



# Microsoft

## Exam Questions 70-775

Perform Data Engineering on Microsoft Azure HDInsight (beta)

### NEW QUESTION 1

#### HOTSPOT

You install the Microsoft Hive ODBC Driver on a computer that runs Windows 10 and has the 64-bit version of Microsoft Office 2016 installed.

You deploy a new Apache Interactive Hive cluster in Azure HDInsight. The cluster is hosted at myHDICluster.azurehdinsight.net and contains a Hive table named hivesampletable that has 200,000 rows.

You plan to use HiveQL exclusively for the queries. The queries will return from 6,000 to 10,000 rows 90 percent of the time.

You need to configure a data source to ensure that you can use Microsoft Excel to access the data. The solution must ensure that the Hive queries execute as quickly as possible.

How should you configure the Advanced Options from the Microsoft Hive ODBC Driver DSN Setup dialog box? To answer select the appropriate options in the answer area.

NOTE:

Each correct selection is worth one point.

## Answer Area

Rows fetched per block:

	▼
3,000	
6,000	
10,000	
20,000	

Fast SQLPrepare:

	▼
Disabled	
Enabled	

Use Native Query:

	▼
Disabled	
Enabled	

**Answer:**

**Explanation:** References: <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-connect-excel-hiveodbc-driver>

### NEW QUESTION 2

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:** <https://docs.microsoft.com/en-us/azure/data-factory/data-factory-supported-file-and-compression-formats>

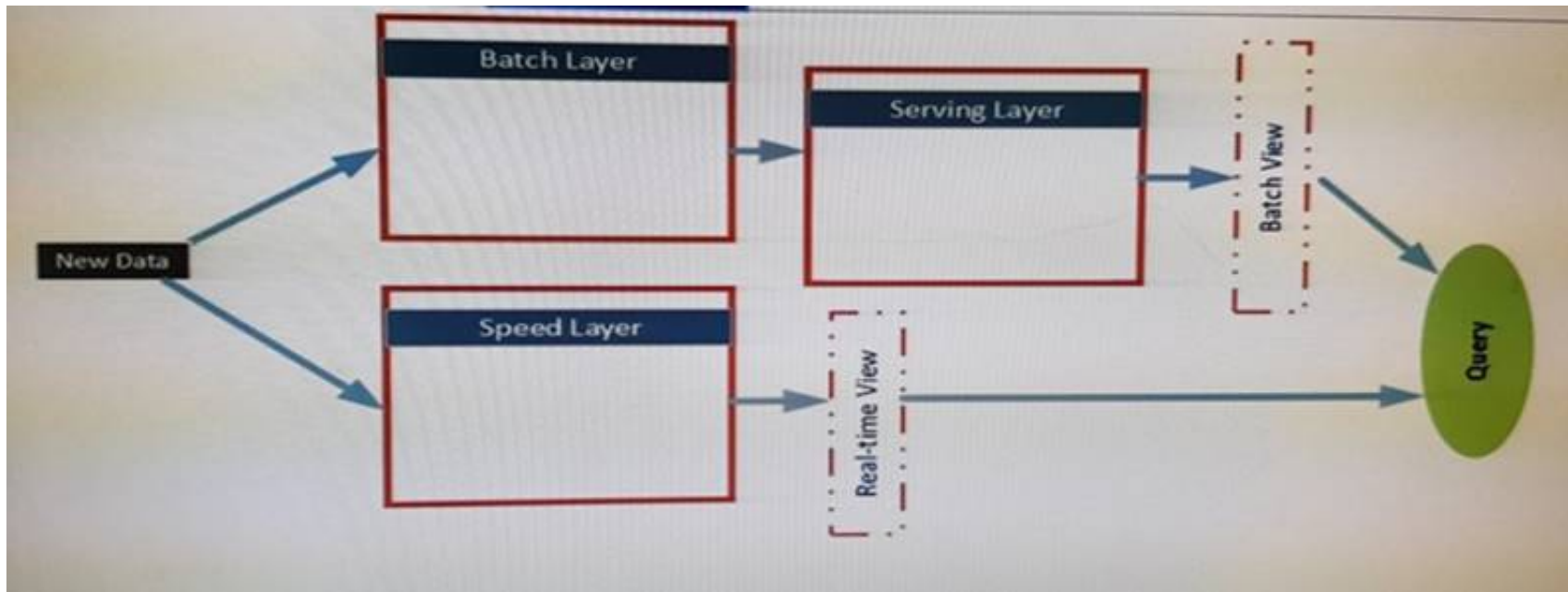
### NEW QUESTION 3

#### DRAG DROP

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 correct solution.

Start of Repeated Scenario:

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:



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 result for 15 minutes. The result 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:

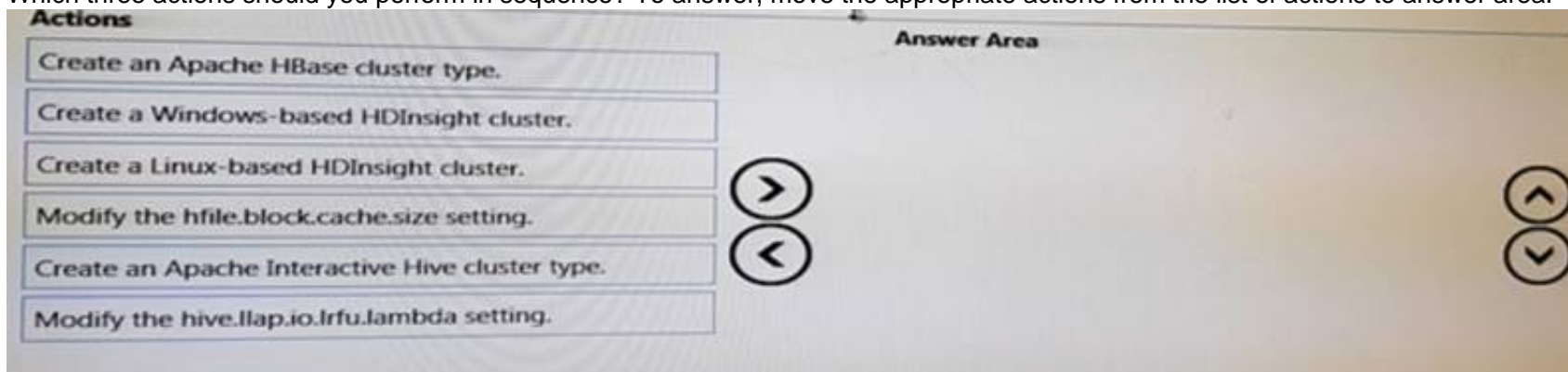
- \* Hive for sales data
- \* Apache HBase for operations data
- \* HBase for logistics data by using a single region server.

End of Repeated scenario.

The business analysts require to monitor the sales data. The queries must be faster and more interactive than the batch layer queries.

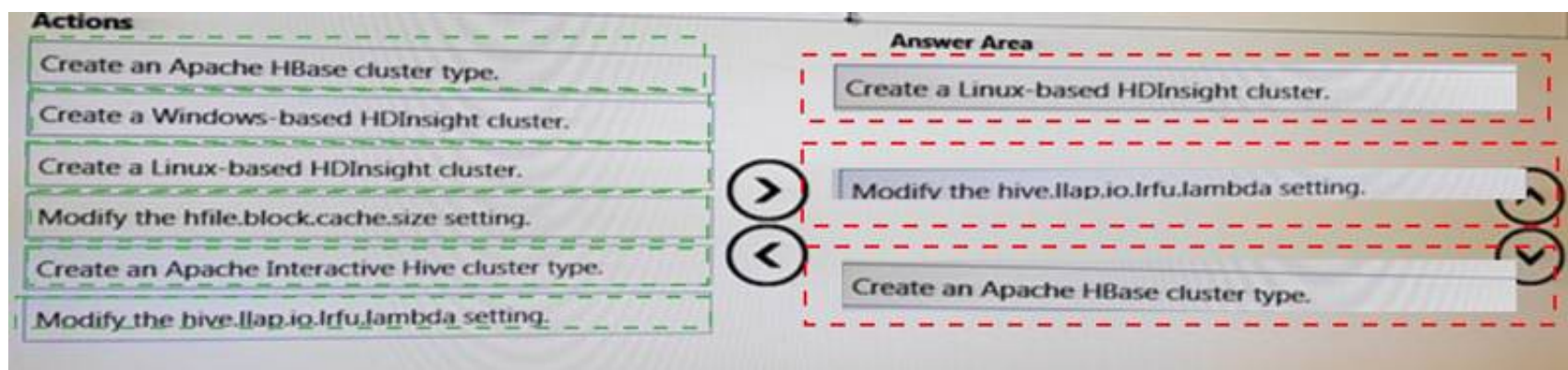
You need to create a new infrastructure to support the queries. The solution must ensure that you can tune the cache policies of the queries.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to answer area.



Answer:

Explanation:



#### NEW QUESTION 4

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 correct solution. Start of Repeated Scenario:

You have an initial data 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 training 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.

End of repeated Scenario.

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

#### NEW QUESTION 5

You have on 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

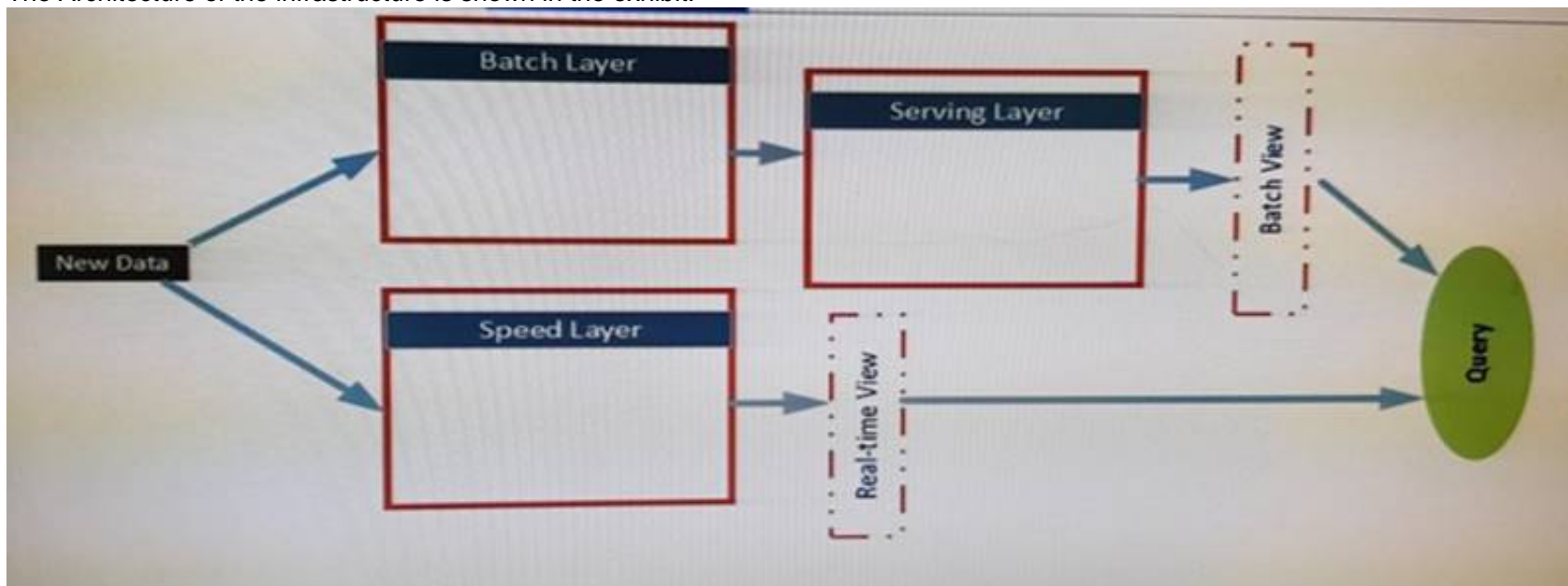
**Explanation:** References: <https://www.qubole.com/blog/5-tips-for-efficient-hive-queries/>

#### NEW QUESTION 6

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 correct solution.

Start of Repeated Scenario:

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:



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 result for 15 minutes. The result 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.

End of Repeated scenario.

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

**Answer:** D

#### NEW 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.

**Answer:** B

#### NEW QUESTION 8

DRAG DROP

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.

#### Code Segments

```
sqlContext.sql("CREATE TABLE IF NOT EXISTS product (productid INT, productname STRING)")
```

```
sqlContext.sql("SELECT productid, productname FROM product LIMIT 100").collect().foreach (println)
```

```
sqlContext.sql("LOAD DATA LOCAL INPATH 'data/examples/product.txt' INTO TABLE product")
```

```
sqlContext.sql("DROP TABLE [IF EXISTS] product")
```

```
val sqlContext = new org.apache.spark.sql.hive.HiveContext(sc)
```

```
sqlContext.sql("SELECT productid, productname FROM product WHERE productid < '100'").collect().foreach (println)
```

#### Answer Area



**Answer:**

#### Explanation:

```
val sqlContext = new org.apache.spark.sql.hive.HiveContext(sc)
sqlContext.sql("CREATE TABLE IF NOT EXISTS productid INT, productname STRING")
sqlContext.sql("LOAD DATA LOCAL INPATH 'Data/examples/product.txt' INTO TABLE product")
sqlContext.sql("SELECT productid, productname FROM product LIMIT 100").collect().foreach (println)
References: https://www.tutorialspoint.com/spark\_sql/spark\_sql\_hive\_tables.htm
```

#### NEW QUESTION 9

DRAG DROP

You have a domain-joined Azure HDInsight cluster. You plan to assign permissions to several support staff.

You need to assign roles to the staff so that they can perform specific tasks.

The solution must use the principle of least privilege.

Which role should you assign for each task? To answer, drag the appropriate roles to the correct tasks. Each role may be used once, more than once, or not at all.

You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

#### Roles

Cluster Administrator

Cluster Operator

Cluster User

Service Administrator

Service Operator

#### Answer Area

##### Task

Configure services:

View service-level alerts:

View cluster configurations:

##### Roles

Role

Role

Role

Answer:

Explanation:

Roles

- Cluster Administrator
- Cluster Operator
- Cluster User
- Service Administrator
- Service Operator

Answer Area

Task

- Configure services:
- View service-level alerts:
- View cluster configurations:

Roles

- Service Administrator
- Cluster User
- Cluster User

NEW QUESTION 10

DRAG DROP

You have a domain joined Apache Hadoop cluster in Azure HDInsight named hdicluster. The Linux account for hdicluster is named Inxuser. Your Active Directory account is names user1@fabrikam.com. You need to run Hadoop commands from an SSH session. Which credentials should you use? To answer, drag the appropriate credentials to the correct commands. Each credential may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

Credentials

- Inxuser@hdicluster
- Inxuser@hdicluster-ssh.azurehdinsight.net
- user1@fabrikam.com
- user1@fabrikam.com.hdinsight.net

Answer Area

- SSH: Credential
- Kinit: Credential

Answer:

Explanation: References: <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-linux-usessh-unix>

NEW QUESTION 10

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

Answer: AC

Explanation: References: <https://stackoverflow.com/questions/37871194/how-to-tune-spark-executornumber-cores-and-executor-memory>

NEW QUESTION 14

DRAG DROP

You have an Apache Hive cluster in Azure HDInsight. You need to tune a Hive query to meet the following requirements:

- Use the Tez engine.
- Process 1,024 rows in a batch.

How should you complete this query? To answer, drag the appropriate values to the correct targets.

Values

- hive.execution.engine
- hive.prewarm.enabled
- hive.tez.auto.reducer.parallelism
- hive.tez.container.size
- hive.vectorized.execution.enabled
- tez.runtime.compress.codec

Answer Area

Set Value = Tez

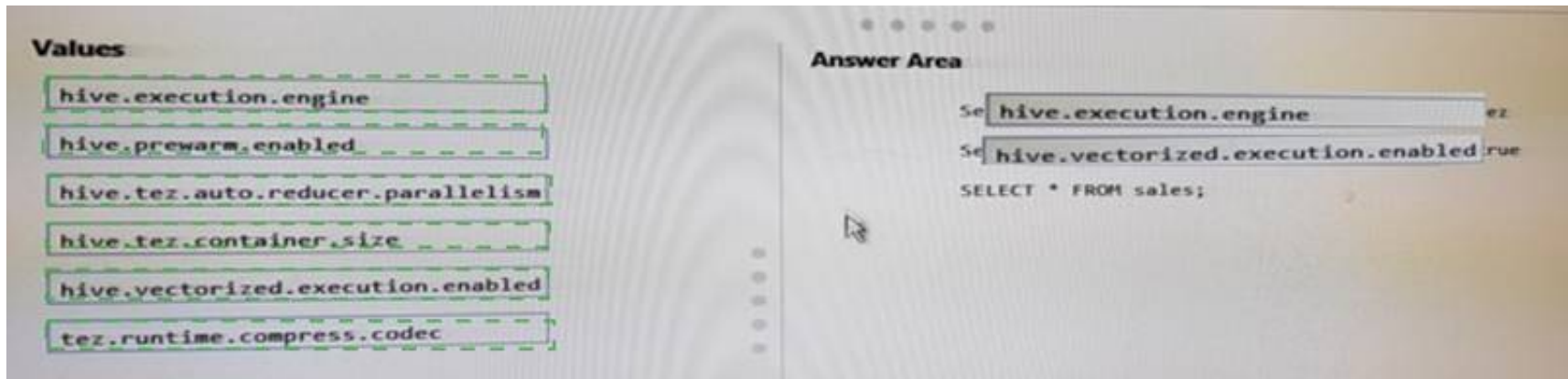
Set Value = true

SELECT \* FROM sales;



Answer:

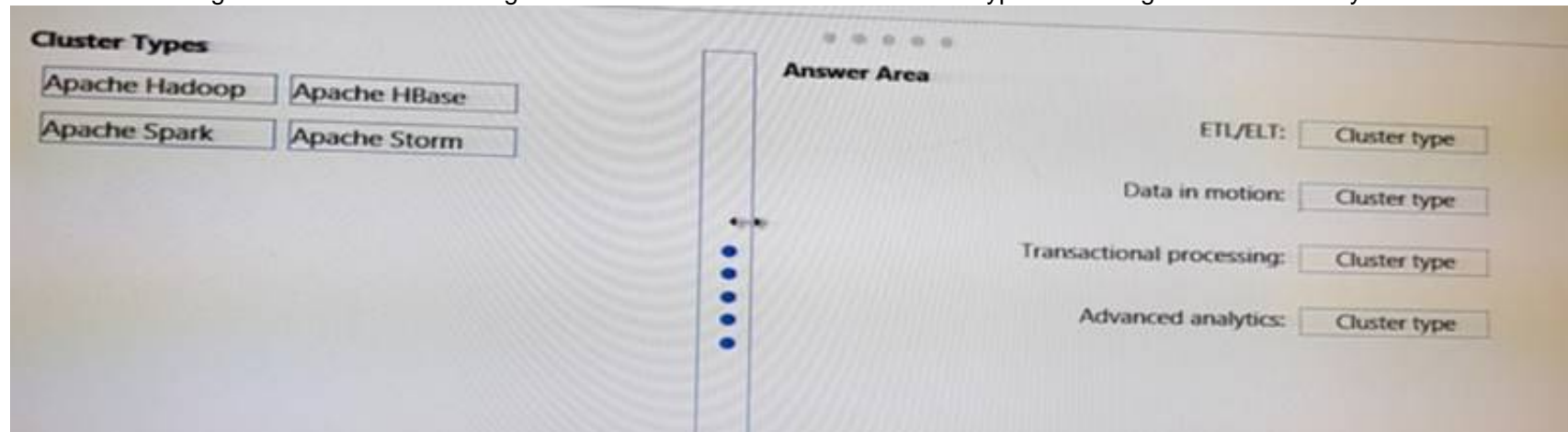
Explanation:



#### NEW QUESTION 16

DRAG DROP

You are evaluating the use of Azure HDInsight clusters for various workloads. Which type of HDInsight cluster should you create for each workloads?



Answer:

Explanation: <https://www.blue-granite.com/blog/how-to-choose-the-right-hdinsight-cluster>

#### NEW QUESTION 18

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 correct solution.

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?

- 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.
- Use the Azure portal to create a standard HDInsight cluster
- Specify Apache Spark as the cluster type and use Linux as the operating system.
- 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.
- Use an Azure PowerShell script to create a standard HDInsight cluster
- Specify ApacheStorm as the cluster type and use Windows as the operating system.
- Use an Azure PowerShell script to create a premium HDInsight cluster
- Specify Apache HBase as the cluster type and use Windows as the operating system.
- Use an Azure portal to create a standard HDInsight cluster
- Specify Apache Interactive Hive as the cluster type and use Windows as the operating system.
- Use an Azure portal to create a standard HDInsight cluster
- Specify Apache HBase as the cluster type and use Windows as the operating system

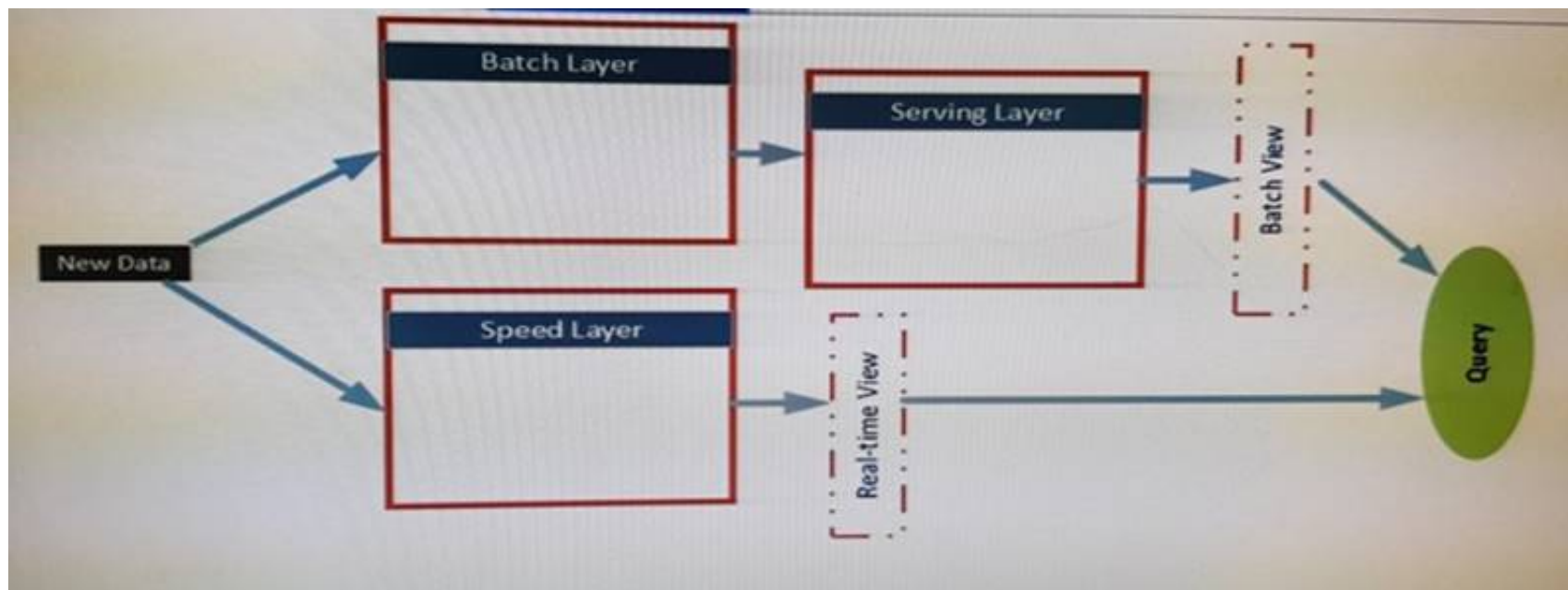
Answer: B

#### NEW QUESTION 23

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 correct solution.

Start of Repeated Scenario:

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:



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 result for 15 minutes. The result 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.

End of Repeated scenario.

You need to ensure that the analysts can query the logistics data by using JDBC APIs and SQL APIs. Which technology should you implement?

- A. Apache Phoenix
- B. Apache Spark
- C. Apache Storm
- D. Apache Hive

**Answer: D**

#### NEW QUESTION 24

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 correct solution.

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 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**

#### NEW QUESTION 25

You have an Azure HDInsight cluster.

You need to build a solution to ingest real-time streaming data into 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 30

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 Ambari 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.
- C. Specify Apache Spark as the cluster type and use Linux as the operating system.



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

**Answer:** F

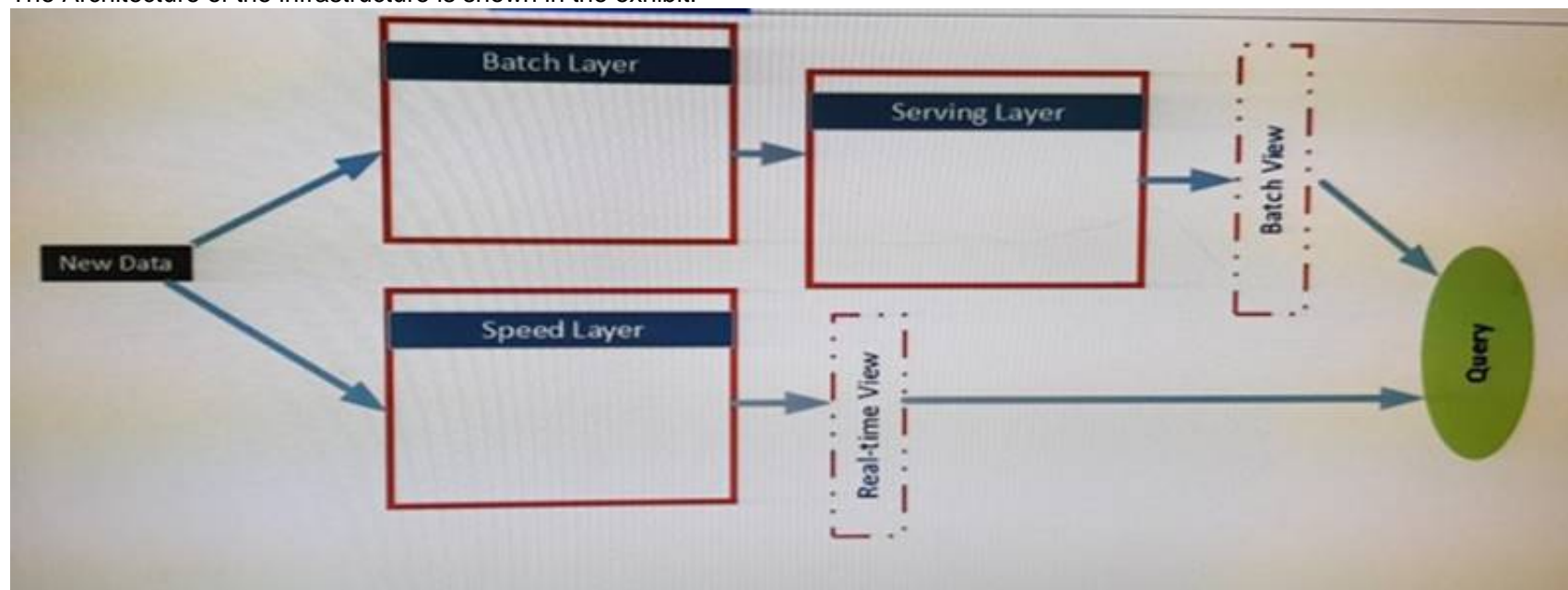
**Explanation:** References: <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-useinteractive-hive>

### NEW QUESTION 35

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 correct solution.

Start of Repeated Scenario:

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:



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 result for 15 minutes. The result 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.

End of Repeated scenario.

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

**Answer:** D

**Explanation:** References: <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-apache-spark-livyrest-interface>

### NEW QUESTION 37

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 correct solution.

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 brokers. Create a file import process to send messages. Run the producer.

Does this meet the goal?

- A. Yes
- B. No

**Answer:** B

#### NEW QUESTION 40

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 correct solution.

Start of Repeated Scenario:

You have an initial data 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 training 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.

End of repeated Scenario.

You plan to consolidate all of the stream 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 align with the time interval stream. The result must be the sum of all values for each 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

**Answer: E**

#### NEW QUESTION 45

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 correct solution.

You need to deploy a NoSQL database to an HDInsight cluster. You will manage the servers 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
- B. Specify Apache Hadoop as the cluster type and use Linux as the operating system.
- C. Use the Azure portal to create a standard HDInsight cluster
- D. Specify Apache Spark as the cluster type and use Linux as the operating system.
- E. Use an Azure PowerShell script to create a standard HDInsight cluster
- F. Specify Apache HBase as the cluster type and use Windows as the operating system.
- G. Use an Azure PowerShell script to create a standard HDInsight cluster
- H. Specify Apache Storm as the cluster type and use Windows as the operating system.
- I. Use an Azure PowerShell script to create a premium HDInsight cluster
- J. Specify Apache HBase as the cluster type and use Windows as the operating system.
- K. Use an Azure portal to create a standard HDInsight cluster
- L. Specify Apache Interactive Hive as the cluster type and use Windows as the operating system.
- M. Use an Azure portal to create a standard HDInsight cluster
- N. Specify Apache HBase as the cluster type and use Windows as the operating system.

**Answer: G**

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

#### NEW QUESTION 49

.....

## Thank You for Trying Our Product

### We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

### 70-775 Practice Exam Features:

- \* 70-775 Questions and Answers Updated Frequently
- \* 70-775 Practice Questions Verified by Expert Senior Certified Staff
- \* 70-775 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* 70-775 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

**100% Actual & Verified — Instant Download, Please Click**  
**[Order The 70-775 Practice Test Here](#)**