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.pass4itsure.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

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

QUESTION 2

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 plan to import 300 TB of data.

You plan to use one job that has many concurrent tasks to import the data in memory.

You need to maximize the amount of concurrent tasks for the job.

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: C

References: https://blog.cloudera.com/blog/2015/03/how-to-tune-your-apache-spark-jobs- part-2/

QUESTION 3

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 designing the real-time portion of the input stream processing. The input will be a continuous stream of data and each record will be processed one at a time. The data will come from an Apache Kafka producer.

You need to identify which HDInsight cluster to use for the final processing of the input data. This will be used to generate continuous statistics and real-time analytics. The latency to process each record must be less than one millisecond and

tasks must be performed in parallel.

Which type of cluster should you identify?

A. Apache Storm

B. Apache Hadoop

C. Apache HBase

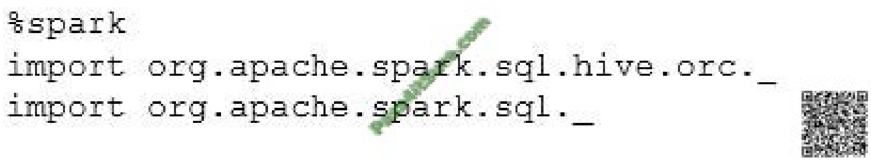
D. Apache Spark

Correct Answer: A

References: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-storm-overview

QUESTION 4

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

QUESTION 5

You have an Azure HDInsight cluster.

You need to store data in a file format that maximizes compression and increases read performance.

Which type of file format should you use?

A. ORC

B. Apache Parquet

C. Apache Avro

D. Apache Sequence

Correct Answer: A

References: http://www.semantikoz.com/blog/orc-intelligent-big-data-file-format-hadoop- hive/

QUESTION 6

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.

Correct Answer: B

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

QUESTION 7

You are configuring the Hive views on an Azure HDInsight cluster that is configured to use Kerberos.

You plan to use the YARN logs to troubleshoot a query that runs against Apache Hadoop.

You need to view the method, the service, and the authenticated account used to run the query.

Which method call should you view in the YARN logs?

A. HQL

B. WebHDFS

C. HDFS C* API

D. Ambari RESR API

Correct Answer: D

QUESTION 8

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

QUESTION 9

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

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

QUESTION 10

You have an Apache Hadoop cluster in Azure HDInsight that has a head node and three data nodes. You have a MapReduce job.

You receive a notification that a data node failed.

You need to identify which component cause the failure.

Which tool should you use?

A. JobTracker

B. TaskTracker

C. ResourceManager

D. ApplicationMaster

Correct Answer: C

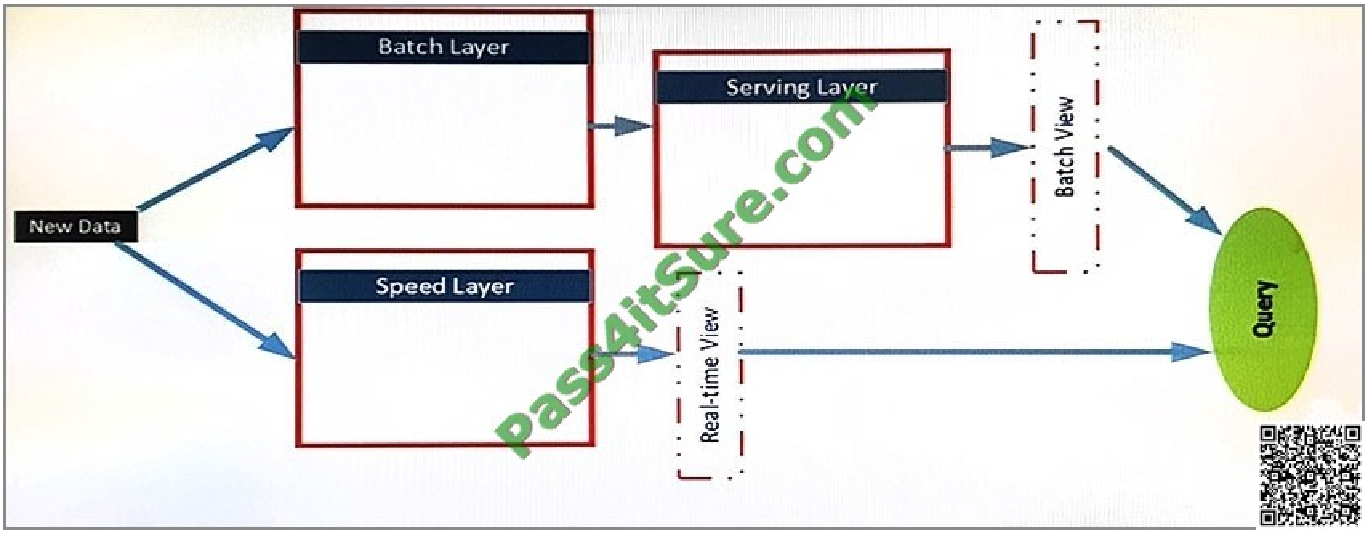
QUESTION 11

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:

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 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 an HDInsight cluster that will provide in-memory processing, interactive queries, and micro-batch stream processing. The cluster has the following requirements:

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.

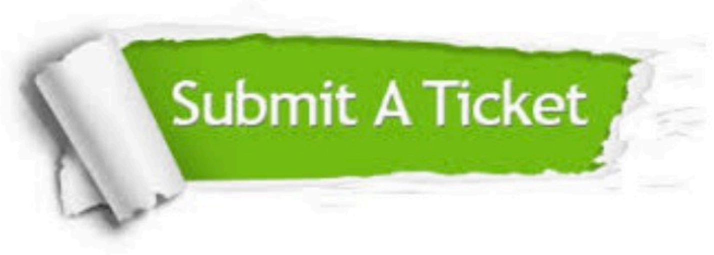
F. Use an Azure portal to create a standard HDInsight cluster. Specify Apache Interactive Hive as the cluster type and use Linux as the operating system.

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

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

Latest 70-775 Dumps 70-775 Study Guide 70-775 Exam Questions


Submit A Ticket