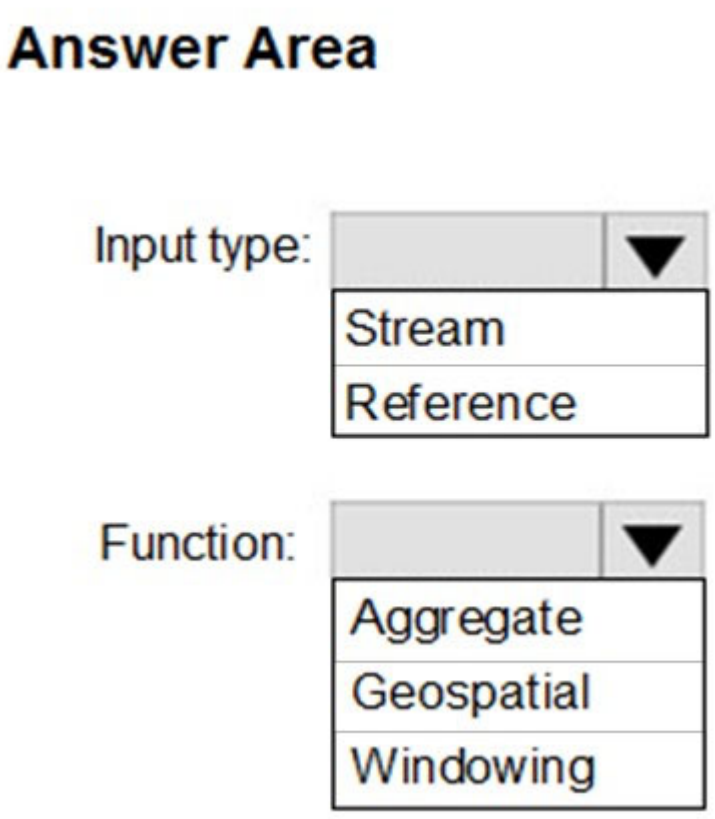
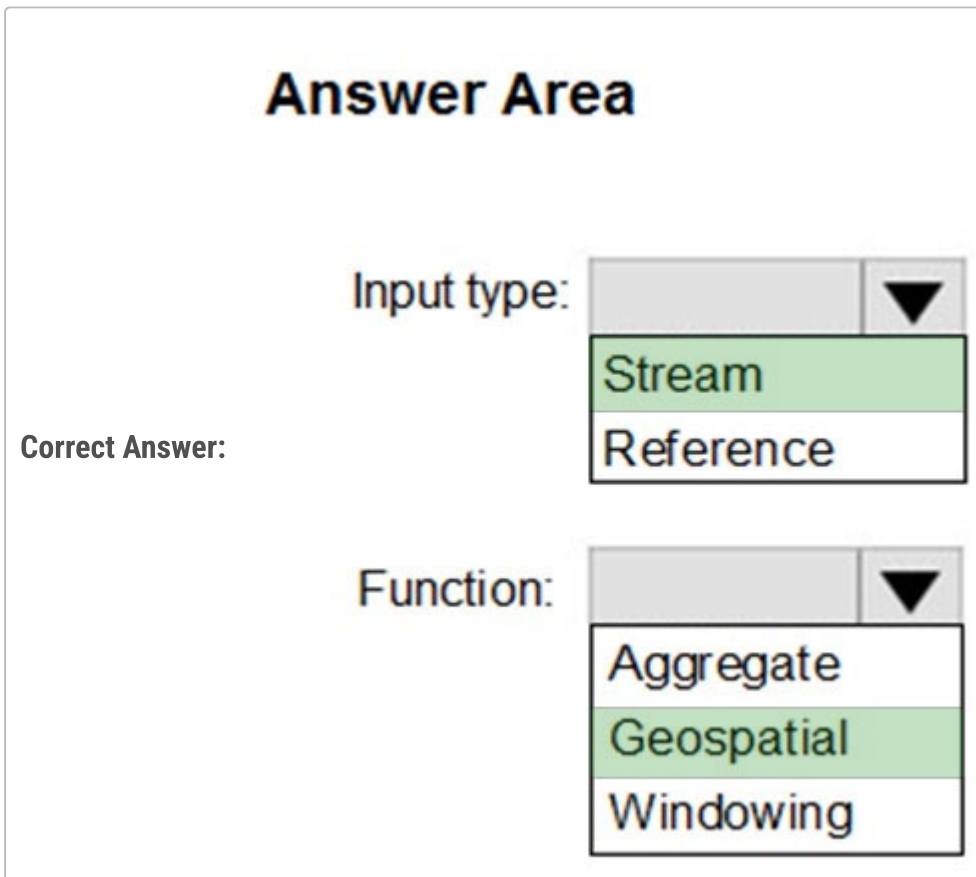
Topic 2 - Question Set 2

Question #1 *Topic 2*

HOTSPOT -
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.
You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.
What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:



Correct Answer:



Input type: Stream -
You can process real-time IoT data streams with Azure Stream Analytics.

Function: Geospatial -
With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride sharing, connected cars, and asset tracking.
Note: In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.

Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios

Question #2 *Topic 2*

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream
Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Implement event ordering.

B. Implement Azure Stream Analytics user-defined functions (UDF).

C. Implement query parallelization by partitioning the data output. Most Voted

D. Scale the SU count for the job up. Most Voted

E. Scale the SU count for the job down.

F. Implement query parallelization by partitioning the data input. Most Voted Most Voted

Correct Answer: *DF*
D: Scale out the query by allowing the system to process each input partition separately.
F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

Community vote distribution
CF (45%) DF (45%) 6%

Question #3 *Topic 2*

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.
Which resource provider should you enable?

A. Microsoft.Sql

B. Microsoft.Automation

C. Microsoft.EventGrid Most Voted

D. Microsoft.EventHub

Correct Answer: *C*
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events.
Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure
Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

Community vote distribution
C (100%)

Question #4 *Topic 2*

You plan to perform batch processing in Azure Databricks once daily.

Which type of Databricks cluster should you use?

A. High Concurrency

B. automated Most Voted

C. interactive

Correct Answer: *B*

Automated Databricks clusters are the best for jobs and automated batch processing.

Note: Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.

Example: Scheduled batch workloads (data engineers running ETL jobs)

This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform.

The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.

Reference:

https://docs.microsoft.com/en-us/azure/databricks/clusters/create https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs

Community vote distribution

B (100%)

Question #5 *Topic 2*

HOTSPOT -

You are processing streaming data from vehicles that pass through a toll booth.

You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [ ▼ ] (Time) AS LastEventTime
                  COUNT
                  MAX
                  MIN
                  TOPONE
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [ ▼ ] (minute, 10)
                  HoppingWindow
                  SessionWindow
                  SlidingWindow
                  TumblingWindow
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [ ▼ ] (minute, Input, LastInWindow) BETWEEN 0 AND 10
                  DATEADD
                  DATEDIFF
                  DATENAME
                  DATEPART
AND Input.Time = LastInWindow.LastEventTime
```

Correct Answer:

Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [COUNT | MAX | MIN | TOPONE] (Time) AS LastEventTime
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [HoppingWindow | SessionWindow | SlidingWindow | TumblingWindow] (minute, 10)
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [DATEADD | DATEDIFF | DATENAME | DATEPART] (minute, Input, LastInWindow) BETWEEN 0 AND 10
AND Input.Time = LastInWindow.LastEventTime
```

Box 1: MAX -

The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window.

Query:

```
WITH LastInWindow AS -
(

SELECT -

MAX(Time) AS LastEventTime -

FROM -

Input TIMESTAMP BY Time -

GROUP BY -
TumblingWindow(minute, 10)
)

SELECT -
Input.License_plate,
Input.Make,
```

Input.Time -

FROM -

Input TIMESTAMP BY Time -

INNER JOIN LastInWindow -
ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10
AND Input.Time = LastInWindow.LastEventTime

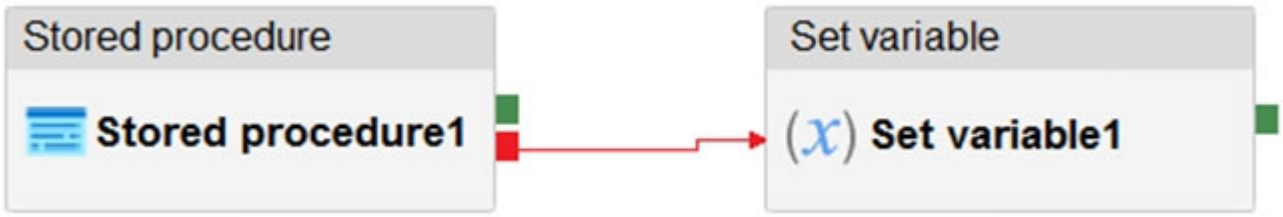
Box 2: TumblingWindow -
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Box 3: DATEDIFF -
DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

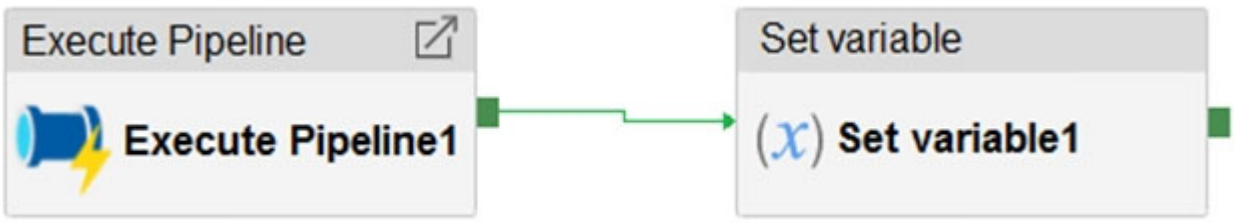
Question #6 *Topic 2*

You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.

Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.

What is the status of the pipeline runs?

A. Pipeline1 and Pipeline2 succeeded. Most Voted

B. Pipeline1 and Pipeline2 failed.

C. Pipeline1 succeeded and Pipeline2 failed.

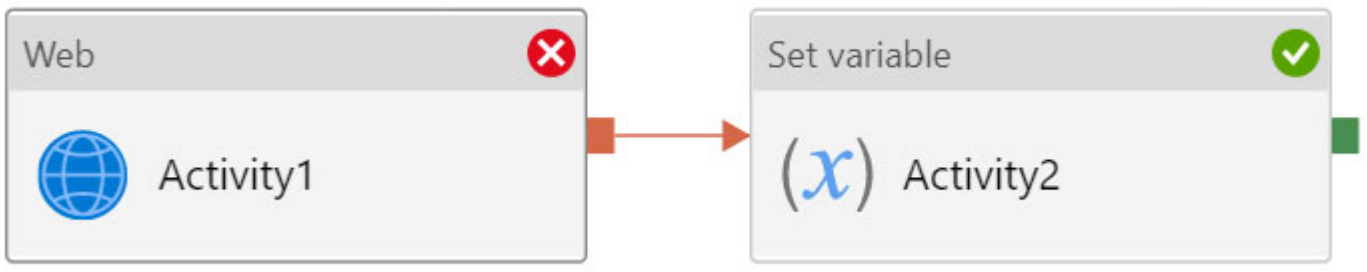
D. Pipeline1 failed and Pipeline2 succeeded.

Correct Answer: *A*

Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed.

Consider Pipeline1:

If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.



The failure dependency means this pipeline reports success.

Note:

If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.

Reference:

https://datasavvy.me/category/azure-data-factory/

Community vote distribution

A (89%) 11%

Question #7 *Topic 2*

HOTSPOT -

A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:

Ingest:

- Access multiple data sources.
- Provide the ability to orchestrate workflow.
- Provide the capability to run SQL Server Integration Services packages.

Store:

- Optimize storage for big data workloads.
- Provide encryption of data at rest.
- Operate with no size limits.

Prepare and Train:

- Provide a fully-managed and interactive workspace for exploration and visualization.
- Provide the ability to program in R, SQL, Python, Scala, and Java.

Provide seamless user authentication with Azure Active Directory.

▪

Model & Serve:

- Implement native columnar storage.
- Support for the SQL language
- Provide support for structured streaming.

You need to build the data integration pipeline.

Which technologies should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Architecture requirement	Technology
Ingest	Logic Apps / Azure Data Factory / Azure Automation
Store	Azure Data Lake Storage / Azure Blob storage / Azure files
Prepare and Train	HDInsight Apache Spark cluster / Azure Databricks / HDInsight Apache Storm cluster
Model and Serve	HDInsight Apache Kafka cluster / Azure Synapse Analytics / Azure Data Lake Storage

Correct Answer:

Answer Area

Architecture requirement	Technology
Ingest	Logic Apps / **Azure Data Factory** / Azure Automation
Store	**Azure Data Lake Storage** / Azure Blob storage / Azure files
Prepare and Train	HDInsight Apache Spark cluster / **Azure Databricks** / HDInsight Apache Storm cluster
Model and Serve	HDInsight Apache Kafka cluster / **Azure Synapse Analytics** / Azure Data Lake Storage

Ingest: Azure Data Factory -
Azure Data Factory pipelines can execute SSIS packages.
In Azure, the following services and tools will meet the core requirements for pipeline orchestration, control flow, and data movement: Azure Data Factory, Oozie on HDInsight, and SQL Server Integration Services (SSIS).

Store: Data Lake Storage -
Data Lake Storage Gen1 provides unlimited storage.
Note: Data at rest includes information that resides in persistent storage on physical media, in any digital format. Microsoft Azure offers a variety of data storage solutions to meet different needs, including file, disk, blob, and table storage. Microsoft also provides encryption to protect Azure SQL Database, Azure Cosmos
DB, and Azure Data Lake.
Prepare and Train: Azure Databricks
Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.
With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace.
Azure Databricks supports Python, Scala, R, Java and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch and scikit-learn.
Model and Serve: Azure Synapse Analytics
Azure Synapse Analytics/ SQL Data Warehouse stores data into relational tables with columnar storage.
Azure SQL Data Warehouse connector now offers efficient and scalable structured streaming write support for SQL Data Warehouse. Access SQL Data
Warehouse from Azure Databricks using the SQL Data Warehouse connector.
Note: As of November 2019, Azure SQL Data Warehouse is now Azure Synapse Analytics.
Reference:
https://docs.microsoft.com/bs-latn-ba/azure/architecture/data-guide/technology-choices/pipeline-orchestration-data-movement
https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks

Question #8 *Topic 2*

DRAG DROP -

You have the following table named Employees.

first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

You need to calculate the employee_type value based on the hire_date value.

How should you complete the Transact-SQL statement? 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.

Select and Place:

Values

CASE
ELSE
OVER
PARTITION BY
ROW_NUMBER

Answer Area

```
SELECT
    *,
    [          ]
      WHEN hire_date >= '2019-01-01' THEN 'New'
      [          ] 'Standard'
    END AS employee_type
FROM
    employees
```

Correct Answer:

Values

CASE
ELSE
OVER
PARTITION BY
ROW_NUMBER

Answer Area

```
SELECT
    *,
    CASE
      WHEN hire_date >= '2019-01-01' THEN 'New'
      ELSE 'Standard'
    END AS employee_type
FROM
    employees
```

Box 1: CASE -

CASE evaluates a list of conditions and returns one of multiple possible result expressions.

CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE,

DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.

Syntax: Simple CASE expression:

CASE input_expression -

WHEN when_expression THEN result_expression [...n]

[ELSE else_result_expression]

END -

Box 2: ELSE -

Reference:

https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

Question #9 *Topic 2*

DRAG DROP -

You have an Azure Synapse Analytics workspace named WS1.

You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.

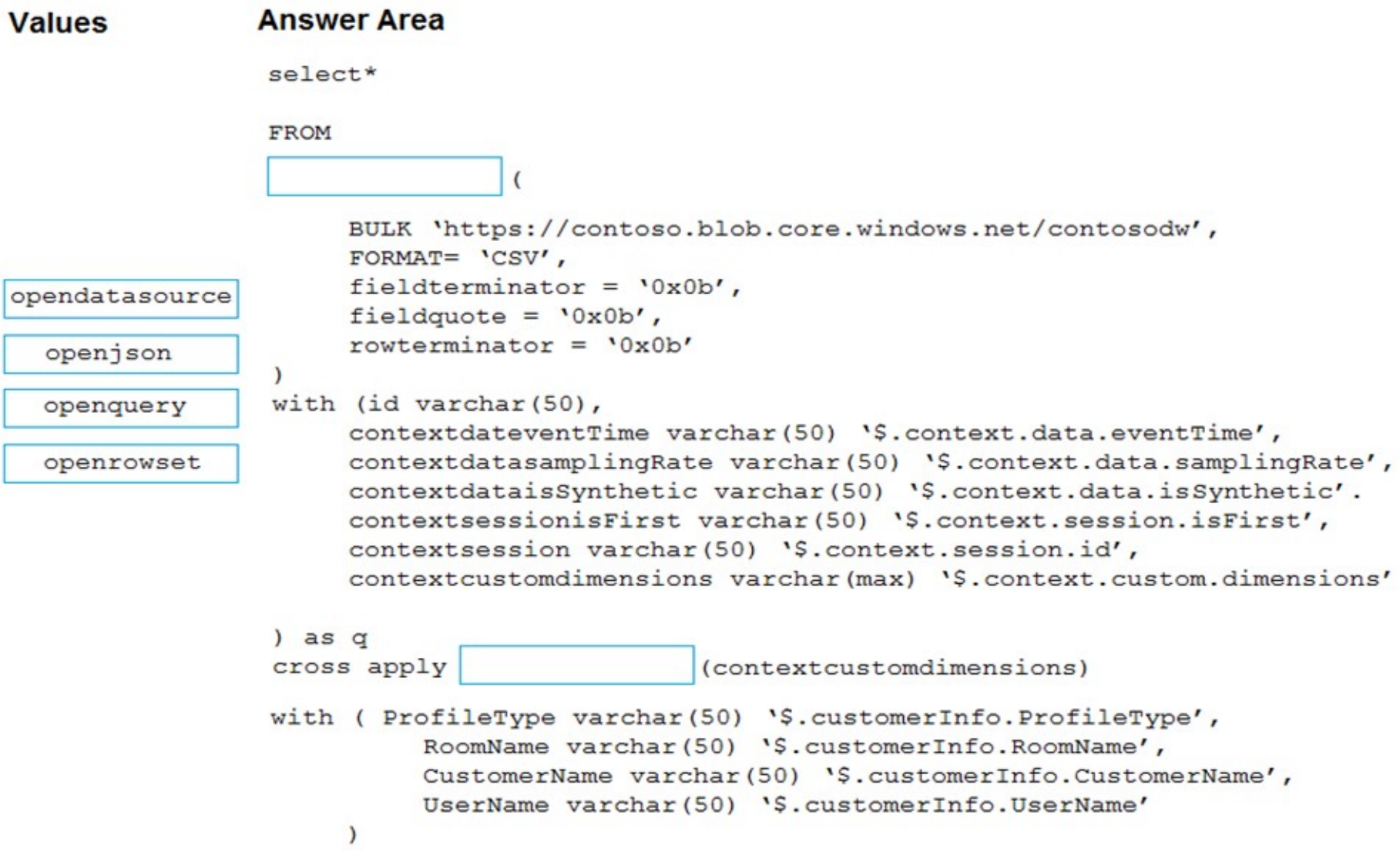
```
{
     "id": "66532691-ab20-11ea-8b1d-936b3ec64e54",
     "context": {
          "data": {
               "eventTime": "2020-06-10T13:43:34.553Z",
               "samplingRate": "100.0",
               "isSynthetic": "false"
          },
          "session": {
               "isFirst": "false",
               "id": "38619c14-7a23-4687-8268-95862c5326b1"
          },
          "custom": {
               "dimensions": [
                    {
                         "customerInfo": {
                              "ProfileType": "ExpertUser",
                              "RoomName": "",
                              "CustomerName": "diamond",
                              "UserName": "XXXX@yahoo.com"
                    }
               },
               {
                         "customerInfo" {
                              "ProfileType": "Novice",
                              "RoomName": "",
                              "CustomerName": "topaz",
                              "UserName": "XXXX@outlook.com"
                         }
                    }
               ]
          }
     }
}
```

You need to use the serverless SQL pool in WS1 to read the files.

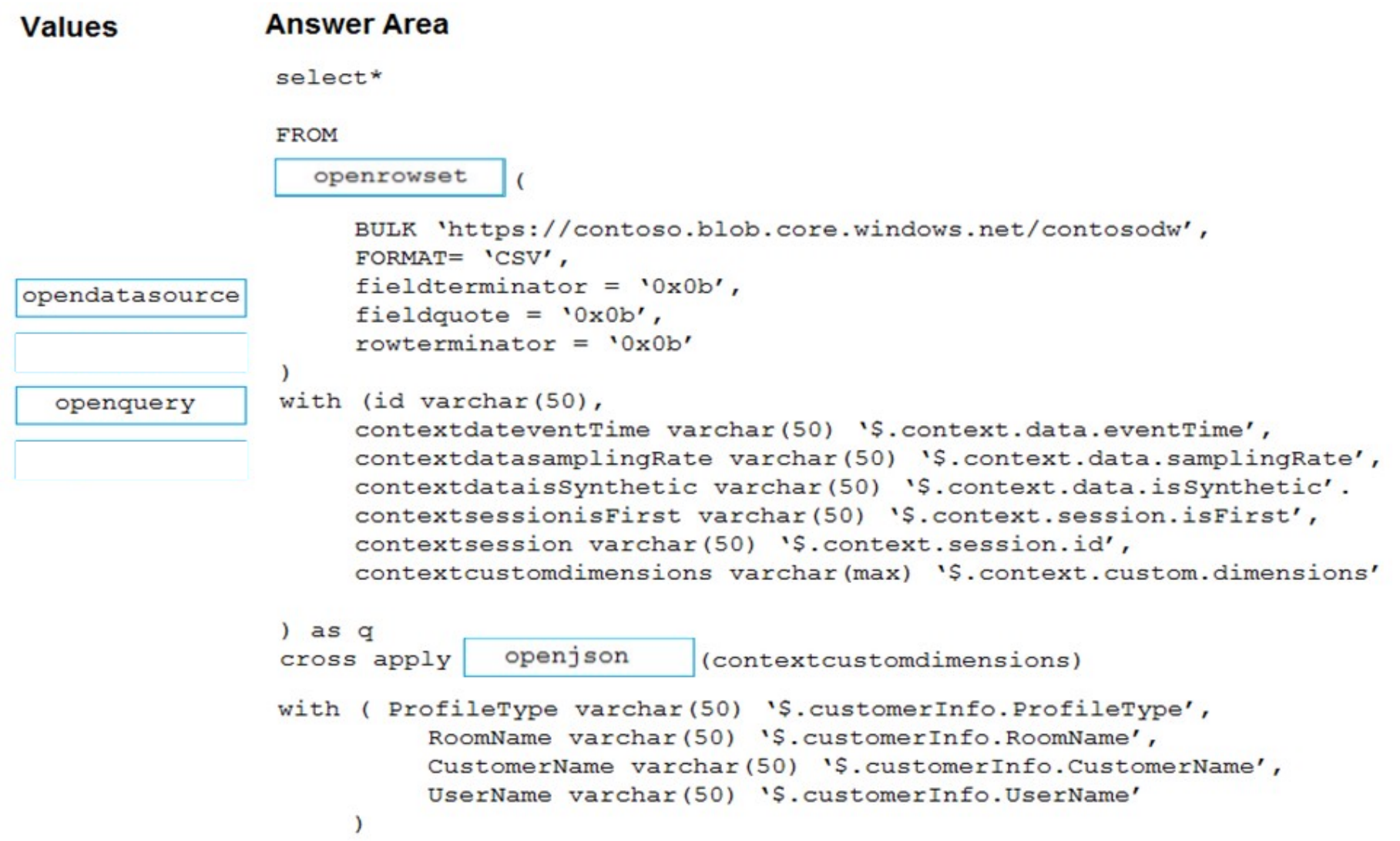
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Correct Answer:



Box 1: openrowset -

The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.

Example:

SELECT *

FROM OPENROWSET(

BULK 'csv/population/population.csv',

DATA_SOURCE = 'SqlOnDemandDemo',

FORMAT = 'CSV', PARSER_VERSION = '2.0',

FIELDTERMINATOR =',',

ROWTERMINATOR = '\n'

Box 2: openjson -

You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:

```
SELECT book.* FROM -
OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json
CROSS APPLY OPENJSON(BulkColumn)
WITH( id nvarchar(100), name nvarchar(100), price float,
pages_i int, author nvarchar(100)) AS book
```

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server

Question #10 *Topic 2*

DRAG DROP -

You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

Date	Temp
...	...
18-01-2021	3
19-01-2021	4
20-01-2021	2
21-01-2021	2
...	...

You need to produce the following table by using a Spark SQL query.

Year	JAN	FEB	MAR	APR	MAY
2019	2.3	4.1	5.2	7.6	9.2
2020	2.4	4.2	4.9	7.8	9.1
2021	2.6	5.3	3.4	7.9	9.5

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.

Select and Place:

Values

CAST

COLLATE

CONVERT

FLATTEN

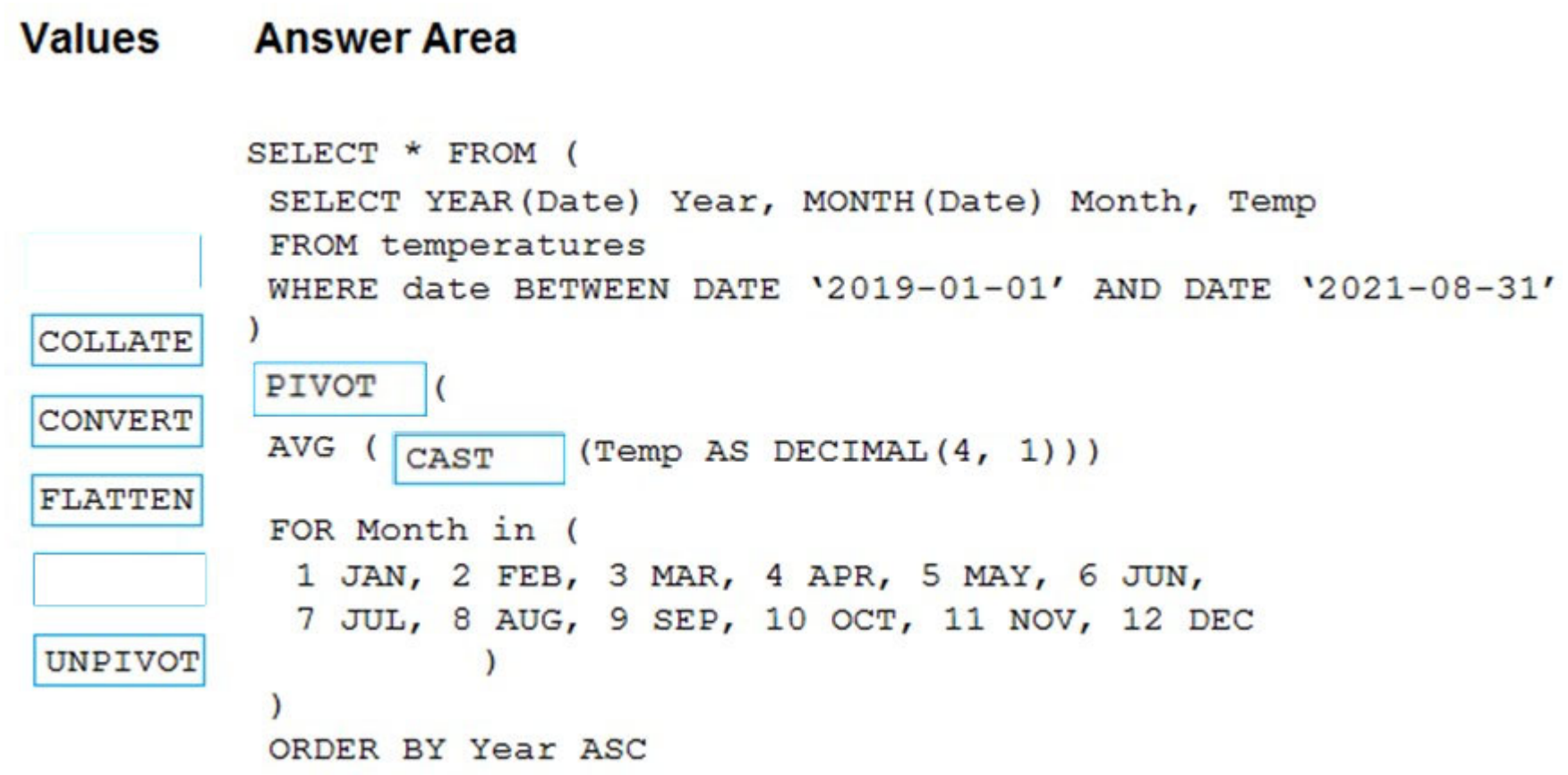
PIVOT

UNPIVOT

Answer Area

```
SELECT * FROM (
 SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
 FROM temperatures
 WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
[          ](
 AVG ([          ](Temp AS DECIMAL(4, 1)))
 FOR Month in (
  1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
  7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
            )
)
ORDER BY Year ASC
```

Correct Answer:



Box 1: PIVOT -

PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.

Incorrect Answers:

UNPIVOT carries out the opposite operation to PIVOT by rotating columns of a table-valued expression into column values.

Box 2: CAST -

If you want to convert an integer value to a DECIMAL data type in SQL Server use the CAST() function.

Example:

SELECT -

CAST(12 AS DECIMAL(7,2)) AS decimal_value;

Here is the result:

decimal_value

12.00

Reference:

https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/ https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot

Previous Questions Next Questions

Custom View Settings

Question #11 *Topic 2*

You have an Azure Data Factory that contains 10 pipelines.
You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.
What should you add to each pipeline?

A. a resource tag

B. a correlation ID

C. a run group ID

D. an annotation Most Voted

Correct Answer: *D*
Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.
Reference:
https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/

Community vote distribution

D (100%)

Question #12 *Topic 2*

HOTSPOT -

The following code segment is used to create an Azure Databricks cluster.

```
{
     "num_workers": null,
     "autoscale": {
          "min_workers": 2,
          "max_workers": 8
     },
     "cluster_name": "MyCluster",
     "spark_version": "latest-stable-scala2.11",
     "spark_conf": {
          "spark.databricks.cluster.profile": "serverless",
          "spark.databricks.repl.allowedLanguages": "sql,python,r"
     },
     "node_type_id": "Standard_DS13_v2",
     "ssh_public_keys": [],
     "custom_tags": {
          "ResourceClass": "Serverless"
     },
     "spark_env_vars": {
          "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
     },
     "autotermination_minutes": 90,
     "enable_elastic_disk": true,
     "init_scripts": []
}
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

Answer Area

	Statements	Yes	No
Correct Answer:	The Databricks cluster supports multiple concurrent users.	○	○
	The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
	The Databricks cluster supports the creation of a Delta Lake table.	○	○

Box 1: Yes -

A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2.

Box 2: No -

When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.

Box 3: Yes -

Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns.

Reference:
https://adatis.co.uk/databricks-cluster-sizing/
https://docs.microsoft.com/en-us/azure/databricks/jobs
https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html https://docs.databricks.com/delta/index.html

Question #13 *Topic 2*

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs.
You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.
What should you recommend?

A. Azure Synapse Analytics

B. Azure Databricks Most Voted

C. Azure Stream Analytics

D. Azure SQL Database

Correct Answer: *C*
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

Community vote distribution
B (71%) C (29%)

Question #14 *Topic 2*

HOTSPOT -

You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.

You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

- Create four partitions based on the order date.
- Ensure that each partition contains all the orders placed during a given calendar year.

How should you complete the T-SQL command? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int]    NOT NULL,
[CustomerKey] [int]   NOT NULL,
[SalesOrderNumber] [nvarchar](20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money]   NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [ ▼ ] FOR VALUES
```

RIGHT
LEFT

```
(  [ ▼ ]  )
```

20090101,20121231
20100101,20110101,20120101
20090101,20100101,20110101,20120101

Correct Answer:

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int]    NOT NULL,
[CustomerKey] [int]   NOT NULL,
[SalesOrderNumber] [nvarchar](20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money]   NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [ RIGHT / LEFT ] FOR VALUES
( [ 20090101,20121231 / 20100101,20110101,20120101 / 20090101,20100101,20110101,20120101 ] )
```

Range Left or Right, both are creating similar partition but there is difference in comparison

For example: in this scenario, when you use LEFT and 20100101,20110101,20120101

Partition will be, datecol<=20100101, datecol>20100101 and datecol<=20110101, datecol>20110101 and datecol<=20120101, datecol>20120101

But if you use range RIGHT and 20100101,20110101,20120101

Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, datecol>=20120101

In this example, Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st

Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15

Question #15 *Topic 2*

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement.

```
CREATE TABLE [DBO].[DimProduct] (
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductName] [nvarchar] (100) NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEndDate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowUpdatedDateTime] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

Which two columns should you add to the table? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. [EffectiveStartDate] [datetime] NOT NULL,

B. [CurrentProductCategory] [nvarchar] (100) NOT NULL, Most Voted

C. [EffectiveEndDate] [datetime] NULL,

D. [ProductCategory] [nvarchar] (100) NOT NULL,

E. [OriginalProductCategory] [nvarchar] (100) NOT NULL, Most Voted

Correct Answer: *BE*

A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.

This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

Reference:

https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/

Community vote distribution

BE (100%)

Question #16 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds.
Does this meet the goal?

A. Yes Most Voted

B. No

Correct Answer: *B*
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Community vote distribution

A (69%) B (31%)

Question #17 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.
Does this meet the goal?

A. Yes

B. No Most Voted

Correct Answer: *B*
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Community vote distribution

B (100%)

Question #18 *Topic 2*

HOTSPOT -

You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.

The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.

You need to calculate the duration between start and end events.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT
    [user],
    feature,
    [DATEADD( | DATEDIFF( | DATEPART(]
        second,
        [ISFIRST | LAST | TOPONE](Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

Correct Answer:

Answer Area

```
SELECT
    [user],
    feature,
    [DATEADD( | DATEDIFF( | DATEPART(]
        second,
        [ISFIRST | LAST | TOPONE](Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

Box 1: DATEDIFF -

DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate.

Syntax: DATEDIFF (datepart , startdate, enddate)

Box 2: LAST -

The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.

Example:

SELECT -

[user],

feature,

DATEDIFF(

second,

LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,

1) WHEN Event = 'start'),

Time) as duration -

FROM input TIMESTAMP BY Time -

WHERE -

Event = 'end'

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

Question #19 *Topic 2*

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an
Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.
The data flow already contains the following:
- A source transformation.
- A Derived Column transformation to set the appropriate types of data.
- A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements:
- All valid rows must be written to the destination table.
- Truncation errors in the comment column must be avoided proactively.
- Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

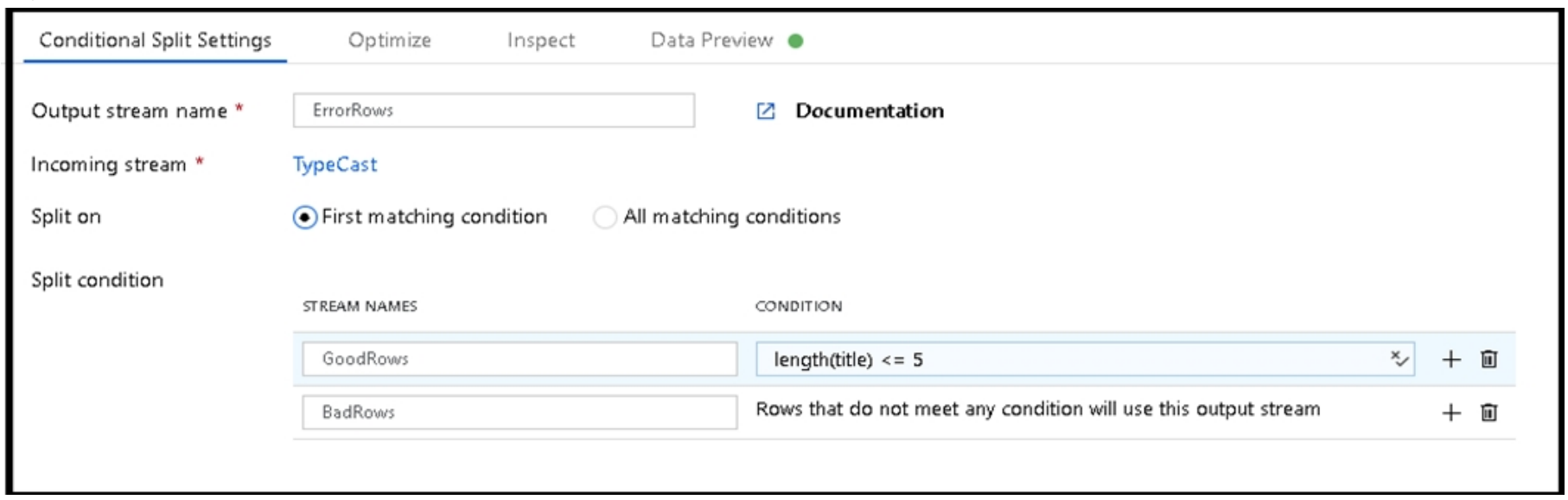
A. To the data flow, add a sink transformation to write the rows to a file in blob storage. Most Voted

B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors. Most Voted

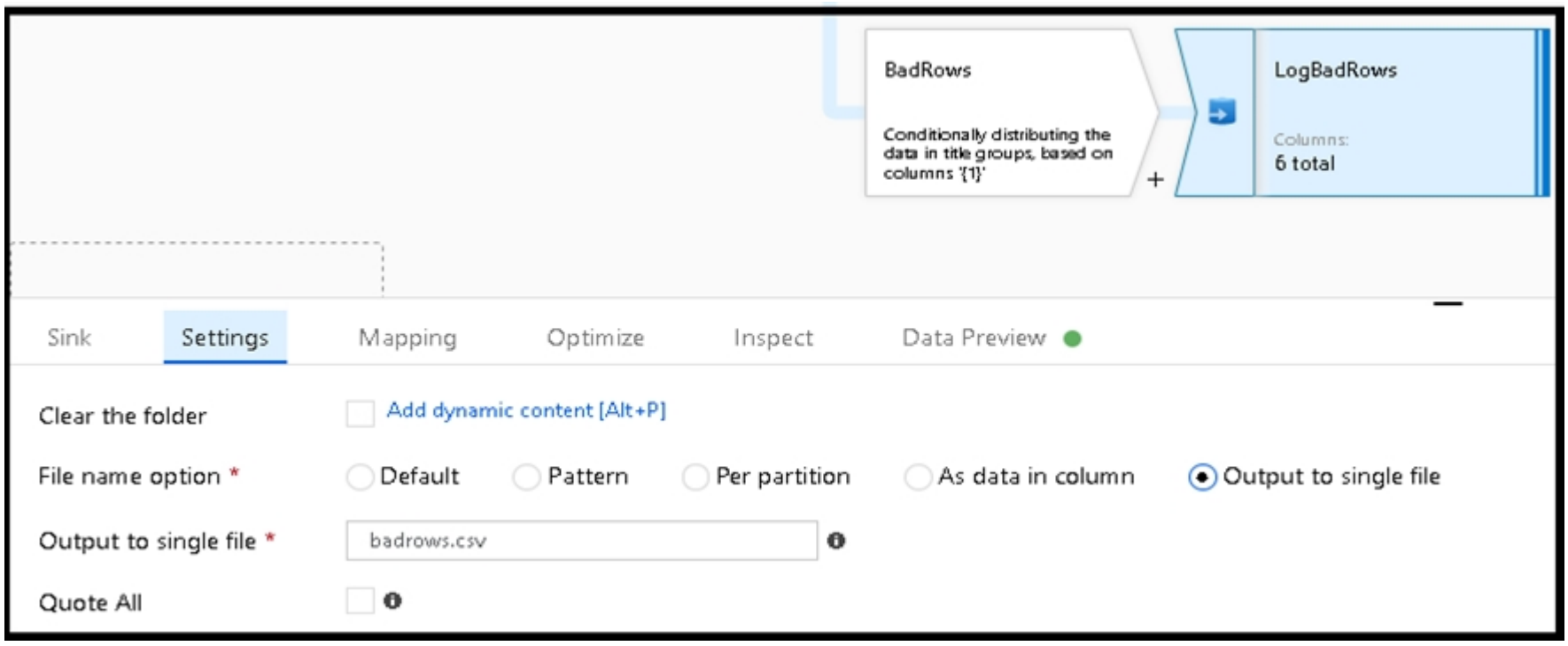
C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.

D. Add a select transformation to select only the rows that will cause truncation errors.

Correct Answer: *AB*
B: Example:
1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream.
Any row that is larger than five will go into the BadRows stream.

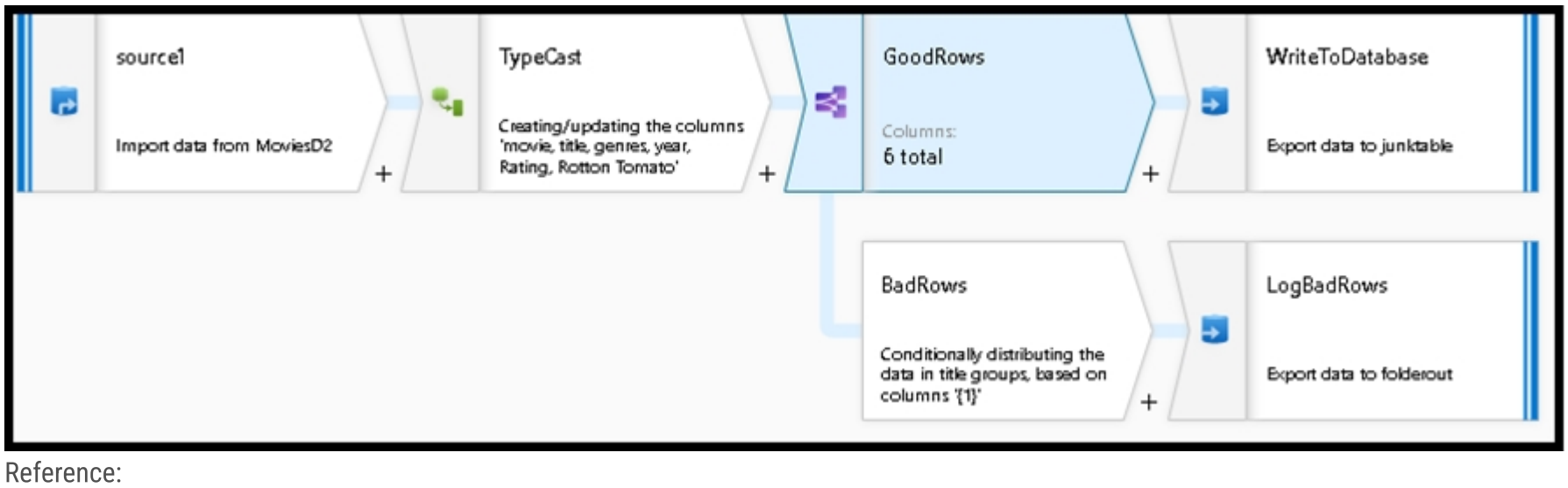


2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream.
Any row that is larger than five will go into the BadRows stream.
A:
3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file.

Meanwhile, successful rows can continue to write to our target database.



Reference:

https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows

Community vote distribution

AB (100%)

Question #20 *Topic 2*

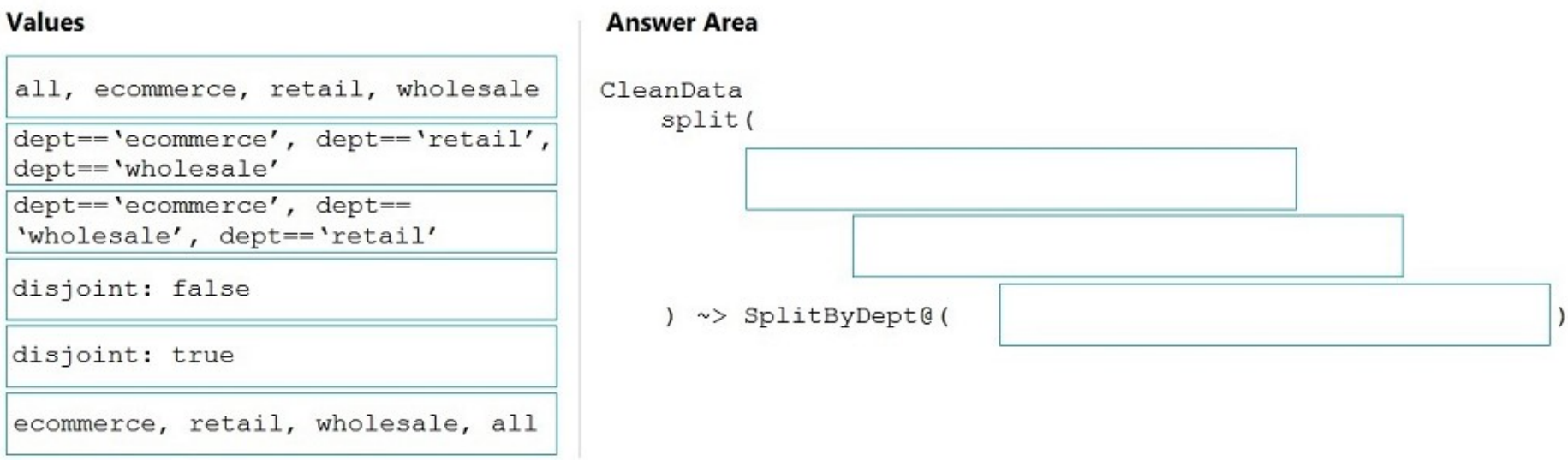
DRAG DROP -

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

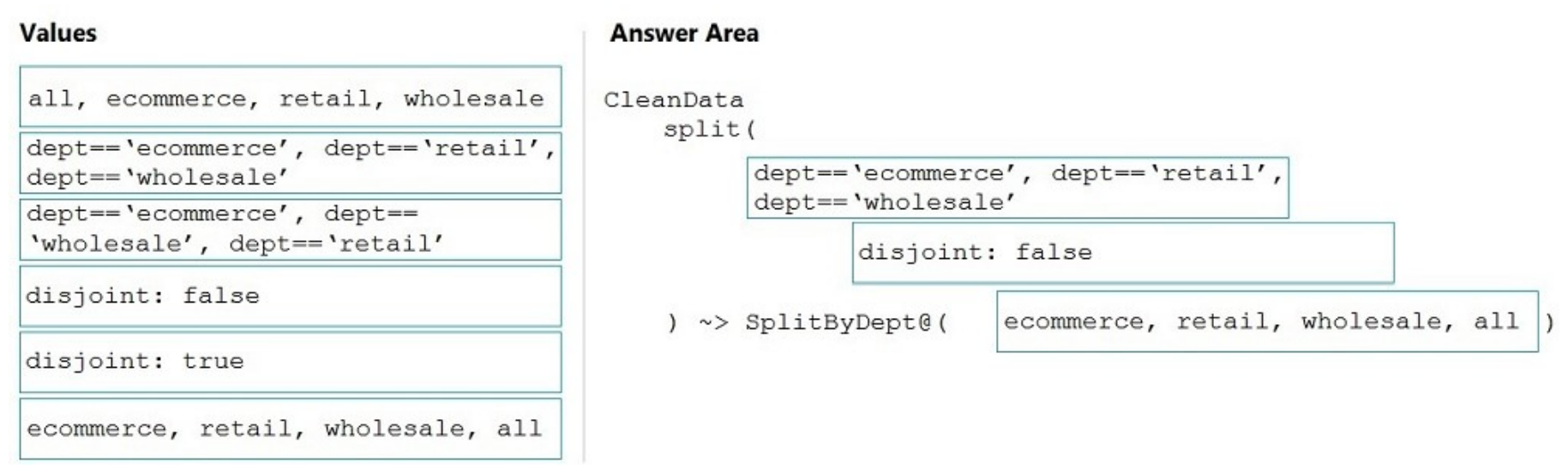
How should you complete the Data Factory data flow script? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Correct Answer:



The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.

Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'

First we put the condition. The order must match the stream labeling we define in Box 3.

Syntax:

<incomingStream>
split(
<conditionalExpression1>
<conditionalExpression2>
...
disjoint: {true | false}
) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)

Box 2: discount : false -

disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.

Box 3: ecommerce, retail, wholesale, all

Label the streams -

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split

Previous Questions Next Questions

Question #21 *Topic 2*

DRAG DROP -

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

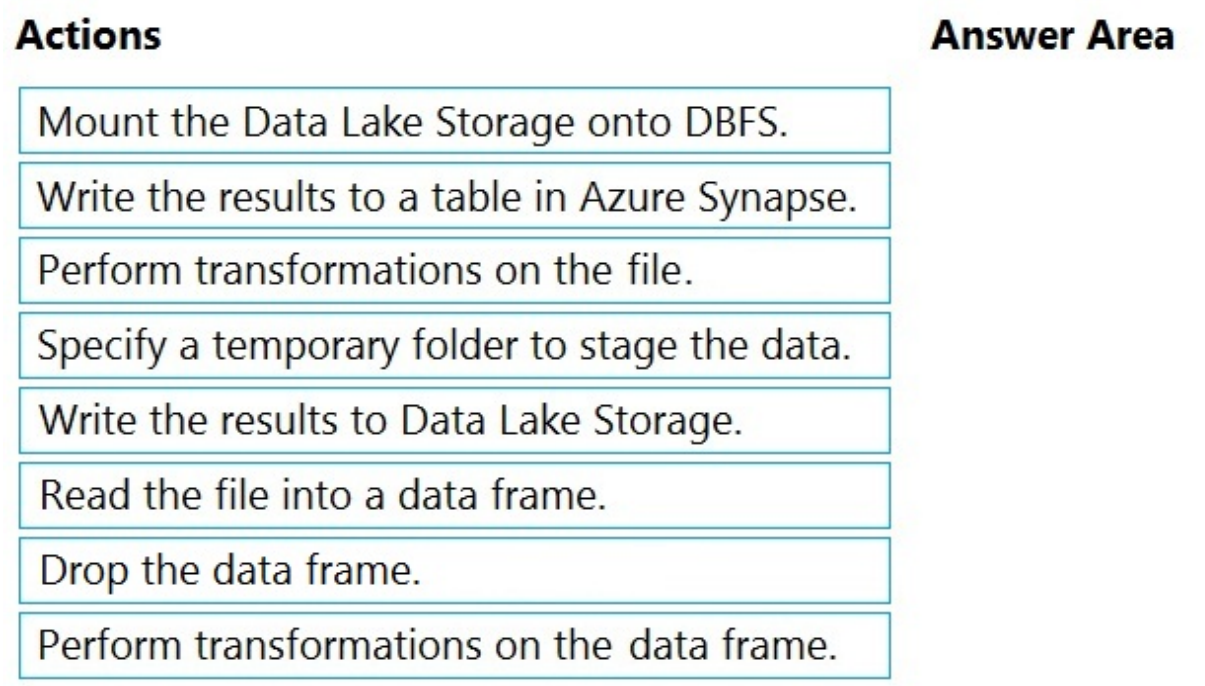
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

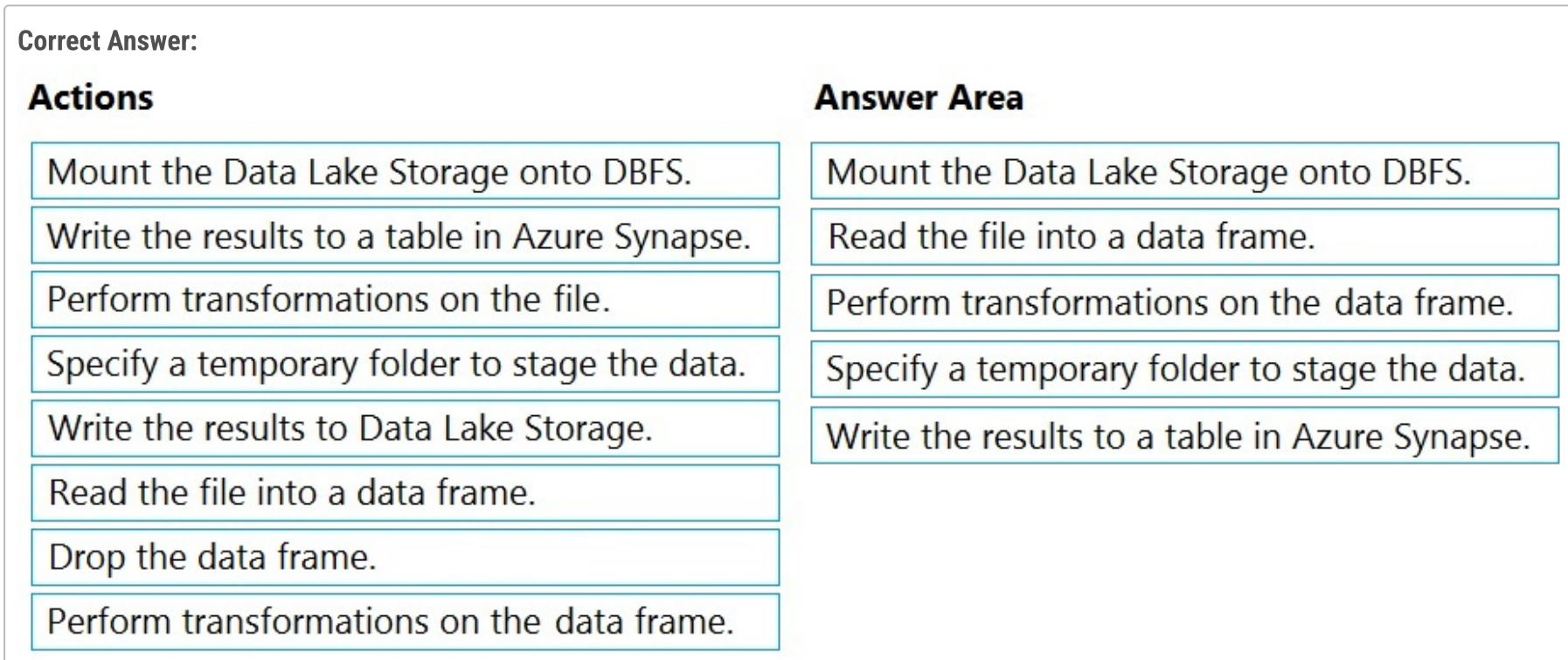
- A destination table in Azure Synapse
- An Azure Blob storage container
- A service principal

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

Select and Place:



Correct Answer:



Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account.

Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 5: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Reference:

https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

Question #22 *Topic 2*

HOTSPOT -

You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated

SQL pool.

Data in the container is stored in the following folder structure.

/in/{YYYY}/{MM}/{DD}/{HH}/{mm}

The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45.

You need to configure a pipeline trigger to meet the following requirements:

- Existing data must be loaded.
- Data must be loaded every 30 minutes.
- Late-arriving data of up to two minutes must be included in the load for the time at which the data should have arrived.

How should you configure the pipeline trigger? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Type: ▼
- Event
- On-demand
- Schedule
- Tumbling window

Additional properties: ▼
- Prefix: /in/, Event: Blob created
- Recurrence: 30 minutes, Start time: 2021-01-01T00:00
- Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
- Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Correct Answer:

Answer Area

Type: ▼
- Event
- On-demand
- Schedule
- Tumbling window

Additional properties: ▼
- Prefix: /in/, Event: Blob created
- Recurrence: 30 minutes, Start time: 2021-01-01T00:00
- Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
- Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Box 1: Tumbling window -

To be able to use the Delay parameter we select Tumbling window.

Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay.

The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

Question #23 *Topic 2*

HOTSPOT -

You are designing a near real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.

The solution will use Azure Stream Analytics and must meet the following requirements:

- Minimize latency from an Azure Event hub to the dashboard.
- Minimize the required storage.
- Minimize development effort.

What should you include in the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Azure Stream Analytics input type:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Azure Stream Analytics output type:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Aggregation query location:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Answer Area

Azure Stream Analytics input type:

- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Azure Stream Analytics output type:

- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Aggregation query location:

- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Correct Answer:

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard

Question #24 *Topic 2*

DRAG DROP -

You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.

You need to modify the job to accept data generated by the IoT devices in the Protobuf format.

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

Select and Place:

Actions	Answer Area
Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.	
Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.	
Add .NET deserializer code for Protobuf to the custom deserializer project.	
Add .NET deserializer code for Protobuf to the Stream Analytics project.	
Add an Azure Stream Analytics Application project to the solution.	

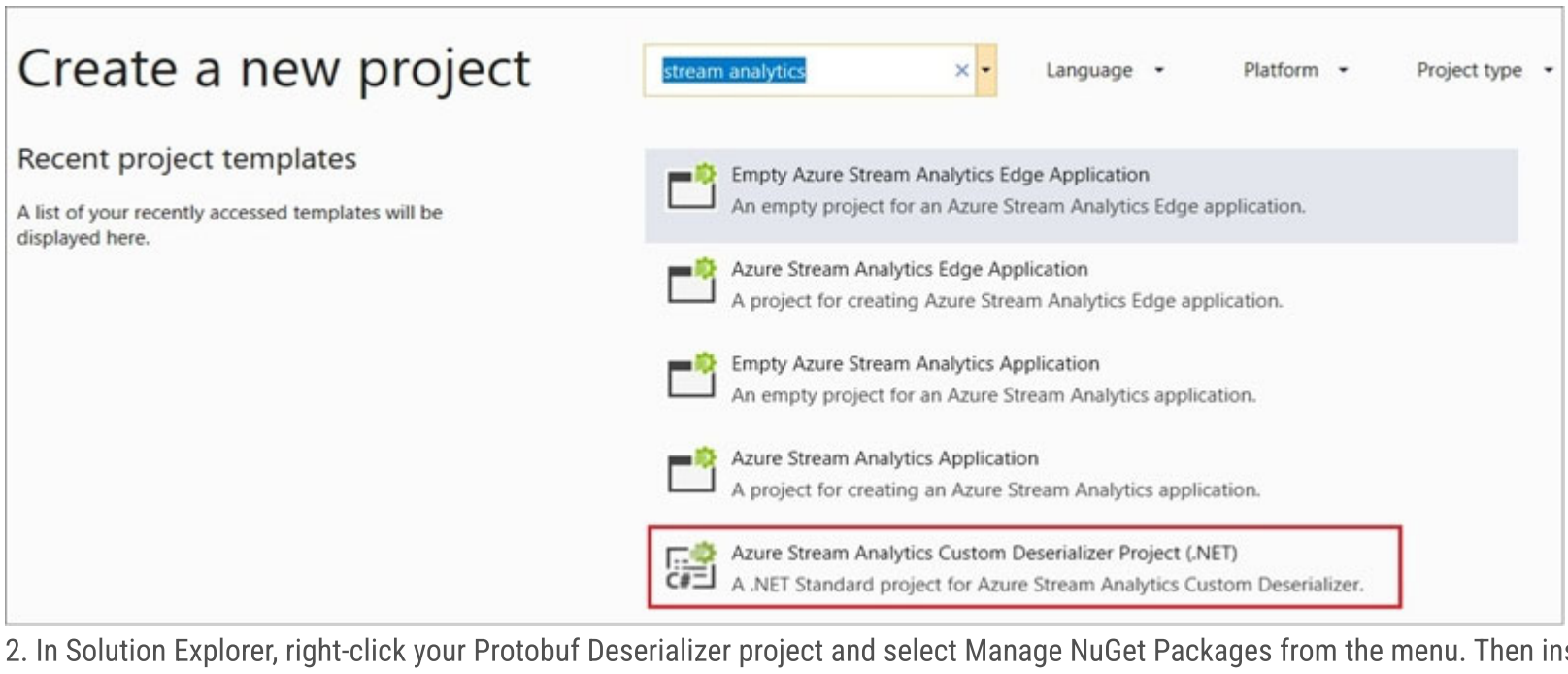
Correct Answer:

Actions	Answer Area
Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.	Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.	Add .NET deserializer code for Protobuf to the custom deserializer project.
Add .NET deserializer code for Protobuf to the custom deserializer project.	Add an Azure Stream Analytics Application project to the solution.
Add .NET deserializer code for Protobuf to the Stream Analytics project.	
Add an Azure Stream Analytics Application project to the solution.	

Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

Create a custom deserializer -

1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.



2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.

3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.
4. Build the Protobuf Deserializer project.
Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project
Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.
Step 3: Add an Azure Stream Analytics Application project to the solution
Add an Azure Stream Analytics project
1. In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose
Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.
2. Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer

Question #25 *Topic 2*

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region.
You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:
- Ensure that the data remains in the UK South region at all times.
- Minimize administrative effort.

Which type of integration runtime should you use?

A. Azure integration runtime Most Voted
B. Azure-SSIS integration runtime
C. Self-hosted integration runtime

Correct Answer: *A*

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Incorrect Answers:
C: Self-hosted integration runtime is to be used On-premises.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

Community vote distribution

A (100%)

Question #26 *Topic 2*

HOTSPOT -

You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

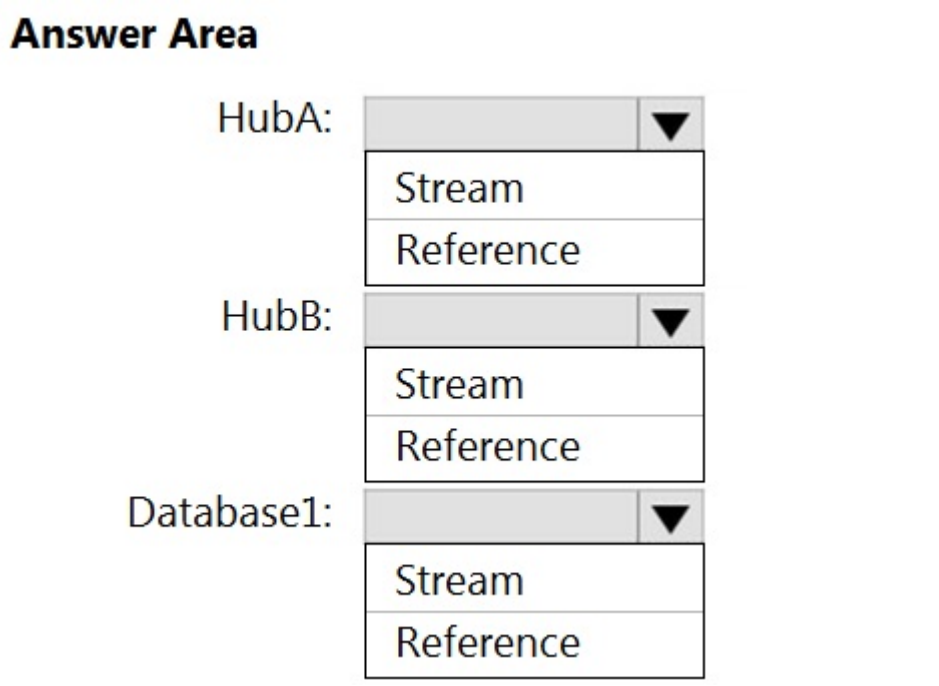
Source	Data
Database1	Driver's name Driver's license number
HubA	Ride route Ride distance Ride duration
HubB	Ride fare Ride payment

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.

How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.

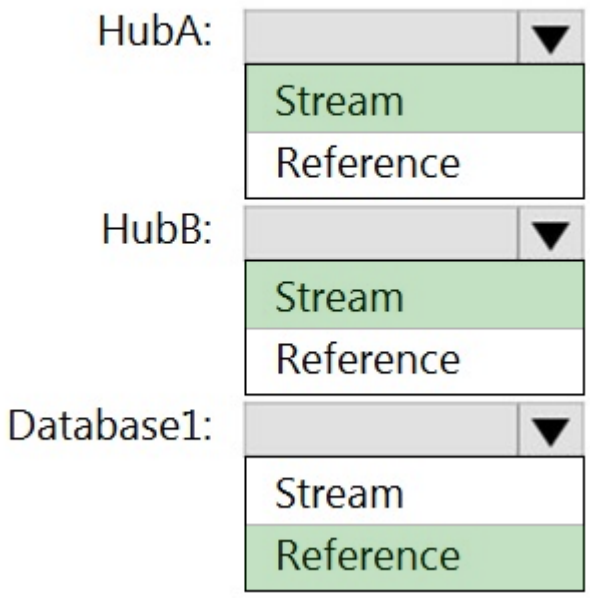
NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

Correct Answer:



HubA: Stream -

HubB: Stream -

Database1: Reference -

Reference data (also known as a lookup table) is a finite data set that is static or slowly changing in nature, used to perform a lookup or to augment your data streams. For example, in an IoT scenario, you could store metadata about sensors (which don't change often) in reference data and join it with real time IoT data streams. Azure Stream Analytics loads reference data in memory to achieve low latency stream processing

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

Question #27 *Topic 2*

You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.

You need to define a query in the Stream Analytics job. The query must meet the following requirements:

- Count the number of clicks within each 10-second window based on the country of a visitor.
- Ensure that each click is NOT counted more than once.

How should you define the Query?

A. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)

B. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
Most Voted

C. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)

D. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

Correct Answer: *B*

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Example:

Incorrect Answers:

A: Sliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window actually changes. In other words, when an event enters or exits the window. Every window has at least one event, like in the case of Hopping windows, events can belong to more than one sliding window.

C: Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap, so events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

D: Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

Community vote distribution

B (100%)

Question #28 *Topic 2*

HOTSPOT -

You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.

You need to calculate the difference in the number of readings per sensor per hour.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT sensorId,
       growth = reading -
                [ ▼ ] (reading) OVER (PARTITION BY sensorId [ ▼ ] (hour,1))
FROM input
```

LAG / LAST / LEAD

LIMIT DURATION / OFFSET / WHEN

Correct Answer:

Answer Area

```
SELECT sensorId,
       growth = reading -
                [ LAG ▼ ] (reading) OVER (PARTITION BY sensorId [ LIMIT DURATION ▼ ] (hour,1))
FROM input
```

Box 1: LAG -

The LAG analytic operator allows one to look up a ג€previousג€ event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.

Box 2: LIMIT DURATION -

Example: Compute the rate of growth, per sensor:

SELECT sensorId,

growth = reading -

LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1))

FROM input -

Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics

Question #29 *Topic 2*

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.
Which type of trigger should you use?

A. on-demand

B. tumbling window

C. schedule

D. event Most Voted

Correct Answer: *D*
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events.
Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure
Blob Storage account.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

Community vote distribution

D (100%)

Question #30 *Topic 2*

You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.
You publish changes from the main branch of the Git repository to ADFdev.
You need to deploy the artifacts from ADFdev to ADFprod.
What should you do first?

A. From ADFdev, modify the Git configuration.

B. From ADFdev, create a linked service.

C. From Azure DevOps, create a release pipeline. Most Voted

D. From Azure DevOps, update the main branch.

Correct Answer: *C*
In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.
Note: The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.
1. In Azure DevOps, open the project that's configured with your data factory.
2. On the left side of the page, select Pipelines, and then select Releases.
3. Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
4. In the Stage name box, enter the name of your environment.
5. Select Add artifact, and then select the git repository configured with your development data factory. Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.
6. Select the Empty job template.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment

Community vote distribution

C (100%)

Previous Questions Next Questions

Custom View Settings

Question #31 *Topic 2*

You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.
Which input type should you use for the reference data?

A. Azure Cosmos DB

B. Azure Blob storage Most Voted

C. Azure IoT Hub

D. Azure Event Hubs

Correct Answer: *B*
Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

Community vote distribution

B (100%)

Question #32 *Topic 2*

You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.
You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.
Which type of window should you use?

A. snapshot

B. tumbling

C. hopping Most Voted

D. sliding

Correct Answer: *C*
Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consist of three parameters: the timeunit, the windowsize (how long each window lasts) and the hopsize (by how much each window moves forward relative to the previous one).
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics

Community vote distribution

C (100%)

Question #33 *Topic 2*

HOTSPOT -

You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.

You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.

You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.

What should you include in the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

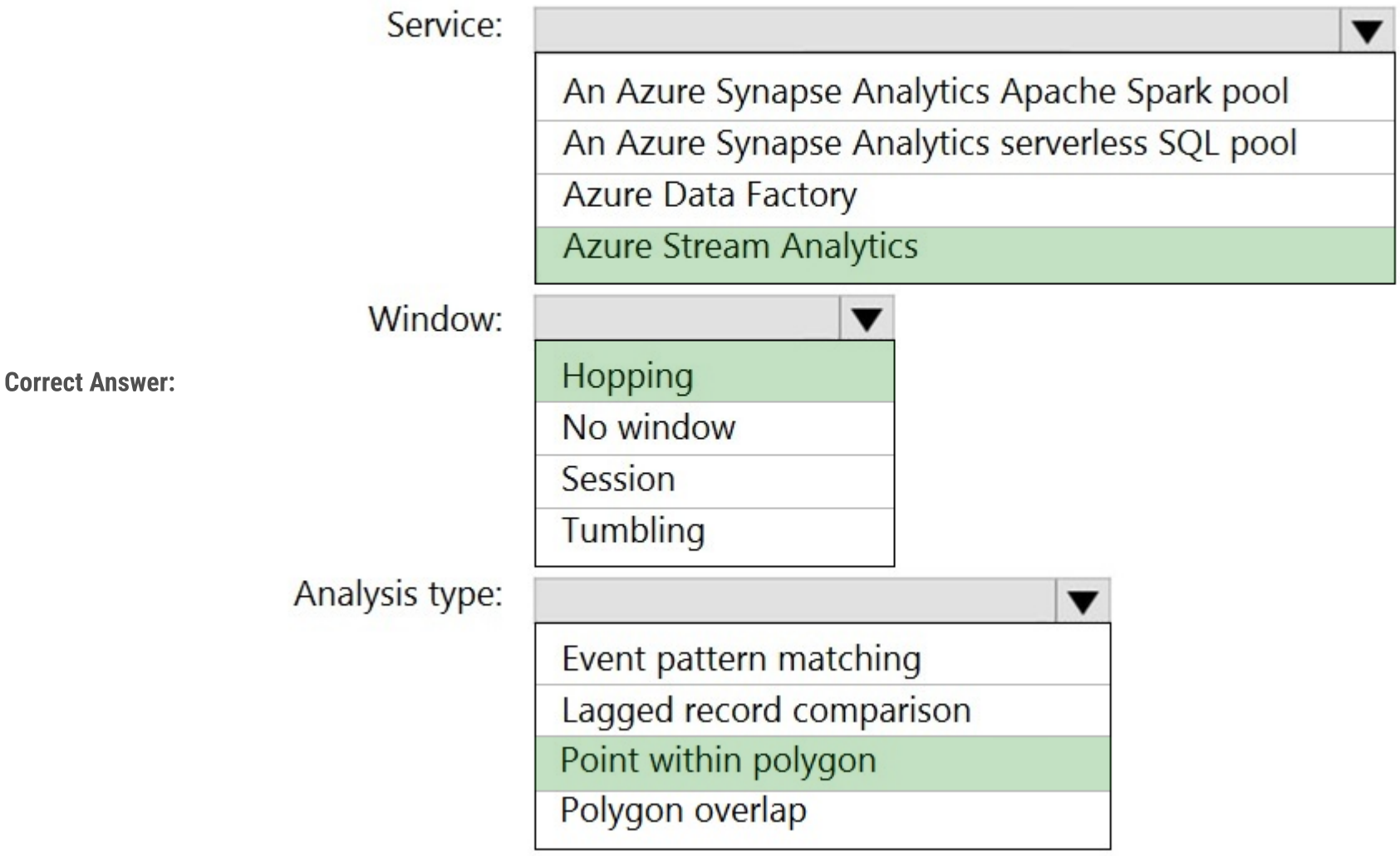
Service:
- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window:
- Hopping
- No window
- Session
- Tumbling

Analysis type:
- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

Answer Area

Correct Answer:



Box 1: Azure Stream Analytics -

Box 2: Hopping -
Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Box 3: Point within polygon -
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

Question #34 *Topic 2*

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.
You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.
What should you include in the solution?

A. Partition by DateTime fields.

B. Sink to Azure Queue storage. Most Voted

C. Include a watermark column.

D. Use a JSON format for physical data storage.

Correct Answer: *B*
The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files. This provides two major advantages:
- Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive.
- Lower costs: no more costly LIST API requests made to ABS.

Reference:
https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs

Community vote distribution

B (85%) A (15%)

Question #35 *Topic 2*

HOTSPOT -

You have a self-hosted integration runtime in Azure Data Factory.

The current status of the integration runtime has the following configurations:

- Status: Running
- Type: Self-Hosted
- Version: 4.4.7292.1
- Running / Registered Node(s): 1/1
- High Availability Enabled: False
- Linked Count: 0
- Queue Length: 0
- Average Queue Duration. 0.00s

The integration runtime has the following node details:

- Name: X-M
- Status: Running
- Version: 4.4.7292.1
- Available Memory: 7697MB
- CPU Utilization: 6%
- Network (In/Out): 1.21KBps/0.83KBps
- Concurrent Jobs (Running/Limit): 2/14
- Role: Dispatcher/Worker
- Credential Status: In Sync

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

- fail until the node comes back online
- switch to another integration runtime
- exceed the CPU limit

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

- raised
- lowered
- left as is

Correct Answer:

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

- fail until the node comes back online
- switch to another integration runtime
- exceed the CPU limit

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

- raised
- lowered
- left as is

Box 1: fail until the node comes back online

We see: High Availability Enabled: False

Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with

Data Factory.

Box 2: lowered -

We see:

Concurrent Jobs (Running/Limit): 2/14

CPU Utilization: 6%

Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

Question #36 *Topic 2*

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.

You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:

- Automatically scale down workers when the cluster is underutilized for three minutes.
- Minimize the time it takes to scale to the maximum number of workers.
- Minimize costs.

What should you do first?

A. Enable container services for workspace1.

B. Upgrade workspace1 to the Premium pricing tier. Most Voted

C. Set Cluster Mode to High Concurrency.

D. Create a cluster policy in workspace1.

Correct Answer: *B*

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan

Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state.

Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds.

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling -

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.

Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes.

Scales down exponentially, starting with 1 node.

Reference:

https://docs.databricks.com/clusters/configure.html

Community vote distribution

B (93%) 7%

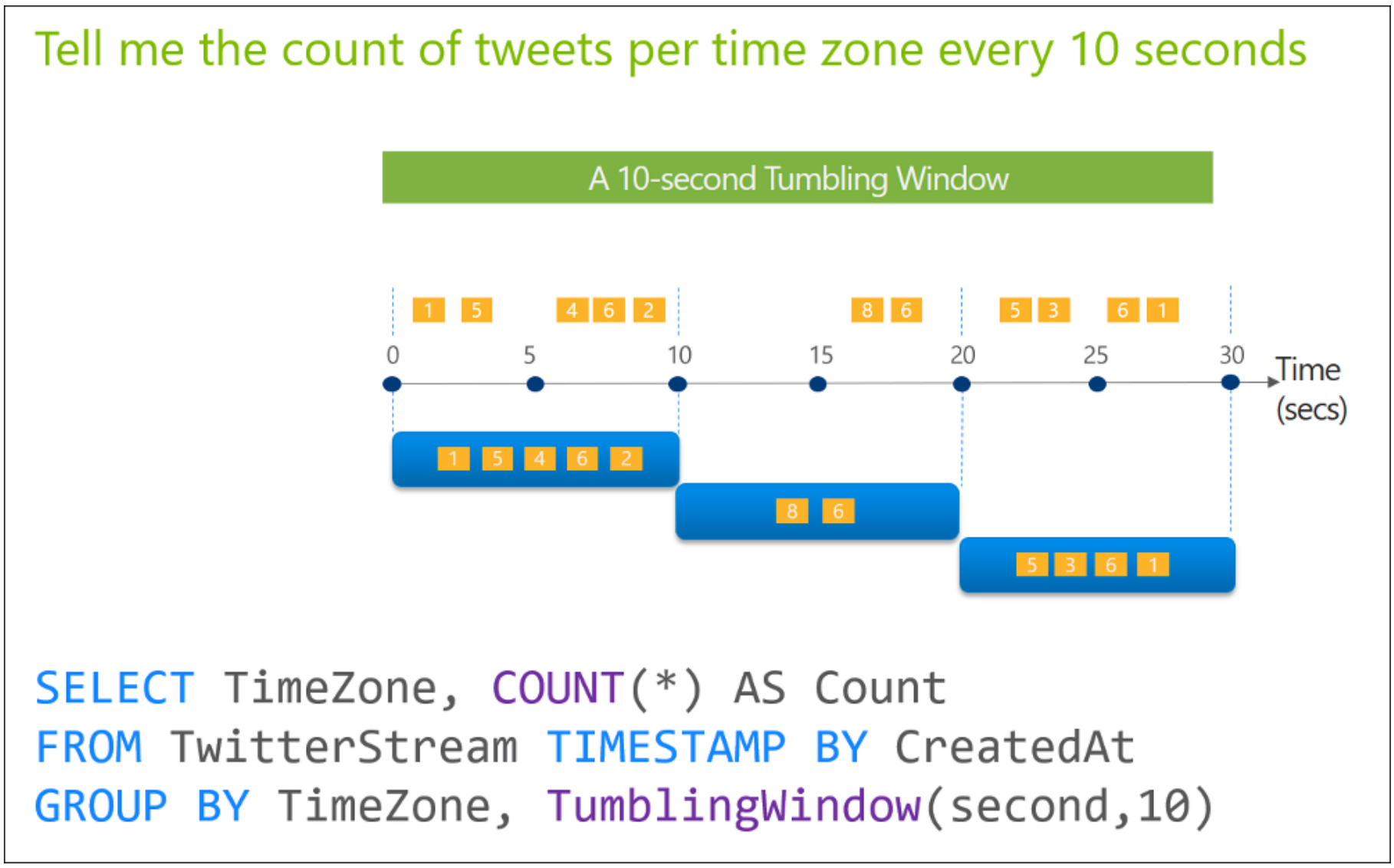
Question #37 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a tumbling window, and you set the window size to 10 seconds.
Does this meet the goal?

A. Yes Most Voted

B. No

Correct Answer: *A*
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.



Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Community vote distribution

A (100%)

Question #38 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a session window that uses a timeout size of 10 seconds.
Does this meet the goal?

A. Yes

B. No Most Voted

Correct Answer: *B*
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Community vote distribution

B (100%)

Question #39 *Topic 2*

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account.
You need to output the count of records received from the last five minutes every minute.
Which windowing function should you use?

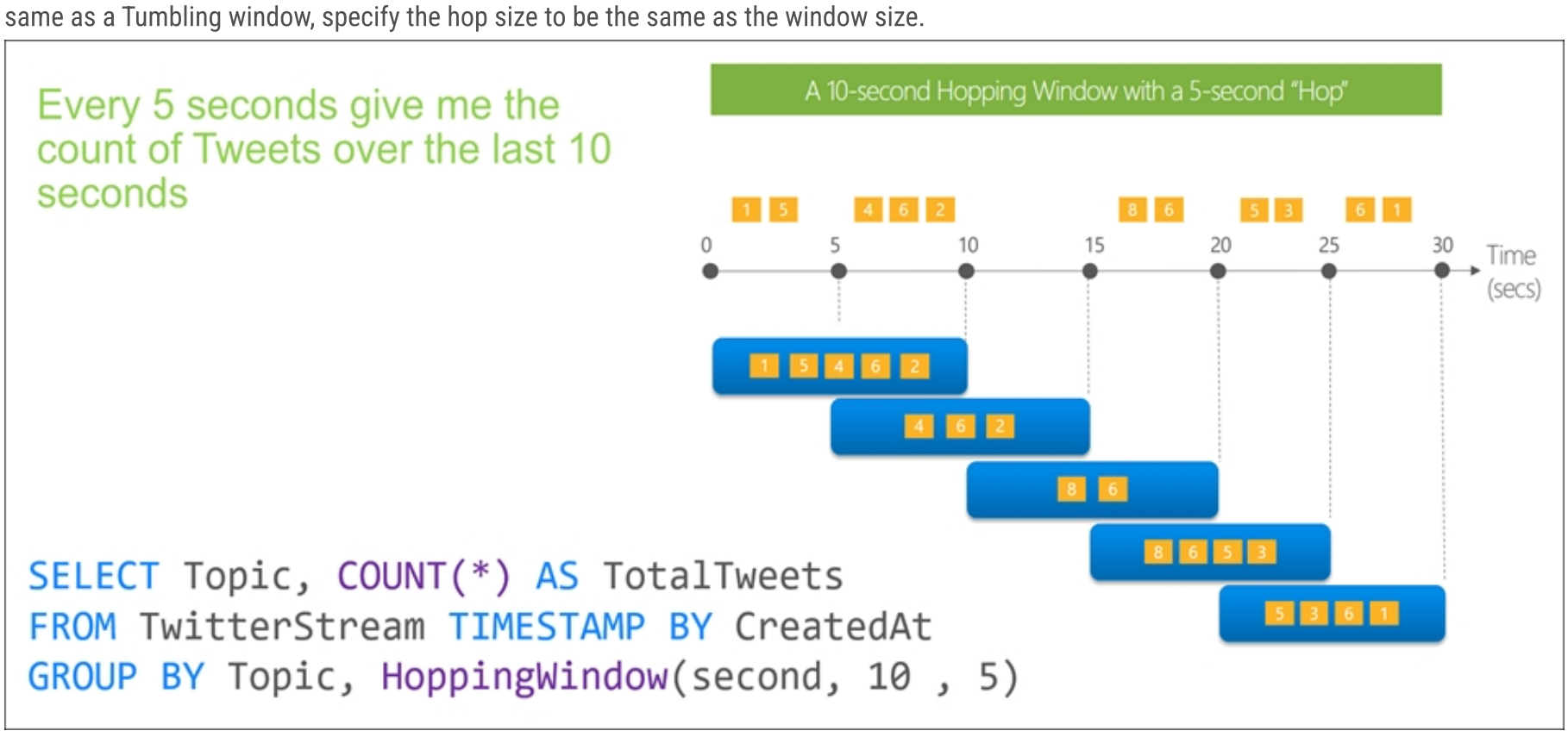
A. Session

B. Tumbling

C. Sliding

D. Hopping Most Voted

Correct Answer: *D*
Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.



Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

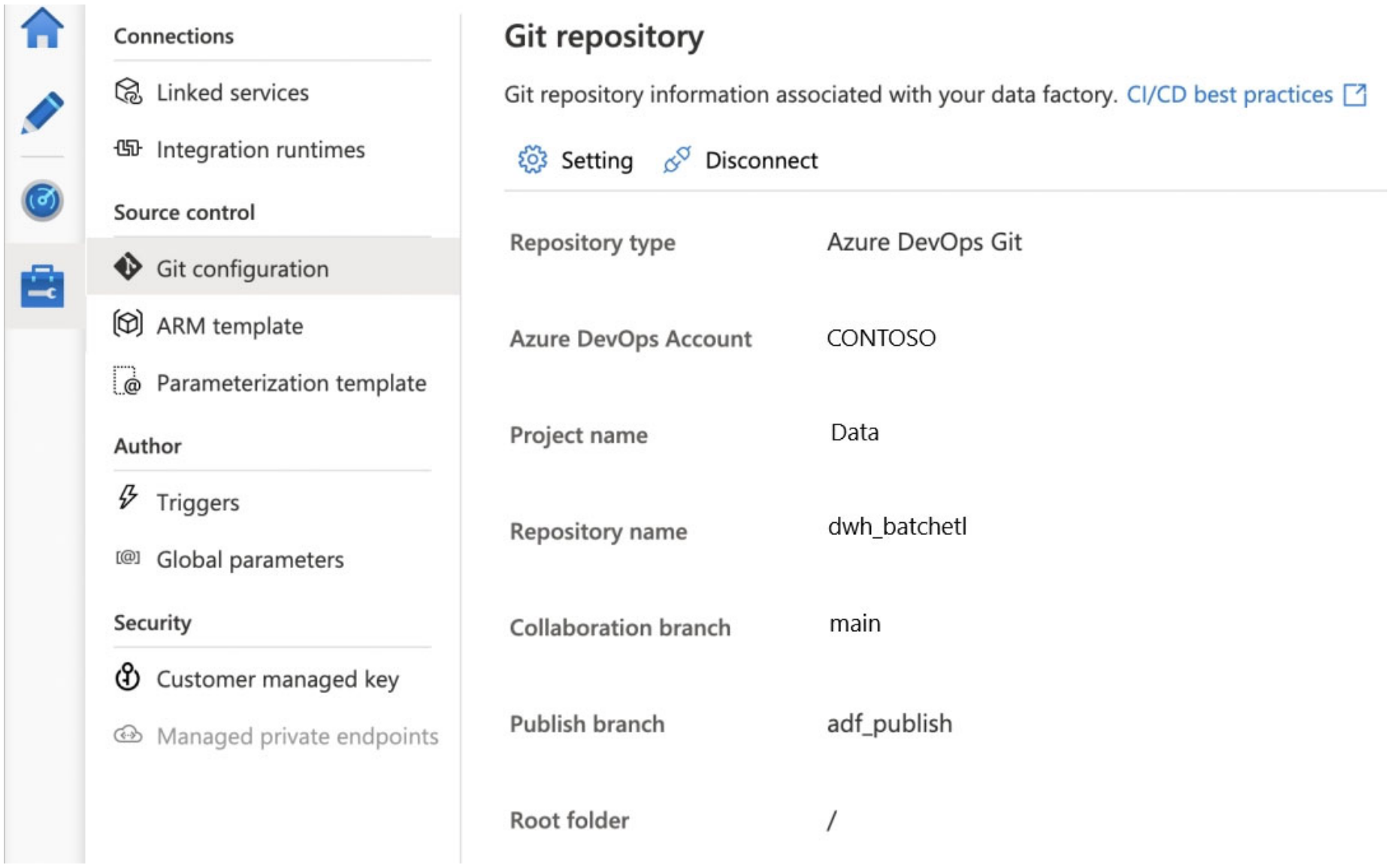
Community vote distribution

D (100%)

Question #40 *Topic 2*

HOTSPOT -

You configure version control for an Azure Data Factory instance as shown in the following exhibit.



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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in **[answer choice]**

- /
- adf_publish
- main
- Parameterization template

A Data Factory Azure Resource Manager (ARM) template named `contososales` can be found in **[answer choice]**

- /
- /contososales
- /dwh_batchetl/adf_publish/contososales
- /main

Correct Answer:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in **[answer choice]**

- /
- adf_publish
- main
- Parameterization template

A Data Factory Azure Resource Manager (ARM) template named `contososales` can be found in **[answer choice]**

- /
- /contososales
- /dwh_batchetl/adf_publish/contososales
- /main

Box 1: adf_publish -

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.
Box 2: / dwh_batchetl/adf_publish/contososales
Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

Previous Questions Next Questions

Question #41 *Topic 2*

HOTSPOT -

You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.

You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
Select TimeZone, count (*) AS MessageCount

FROM MessageStream  [ ▼ ]  CreatedAt

GROUP BY TimeZone,  [ ▼ ]  (second,15)
```

FROM MessageStream: LAST / OVER / SYSTEM.TIMESTAMP() / TIMESTAMP BY

GROUP BY TimeZone,: HOPPINGWINDOW / SESSIONWINDOW / SLIDINGWINDOW / TUMBLINGWINDOW

Correct Answer:

Answer Area

```
Select TimeZone, count (*) AS MessageCount

FROM MessageStream  [ ▼ ]  CreatedAt

GROUP BY TimeZone,  [ ▼ ]  (second,15)
```

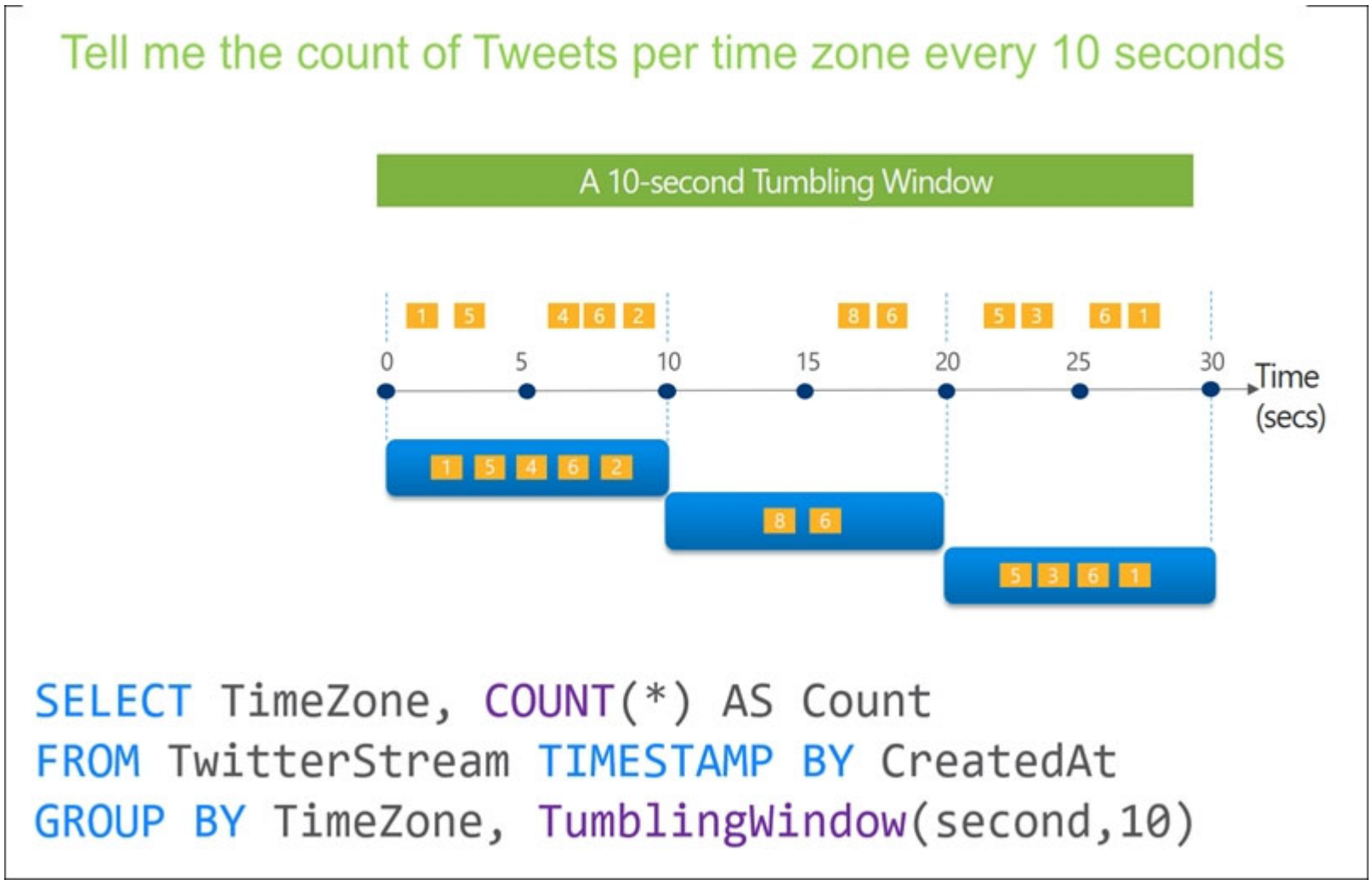
FROM MessageStream: LAST / OVER / SYSTEM.TIMESTAMP() / **TIMESTAMP BY**

GROUP BY TimeZone,: HOPPINGWINDOW / SESSIONWINDOW / SLIDINGWINDOW / **TUMBLINGWINDOW**

Box 1: timestamp by -

Box 2: TUMBLINGWINDOW -

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.



Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

Question #42 *Topic 2*

HOTSPOT -

You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.

ADF1 contains the following pipelines:

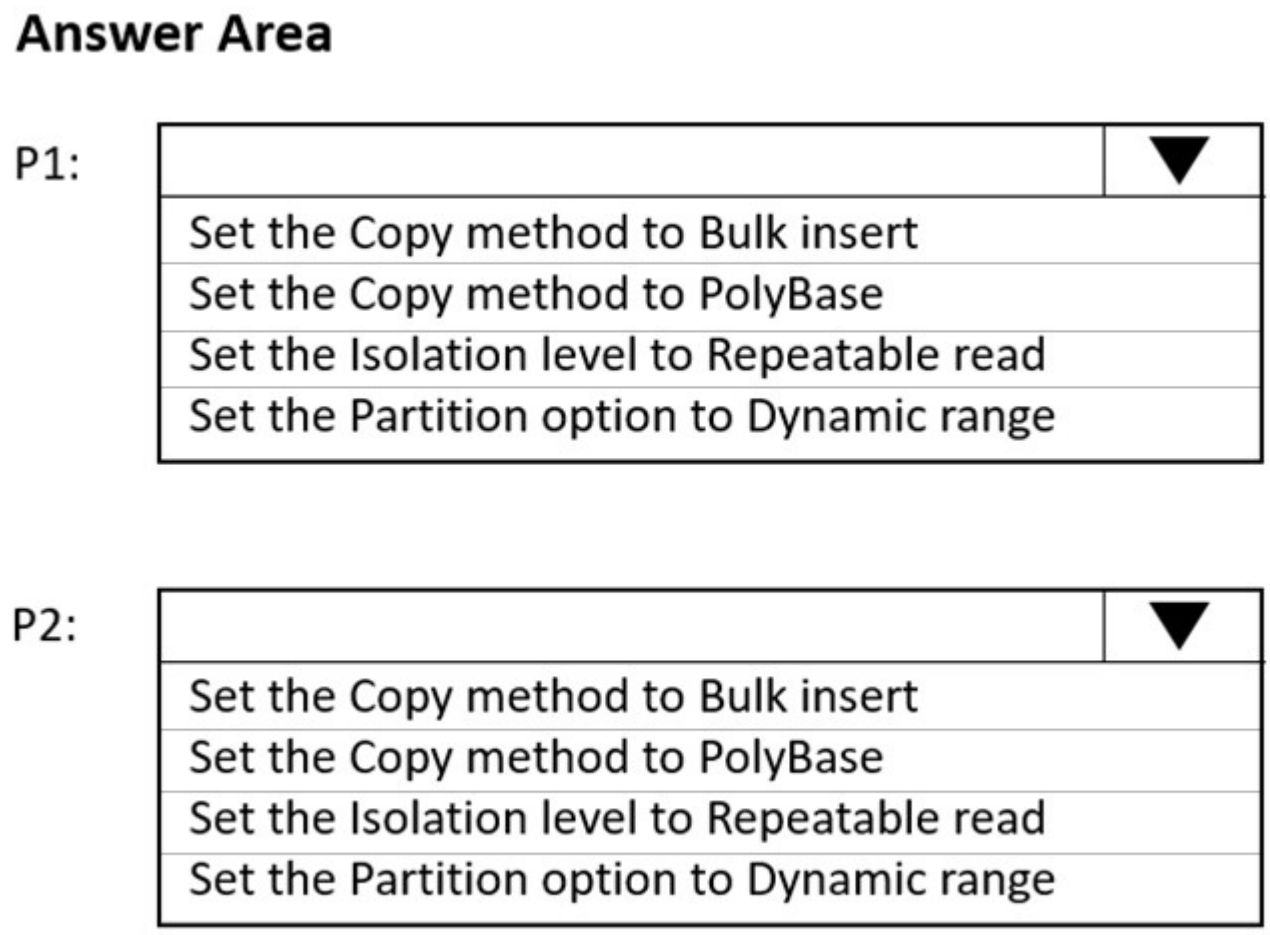
- P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account
- P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2

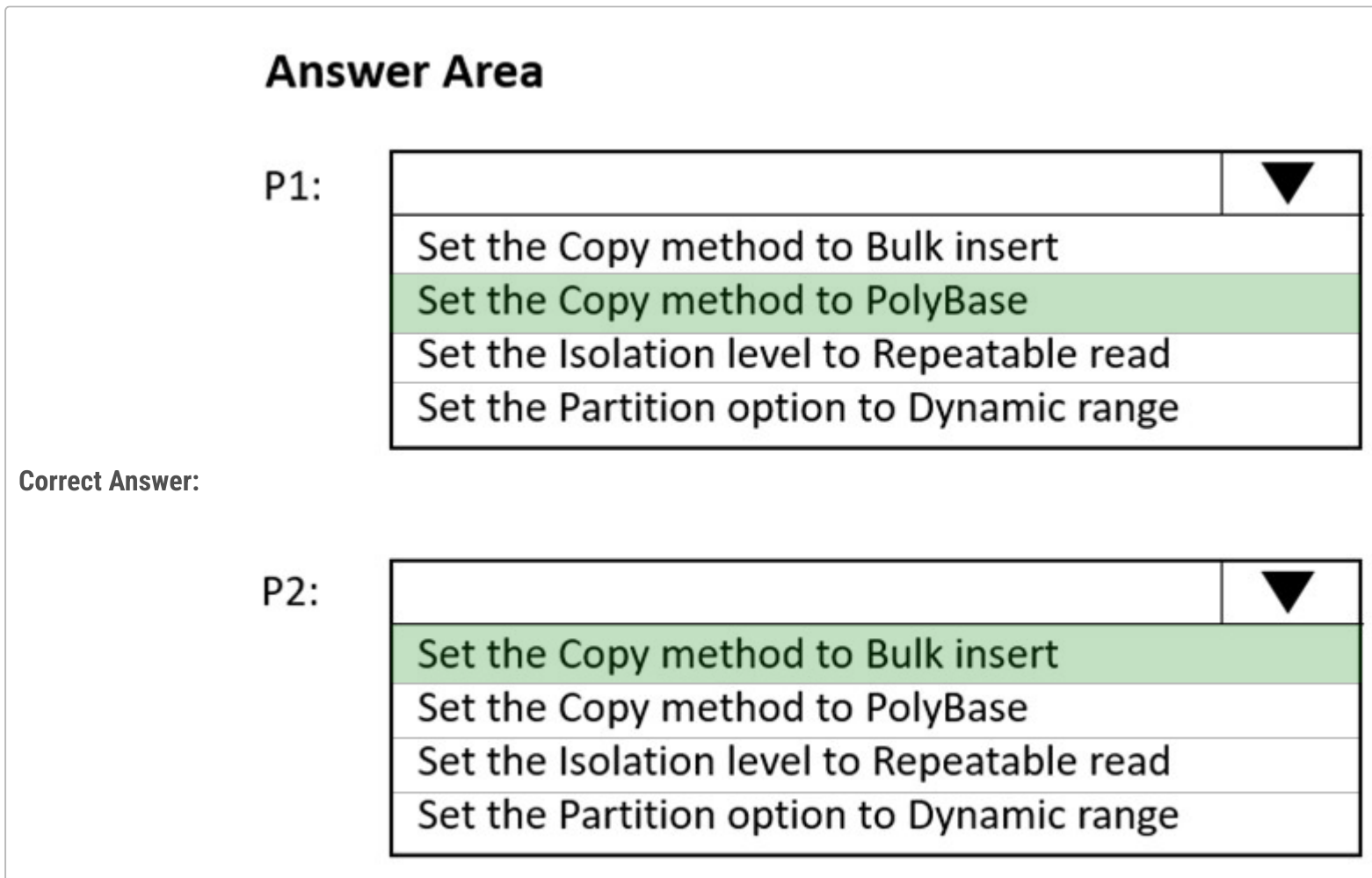
You need to configure P1 and P2 to maximize parallelism and performance.

Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:





Correct Answer:

Box 1: Set the Copy method to PolyBase

While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data

Lake Store via the T-SQL language.

Box 2: Set the Copy method to Bulk insert

Polybase not possible for text files. Have to use Bulk insert.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview

Question #43 *Topic 2*

HOTSPOT -

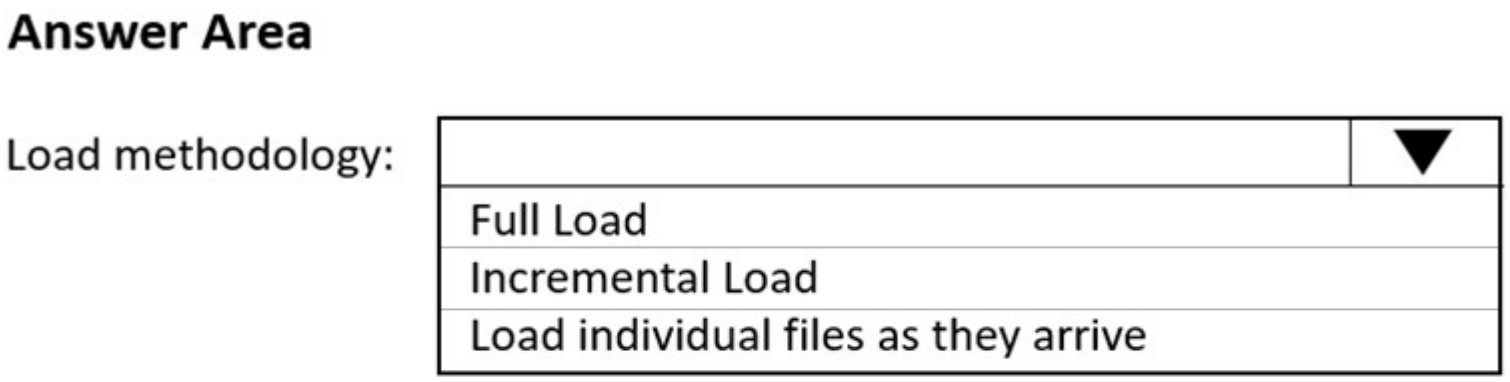
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.

You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs.

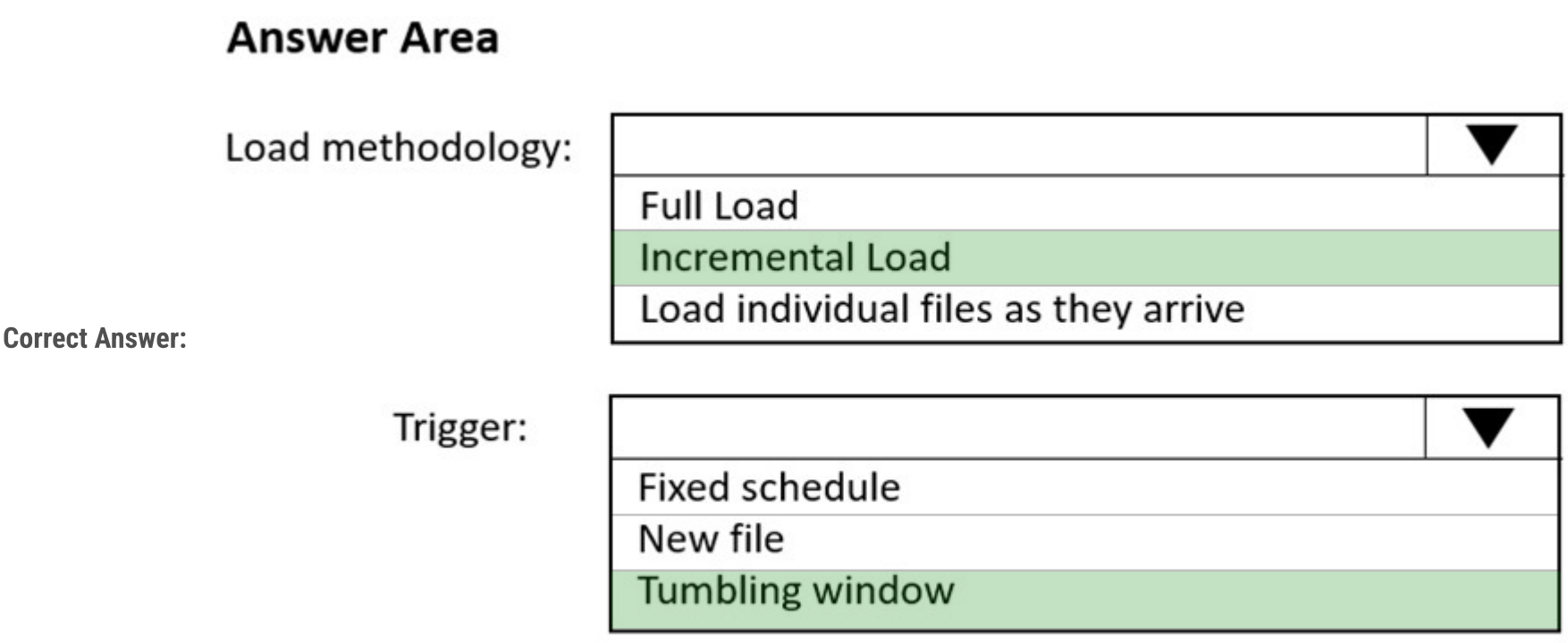
How should you configure the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



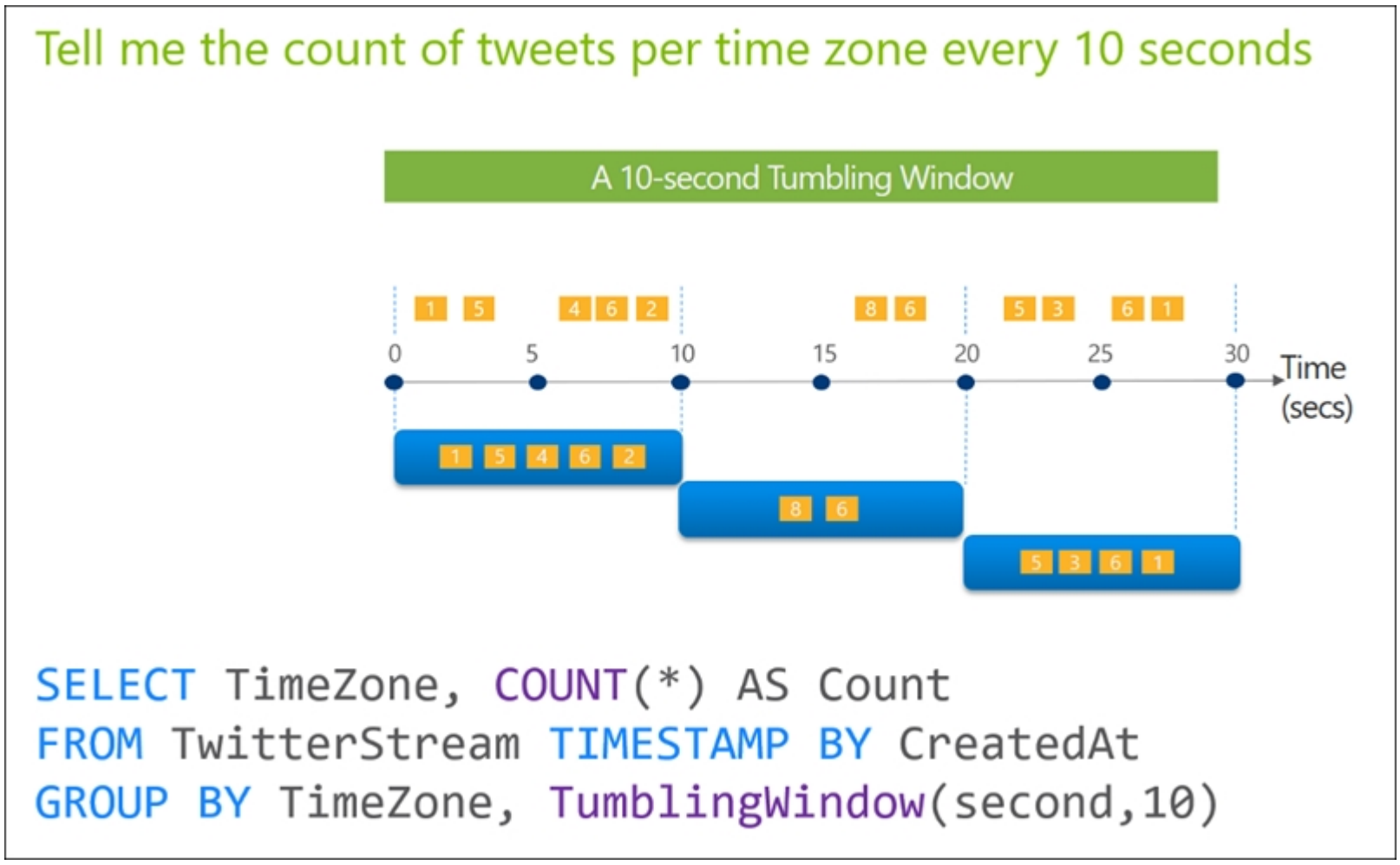
Correct Answer:



Box 1: Incremental load -

Box 2: Tumbling window -

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.



Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Question #44 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this 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 create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.
Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.
Does this meet the goal?

A. Yes

B. No Most Voted

Correct Answer: *B*

We need a High Concurrency cluster for the data engineers and the jobs.
Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

Community vote distribution

B (96%) 4%

Question #45 *Topic 2*

You have the following Azure Data Factory pipelines:

- Ingest Data from System1
- Ingest Data from System2
- Populate Dimensions
- Populate Facts

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest

Data from System2. Populate Facts must execute after Populate Dimensions pipeline. All the pipelines must execute every eight hours.

What should you do to schedule the pipelines for execution?

A. Add an event trigger to all four pipelines.

B. Add a schedule trigger to all four pipelines.

C. Create a patient pipeline that contains the four pipelines and use a schedule trigger. Most Voted

D. Create a patient pipeline that contains the four pipelines and use an event trigger.

Correct Answer: *C*

Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

Community vote distribution

C (100%)

Question #46 *Topic 2*

DRAG DROP -

You are responsible for providing access to an Azure Data Lake Storage Gen2 account.

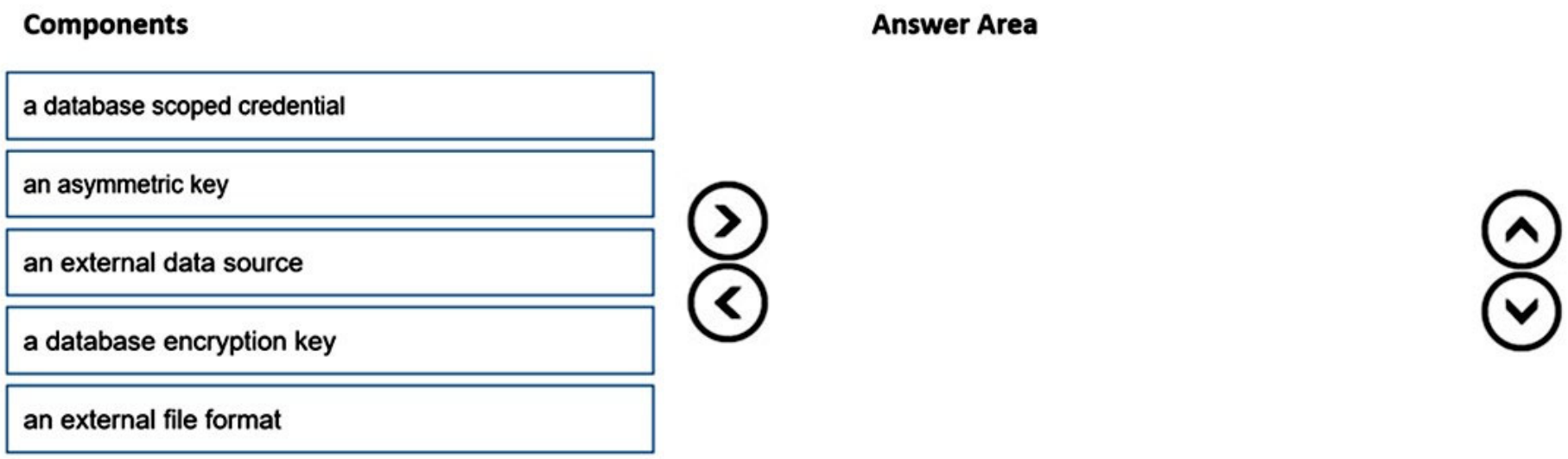
Your user account has contributor access to the storage account, and you have the application ID and access key.

You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics.

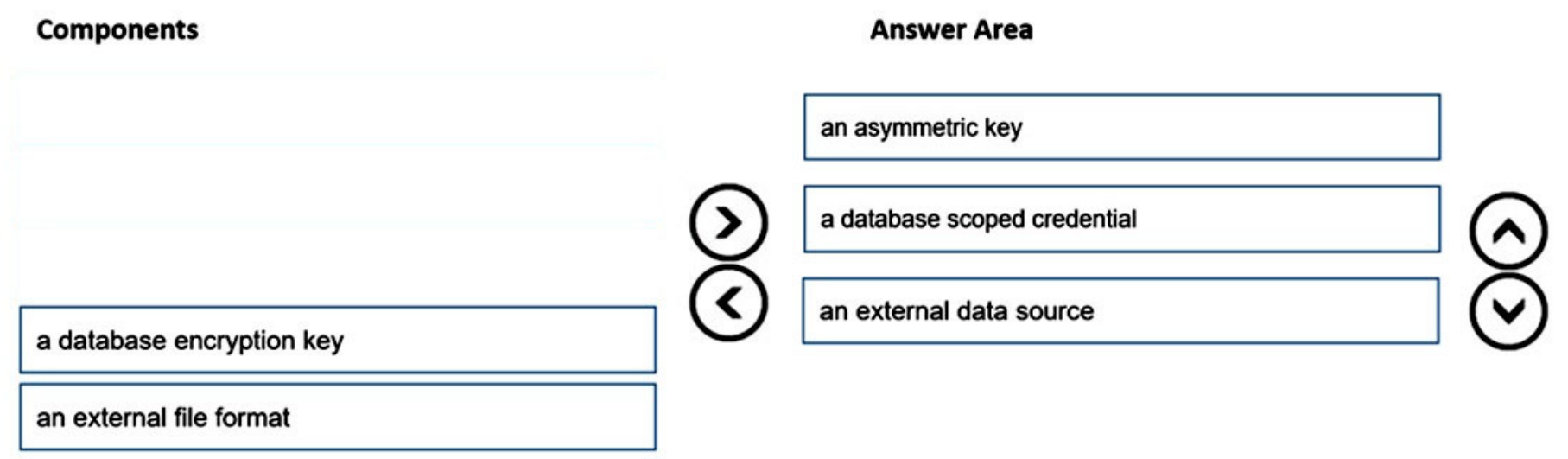
You need to configure PolyBase to connect the data warehouse to storage account.

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

Select and Place:



Correct Answer:



Step 1: an asymmetric key -

A master key should be created only once in a database. The Database Master Key is a symmetric key used to protect the private keys of certificates and asymmetric keys in the database.

Step 2: a database scoped credential

Create a Database Scoped Credential. A Database Scoped Credential is a record that contains the authentication information required to connect an external resource. The master key needs to be created first before creating the database scoped credential.

Step 3: an external data source -

Create an External Data Source. External data sources are used to establish connectivity for data loading using Polybase.

Reference:

https://www.sqlservercentral.com/articles/access-external-data-from-azure-synapse-analytics-using-polybase

Question #47 *Topic 2*

You are monitoring an Azure Stream Analytics job by using metrics in Azure.
You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.
What is a possible cause of this behavior?

A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.

B. There are errors in the input data.

C. The late arrival policy causes events to be dropped.

D. The job lacks the resources to process the volume of incoming data. Most Voted

Correct Answer: *D*
Watermark Delay indicates the delay of the streaming data processing job.
There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:
1. Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming
Units.
2. Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
3. Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling

Community vote distribution

D (100%)

Question #48 *Topic 2*

HOTSPOT -

You are building an Azure Stream Analytics job to retrieve game data.

You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT [                                                    ▼]  as HighestScore
FROM input TIMESTAMP BY CreatedAt
GROUP BY [                                                  ▼]
```

SELECT:
- Collect(Score)
- CollectTop(1) OVER(ORDER BY Score Desc)
- Game, MAX(Score)
- TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)

GROUP BY:
- Game
- Hopping(minute,5)
- Tumbling(minute,5)
- Windows(TumblingWindow(minute,5),Hopping(minute,5))

Correct Answer:

Answer Area

```
SELECT [TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)▼]  as HighestScore
FROM input TIMESTAMP BY CreatedAt
GROUP BY [Hopping(minute,5)                                 ▼]
```

SELECT:
- Collect(Score)
- CollectTop(1) OVER(ORDER BY Score Desc)
- Game, MAX(Score)
- TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)

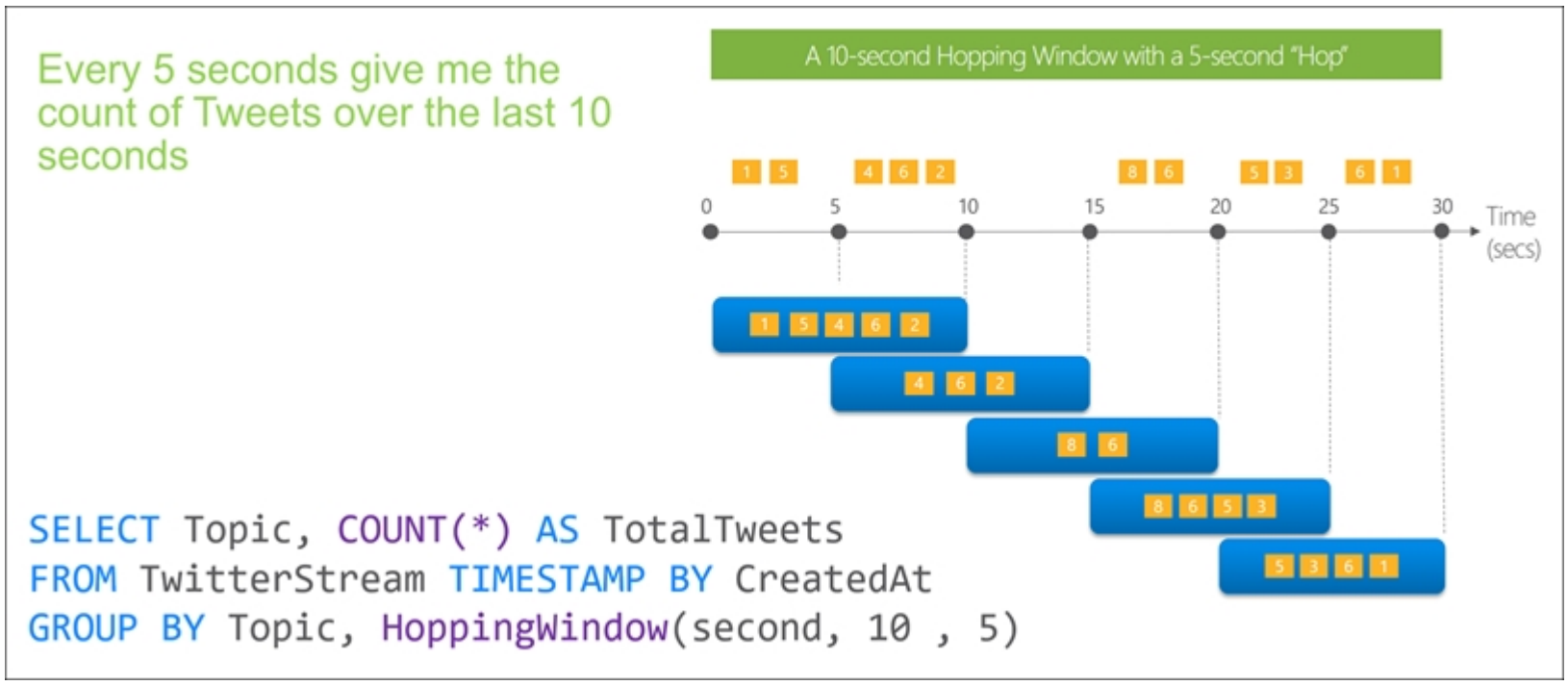
GROUP BY:
- Game
- Hopping(minute,5)
- Tumbling(minute,5)
- Windows(TumblingWindow(minute,5),Hopping(minute,5))

Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)

TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.

Box 2: Hopping(minute,5)

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.



Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/topone-azure-stream-analytics https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

Question #49 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this 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 Azure Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.
Does this meet the goal?

A. Yes

B. No Most Voted

Correct Answer: *A*
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.
Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

Community vote distribution

B (93%) 7%

Question #50 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this 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 create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:
- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:
- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.
Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.
Does this meet the goal?

A. Yes

B. No Most Voted

Correct Answer: *B*
Need a High Concurrency cluster for the jobs.
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

Community vote distribution

B (100%)

Previous Questions Next Questions

Custom View Settings

Question #51 *Topic 2*

You are designing an Azure Databricks cluster that runs user-defined local processes.

You need to recommend a cluster configuration that meets the following requirements:

- Minimize query latency.
- Maximize the number of users that can run queries on the cluster at the same time.
- Reduce overall costs without compromising other requirements.

Which cluster type should you recommend?

A. Standard with Auto Termination

B. High Concurrency with Autoscaling Most Voted

C. High Concurrency with Auto Termination

D. Standard with Autoscaling

Correct Answer: *B*

A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies.

Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.

Incorrect Answers:

C: The cluster configuration includes an auto terminate setting whose default value depends on cluster mode:

Standard and Single Node clusters terminate automatically after 120 minutes by default.

High Concurrency clusters do not terminate automatically by default.

Reference:

https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

Community vote distribution

B (71%) C (29%)

Question #52 *Topic 2*

HOTSPOT -

You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.

The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.

/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}HH}{mm}.json

You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.

How should you configure the sink for the copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

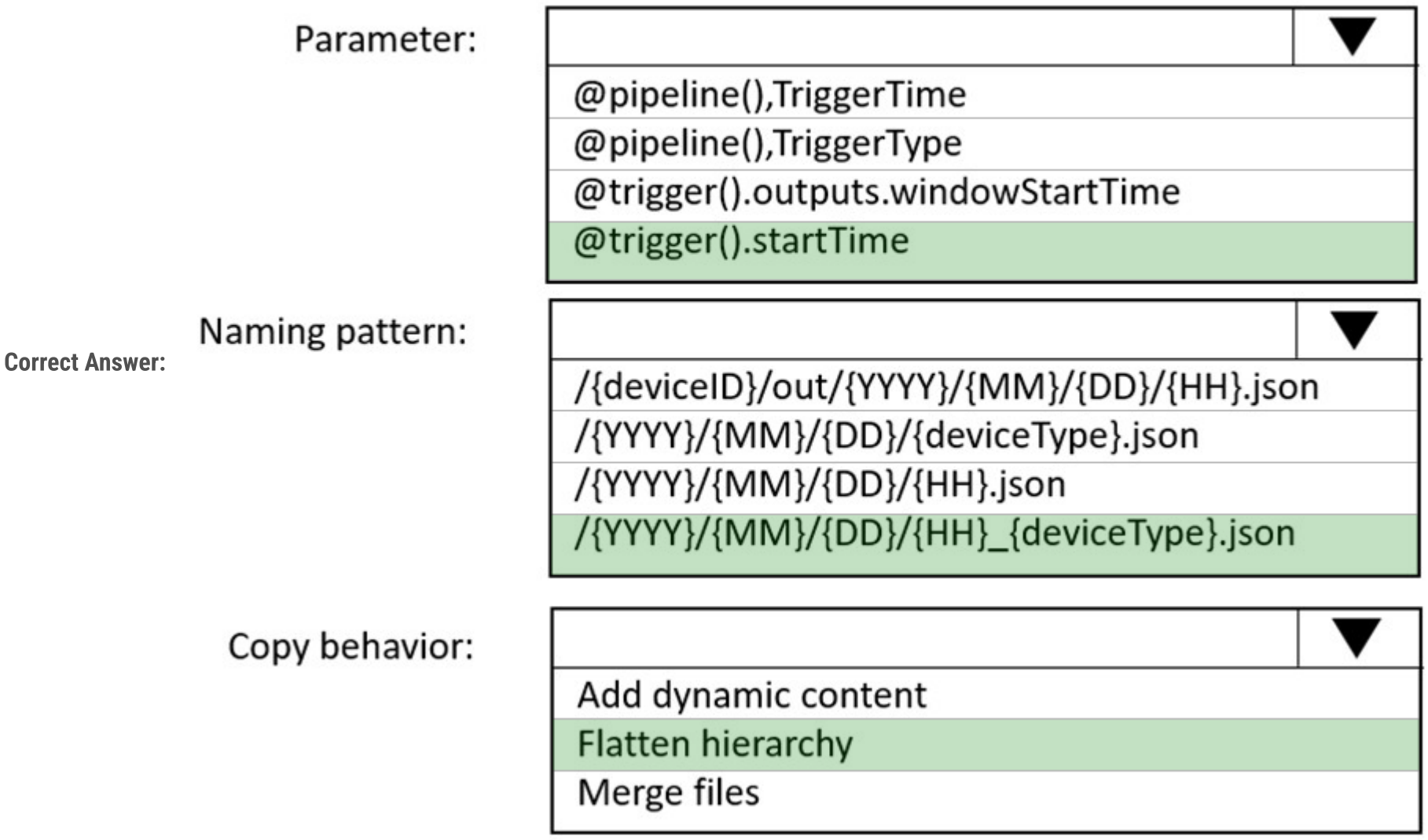
Parameter:
- @pipeline(),TriggerTime
- @pipeline(),TriggerType
- @trigger().outputs.windowStartTime
- @trigger().startTime

Naming pattern:
- /{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json
- /{YYYY}/{MM}/{DD}/{deviceType}.json
- /{YYYY}/{MM}/{DD}/{HH}.json
- /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

Copy behavior:
- Add dynamic content
- Flatten hierarchy
- Merge files

Correct Answer:

Answer Area



Box 1: @trigger().startTime -

startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.

Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

One dataset per hour per deviceType.

Box 3: Flatten hierarchy -

- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system

Question #53 *Topic 2*

DRAG DROP -

You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.

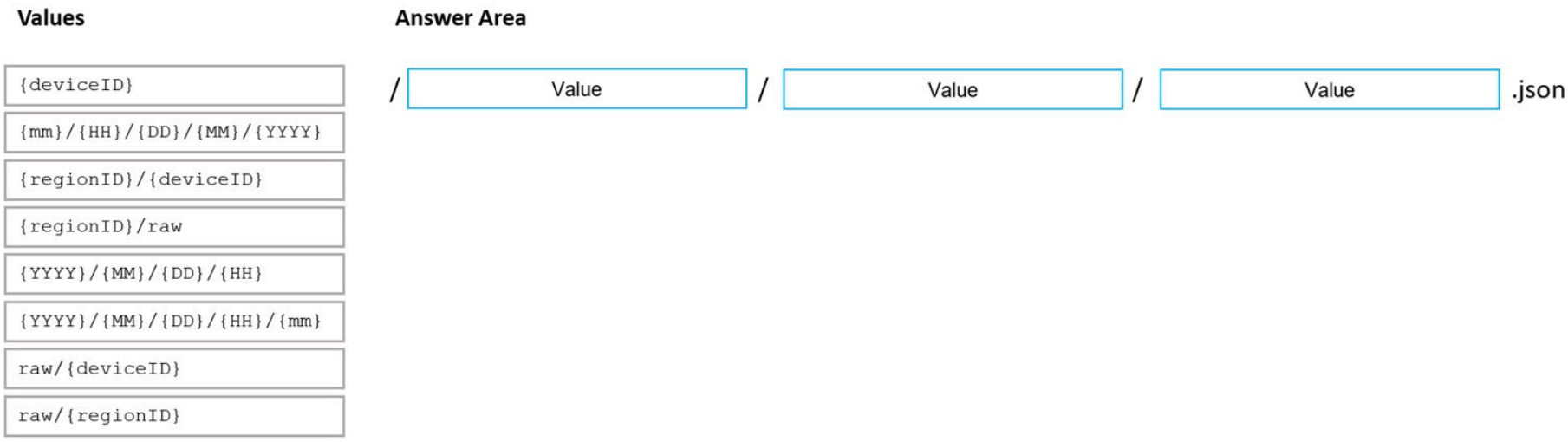
You need to recommend a folder structure for the data. The solution must meet the following requirements:

- Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
- The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.

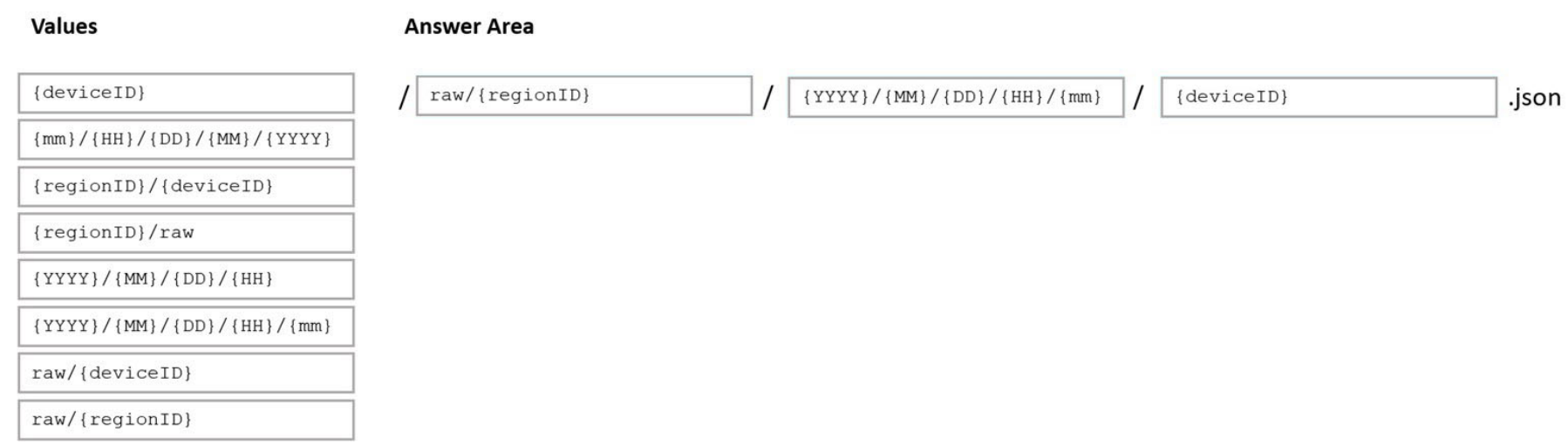
How should you recommend completing the structure? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Correct Answer:



Box 1: {raw/regionID}

Box 2: {YYYY}/{MM}/{DD}/{HH}/{mm}

Box 3: {deviceID}

Reference:

https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md

Question #54 *Topic 2*

HOTSPOT -

You are implementing an Azure Stream Analytics solution to process event data from devices.

The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

A sample of the events is shown in the following table.

DeviceID	EventType	EventTime
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:00.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:05.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	TemperatureSensorFault	2020-12-01T19:07.000Z

You need to calculate the uptime between the faults.

How should you complete the Stream Analytics SQL query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds

FROM input TIMESTAMP BY EventTime
```

[▼]
- WHERE EventType='HeartBeat'
- WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType
- WHERE IsFirst(second,5) = 1

```
GROUP BY

DeviceID
```

[▼]
- ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)
- ,TumblingWindow(second,5)
- HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

Answer Area

Correct Answer:

```
SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds

FROM input TIMESTAMP BY EventTime
```

[▼]
- WHERE EventType='HeartBeat' ✓
- WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType
- WHERE IsFirst(second,5) = 1

```
GROUP BY

DeviceID
```

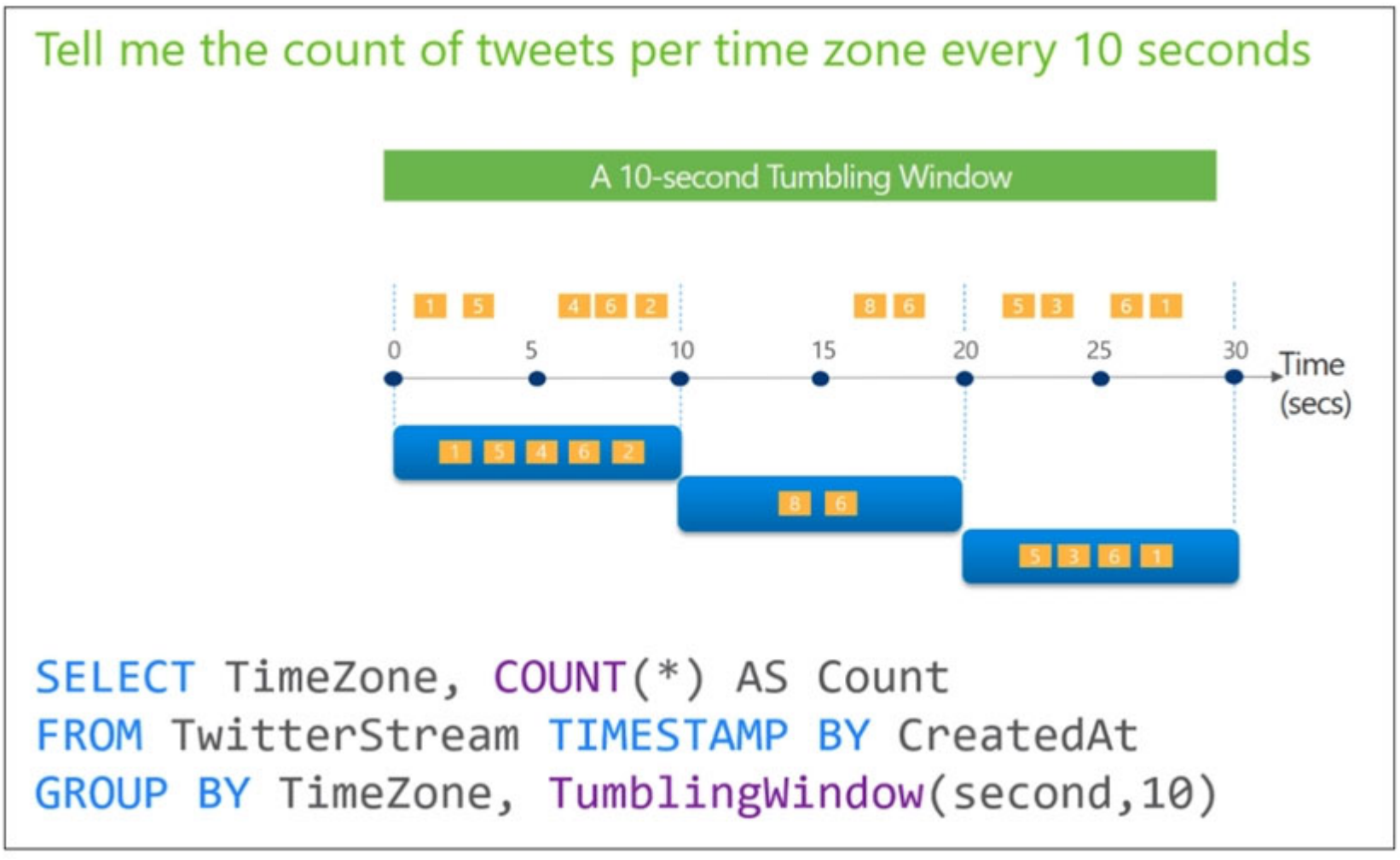
[▼]
- ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)
- ,TumblingWindow(second,5) ✓
- HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

Box 1: WHERE EventType='HeartBeat'

Box 2: ,TumblingWindow(Second, 5)

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.



Incorrect Answers:

,SessionWindow.. : Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Question #55 *Topic 2*

You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL.
Which switch should you use to switch between languages?

A. %<language> Most Voted

B. @<Language >

C. \\[<language >]

D. \\(<language >)

Correct Answer: *A*
To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.
%python //or r, scala, sql
Reference:
https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure-databricks

Community vote distribution

A (100%)

Question #56 *Topic 2*

You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account.
Data to be loaded is identified by a column named LastUpdatedDate in the source table.
You plan to execute the pipeline every four hours.
You need to ensure that the pipeline execution meets the following requirements:
- Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits.
- Supports backfilling existing data in the table.

Which type of trigger should you use?

A. event

B. on-demand

C. schedule

D. tumbling window Most Voted

Correct Answer: *D*
In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

Community vote distribution

D (100%)

Question #57 *Topic 2*

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.
You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point

A. Specify a file naming pattern for the destination. Most Voted

B. Delete the files in the destination before loading the data.

C. Filter by the last modified date of the source files. Most Voted

D. Delete the source files after they are copied.

Correct Answer: *AC*
Copy only the daily files by using filtering.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

Community vote distribution
AC (76%) AD (15%) 9%

Question #58 *Topic 2*

You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.
Which output mode should you use?

A. update

B. complete

C. append

Correct Answer: *C*
Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.
Incorrect Answers:
B: Complete Mode: The entire updated result table is written to external storage. It is up to the storage connector to decide how to handle the writing of the entire table.
A: Update Mode: Only the rows that were updated in the result table since the last trigger are written to external storage. This is different from Complete Mode in that Update Mode outputs only the rows that have changed since the last trigger. If the query doesn't contain aggregations, it is equivalent to Append mode.
Reference:
https://docs.databricks.com/getting-started/spark/streaming.html

Community vote distribution
C (100%)

Question #59 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this 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 Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.
You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of
Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.
Does this meet the goal?

A. Yes Most Voted

B. No

Correct Answer: *A*
Use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

Community vote distribution
A (80%) B (20%)

Question #60 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this 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 Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.
You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of
Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column.
Does this meet the goal?

A. Yes

B. No Most Voted

Correct Answer: *B*
Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

Community vote distribution
B (100%)

Previous Questions Next Questions