70-775.exam.34q

Number: 70-775
Passing Score: 800
Time Limit: 120 min
File Version: 1

Microsoft 70-775

Perform Data Engineering on Microsoft Azure HDInsight

Exam A

QUESTION 1
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this sections, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are building a security tracking solution in Apache Kafka to parse security logs. The security logs record an entry each time a user attempts to access an application. Each log entry contains the IP address used to make the attempt and the country from which the attempt originated.

You need to receive notifications when an IP address from outside of the United States is used to access the application.

Solution: Create two new consumers. Create a file import process to send messages. Start the producer.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 2
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this sections, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are building a security tracking solution in Apache Kafka to parse security logs. The security logs record an entry each time a user attempts to access an application. Each log entry contains the IP address used to make the attempt and the country from which the attempt originated.

You need to receive notifications when an IP address from outside of the United States is used to access the application.

Solution: Create new topics. Create a file import process to send messages. Start the consumer and run the producer.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 3
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this sections, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are building a security tracking solution in Apache Kafka to parse security logs. The security logs record an entry each time a user attempts to access an application. Each log entry contains the IP address used to make the attempt and the country from which the attempt originated.

You need to receive notifications when an IP address from outside of the United States is used to access the application.

Solution: Create a consumer and a broker. Create a file import process to send messages. Run the producer.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 4
You have an Azure HDInsight cluster.

You need a build a solution to ingest real-time streaming data into a nonrelational distributed database.

What should you use to build the solution?

A. Apache Hive and Apache Kafka
B. Spark and Phoenix
C. Apache Storm and Apache HBase
D. Apache Pig and Apache HCatalog

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
References:
http://storm.apache.org/
http://hbase.apache.org/

QUESTION 5
You have an Apache Hive table that contains one billion rows.

You plan to use queries that will filter the data by using the WHERE clause. The values of the columns will be known only while the data loads into a Hive table.

You need to decrease the query runtime.

What should you configure?

A. static partitioning
B. bucket sampling
C. parallel execution
D. dynamic partitioning

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
References: https://www.qubole.com/blog/5-tips-for-efficient-hive-queries/

QUESTION 6
You plan to copy data from Azure Blob storage to an Azure SQL database by using Azure Data Factory.

Which file formats can you use?

A. binary, JSON, Apache Parquet, and ORC
B. OXPS, binary, text and JSON
C. XML, Apache Avro, text, and ORC
D. text, JSON, Apache Avro, and Apache Parquet

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
References: https://docs.microsoft.com/en-us/azure/data-factory/supported-file-formats-and-compression-codecs

QUESTION 7
You have an Apache Spark cluster in Azure HDInsight.

You plan to join a large table and a lookup table.

You need to minimize data transfers during the join operation.

What should you do?

A. Use the reduceByKey function.
B. Use a Broadcast variable.
C. Repartition the data.
D. Use the DISK_ONLY storage level.
E. Store the lookup table to a disk.
F. Store the lookup table to Azure Blob storage.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:

References: https://www.dezyre.com/article/top-50-spark-interview-questions-and-answers-for-2017/208

QUESTION 8
You have an Apache Spark cluster in Azure HDInsight.

You execute the following command.

```
%spark
import org.apache.spark.sql.hive.orc._
import org.apache.spark.sql._
```

What is the result of running the command?

A. the Hive ORC library is imported to Spark and external tables in ORC format are created
B. the Spark library is imported and the data is loaded to an Apache Hive table
C. the Hive ORC library is imported to Spark and the ORC-formatted data stored in Apache Hive tables becomes accessible
D. the Spark library is imported and Scala functions are executed

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 9
You use YARN to manage the resources for a Spark Thrift Server running on a Linux-based Apache Spark cluster in Azure HDInsight.

You discover that the cluster does not fully utilize the resources. You want to increase resource allocation.

You need to increase the number of executors and the allocation of memory to the Spark Thrift Server driver.

Which two parameters should you modify? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. spark.dynamicAllocation.maxExecutors
B. spark.cores.max
C. spark.executor.memory
D. spark_thrift_cmd_opts
E. spark.executor.instances

Correct Answer: AC
Section: (none)
Explanation

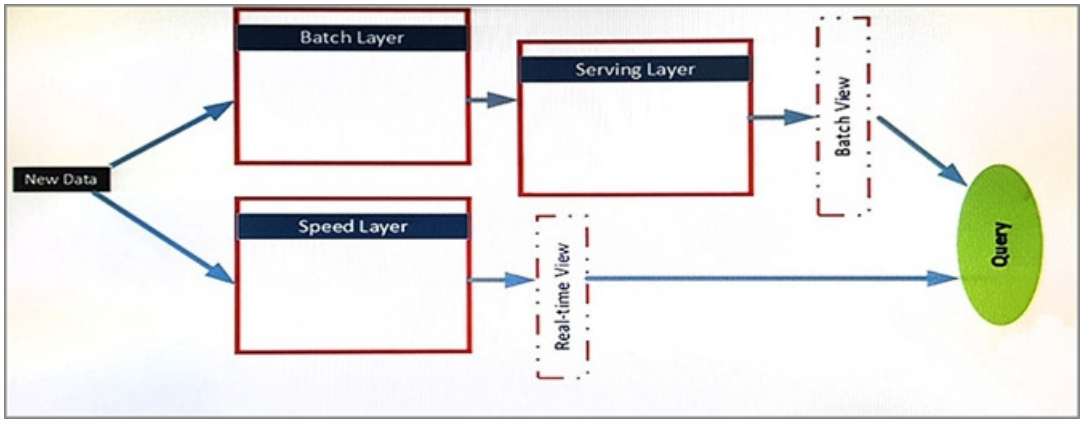
Explanation/Reference:
Explanation:
References: https://stackoverflow.com/questions/37871194/how-to-tune-spark-executor-number-cores-and-executor-memory

QUESTION 10
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:
- Support analysts who run applications that will use REST to submit Spark jobs.
- Business analysts who use JDBC and ODBC client applications from a real-time view. The business analysts run monitoring queries to access aggregate results for 15 minutes. The results will be referenced by subsequent queries.
- Data analysts who publish notebooks drawn from batch layer, serving layer, and speed layer queries. All of the notebooks must support native interpreters for data sources that are batch processed. The serving layer queries are written in Apache Hive and must support multiple sessions. Unique GUIDs are used across the data sources, which allow the data analysts to use Spark SQL.

The data sources in the batch layer share a common storage container. The following data sources are used:
- Hive for sales data
- Apache HBase for operations data
- HBase for logistics data by using a single region server

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
Section: (none)
Explanation

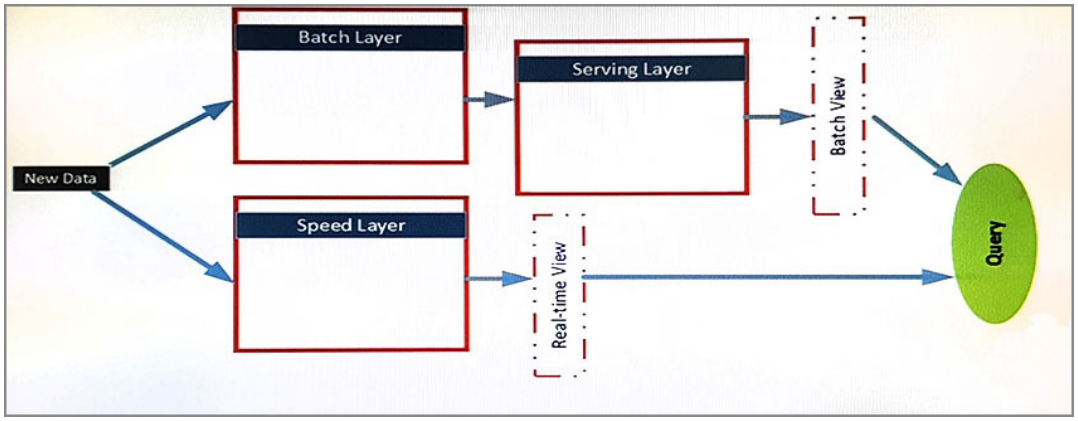
Explanation/Reference:
Explanation:

QUESTION 11
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:

- Support analysts who run applications that will use REST to submit Spark jobs.
- Business analysts who use JDBC and ODBC client applications from a real-time view. The business analysts run monitoring queries to access aggregate results for 15 minutes. The results will be referenced by subsequent queries.
- Data analysts who publish notebooks drawn from batch layer, serving layer, and speed layer queries. All of the notebooks must support native interpreters for data sources that are batch processed. The serving layer queries are written in Apache Hive and must support multiple sessions. Unique GUIDs are used across the data sources, which allow the data analysts to use Spark SQL.

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

- Hive for sales data
- Apache HBase for operations data
- HBase for logistics data by using a single region server

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
Section: (none)
Explanation

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

QUESTION 12
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 need to deploy a NoSQL database to an HDInsight cluster. You will manage the server that host the database by using Remote Desktop. The database must use the key/value pair format in a columnar model.

What should you do?

A. Use an Azure PowerShell script to create and configure a premium HDInsight cluster. Specify Apache Hadoop as the cluster type and use Linux as the operating system.
B. Use the Azure portal to create a standard HDInsight cluster. Specify Apache Spark as the cluster type and use Linux as the operating system.
C. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Windows as the operating system.
D. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache Storm as the cluster type and use Windows as the operating system.
E. Use an Azure PowerShell script to create a premium HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.
F. Use an Azure portal to create a standard HDInsight cluster. Specify Apache Interactive Hive as the cluster type and use Linux as the operating system.
G. Use an Azure portal to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.

Correct Answer: G
Section: (none)
Explanation

Explanation/Reference:
Explanation:
References: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hbase-overview

QUESTION 13
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 need to deploy an enterprise data warehouse that will support in-memory analytics. The data warehouse must support connections that use the Microsoft Hive ODBC Driver and Beeline. The data warehouse will be managed by using Apache Amrabi only.

What should you do?

A. Use an Azure PowerShell script to create and configure a premium HDInsight cluster. Specify Apache Hadoop as the cluster type and use Linux as the operating system.
B. Use the Azure portal to create a standard HDInsight cluster. Specify Apache Spark as the cluster type and use Linux as the operating system.
C. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Windows as the operating system.
D. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache Storm as the cluster type and use Windows as the operating system.
E. Use an Azure PowerShell script to create a premium HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.
F. Use an Azure portal to create a standard HDInsight cluster. Specify Apache Interactive Hive as the cluster type and use Linux as the operating system.
G. Use an Azure portal to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.

Correct Answer: F

Section: (none)
Explanation

Explanation/Reference:
Explanation:
References: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-use-interactive-hive

QUESTION 14
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 need to deploy an HDInsight cluster that will provide in-memory processing, interactive queries, and micro-batch stream processing. The cluster has the following requirements:
- Uses Azure Data Lake Store as the primary storage
- Can be used by HDInsight applications

What should you do?

A. Use an Azure PowerShell script to create and configure a premium HDInsight cluster. Specify Apache Hadoop as the cluster type and use Linux as the operating system.
B. Use the Azure portal to create a standard HDInsight cluster. Specify Apache Spark as the cluster type and use Linux as the operating system.
C. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Windows as the operating system.
D. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache Storm as the cluster type and use Windows as the operating system.
E. Use an Azure PowerShell script to create a premium HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.
F. Use an Azure portal to create a standard HDInsight cluster. Specify Apache Interactive Hive as the cluster type and use Linux as the operating system.
G. Use an Azure portal to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
References: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-apache-spark-overview

QUESTION 15
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 need to deploy an HDInsight cluster that will have a custom Apache Ambari configuration. The cluster will be joined to a domain and must perform the following:
- Fast data analytics and cluster computing by using in-memory processing
- Interactive queries and micro-batch stream processing

What should you do?

A. Use an Azure PowerShell script to create and configure a premium HDInsight cluster. Specify Apache Hadoop as the cluster type and use Linux as the operating system.
B. Use the Azure portal to create a standard HDInsight cluster. Specify Apache Spark as the cluster type and use Linux as the operating system.

C. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Windows as the operating system.
D. Use an Azure PowerShell script to create a standard HDInsight cluster. Specify Apache Storm as the cluster type and use Windows as the operating system.
E. Use an Azure PowerShell script to create a premium HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.
F. Use an Azure portal to create a standard HDInsight cluster. Specify Apache Interactive Hive as the cluster type and use Linux as the operating system.
G. Use an Azure portal to create a standard HDInsight cluster. Specify Apache HBase as the cluster type and use Linux as the operating system.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
References: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-introduction

QUESTION 16
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 plan to consolidate all of the streams into a single timeline, even though none of the streams report events at the same interval.

You need to aggregate the data from the feeds to alight with the time interval stream. The result must be the sum of all the values for each key within a 10 second interval, with the keys being the hashtags.

Which function should you use?

A. countByWindow
B. reduceByWindow
C. reduceByKeyAndWindow
D. countByValueAndWindow
E. updateStateByKey

Correct Answer: E
Section: (none)