VCE & PDF
GeekCert.com

DP-203^{Q&As}

Data Engineering on Microsoft Azure

Pass Microsoft DP-203 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.geekcert.com/dp-203.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Microsoft Official Exam Center

⚙ **Instant Download** After Purchase

⚙ **100% Money Back** Guarantee

⚙ **365 Days** Free Update

⚙ **800,000+** Satisfied Customers

QUESTION 1

You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool.

You plan to create a table named DimProduct.

DimProduct must be a Type 3 slowly changing dimension (SCO) table that meets the following requirements:

The values in two columns named ProductKey and ProductSourceID will remain the same.

The values in three columns named ProductName, ProductDescription, and Color can change.

You need to add additional columns to complete the following table definition.

```
CREATE TABLE [dbo].[dimproduct]
  (
      [ProductKey]          INT NOT NULL,
      [ProductSourceID]     INT NOT NULL,
      [ProductName]         NVARCHAR(100) NOT NULL,
      [ProductDescription]  NVARCHAR(2000) NOT NULL,
      [Color]               NVARCHAR(50) NOT NULL
  )
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
```

A. `[OriginalProductDescription] NVARCHAR(2000) NOT NULL`

B. `[IsCurrentRow] [bit] NOT NULL`

C. `[EffectiveStartDate] [datetime] NOT NULL`

D. `[EffectiveEndDate] [datetime] NOT NULL`

E. `[OriginalProductName] NVARCHAR(100) NULL`

F. `[OriginalColor] NVARCHAR(50) NOT NULL`

A. Option A

B. Option B

C. Option C

D. Option D

E. Option E

F. Option F

Correct Answer: ABC

QUESTION 2

HOTSPOT

You need to output files from Azure Data Factory.

Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Columnar format:

| Avro |
| GZip |
| Parquet |
| TXT |

JSON with a timestamp:

| Avro |
| GZip |
| Parquet |
| TXT |

Correct Answer:

Answer Area

Columnar format:

- Avro
- GZip
- **Parquet**
- TXT

JSON with a timestamp:

- **Avro**
- GZip
- Parquet
- TXT

Box 1: Parquet

Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy

transactional workloads.

Box 2: Avro

An Avro schema is created using JSON format.

AVRO supports timestamps.

Note: Azure Data Factory supports the following file formats (not GZip or TXT).

Avro format

Binary format

Delimited text format

Excel format

JSON format

ORC format

Parquet format

XML format

Reference:

https://www.datanami.com/2018/05/16/big-data-file-formats-demystified

---

QUESTION 3

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 he 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 4

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:

Answer Area

Load methodology:

Full Load
Incremental Load
Load individual files as they arrive

Trigger:

Fixed schedule
New file
Tumbling window

Correct Answer:

Answer Area

Load methodology:

Full Load
Incremental Load
Load individual files as they arrive

Trigger:

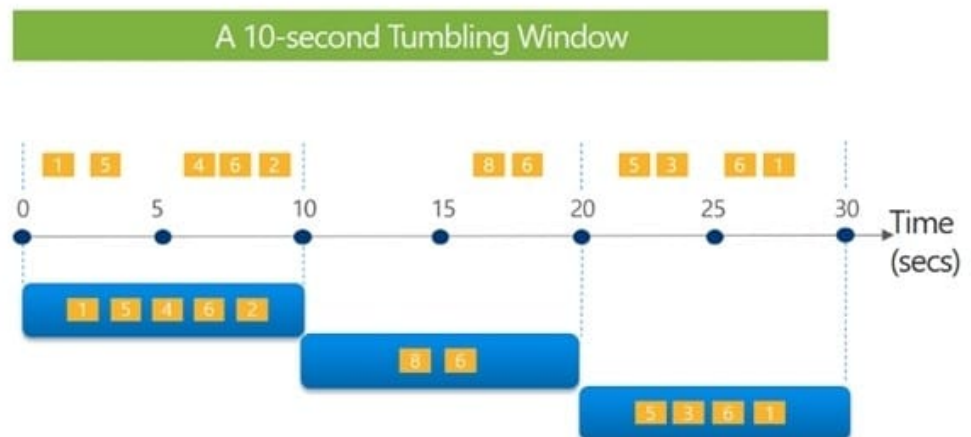| |
|---|
| Fixed schedule |
| New file |
| Tumbling window |

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.

Tell me the count of tweets per time zone every 10 seconds

A 10-second Tumbling Window

```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

QUESTION 5

HOTSPOT

You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytics in a star schema.

You plan to have a fact table for website visits. The table will be approximately 5 GB. You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Distribution:

Hash
Round robin
Replicated

Index:

Clustered columnstore
Clustered
Nonclustered

Correct Answer:

Distribution:

Hash
Round robin
Replicated

Index:

Clustered columnstore
Clustered
Nonclustered

Box 1: Hash

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations.

Box 2: Clustered columnstore

Clustered columnstore tables offer both the highest level of data compression and the best overall query performance.

[DP-203 VCE Dumps](https://www.geekcert.com/dp-203.html)   [DP-203 Practice Test](https://www.geekcert.com/dp-203.html)   [DP-203 Exam Questions](https://www.geekcert.com/dp-203.html)