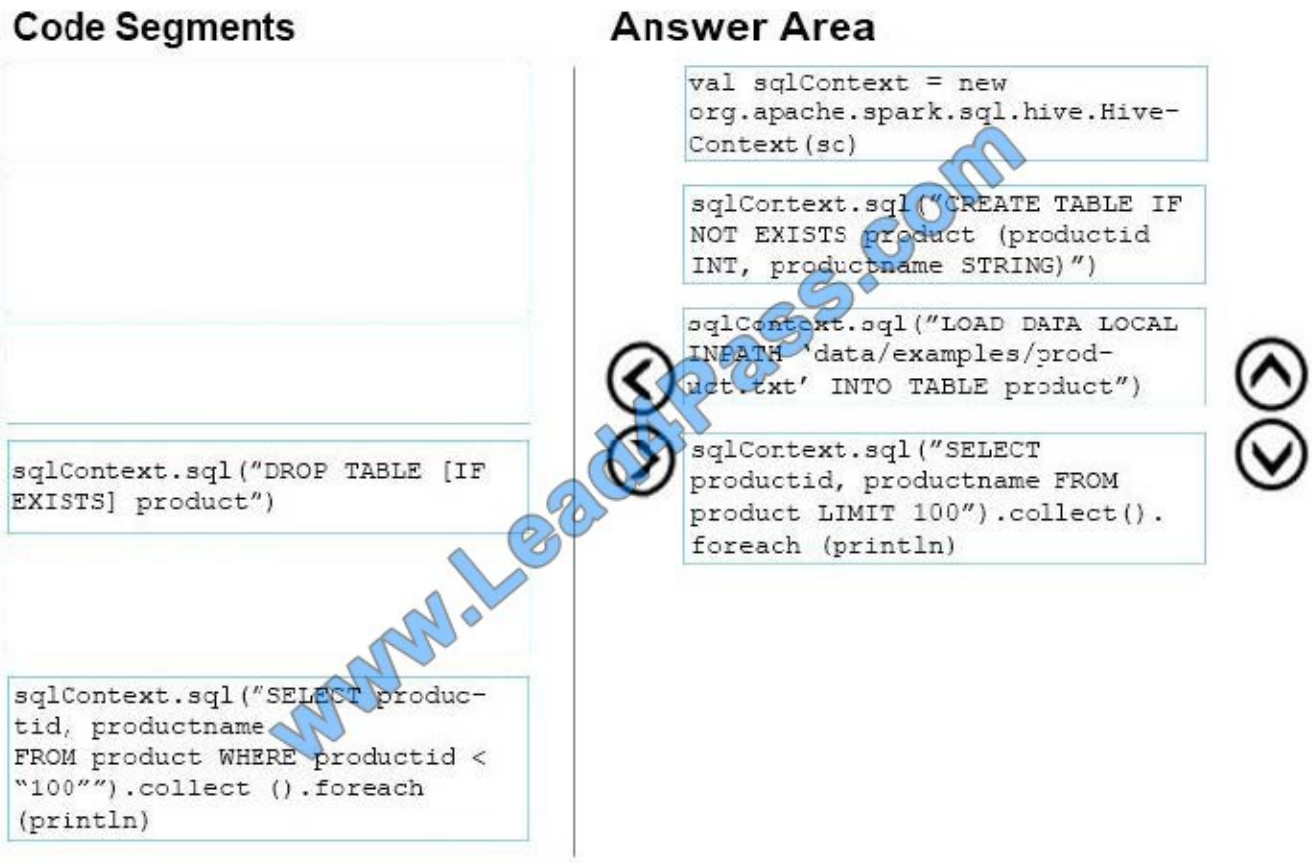
You need to create a new Apache Hive table, import the product information to the table, and then read the top 100 rows of the table.

Which four code segments should you use in sequence? To answer, move the appropriate code segments from the list of code segments to the answer area and arrange them in the correct order.

Select and Place:



Correct Answer:



QUESTION 2

Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is

exactly the same in each question in this series.

You have an initial dataset that contains the crime data from major cities.

You plan to build training models from the training data. You plan to automate the process of adding more data to the training models and to constantly tune the models by using the additional data, including data that is collected in near real-

time. The system will be used to analyze event data gathered from many different sources, such as Internet of Things (IoT) devices, live video surveillance, and traffic activities, and to generate predictions of an increased crime risk at a

particular time and place.

You have an incoming data stream from Twitter and an incoming data stream from Facebook, which are event-based only, rather than time-based. You also have a time interval stream every 10 seconds.

The data is in a key/value pair format. The value field represents a number that defines how many times a hashtag occurs within a Facebook post, or how many times a Tweet that contains a specific hashtag is retweeted.

You must use the appropriate data storage, stream analytics techniques, and Azure HDInsight cluster types for the various tasks associated to the processing pipeline.

You are planning a storage strategy for a large amount of analytic data used for the crime data analytics system. The initial data load involves over 100 billion records, and more than two billion records will be added daily.

You already created an Apache Hadoop cluster in HDInsight premium.

You need to implement the storage strategy to meet the following requirements:

What should you create?

A. a virtual machine (VM) by using the Data Science Virtual Machine template for Windows that has premium storage, a G-series size, and uses Microsoft SQL Server 2016 to store the data

B. an Azure Data Lake Analytics service by using Azure PowerShell

C. an Azure Data Lake Store account by using the Azure portal

D. an Azure Blob storage account by using the Azure portal

Correct Answer: C

References: https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-get- started-portal

QUESTION 3

Note: This question is part of a series of questions that use the same or similar answer choices. An answer choice may be correct for more than one question in the series. Each question is independent of the other questions in this series.

Information and details provided in a question apply only to that question. You are implementing a batch processing solution by using Azure HDInsight.

You need to integrate Apache Sqoop data and to chain complex jobs. The data and the jobs will implement MapReduce.

What should you do?

A. Use a shuffle join in an Apache Hive query that stores the data in a JSON format.

B. Use a broadcast join in an Apache Hive query that stores the data in an ORC format.

C. Increase the number of spark.executor.cores in an Apache Spark job that stores the data in a text format.

D. Increase the number of spark.executor.instances in an Apache Spark job that stores the data in a text format.

E. Decrease the level of parallelism in an Apache Spark job that stores the data in a text format.

F. Use an action in an Apache Oozie workflow that stores the data in a text format.

G. Use an Azure Data Factory linked service that stores the data in Azure Data Lake.

H. Use an Azure Data Factory linked service that stores the data in an Azure DocumentDB database.

Correct Answer: F

References: https://www.ibm.com/developerworks/library/bd-ooziehadoop/index.html

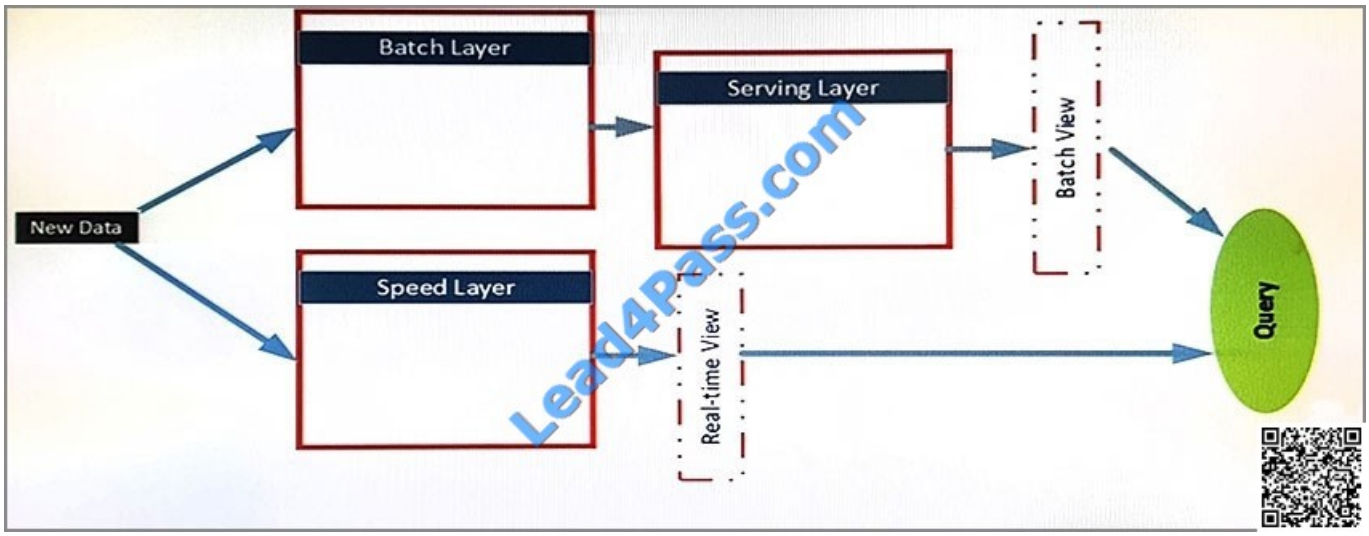
QUESTION 4

Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is

exactly the same in each question in this series.

You are planning a big data infrastructure by using an Apache Spark cluster in Azure HDInsight. The cluster has 24 processor cores and 512 GB of memory.

The architecture of the infrastructure is shown in the exhibit. (Click the Exhibit button.)



The architecture will be used by the following users:

The data sources in the batch layer share a common storage container. The following data sources are used:

The business analysts report that they experience performance issues when they run the monitoring queries.

You troubleshoot the performance issues and discover that the intermediate tables generated when the analysts run the queries cause pressure for the Java Virtual Machine (JVM) garbage collection per job.

Which configuration settings should you modify to alleviate the performance issues?

A. spark.sql.inMemoryColumnarStorage.batchSize

B. spark.sql.broadcastTimeout

C. spark.sql.files.openCostInBytes

D. spark.sql.shuffle.partitions

Correct Answer: D

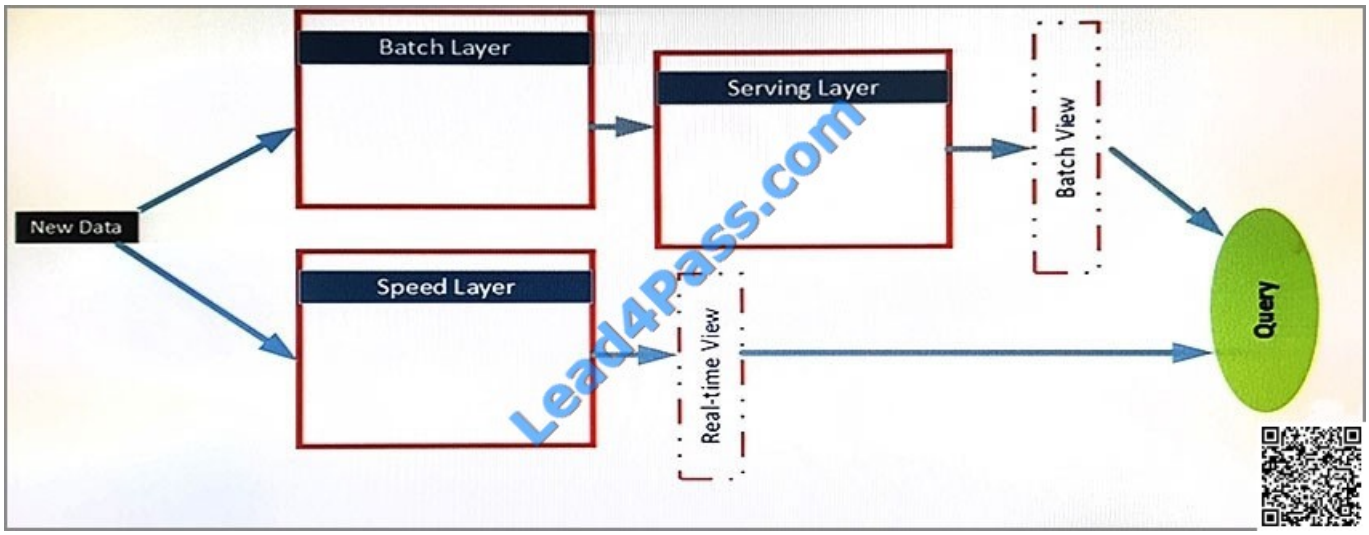
QUESTION 5

Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is

exactly the same in each question in this series.

You are planning a big data infrastructure by using an Apache Spark cluster in Azure HDInsight. The cluster has 24 processor cores and 512 GB of memory.

The architecture of the infrastructure is shown in the exhibit. (Click the Exhibit button.)



The architecture will be used by the following users:

The data sources in the batch layer share a common storage container. The following data sources are used:

You need to ensure that the support analysts can develop embedded analytics applications by using the least amount of development effort.

Which technology should you implement?

A. Zeppelin

B. Jupyter

C. Apache Ambari

D. Livy

Correct Answer: D

References: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-apache-spark-livy- rest-interface