

70-475 Dumps

Designing and Implementing Big Data Analytics Solutions

<https://www.certleader.com/70-475-dumps.html>



NEW QUESTION 1

You need to configure the alert to meet the requirements for ETL.

Which settings should you use for the alert? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Event:

Activity Run Finished
Activity Run Started
On-Demand HDI Cluster Create Start
On-Demand HDI Cluster Created Successfully
On-Demand HDI Cluster Deleted

Status:

Failed
Succeeded

Substatus:

--
Abandoned
Failed Execution
Failed Resource Allocation
Failed Validation
Timed Out

Answer:

Explanation: Scenario: Relecloud identifies the following requirements for extract, transformation, and load (ETL): An email alert must be generated when a failure of any type occurs during ETL processing.

NEW QUESTION 2

You need to implement rls_table1.

Which code should you execute? To answer, drag the appropriate values to the correct targets. Each value 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.

Values

Block

Filter

Grant

Security

Server

Answer Area

CREATE

Value

POLICY

dbo.rls_table1_policy

ADD

Value

PREDICATE

dbo.rls_table1(CustomerId, salespersonid)

on

dbo.table1,

ADD

Value

PREDICATE

dbo.rls_table1(CustomerId, salespersonid)

on

dbo.table1

BEFORE

UPDATE,

ADD

Value

PREDICATE

dbo.rls_table1(CustomerId, salespersonid)

on

dbo.table1

BEFORE

DELETE,

ADD

Value

PREDICATE

dbo.rls_table1(CustomerId, salespersonid)

on

dbo.table1

AFTER

INSERT

with

(

state = on

)

Answer:

Explanation: Box 1: Security Security Policy

Example: After we have created Predicate function, we have to bind it to the table, using Security Policy. We will be using CREATE SECURITY POLICY command to set the security policy in place.

CREATE SECURITY POLICY DepartmentSecurityPolicy

ADD FILTER PREDICATE dbo.DepartmentPredicateFunction(UserDepartment) ON dbo.Department WITH(STATE = ON)

Box 2: Filter

[FILTER | BLOCK]

The type of security predicate for the function being bound to the target table. FILTER predicates silently filter the rows that are available to read operations.

BLOCK predicates explicitly block write operations that violate the predicate function.

Box 3: Block

Box 4: Block

Box 5: Filter

NEW QUESTION 3

Which technology should you recommend to meet the technical requirement for analyzing the social media data?

- A. Azure Stream Analytics
- B. Azure Data Lake Analytics
- C. Azure Machine Learning
- D. Azure HDInsight Storm clusters

Answer: A

Explanation: Azure Stream Analytics is a fully managed event-processing engine that lets you set up real-time analytic computations on streaming data.

Scalability

Stream Analytics can handle up to 1 GB of incoming data per second. Integration with Azure Event Hubs and Azure IoT Hub allows jobs to ingest millions of events per second coming from connected devices, clickstreams, and log files, to name a few. Using the partition feature of event hubs, you can partition computations into logical steps, each with the ability to be further partitioned to increase scalability.

NEW QUESTION 4

You have a Microsoft Azure Data Factory pipeline.

You discover that the pipeline fails to execute because data is missing. You need to rerun the failure in the pipeline.

Which cmdlet should you use?

- A. Set-AzureRmAutomationJob
- B. Set-AzureRmDataFactorySliceStatus
- C. Resume-AzureRmDataFactoryPipeline
- D. Resume-AzureRmAutomationJob

Answer: B

Explanation: Use some PowerShell to inspect the ADF activity for the missing file error. Then simply set the dataset slice to either skipped or ready using the cmdlet to override the status.

For example:

```
Set-AzureRmDataFactorySliceStatus `
-ResourceGroupName $ResourceGroup `
-DataFactoryName $ADFName.DataFactoryName `
-DatasetName $Dataset.OutputDatasets `
-StartDateTime $Dataset.WindowStart `
-EndDateTime $Dataset.WindowEnd `
-Status "Ready" `
-UpdateType "Individual" References:
```

<https://stackoverflow.com/questions/42723269/azure-data-factory-pipelines-are-failing-when-no-files-available->

NEW QUESTION 5

You have data pushed to Microsoft Azure Blob storage every few minutes.

You want to use an Azure Machine Learning web service to score the data hourly. You plan to deploy the data factory pipeline by using a Microsoft.NET application. You need to create an output dataset for the web service.

Which three properties should you define? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Source
- B. LinkedServiceName
- C. TypeProperties
- D. Availability
- E. External

Answer: ABC

NEW QUESTION 6

You plan to implement a Microsoft Azure Data Factory pipeline. The pipeline will have custom business logic that requires a custom processing step.

You need to implement the custom processing step by using C#.

Which interface and method should you use? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Interface:

	▼
ICustomActivity	
IDotNetActivity	
IGenericActivity	

Method:

	▼
Copy	
Execute	
Run	
Update	

Answer:

Explanation: References:

<https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/data-factory/v1/data-factory-use-custom-activ>

NEW QUESTION 7

You need to recommend a platform architecture for a big data solution that meets the following requirements: Supports batch processing

Provides a holding area for a 3-petabyte (PB) dataset

Minimizes the development effort to implement the solution

Provides near real time relational querying across a multi-terabyte (TB) dataset

Which two platform architectures should you include in the recommendation? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a Microsoft Azure SQL data warehouse
- B. a Microsoft Azure HDInsight Hadoop cluster
- C. a Microsoft SQL Server database
- D. a Microsoft Azure HDInsight Storm cluster
- E. Microsoft Azure Table Storage

Answer: AE

Explanation: A: Azure SQL Data Warehouse is a SQL-based, fully-managed, petabyte-scale cloud data warehouse. It's highly elastic, and it enables you to set up in minutes and scale capacity in seconds. Scale compute and storage independently, which allows you to burst compute for complex analytical workloads, or scale down your warehouse for archival scenarios, and pay based on what you're using instead of being locked into predefined cluster configurations—and get more cost efficiency versus traditional data warehouse solutions.

E: Use Azure Table storage to store petabytes of semi-structured data and keep costs down. Unlike many data stores—on-premises or cloud-based—Table storage lets you scale up without having to manually shard your dataset. Perform OData-based queries.

NEW QUESTION 8

You have a Microsoft Azure Stream Analytics solution.

You need to identify which types of windows must be used to group like following types of events:

- ▶ Events that have random time intervals and are captured in a single fixed-size window
- ▶ Events that have random time intervals and are captured in overlapping windows

Which window type should you identify for each event type? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Events that have random time intervals and are captured
in a single fixed-size window:

A hopping window
A sliding window
A tumbling window

Events that have random time intervals and are captured
in overlapping windows:

A hopping window
A sliding window
A tumbling window

Answer:

Explanation: Box 1. A sliding Window Box 2: A sliding Window

With a Sliding Window, the system is asked to logically consider all possible windows of a given length and output events for cases when the content of the window actually changes – that is, when an event entered or existed the window.

NEW QUESTION 9

You have an Apache Hadoop system that contains 5 TB of data.

You need to create queries to analyze the data in the system. The solution must ensure that the queries execute as quickly as possible.

Which language should you use to create the queries?

- A. Apache Pig
- B. Java
- C. Apache Hive
- D. MapReduce

Answer: D

NEW QUESTION 10

You are designing an application that will perform real-time processing by using Microsoft Azure Stream Analytics.

You need to identify the valid outputs of a Stream Analytics job.

What are three possible outputs that you can use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Microsoft Power BI
- B. Azure SQL Database
- C. a Hive table in Azure HDInsight
- D. Azure Blob storage
- E. Azure Redis Cache

Answer: ABD

Explanation: <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs>

NEW QUESTION 10

You need to design the data load process from DB1 to DB2. Which data import technique should you use in the design?

- A. PolyBase
- B. SQL Server Integration Services (SSIS)
- C. the Bulk Copy Program (BCP)
- D. the BULK INSERT statement

Answer: C

NEW QUESTION 15

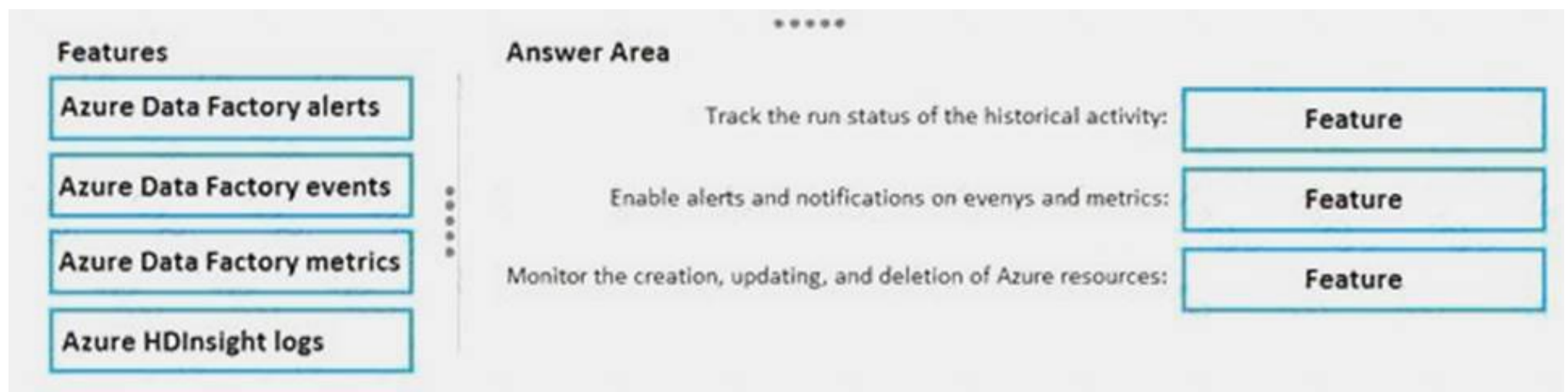
You plan to deploy a Microsoft Azure Data Factory pipeline to run an end-to-end data processing workflow. You need to recommend which Azure Data Factory features must be used to meet the Following requirements: Track the run status of the historical activity.

Enable alerts and notifications on events and metrics.

Monitor the creation, updating, and deletion of Azure resources.

Which features should you recommend? To answer, drag the appropriate features to the correct requirements. Each feature 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.



Answer:

Explanation: Box 1: Azure Hdinsight logs Logs contain historical activities. Box 2: Azure Data Factory alerts Box 3: Azure Data Factory events

NEW QUESTION 16

You have a Microsoft Azure SQL database that contains Personally Identifiable Information (PII).

To mitigate the PII risk, you need to ensure that data is encrypted while the data is at rest. The solution must minimize any changes to front-end applications. What should you use?

- A. Transport Layer Security (TLS)
- B. transparent data encryption (TDE)
- C. a shared access signature (SAS)
- D. the ENCRYPTBYPASSPHRASE T-SQL function

Answer: B

Explanation: Transparent data encryption (TDE) helps protect Azure SQL Database, Azure SQL Managed Instance, and Azure Data Warehouse against the threat of malicious activity. It performs real-time encryption and decryption of the database, associated backups, and transaction log files at rest without requiring changes to the application.

References: <https://docs.microsoft.com/en-us/azure/sql-database/transparent-data-encryption-azure-sql>

NEW QUESTION 19

You have an application that displays data from a Microsoft Azure SQL database. The database contains credit card numbers.

You need to ensure that the application only displays the last four digits of each credit card number when a credit card number is returned from a query. The solution must NOT require any changes to the data in the database. What should you use?

- A. Dynamic Data Masking
- B. cell-level security
- C. Transparent Data Encryption (TDE)
- D. row-level security

Answer: A

NEW QUESTION 21

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 states goals. Some question sets might have more than one correct solution, while the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Apache Spark system that contains 5 TB of data.

You need to write queries that analyze the data in the system. The queries must meet the following requirements:

- ▶ Use static data typing.
- ▶ Execute queries as quickly as possible.
- ▶ Have access to the latest language features. Solution: You write the queries by using Scala.

- A. Yes
- B. No

Answer: A

NEW QUESTION 22

You have data generated by sensors. The data is sent to Microsoft Azure Event Hubs.

You need to have an aggregated view of the data in near real-time by using five minute tumbling windows to identity short-term trends. You must also have hourly and a daily aggregated views of the data.

Which technology should you use for each task? To answer, drag the appropriate technologies to the correct tasks. Each technology 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.

• • • • •

Technologies

Azure Event Hubs

Azure HDInsight MapReduce

Azure Stream Analytics

Answer Area

Create a near real-time tumbling window job:

Technology

Create hourly and daily aggregated views of the data stored in Azure Blob storage:

Technology

Write data to Azure Blob storage in near real-time:

Technology

Answer:

Explanation: Box 1: Azure HDInsight MapReduce

Azure Event Hubs allows you to process massive amounts of data from websites, apps, and devices. The Event Hubs spout makes it easy to use Apache Storm on HDInsight to analyze this data in real time.

Box 2: Azure Event Hub

Box 3: Azure Stream Analytics

Stream Analytics is a new service that enables near real time complex event processing over streaming data. Combining Stream Analytics with Azure Event Hubs enables near real time processing of millions of events per second. This enables you to do things such as augment stream data with reference data and output to storage (or even output to another Azure Event Hub for additional processing).

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 states goals. Some question sets might have more than one correct solution, while the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Apache Spark system that contains 5 TB of data.

You need to write queries that analyze the data in the system. The queries must meet the following requirements:

- ☐ Use static data typing.
- ☐ Execute queries as quickly as possible.
- ☐ Have access to the latest language features. Solution: You write the queries by using Java.

A. Yes

B. No

Answer: B

NEW QUESTION 29

You are designing a data-driven data flow in Microsoft Azure Data Factory to copy data from Azure Blob storage to Azure SQL Database.

You need to create the copy activity.

How should you complete the JSON code? To answer, drag the appropriate code elements to the correct targets. Each element 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.

Values

AzureSQLInput

AzureTableSink

BlobSink

BlobSource

SQLSink

Answer Area

```

{
  "name": "SamplePipeline",
  "properties": {
    "start": "2017-08-01T12:00:00",
    "end": "2017-08-01T14:00:00",
    "description": "Copy",
    "activities": [
      {
        "name": "StorageToSQL",
        "description": "Copy Activity",
        "type": "Copy",
        "inputs": [
          {
            "name": "Input"
          }
        ],
        "outputs": [
          {
            "name": "Output"
          }
        ],
        "typeProperties": {
          "source": {
            "type": Value
          },
          "sink": {
            "type": Value
          }
        },
        "scheduler": {
          "frequency": "Hour",
          "interval": 1
        },
        "policy": {
          "concurrency": 1,
          "executionPriorityOrder": "Oldest",
          "retry": 0,
          "timeout": "01:00:00"
        }
      }
    ]
  }
}

```

Answer:

Explanation:

Values	Answer Area
AzureSQLInput	<pre>{ "name": "SamplePipeline", "properties": { "start": "2017-08-01T12:00:00", "end": "2017-08-01T14:00:00", "description": "Copy", "activities": [{ "name": "StorageToSQL", "description": "Copy Activity", "type": "Copy", "inputs": [{ "name": "Input" }], "outputs": [{ "name": "Output" }], "typeProperties": { "source": { "type": "BlobSource", "sink": { "type": "AzureTableSink" } } }, "scheduler": { "frequency": "Hour", "interval": 1 }, "policy": { "concurrency": 1, "executionPriorityOrder": "Oldest", "retry": 0, "timeout": "01:00:00" } }] } }</pre>
AzureTableSink	
BlobSink	
BlobSource	
SQLSink	

NEW QUESTION 30

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 the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a Microsoft Azure deployment that contains the following services:

- ▶ Azure Data Lake
- ▶

Azure Cosmos DB



Azure Data Factory



Azure SQL Database

You load several types of data to Azure Data Lake.

You need to load data from Azure SQL Database to Azure Data Lake. Solution: You use a stored procedure.

Does this meet the goal?

A. Yes

B. No

Answer: B

Explanation: Note: You can use the Copy Activity in Azure Data Factory to copy data to and from Azure Data Lake Storage Gen1 (previously known as Azure Data Lake Store). Azure SQL database is supported as source.

References: <https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-store>

NEW QUESTION 34

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 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 section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to deploy a Microsoft Azure SQL data warehouse and a web application.

The data warehouse will ingest 5 TB of data from an on-premises Microsoft SQL Server database daily. The web application will query the data warehouse.

You need to design a solution to ingest data into the data warehouse.

Solution: You use AzCopy to transfer the data as text files from SQL Server to Azure Blob storage, and then you use Azure Data Factory to refresh the data warehouse database.

Does this meet the goal?

A. Yes

B. No

Answer: B

NEW QUESTION 37

You need to ingest data from various data stores into a Microsoft Azure SQL data warehouse by using PolyBase.

You create an Azure Data Factory.

Which three components should you create next? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. an Azure Function

B. datasets

C. a pipeline

D. an Azure Batch account

E. linked services

Answer: AE

NEW QUESTION 41

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 states goals. Some question sets might have more than one correct solution, while the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to implement a new data warehouse.

You have the following information regarding the data warehouse:



The first data files for the data warehouse will be available in a few days.



Most queries that will be executed against the data warehouse are ad-hoc.



The schemas of data files that will be loaded to the data warehouse change often.



One month after the planned implementation, the data warehouse will contain 15 TB of data. You need to recommend a database solution to support the planned implementation.

Solution: You recommend an Apache Hadoop system. Does this meet the goal?

A. Yes

B. No

Answer: A

NEW QUESTION 44

You plan to analyze the execution logs of a pipeline to identify failures by using Microsoft power BI. You need to automate the collection of monitoring data for the planned analysis.

What should you do from Microsoft Azure?

A. Create a Data Factory Set

B. Save a Data Factory Log

C. Add a Log Profile

D. Create an Alert Rule Email

Answer: A

Explanation: You can import the results of a Log Analytics log search into a Power BI dataset so you can take advantage of its features such as combining data from different sources and sharing reports on the web and mobile devices.

To import data from a Log Analytics workspace into Power BI, you create a dataset in Power BI based on a log search query in Log Analytics. The query is run each time the dataset is refreshed. You can then build Power BI reports that use data from the dataset.

References: <https://docs.microsoft.com/en-us/azure/azure-monitor/platform/powerbi>

NEW QUESTION 47

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 section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

Your company has multiple databases that contain millions of sales transactions. You plan to implement a data mining solution to identify purchasing fraud.

You need to design a solution that mines 10 terabytes (TB) of sales data. The solution must meet the following requirements:

- ▶ Run the analysis to identify fraud once per week.
- ▶ Continue to receive new sales transactions while the analysis runs.
- ▶ Be able to stop computing services when the analysis is NOT running. Solution: You create a Microsoft Azure HDInsight cluster.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation: HDInsight cluster billing starts once a cluster is created and stops when the cluster is deleted. Billing is pro-rated per minute, so you should always delete your cluster when it is no longer in use.

NEW QUESTION 50

Your company plans to deploy a web application that will display marketing data to its customers. You create an Apache Hadoop cluster in Microsoft Azure HDInsight and an Azure data factory. You need to implement a linked service to the cluster.

Which JSON specification should you use to create the linked service?

- A. {
 "name": "AzureBlobOutput",
 "properties": {
 "type": "AzureBlob",
 "linkedServiceName": "StorageLinkedService",
 "typeProperties": {
 "folderPath": "adfgetstarted/partitioneddata",
 "format": {
 "type": "TextFormat",
 "columnDelimiter": ",",
 }
 }
 "availability": {
 "frequency": "Month",
 "interval": 1
 }
 }
}
- B. {
 "name": "HDInsightOnDemandLinkedService",
 "properties": {
 "type": "HDInsightOnDemand",
 "typeProperties": {
 "version": "3.2",
 "clusterSize": 1,
 "timeToLive": "00:30:00",
 "linkedServiceName": "StorageLinkedService"
 }
 }
}

B. {
 "name": "HDInsightOnDemandLinkedService",
 "properties": {
 "type": "HDInsightOnDemand",
 "typeProperties": {
 "version": "3.2",
 "clusterSize": 1,
 "timeToLive": "00:30:00",
 "linkedServiceName": "StorageLinkedService"
 }
 }
}

C. {
 "name": "StorageLinkedService",
 "properties": {
 "type": "AzureStorage",
 "description": "",
 "typeProperties": {
 "connectionString":
 "DefaultEndpointProtocol=https;AccountName=<account>;
 AccountKey=<accountkey>"
 }
 }
}

D. {
 "name": "AzureBlobInput",
 "properties": {
 "type": "AzureBlob",
 "linkedServiceName": "StorageLinkedService",
 "typeProperties": {
 "folderPath": "adfgetstarted/inputdata",
 "format": {
 "type": "TextFormat",
 "columnDelimiter": ",",
 }
 }
 "availability": {
 "frequency": "Month",
 "interval": 1
 },
 "external": true,
 "policy": {}
 }
}

- A. Option A
- B. Option B
- C. Option C
- D. Option D

Answer: B

NEW QUESTION 54

You plan to design a solution to gather data from 5,000 sensors that are deployed to multiple machines. The sensors generate events that contain data on the health status of the machines.

You need to create a new Microsoft Azure event hub to collect the event data.

Which command should you run? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

The screenshot shows the following commands in the command editor:

- Add-Type -Path "C:\temp\Microsoft.ServiceBus.dll"
- Get-AzureSBNamespace -Name \$Namespace -NameSpaceType Messaging
- Set-AzureSBNamespace -Name \$Namespace** (Highlighted with a blue box and a callout: "These are the selections for the third part of the command.")
- New-Object -TypeName Microsoft.ServiceBus.Messaging.EventHubDescription** (Highlighted with a blue box and a callout: "These are the selections for the fourth part of the command.")

Answer:

Explanation:



NEW QUESTION 56

A Company named Fabrikam, Inc. has a web app. Millions of users visit the app daily. Fabrikam performs a daily analysis of the previous day's logs by scheduling the following Hive query.

```
CREATE EXTERNAL TABLE IF NOT EXISTS UserActivity (...) PARTITIONED BY (LogDate string) LOCATION 'wasb:///logs';
MSCK REPAIR TABLE UserActivity;
Select ... From UserActivity where LogDate = "{date}";
```

You need to recommend a solution to gather the log collections from the web app. What should you recommend?

- A. Generate a single directory that contains multiple files for each da
- B. Name the file by using the syntax of {date}_{randomsuffix}.txt.
- C. Generate a directory that is named by using the syntax of "LogDate={date}" and generate a set of files for that day.
- D. Generate a directory each day that has a single file.
- E. Generate a single directory that has a single file for each day.

Answer: B

NEW QUESTION 57

A company named Fabrikam, Inc. has a Microsoft Azure web app. Billions of users visit the app daily. The web app logs all user activity by using text files in Azure Blob storage. Each day, approximately 200 GB of text files are created. Fabrikam uses the log files from an Apache Hadoop cluster on Azure DHInsight. You need to recommend a solution to optimize the storage of the log files for later Hive use. What is the best property to recommend adding to the Hive table definition to achieve the goal? More than one answer choice may achieve the goal. Select the BEST answer.

- A. STORED AS RCFILE
- B. STORED AS GZIP
- C. STORED AS ORC
- D. STORED AS TEXTFILE

Answer: C

Explanation: The Optimized Row Columnar (ORC) file format provides a highly efficient way to store Hive data. It was designed to overcome limitations of the other Hive file formats. Using ORC files improves performance when Hive is reading, writing, and processing data. Compared with RCFile format, for example, ORC file format has many advantages such as:

- ▶ a single file as the output of each task, which reduces the NameNode's load
- ▶ Hive type support including datetime, decimal, and the complex types (struct, list, map, and union)
- ▶ light-weight indexes stored within the file
- ▶ skip row groups that don't pass predicate filtering
- ▶ seek to a given row
- ▶ block-mode compression based on data type
- ▶ run-length encoding for integer columns
- ▶ dictionary encoding for string columns
- ▶ concurrent reads of the same file using separate RecordReaders
- ▶ ability to split files without scanning for markers
- ▶ bound the amount of memory needed for reading or writing
- ▶ metadata stored using Protocol Buffers, which allows addition and removal of fields

NEW QUESTION 61

You are designing a solution based on the lambda architecture. The solution has the following layers;

- ▶ Batch
- ▶ Speed
- ▶ Serving

You are planning the data ingestion process and the query execution.

For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.

Answer Area:

The data ingestion process must only communicate with the batch layer:

▼

Yes

No

The query execution must communicate with both the serving layer and the speed layer:

▼

Yes

No

You can use Kafka to execute the queries:

▼

Yes

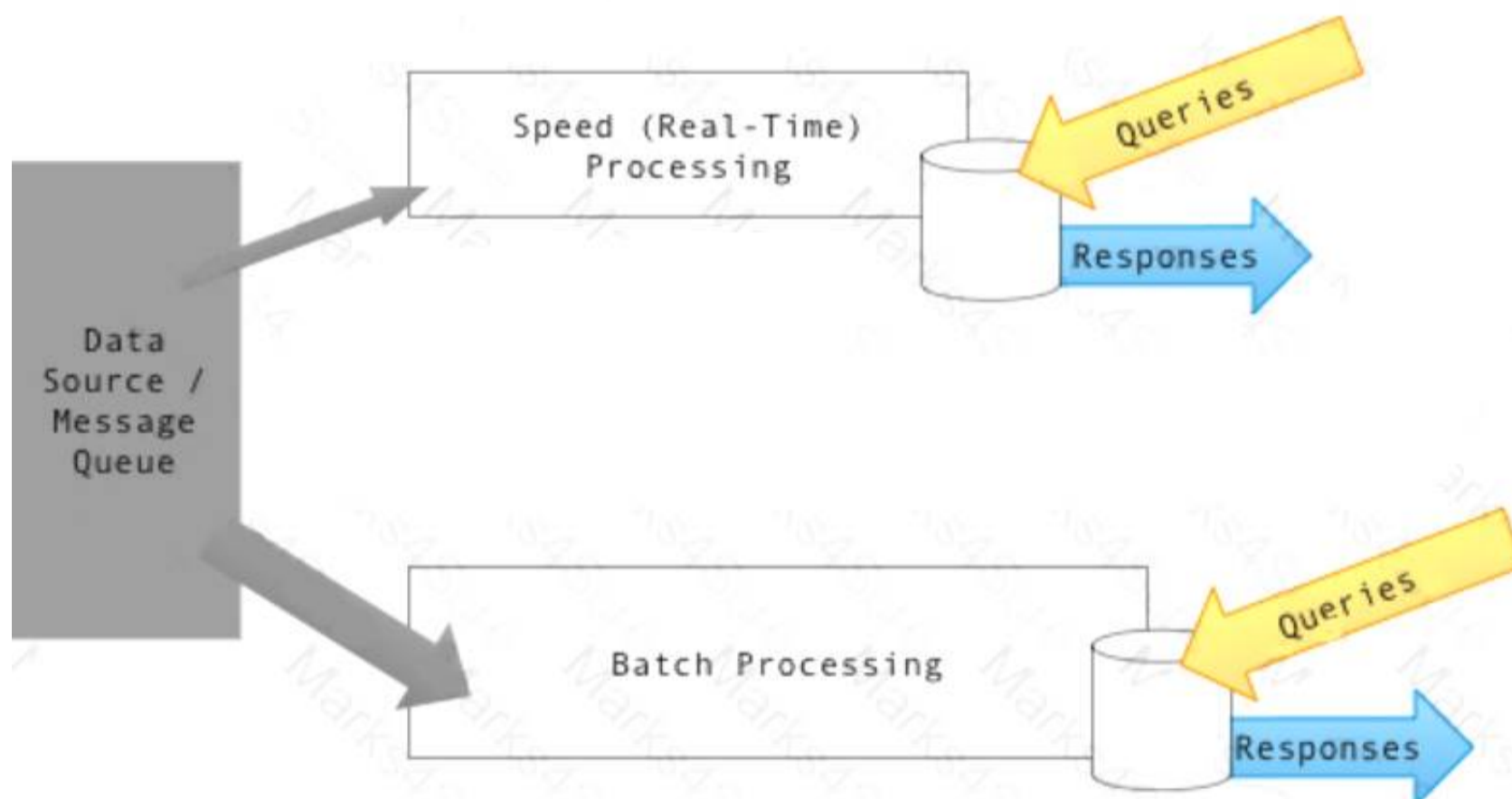
No

Answer:

Explanation: Box 1: No

Box 2: No

Output from the batch and speed layers are stored in the serving layer, which responds to ad-hoc queries by returning precomputed views or building views from the processed data.



Box 3: Yes.

We are excited to announce Interactive Queries, a new feature for stream processing with Apache Kafka. Interactive Queries allows you to get more than just processing from streaming.

Note: Lambda architecture is a popular choice where you see stream data pipelines applied (speed layer). Architects can combine Apache Kafka or Azure Event Hubs (ingest) with Apache Storm (event processing), Apache HBase (speed layer), Hadoop for storing the master dataset (batch layer), and, finally, Microsoft Power BI for reporting and visualization (serving layer).

NEW QUESTION 66

You have a Microsoft Azure data factory.

You assign administrative roles to the users in the following table.

User name	Role
User1	Contributor
User2	Administrator
User3	Reader
User4	Automation Operator
User5	Owner

You discover that several new data factory instances were created.

You need to ensure that only User5 can create a new data factory instance.

Which two roles should you change? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. User2 to Reader
- B. User3 to Contributor
- C. User1 to Reader
- D. User4 to Contributor
- E. User5 to Administrator

Answer: AC

NEW QUESTION 71

You are developing a solution to ingest data in real-time from manufacturing sensors. The data will be archived. The archived data might be monitored after it is written.

You need to recommend a solution to ingest and archive the sensor data. The solution must allow alerts to be sent to specific users as the data is ingested.

What should you include in the recommendation?

- A. a Microsoft Azure notification hub and an Azure function
- B. a Microsoft Azure notification hub an Azure logic app
- C. a Microsoft Azure Stream Analytics job that outputs data to an Apache Storm cluster in AzureHDInsight
- D. a Microsoft Azure Stream Analytics job that outputs data to Azure Cosmos DB

Answer: C

NEW QUESTION 73

Your company has a Microsoft Azure environment that contains an Azure HDInsight Hadoop cluster and an Azure SQL data warehouse. The Hadoop cluster contains text files that are formatted by using UTF-8 character encoding.

You need to implement a solution to ingest the data to the SQL data warehouse from the Hadoop cluster. The solution must provide optimal read performance for the data after ingestion.

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

Actions

From the SQL data warehouse, create external objects.

From Apache Hive, create a stored procedure.

From the SQL data warehouse, create statistics on the data.

From Apache Hive, create external objects.

From Apache Hive, create statistics on the data.

From the SQL data warehouse, create a stored procedure.

>

<

Answer Area

^

v

Answer:

Explanation: SQL Data Warehouse supports loading data from HDInsight via PolyBase. The process is the same as loading data from Azure Blob Storage - using PolyBase to connect to HDInsight to load data.

Use PolyBase and T-SQL Summary of loading process: Recommendations

Create statistics on newly loaded data. Azure SQL Data Warehouse does not yet support auto create or auto update statistics. In order to get the best performance from your queries, it's important to create statistics on all columns of all tables after the first load or any substantial changes occur in the data.





NEW QUESTION 78

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 states goals. Some question sets might have more than one correct solution, while the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to implement a new data warehouse.

You have the following information regarding the data warehouse:

-  The first data files for the data warehouse will be available in a few days.
-  Most queries that will be executed against the data warehouse are ad-hoc.
-  The schemas of data files that will be loaded to the data warehouse change often.
-  One month after the planned implementation, the data warehouse will contain 15 TB of data. You need to recommend a database solution to support the planned implementation.

Solution: You recommend a Microsoft SQL server on a Microsoft Azure virtual machine. Does this meet the goal?

- A. Yes
- B. No

Answer: B

NEW QUESTION 80

You are designing an Internet of Things (IoT) solution intended to identify trends. The solution requires the real-time analysis of data originating from sensors. The results of the analysis will be stored in a SQL database.

You need to recommend a data processing solution that uses the Transact-SQL language. Which data processing solution should you recommend?

- A. Microsoft Azure Stream Analytics
- B. Microsoft Azure HDInsight Spark clusters
- C. Microsoft Azure Event Hubs
- D. Microsoft Azure HDInsight Hadoop clusters

Answer: A





Explanation: For your Internet of Things (IoT) scenarios that use Event Hubs, Azure Stream Analytics can serve as a possible first step to perform near real-time analytics on telemetry data. Just like Event Hubs, Steam Analytics supports the streaming of millions of event per second. Unlike a standard database, analysis is performed on data in motion. This streaming input data can also be combined with reference data inputs to perform lookups or do correlation to assist in unlocking business insights. It uses a SQL-like language to simplify the analysis of data inputs and detect anomalies, trigger alerts or transform the data in order to create valuable outputs

NEW QUESTION 81

You have an Apache Spark cluster on Microsoft Azure HDInsight for all analytics workloads.

You plan to build a Spark streaming application that processes events ingested by using Azure Event Hubs. You need to implement checkpointing in the Spark streaming application for high availability of the event data.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

Actions		Answer Area
Create a SparkConf object.		
Set the checkpoint interval for the stream.	 	 
Set the checkpoint directory for the StreamingContext.		
Create a StreamingContext.		
Create a stream.		

Answer:

Explanation:



NEW QUESTION 84

You have a Microsoft Azure SQL data warehouse named DW1.

A department in your company creates an Azure SQL database named DB1. DB1 is a data mart.

Each night, you need to insert new rows into 9,000 tables in DB1 from changed data in DW1. The solution must minimize costs.

What should you use to move the data from DW1 to DB1, and then to import the changed data to DB1? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Move the data from DW1 to DB1:

▼

Azure Data Factory

Azure Stream Analytics

Microsoft SQL Server Integration Services

Import the data to DB1:

▼

The BULK INSERT statement

PolyBase

Lazy Loading

Answer:

Explanation: Box 1: Azure Data Factory

Use the Copy Activity in Azure Data Factory to move data to/from Azure SQL Data Warehouse. Box 2: The BULK INSERT statement

NEW QUESTION 85

You have a data warehouse that contains the sales data of several customers.

You plan to deploy a Microsoft Azure data factory to move additional sales data to the data warehouse. You need to develop a data factory job that reads reference data from a table in the source data.

Which type of activity should you add to the control flow of the job?

- A. a ForEach activity
- B. a lookup activity
- C. a web activity
- D. a GetMetadata activity

Answer: B

Explanation: References:

<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-lookup-activity>

NEW QUESTION 89

You manage a Microsoft Azure HDInsight Hadoop cluster. All of the data for the cluster is stored in Azure Premium Storage.

You need to prevent all users from accessing the data directly. The solution must allow only the HDInsight service to access the data.

Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions		Answer Area
From the Azure portal, create a copy of the storage account key.		
From the Azure portal, create a stored access policy.		
From the HDInsight Hadoop cluster, restart all of the Hadoop services.		
From the HDInsight Hadoop cluster, modify the properties of the custom core-site.		
From the HDInsight Hadoop cluster, enable maintenance mode.		
From the Azure portal, create a copy of the shared access signature (SAS) token.		

Answer:

Explanation: 1. Create Shared Access Signature policy2. Save the SAS policy token, storage account name, and container name. These values are used when associating the storage account with your HDInsight cluster.3. Update property of core-site4. Maintenance mode5. Restart all services<https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-storage-sharedaccesssignature-permissions>

NEW QUESTION 94

You plan to deploy a Hadoop cluster that includes a Hive installation.
Your company identifies the following requirements for the planned deployment:

- ▶ During the creation of the cluster nodes, place JAR files in the clusters.
- ▶ Decouple the Hive metastore lifetime from the cluster lifetime.
- ▶ Provide anonymous access to the cluster nodes.

You need to identify which technology must be used for each requirement.

Which technology should you identify for each requirement? To answer, drag the appropriate technologies to the correct requirements. Each technology 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.

Technologies		Answer Area
An Azure SQL Database External Metastore	Provide anonymous access to the cluster nodes:	
An Azure Table Storage External Metastore	During the creation of the cluster nodes, place the JAR files in the clusters:	
An Azure virtual network	Decouple the Hive metastore lifetime from the cluster lifetime:	
Script Actions		

Answer:

Explanation:

Technologies

Answer Area

An Azure SQL Database External Metastore	Provide anonymous access to the cluster nodes:	An Azure virtual network
An Azure Table Storage External Metastore	During the creation of the cluster nodes, place the JAR files in the clusters:	Script Actions
An Azure virtual network	Decouple the Hive metastore lifetime from the cluster lifetime:	An Azure SQL Database External Metastore
Script Actions		

NEW QUESTION 97

You have an Apache Storm cluster.

You need to ingest data from a Kafka queue.

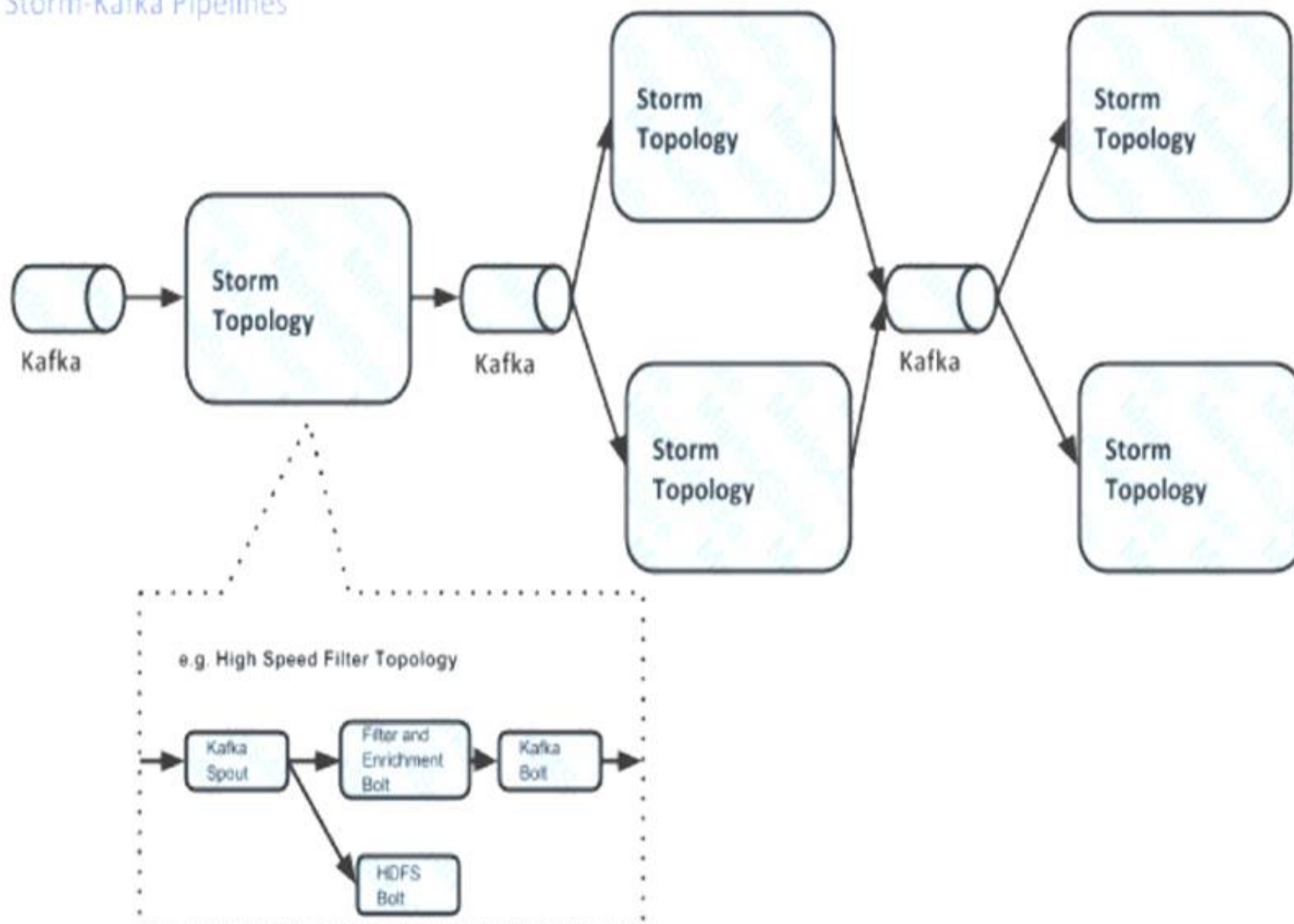
Which component should you use to consume data emitted from Kaka?

- A. Flume
- B. a bolt
- C. a spout
- D. a Microsoft Azure Service Bus queue

Answer: C

Explanation: To perform real-time computation on Storm, we create “topologies.” A topology is a graph of a computation, containing a network of nodes called “Spouts” and “Bolts.” In a Storm topology, a Spout is the source of data streams and a Bolt holds the business logic for analyzing and processing those streams. The org.apache.storm.kafka.KafkaSpout component reads data from Kafka. Example:

Storm-Kafka Pipelines



References:

<https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-apache-storm-with-kafka> <https://hortonworks.com/blog/storm-kafka-together-real-time-data-refinery/>

NEW QUESTION 98

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 section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

Your company has multiple databases that contain millions of sales transactions. You plan to implement a data mining solution to identity purchasing fraud.

You need to design a solution that mines 10 terabytes (TB) of sales data. The solution must meet the following requirements:

- Run the analysis to identify fraud once per week.
- Continue to receive new sales transactions while the analysis runs.
- Be able to stop computing services when the analysis is NOT running.

Solution: You create a Cloudera Hadoop cluster on Microsoft Azure virtual machines. Does this meet the goal?

- A. Yes
- B. No

Answer: A

Explanation: Processing large amounts of unstructured data requires serious computing power and also maintenance effort. As load on computing power typically fluctuates due to time and seasonal influences and/or processes running on certain times, a cloud solution like Microsoft Azure is a good option to be able to scale up easily and pay only for what is actually used.

NEW QUESTION 102

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 the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a Microsoft Azure deployment that contains the following services:

-  Azure Data Lake
-  Azure Cosmos DB
-  Azure Data Factory
-  Azure SQL Database

You load several types of data to Azure Data Lake.

You need to load data from Azure SQL Database to Azure Data Lake. Solution: You use the AzCopy utility.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation: Note: You can use the Copy Activity in Azure Data Factory to copy data to and from Azure Data Lake Storage Gen1 (previously known as Azure Data Lake Store). Azure SQL database is supported as source.

References: <https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-store>

NEW QUESTION 107

You are designing an Apache HBase cluster on Microsoft Azure HDInsight. You need to identify which nodes are required for the cluster.

Which three nodes should you identify? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Nimbus
- B. Zookeeper
- C. Region
- D. Supervisor
- E. Falcon
- F. Head

Answer: BCF

Explanation: <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-provision-linux-clusters>




NEW QUESTION 111

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 states goals. Some question sets might have more than one correct solution, while the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Apache Spark system that contains 5 TB of data.

You need to write queries that analyze the data in the system. The queries must meet the following requirements:

-  Use static data typing.
-  Execute queries as quickly as possible.
-  Have access to the latest language features.

Solution: You write the queries by using Python.

- A. Yes
- B. No

Answer: B

NEW QUESTION 116

You are automating the deployment of a Microsoft Azure Data Factory solution. The data factory will interact with a file stored in Azure Blob storage.

You need to use the REST API to create a linked service to interact with the file.

How should you complete the request body? To answer, drag the appropriate code elements to the correct locations. Each code may be used once, more than

once, or not at all. You may need to drag the slit bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Code Elements	Answer Area
accessKey	
AccountKey1	
accountName	
AccountName=Account2;AccountKey1	: " DefaultEndpointsProtocol=https;
AzureBatchLinkedService	
AzureStorageLinkedService	

Answer:

Explanation:

Code Elements	Answer Area
accessKey	
AccountKey1	
accountName	AzureStorageLinkedService
AccountName=Account2;AccountKey1	: " DefaultEndpointsProtocol=https; AccountName=Account2;AccountKey1
AzureBatchLinkedService	
AzureStorageLinkedService	

NEW QUESTION 121

Your company has two Microsoft Azure SQL databases named db1 and db2.

You need to move data from a table in db1 to a table in db2 by using a pipeline in Azure Data Factory. You create an Azure Data Factory named ADF1.

Which two types Of objects Should you create In ADF1 to complete the pipeline? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a linked service
- B. an Azure Service Bus
- C. sources and targets
- D. input and output I datasets
- E. transformations

Answer: AD

Explanation: You perform the following steps to create a pipeline that moves data from a source data store to a sink data store:

- ▶ Create linked services to link input and output data stores to your data factory.
- ▶ Create datasets to represent input and output data for the copy operation.
- ▶ Create a pipeline with a copy activity that takes a dataset as an input and a dataset as an output.

NEW QUESTION 126

Your Microsoft Azure subscription contains several data sources that use the same XML schema. You plan to process the data sources in parallel.

You need to recommend a compute strategy to minimize the cost of processing the data sources. What should you recommend including in the compute strategy?

- A. Microsoft SQL Server Integration Services (SSIS) on an Azure virtual machine
- B. Azure Batch
- C. a Linux HPC cluster in Azure
- D. a Windows HPC cluster in Azure

Answer: A

NEW QUESTION 130

Your company has a data visualization solution that contains a customized Microsoft Azure Stream Analytics solution. The solution provides data to a Microsoft Power BI deployment.

Every 10 seconds, you need to query for instances that have more than three records.

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

Values

[Count] >= 3

[Instance] >= 3

[Records] >= 3

SlidingWindow(second,10)

System.Timestamp(10)

TumblingWindow(second,10)

●●●●●

●
●
●
●

Answer Area

```

SELECT
    Value
    System.Timestamp AS Time,
    COUNT(*) AS [Count]
INTO
    AlertOutput
FROM
    Input TIMESTAMP BY Time
GROUP BY
    Make,
    [Value]
HAVING
    [Count] >= 3
        
```

Answer:

Explanation: Box 1: TumblingWindow(second, 10)

Tumbling Windows define a repeating, non-overlapping window of time. Example: Calculate the count of sensor readings per device every 10 seconds SELECT sensorId, COUNT(*) AS Count

FROM SensorReadings TIMESTAMP BY time GROUP BY sensorId, TumblingWindow(second, 10) Box 2: [Count] >= 3

Count(*) returns the number of items in a group.

NEW QUESTION 134

Your company has 2000 servers.

You plan to aggregate all of the log files from the servers in a central repository that uses Microsoft Azure HDInsight. Each log file contains approximately one million records. All of the files use the .log file name extension.

The following is a sample of the entries in the log files.

20:26:41 SampleClass3 (ERROR) verbose detail for id 1527353937

In Apache Hive, you need to create a data definition and a query capturing tire number of records that have an error level of [ERROR].

What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

```
CREATE EXTERNAL TABLE log4jLogs (t1 string, t2 string,  
t3 string, t4 string, t5 string, t6 string, t7 string)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY /T  
STORED AS TEXTFILE LOCATION 'wasbs:///example/data/';  
SELECT t4 AS sev, COUNT(*) AS count  
FROM log4jLogs  
WHERE t4 = ' [ERROR] '  
AND INPUT_FILE_NAME LIKE '!.log'  
  
GROUP BY
```

Answer:

Explanation: Box 1: table

Box 2: /t

Apache Hive example:

CREATE TABLE raw (line STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n';

Box 3: count(*)

Box 4: '*.log'

NEW QUESTION 135

You plan to deploy a storage solution to store the output of stream analytics. You plan to store the data for the following three types of data streams:

- ☒ Unstructured JSON data
- ☒ Exploratory analytics
- ☒ Pictures

You need to implement a storage solution for the data stream types.

Which storage solution should you implement for each data stream type? To answer, drag the appropriate storage solutions to the correct data stream types. Each storage solution may be used once, more than once, or not at all. You may need to drag the split bar between the panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Storage Solutions	Answer Area
Azure Data Lake	Exploratory analytics
Azure Blob Storage	Unstructured JSON data
Azure Table Storage	Pictures
Azure Service Bus Queue	
Azure Cosmos DB	

Answer:

Explanation: Box 1: Azure Data Lake Store

Stream Analytics supports Azure Data Lake Store. Azure Data Lake Store is an enterprise-wide hyper-scale repository for big data analytic workloads. Data Lake Store enables you to store data of any size, type and ingestion speed for operational and exploratory analytics. Stream Analytics has to be authorized to access the Data Lake Store.

Box 2: Azure Cosmos DB

Stream Analytics can target Azure Cosmos DB for JSON output, enabling data archiving and low-latency queries on unstructured JSON data.

Box 3: Azure Blob Storage

Blob storage offers a cost-effective and scalable solution for storing large amounts of unstructured data in the cloud.

Incorrect Asnwers: Azure SQL Database:

Azure SQL Database can be used as an output for data that is relational in nature or for applications that depend on content being hosted in a relational database. Stream Analytics jobs write to an existing table in an Azure SQL Database.

Azure Service Bus Queue:

Service Bus Queues offer a First In, First Out (FIFO) message delivery to one or more competing consumers. Typically, messages are expected to be received and processed by the receivers in the temporal order in which they were added to the queue, and each message is received and processed by only one message consumer.

Azure Table Storage

Azure Table storage offers highly available, massively scalable storage, so that an application can automatically scale to meet user demand. Table storage is Microsoft's NoSQL key/attribute store, which one can leverage for structured data with fewer constraints on the schema. Azure Table storage can be used to store data for persistence and efficient retrieval.

References: <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs>

NEW QUESTION 136

You are planning a solution that will have multiple data files stored in Microsoft Azure Blob storage every hour. Data processing will occur once a day at midnight only.

You create an Azure data factory that has blob storage as the input source and an Azure HD Insight activity that uses the input to create an output Hive table.

You need to identify a data slicing strategy for the data factory.

What should you identify? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

The processing frequency:

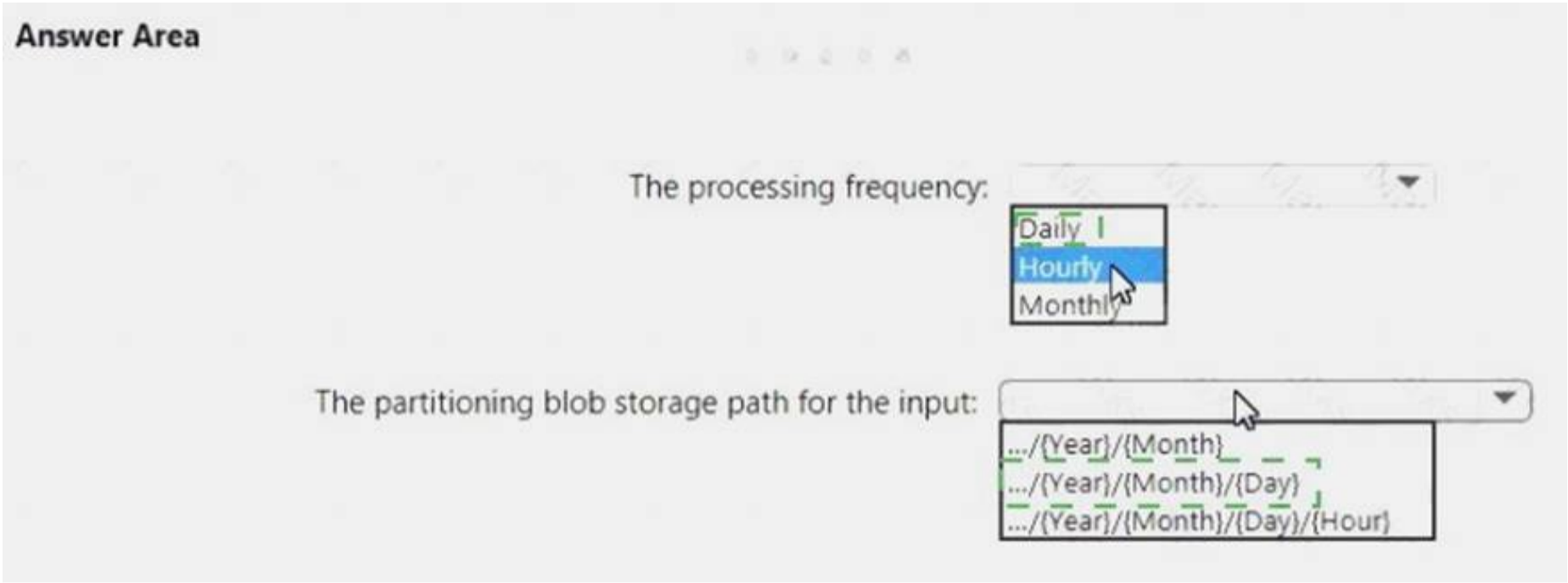
☐ Daily
☒ Hourly
☐ Monthly

The partitioning blob storage path for the input:

☐ .../{Year}/{Month}
☒ .../{Year}/{Month}/{Day}
☐ .../{Year}/{Month}/{Day}/{Hour}

Answer:

Explanation:



NEW QUESTION 139

You have raw data in Microsoft Azure Blob storage. Each data file is 10 KB and is the XML format. You identify the following requirements for the data:

- ▶ The data must be converted into a flat data structure by using a C# MapReduce job.
- ▶ The data must be moved to an Azure SQL database, which will then be used to visualize the data.
- ▶ Additional stored procedures must run against the data once the data is in the database.

You need to create the workflow for the Azure Data Factory pipeline.

Which activity type should you use for each requirement? To answer, drag the appropriate workflow components to the correct requirements. Each workflow component may be used once, more than once, or not at all. You may need to drag the split bar between the panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Workflow Components

Copy

HDInsightHive

HDInsightMapReduce

HDInsightStreaming

SQLServerStoredProcedure

Answer Area

The data must be converted into a flat data strucutre by using a C# MapReduce job:

The data must be moved to an Azure SQL database, which will then be used to visualize the data:

Additioanal stored procedures must run against the data once the data is in the database:

Workflow Component

Workflow Component

Workflow Component

Answer:

Explanation: Box 1: HDInsightMapReduce

The HDInsight MapReduce activity in a Data Factory pipeline invokes MapReduce program on your own or on-demand HDInsight cluster.

Box 2: HDInsightStreaming

Box 3: SQLServerStoredProcedure

NEW QUESTION 142

You have the following script.

```
CREATE TABLE UserVisits (username string, url string, time date) STORED AS TEXTFILE LOCATION "wasb:///Logs";
CREATE TABLE UserVisitsOrc (username string, url string, time date) STORED AS ORC;
INSERT INTO TABLE UserVisitsOrc SELECT * FROM UserVisits
```

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the script.

NOTE: Each correct selection is worth one point.

Answer Area

The INSERT statement [answer choice].

moves the contents of UserVisits to the UserVisitsOrc table directory
inserts data into the UserVisitsOrc table by running a YARN application
inserts data into the UserVisitsOrc table record by record from the Hive command-line interface (CLI)

The UserVisits table type is [answer choice].

dataset
external
managed

Answer:

Explanation: A table created without the EXTERNAL clause is called a managed table because Hive manages its data.

NEW QUESTION 145

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 the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a Microsoft Azure deployment that contains the following services:

- ☒ Azure Data Lake
- ☒ Azure Cosmos DB
- ☒ Azure Data Factory
- ☒ Azure SQL Database

You load several types of data to Azure Data Lake.

You need to load data from Azure SQL Database to Azure Data Lake. Solution: You use the Azure Import/Export service.

Does this meet the goal?

- A. Yes
- B. No

Answer: A

NEW QUESTION 148

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 section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

Your company has multiple databases that contain millions of sales transactions. You plan to implement a data mining solution to identify purchasing fraud.

You need to design a solution that mines 10 terabytes (TB) of sales data. The solution must meet the following requirements:

- ☒ Run the analysis to identify fraud once per week.
- ☒ Continue to receive new sales transactions while the analysis runs.
- ☒ Be able to stop computing services when the analysis is NOT running. Solution: You create a Microsoft Azure Data Lake job.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

NEW QUESTION 150

You have structured data that resides in Microsoft Azure Blob Storage.

You need to perform a rapid interactive analysis of the data and to generate visualizations of the data.

What is the best type of Azure HDInsight cluster to use to achieve the goal? More than one answer choice may achieve the goal. Select the BEST answer.

- A. Apache Storm
- B. Apache HBase
- C. Apache Hadoop
- D. Apache Spark

Answer: D

Explanation:

Reference: <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-provision-linux-clusters>

NEW QUESTION 151

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 states goals. Some question sets might have more than one correct solution, while the others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to implement a new data warehouse.

You have the following information regarding the data warehouse:

- ▶ The first data files for the data warehouse will be available in a few days.
- ▶ Most queries that will be executed against the data warehouse are ad-hoc.
- ▶ The schemas of data files that will be loaded to the data warehouse change often.
- ▶ One month after the planned implementation, the data warehouse will contain 15 TB of data. You need to recommend a database solution to support the planned implementation.

Solution: You recommend an Apache Spark system. Does this meet the goal?

- A. Yes
- B. No

Answer: B

NEW QUESTION 153

You have a Microsoft Azure HDInsight cluster for analytics workloads. You have a C# application on a local computer.

You plan to use Azure Data Factory to run the C# application in Azure.

You need to create a data factory that runs the C# application by using HDInsight.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Actions

- Derive the C# class from the IDotNetActivity interface.
- Zip the build files and upload the ZIP file to an Azure storage account.
- Implement the Execute method of the IDotNetActivity interface.
- Build the C# application in Microsoft Visual Studio.
- Create a pipeline that has a DotNetActivity activity and specify the path to the build files in the Azure storage account.

Answer Area

➤

➤

⬅

⬅

⬆

⬆

⬇

⬇

Answer:

Explanation:

Actions

- Derive the C# class from the IDotNetActivity interface.
- Zip the build files and upload the ZIP file to an Azure storage account.
- Implement the Execute method of the IDotNetActivity interface.
- Build the C# application in Microsoft Visual Studio.
- Create a pipeline that has a DotNetActivity activity and specify the path to the build files in the Azure storage account.

Answer Area

Build the C# application in Microsoft Visual Studio.

Create a pipeline that has a DotNetActivity activity and specify the path to the build files in the Azure storage account.

Derive the C# class from the IDotNetActivity interface.

Implement the Execute method of the IDotNetActivity interface.

Zip the build files and upload the ZIP file to an Azure storage account.

NEW QUESTION 155

You are using a Microsoft Azure Data Factory pipeline to copy data to an Azure SQL database. You need to prevent the insertion of duplicate data for a given dataset slice.

Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Set the External property to true.
- B. Add a column named SliceIdentifierColumnName to the output dataset.
- C. Set the SqlWriterCleanupScript property to true.
- D. Remove the duplicates in post-processing.
- E. Manually delete the duplicate data before running the pipeline activity.

Answer: BC

The Leader of IT Certification

visit - <https://www.certleader.com>

NEW QUESTION 156

You have an Apache Storm cluster.

The cluster will ingest data from a Microsoft Azure event hub.

The event hub has the characteristics described in the following table.

Setting name	Value
Message Retention	1
Namespace	storm1.servicebus.windows.net
Shared access policies	2
Partition Count	16
Region	Central US

You are designing the Storm application topology.

You need to ingest data from all of the partitions. The solution must maximize the throughput of the data ingestion.

Which setting should you use?

- A. Partition Count
- B. Message Retention
- C. Partition Key
- D. Shared access policies

Answer: A

NEW QUESTION 157

You have a pipeline that contains an input dataset in Microsoft Azure Table Storage and an output dataset in Azure Blob storage. You have the following JSON data.

```
availability: { frequency: "Day", interval: 3,
  "anchorDateTime": "2014-10-10T10:00:00Z"
  waitOnExternal: { retryInterval: "00:01:00", retryTimeout: "00:10:00", maximumRetry: 3
  } }
```

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the JSON data.

NOTE: Each correct selection is worth one point.

Answer Area

The pipeline will run [answer choice].

▼

every three days at 10:00
every three days at 22:00
three times a day starting at 10:00
three times a day starting at 22:00

If the pipeline fails to run, it will retry [answer choice].

▼

every minute up to three times
every 10 minutes up to three times
every 10 minutes until the pipeline succeeds
every one minute until the pipeline succeeds

Answer:

Explanation: Box 1: Every three days at 10.00

anchorDateTime defines the absolute position in time used by the scheduler to compute dataset slice boundaries.

"frequency": "<Specifies the time unit for data slice production. Supported frequency: Minute, Hour, Day, Week, Month>",

"interval": "<Specifies the interval within the defined frequency. For example, frequency set to 'Hour' and interval set to 1 indicates that new data slices should be produced hourly>

Box 2: Every minute up to three times.

retryInterval is the wait time between a failure and the next attempt. This setting applies to present time. If the previous try failed, the next try is after the retryInterval period.

Example: 00:01:00 (1 minute)

Example: If it is 1:00 PM right now, we begin the first try. If the duration to complete the first validation check is 1 minute and the operation failed, the next retry is at 1:00 + 1min (duration) + 1min (retry interval) = 1:02 PM.

For slices in the past, there is no delay. The retry happens immediately. retryTimeout is the timeout for each retry attempt.

maximumRetry is the number of times to check for the availability of the external data.

NEW QUESTION 159

You are developing an Apache Storm application by using Microsoft Visual Studio. You need to implement a custom topology that uses a custom bolt. Which type of object should you initialize in the main class?

- A. Stream
- B. TopologyBuilder
- C. StreamInfo

D. Logger

Answer: A

NEW QUESTION 161

You plan to use Microsoft Azure IoT Hub to capture data from medical devices that contain sensors. You need to ensure that each device has its own credentials. The solution must minimize the number of required privileges. Which policy should you apply to the devices?

- A. iothubowner
- B. service
- C. registryReadWrite
- D. device

Answer: D

Explanation: Per-Device Security Credentials. Each IoT Hub contains an identity registry. For each device in this identity registry, you can configure security credentials that grant DeviceConnect permissions scoped to the corresponding device endpoints.

NEW QUESTION 166

You have an analytics solution in Microsoft Azure that must be operationalized. You have the relevant data in Azure Blob storage. You use an Azure HDInsight Cluster to process the data. You plan to process the raw data files by using Azure HDInsight. Azure Data Factory will operationalize the solution. You need to create a data factory to orchestrate the data movement. Output data must be written back to Azure Blob storage. Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions		Answer Area
Create input and output datasets for the files in Azure.	1	
Rename the data factory hub.	2	
Create a data factory.	3	
Create a client to connect to Azure Blob storage.	4	
Create linked services for Azure Blob storage and the Azure HDInsight cluster.		
Create an Azure HDInsight activity in a pipeline to process the data.		

Answer:

Explanation:

Actions	Answer Area
Create input and output datasets for the files in Azure.	1 Create a data factory.
Rename the data factory hub.	2 Create linked services for Azure Blob storage and the Azure HDInsight cluster.
Create a data factory.	3 Create input and output datasets for the files in Azure.
Create a client to connect to Azure Blob storage.	4 Create an Azure HDInsight activity in a pipeline to process the data.
Create linked services for Azure Blob storage and the Azure HDInsight cluster.	
Create an Azure HDInsight activity in a pipeline to process the data.	

NEW QUESTION 167

You have a financial model deployed to an application named finance1. The data from the financial model is stored in several data files. You need to implement a batch processing architecture for the financial model. You upload the data files and finance1 to a Microsoft Azure Storage account. Which three components should you create in sequence next? To answer, move the appropriate components from the list of components to the answer area and arrange them in the correct order.

Components	Answer Area
a pipeline	
a linked service	
a task	
a batch pool of compute nodes	
a batch job	

Answer:

Explanation:

Components	Answer Area
a pipeline	a batch pool of compute nodes
a linked service	a linked service
a task	a pipeline
a batch pool of compute nodes	
a batch job	

NEW QUESTION 172

You need to create a new Microsoft Azure data factory by using Azure PowerShell. The data factory will have a pipeline that copies data to and from Azure Storage. Which four cmdlets should you use in sequence? To answer, move the appropriate cmdlets from the list of cmdlets to the answer area and arrange them in the correct order.

Cmdlets	Answer Area
New-AzureRmDataFactoryLinkedService	
New-AzureDataFactoryRmDataFactoryHub	
New-AzureRmDataFactory	
New-AzureRmDataFactoryDataset	
New-AzureDataFactoryRmDataFactoryGateway	
New-AzureRmDataFactoryPipeline	

Answer:

Explanation: Perform these operations in the following order:

- ▶ Create a data factory.
- ▶ Create linked services.
- ▶ Create datasets.
- ▶ Create a pipeline.

Step 1: New-AzureRmDataFactory Create a data factory

The New-AzureRmDataFactory cmdlet creates a data factory with the specified resource group name and location.

Step 2: New-AzureRmDataFactoryLinkedService

Create linked services in a data factory to link your data stores and compute services to the data factory. The New-AzureRmDataFactoryLinkedService cmdlet links a data store or a cloud service to Azure Data Factory.

Step 3: New-AzureRmDataFactoryDataset

You define a dataset that represents the data to copy from a source to a sink. It refers to the Azure Storage linked service you created in the previous step.

The New-AzureRmDataFactoryDataset cmdlet creates a dataset in Azure Data Factory.

Step 4: New-AzureRmDataFactoryPipeline You create a pipeline.

The New-AzureRmDataFactoryPipeline cmdlet creates a pipeline in Azure Data Factory. References:

<https://docs.microsoft.com/en-us/azure/data-factory/quickstart-create-data-factory-powershell> <https://docs.microsoft.com/en-us/powershell/module/azurerm.datafactories/new-azurermdatafactory>

NEW QUESTION 173

You work for a telecommunications company that uses Microsoft Azure Stream Analytics. You have data related to incoming calls.

You need to group the data in the following ways:

- ▶ Group A: Every five minutes for a duration of five minutes
- ▶ Group B: Every five minutes for a duration of 10 minutes

Which type of window should you use for each group? To answer, drag the appropriate window types to the correct groups. Each window type 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.

Window Types	Answer Area
Hopping	Group A: Window Type
Sliding	Group B: Window Type
Tumbling	

Answer:

Explanation: Group A: Tumbling

Tumbling Windows define a repeating, non-overlapping window of time. Group B: Hopping

Like Tumbling Windows, Hopping Windows move forward in time by a fixed period but they can overlap with one another.

NEW QUESTION 178

You have a Microsoft Azure data factory named ADF1 that contains a pipeline named Pipeline1. You plan to automate updates to Pipeline1.

You need to build the URL that must be called to update the pipeline from the REST API.

How should you complete the URL? To answer, drag the appropriate URL elements to the correct locations. Each URL element 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.

URL Elements	Answer Area
<input type="text" value="datapipelines/adf1"/>	https:// <input type="text" value="URL element"/> /subscriptions/
<input type="text" value="datapipelines/pipeline1"/>	12300000-0000-0000-0000-000000000212/resourcegroups/adf/providers/
<input type="text" value="management.azure.com"/>	<input type="text" value="URL element"/> /
<input type="text" value="Microsoft.DataFactory/datafactories/adf1"/>	<input type="text" value="URL element"/> ?api-version=2015-02-28
<input type="text" value="Microsoft.DataFactory/datafactories/pipeline1"/>	

Answer:

Explanation:

URL Elements	Answer Area
<input type="text" value="datapipelines/adf1"/>	https:// <input type="text" value="Microsoft.DataFactory/datafactories/adf1"/> /subscriptions/
<input type="text" value="datapipelines/pipeline1"/>	12300000-0000-0000-0000-000000000212/resourcegroups/adf/providers/
<input type="text" value="management.azure.com"/>	<input type="text" value="datapipelines/pipeline1"/> /
<input type="text" value="Microsoft.DataFactory/datafactories/adf1"/>	<input type="text" value="management.azure.com"/> ?api-version=2015-02-28
<input type="text" value="Microsoft.DataFactory/datafactories/pipeline1"/>	

NEW QUESTION 179

You have a web application that generates several terabytes (TB) of financial documents each day. The application processes the documents in batches.

You need to store the documents in Microsoft Azure. The solution must ensure that a user can restore the previous version of a document.

Which type of storage should you use for the documents?

- A. Azure Cosmos DB
- B. Azure File Storage
- C. Azure Data Lake
- D. Azure Blob storage

Answer: A

NEW QUESTION 183

You have data in an on-premises Microsoft SQL Server database.

You must ingest the data in Microsoft Azure Blob storage from the on-premises SQL Server database by using Azure Data Factory.

You need to identify which tasks must be performed from Azure.

In which sequence should you perform the actions? To answer, move all of the actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Answer:

Explanation: Step 1: Configure a Microsoft Data Management Gateway Install and configure Azure Data Factory Integration Runtime. The Integration Runtime is a customer managed data integration infrastructure used by Azure Data Factory to provide data integration capabilities across different network environments. This runtime was formerly called "Data Management Gateway".
Step 2: Create a linked service for Azure Blob storage
Create an Azure Storage linked service (destination/sink). You link your Azure storage account to the data factory.
Step 3: Create a linked service for SQL Server
Create and encrypt a SQL Server linked service (source)
In this step, you link your on-premises SQL Server instance to the data factory. Step 4: Create an input dataset and an output dataset.
Create a dataset for the source SQL Server database. In this step, you create input and output datasets. They represent input and output data for the copy operation, which copies data from the on-premises SQL Server database to Azure Blob storage.
Step 5: Create a pipeline..
You create a pipeline with a copy activity. The copy activity uses SqlServerDataset as the input dataset and AzureBlobDataset as the output dataset. The source type is set to SqlSource and the sink type is set to BlobSink.
References: <https://docs.microsoft.com/en-us/azure/data-factory/tutorial-hybrid-copy-powershell>

NEW QUESTION 185

You have a web app that accepts user input, and then uses a Microsoft Azure Machine Learning model to predict a characteristic of the user. You need to perform the following operations:

- ▶ Track the number of web app users from month to month.
- ▶ Track the number of successful predictions made during the last minute.
- ▶ Create a dashboard showcasing the analytics for the predictions and the web app usage.

Which lambda layer should you query for each operation? To answer, drag the appropriate layers to the correct operations. Each layer 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.

Answer:

Explanation: Lambda architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch- and stream-processing methods. This approach to architecture attempts to balance latency, throughput, and fault-tolerance by using batch processing to provide comprehensive and accurate views of batch data, while simultaneously using real-time stream processing to provide views of online data. The two view outputs may be joined before presentation
Box 1: Speed
The speed layer processes data streams in real time and without the requirements of fix-ups or completeness. This layer sacrifices throughput as it aims to minimize latency by providing real-time views into the most recent data.

Box 2: Batch

The batch layer precomputes results using a distributed processing system that can handle very large quantities of data. The batch layer aims at perfect accuracy by being able to process all available data when generating views.

Box 3: Serving

Output from the batch and speed layers are stored in the serving layer, which responds to ad-hoc queries by returning precomputed views or building views from the processed data.

NEW QUESTION 188

You have four on-premises Microsoft SQL Server data sources as described in the following table.

Data source name	Server name
DS1	SQL1
DS2	SQL2
DS3	SQL3
DS4	SQL4

You plan to create three Azure data factories that will interact with the data sources as described in the following table.

Data factory name	Data source used
ADF1	DS1 and DS2
ADF2	DS3
ADF3	DS4

You need to deploy Microsoft Data Management Gateway to support the Azure Data Factory deployment. The solution must use new servers to host the instances of Data Management Gateway.

What is the minimum number of new servers and data management gateways you should you deploy? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Number of servers:

▼

1
2
3
4

Number of Data Management Gateway instances:

▼

1
2
3
4

Answer:

Explanation: Box 1: 3

Box 2: 3

Considerations for using gateway

NEW QUESTION 192

You need to recommend a permanent Azure Storage solution for the activity data. The solution must meet the technical requirements.

What is the best recommendation to achieve the goal? More than one answer choice may achieve the goal. Select the BEST answer.

- A. Azure SQL Database
- B. Azure Queue storage
- C. Azure Blob storage
- D. Azure Event Hubs

Answer: A

NEW QUESTION 196

.....

Thank You for Trying Our Product

* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

* One year free update

You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

We currently serve more than 30,000,000 customers.

* Shop Securely

All transactions are protected by VeriSign!

100% Pass Your 70-475 Exam with Our Prep Materials Via below:

<https://www.certleader.com/70-475-dumps.html>