Microsoft MCSA Certification 70-775 Exam

- **Vendor: Microsoft**
- **Exam Code: 70-775**
- **Exam Name: Perform Data Engineering on Microsoft Azure HDInsight**

Get Complete Version Exam 70-775 Dumps with VCE and PDF Here

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

NEW QUESTION 1
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 have a table that contains sales data. You plan to implement a query that will return the number of orders by zip code. You need to minimize the execution time of the queries and to maximize the compression level of the resulting data. 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.

Answer: B

NEW QUESTION 2
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

Answer: D

NEW 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 two new consumers. Create a file import process to send messages. Start the producer.
Does this meet the goal?

A. Yes
B. No

Answer: B

NEW QUESTION 4
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.

Answer: F
Explanation:
https://www.ibm.com/developerworks/library/bd-ooziehadoop/index.html

NEW 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

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

NEW QUESTION 6
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 have data stored in Azure. You need to ensure that you can access the data by using Azure Active Directory (Azure AD) identities. 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.

Answer: G
Explanation:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services

NEW QUESTION 7
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.

Answer: G
Explanation:
https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hbase-overview

NEW QUESTION 8
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

Answer: A
Explanation:
https://docs.microsoft.com/en-us/azure/hdinsight/storm/apache-storm-overview

NEW QUESTION 9
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

Answer: C

NEW QUESTION 10
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

Answer: C

NEW QUESTION 11
.....