

# Exam Questions DP-203

Data Engineering on Microsoft Azure

<https://www.2passeasy.com/dumps/DP-203/>



### NEW QUESTION 1

- (Exam Topic 1)

You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements. What should you create?

- A. a table that has an IDENTITY property
- B. a system-versioned temporal table
- C. a user-defined SEQUENCE object
- D. a table that has a FOREIGN KEY constraint

**Answer:** A

#### Explanation:

Scenario: Implement a surrogate key to account for changes to the retail store addresses.

A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity>

### NEW QUESTION 2

- (Exam Topic 3)

You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

- > Is partitioned by month
- > Contains one billion rows
- > Has clustered columnstore indexes

At the beginning of each month, you need to remove data from SalesFact that is older than 36 months as quickly as possible.

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

#### Actions

#### Answer Area

Switch the partition containing the stale data from SalesFact to SalesFact\_Work.

Truncate the partition containing the stale data.

Drop the SalesFact\_Work table.

Create an empty table named SalesFact\_Work that has the same schema as SalesFact.

Execute a DELETE statement where the value in the Date column is more than 36 months ago.

Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

- A. Mastered
- B. Not Mastered

**Answer:** A

#### Explanation:

Step 1: Create an empty table named SalesFact\_work that has the same schema as SalesFact. Step 2: Switch the partition containing the stale data from SalesFact to SalesFact\_Work.

SQL Data Warehouse supports partition splitting, merging, and switching. To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions match.

Loading data into partitions with partition switching is a convenient way stage new data in a table that is not visible to users the switch in the new data.

Step 3: Drop the SalesFact\_Work table. Reference:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-partition>

### NEW QUESTION 3

- (Exam Topic 3)

You have several Azure Data Factory pipelines that contain a mix of the following types of activities.

- \* Wrangling data flow
- \* Notebook
- \* Copy
- \* jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution NOTE: Each correct selection is worth one

point.

- A. Azure HDInsight
- B. Azure Databricks
- C. Azure Machine Learning
- D. Azure Data Factory
- E. Azure Synapse Analytics

**Answer:** CE

#### NEW QUESTION 4

- (Exam Topic 3)

What should you recommend using to secure sensitive customer contact information?

- A. data labels
- B. column-level security
- C. row-level security
- D. Transparent Data Encryption (TDE)

**Answer:** B

#### Explanation:

Scenario: All cloud data must be encrypted at rest and in transit.

Always Encrypted is a feature designed to protect sensitive data stored in specific database columns from access (for example, credit card numbers, national identification numbers, or data on a need to know basis). This includes database administrators or other privileged users who are authorized to access the database to perform management tasks, but have no business need to access the particular data in the encrypted columns. The data is always encrypted, which means the encrypted data is decrypted only for processing by client applications with access to the encryption key.

References:

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview>

#### NEW QUESTION 5

- (Exam Topic 3)

You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:

Data storage:

- Serve as a repository (or high volumes of large files in various formats.
- Implement optimized storage for big data analytics workloads.
- Ensure that data can be organized using a hierarchical structure. Batch processing:
- Use a managed solution for in-memory computation processing.
- Natively support Scala, Python, and R programming languages.
- Provide the ability to resize and terminate the cluster automatically. Analytical data store:
- Support parallel processing.
- Use columnar storage.
- Support SQL-based languages.

You need to identify the correct technologies to build the Lambda architecture.

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

Architecture requirement	Technology
Data storage	<div><div></div><div><div>Azure SQL Database</div><div>Azure Blob Storage</div><div>Azure Cosmos DB</div><div>Azure Data Lake Store</div></div></div>
Batch processing	<div><div></div><div><div>HDInsight Spark</div><div>HDInsight Hadoop</div><div>Azure Databricks</div><div>HDInsight Interactive Query</div></div></div>
Analytical data store	<div><div></div><div><div>HDInsight HBase</div><div>Azure SQL Data Warehouse</div><div>Azure Analysis Services</div><div>Azure Cosmos DB</div></div></div>

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Data storage: Azure Data Lake Store

A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.

Batch processing: HD Insight Spark

Apache Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications.

HDInsight is a managed Hadoop service. Use it to deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.

Languages: R, Python, Java, Scala, SQL Analytic data store: SQL Data Warehouse

SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP).

SQL Data Warehouse stores data into relational tables with columnar storage. References:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespaces> <https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing> <https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is>

**NEW QUESTION 6**

- (Exam Topic 3)

What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

- A. a server-level virtual network rule
- B. a database-level virtual network rule
- C. a database-level firewall IP rule
- D. a server-level firewall IP rule

**Answer: A**

**Explanation:**

Virtual network rules are one firewall security feature that controls whether the database server for your single databases and elastic pool in Azure SQL Database or for your databases in SQL Data Warehouse accepts communications that are sent from particular subnets in virtual networks.

Server-level, not database-level: Each virtual network rule applies to your whole Azure SQL Database server, not just to one particular database on the server. In other words, virtual network rule applies at the server level, not at the database level.

References:

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview>

**NEW QUESTION 7**

- (Exam Topic 3)

You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

- Contain sales data for 20,000 products.
- Use hash distribution on a column named ProductID.
- Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance of the clustered columnstore index?

- A. 40
- B. 240
- C. 400
- D. 2,400

**Answer: B**

**NEW QUESTION 8**

- (Exam Topic 3)

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, 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 Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB. Does this meet the goal?

- A. Yes
- B. No

**Answer: B**

**Explanation:**

Instead modify the files to ensure that each row is less than 1 MB. References:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data>

**NEW QUESTION 9**

- (Exam Topic 3)

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this 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 High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.  
Does this meet the goal?

- A. Yes
- B. No

**Answer: B**

**Explanation:**

We would need a High Concurrency cluster for 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>

**NEW QUESTION 10**

- (Exam Topic 3)

You plan to implement an Azure Data Lake Gen2 storage account.  
You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.  
Which type of replication should you use for the storage account?

- A. geo-redundant storage (GRS)
- B. zone-redundant storage (ZRS)
- C. locally-redundant storage (LRS)
- D. geo-zone-redundant storage (GZRS)

**Answer: A**

**Explanation:**

Geo-redundant storage (GRS) copies your data synchronously three times within a single physical location in the primary region using LRS. It then copies your data asynchronously to a single physical location in the secondary region.  
Reference:  
<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

**NEW QUESTION 10**

- (Exam Topic 3)

You plan to create an Azure Synapse Analytics dedicated SQL pool.  
You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queries.  
Which two components should you include in the solution? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. sensitivity-classification labels applied to columns that contain confidential information
- B. resource tags for databases that contain confidential information
- C. audit logs sent to a Log Analytics workspace
- D. dynamic data masking for columns that contain confidential information

**Answer: AC**

**Explanation:**

A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:



Home > MySampleDatabase2 (mydocsamplesqlserver/MySampleDatabase2)

MySampleDatabase2 (mydocsamplesqlserver/MySampleDatabase2) | Data Discovery & Classification

SQL database

Search (Ctrl+/)

Save Discard + Add classification Feedback

Power Platform

- Power BI (preview)
- Power Apps (preview)
- Power Automate (preview)

Settings

- Configure
- Geo-Replication
- Connection strings
- Sync to other databases
- Add Azure Search
- Properties
- Locks

Integrations

- Stream analytics (preview)

Security

- Auditing
- Data Discovery & Classification
- Dynamic Data Masking
- Security Center
- Transparent data encryption

Intelligent Performance

- Performance overview

Overview Classification

15 columns with classification recommendations (Click to minimize)

Accept selected recommendations Dismiss selected recommendations Show dismissed recommendations

Select all Schema: 2 selected Table: 5 selected Filter by column

	Schema	Table	Column
<input type="checkbox"/>	SalesLT	Customer	FirstName
<input type="checkbox"/>	SalesLT	Customer	LastName
<input type="checkbox"/>	SalesLT	Customer	EmailAddress
<input type="checkbox"/>	SalesLT	Customer	Phone
<input type="checkbox"/>	SalesLT	Customer	PasswordHash
<input type="checkbox"/>	SalesLT	Customer	PasswordSalt
<input type="checkbox"/>	dbo	ErrorLog	UserName
<input type="checkbox"/>	SalesLT	Address	AddressLine1
<input type="checkbox"/>	SalesLT	Address	AddressLine2
<input type="checkbox"/>	SalesLT	Address	City
<input type="checkbox"/>	SalesLT	Address	PostalCode
<input type="checkbox"/>	SalesLT	CustomerAddress	AddressType
<input type="checkbox"/>	SalesLT	SalesOrderHeader	AccountNumber
<input type="checkbox"/>	SalesLT	SalesOrderHeader	CreditCardApprovalCode
<input type="checkbox"/>	SalesLT	SalesOrderHeader	TaxAmt

- > Select Add classification in the top menu of the pane.
- > In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
- > Select Add classification at the bottom of the context window.

C: An important aspect of the information-protection paradigm is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called data\_sensitivity\_information. This field logs the sensitivity classifications (labels) of the data that was returned by a query. Here's an example:

d	client_ip	application_name	duration_milliseconds	response_rows	affected_rows	connection_id	data_sensitivity_information
	7.125	Microsoft SQL Server Management Studio - Query	1	847	847	C244A066-2271-...	Confidential - GDPR
	7.125	Microsoft SQL Server Management Studio - Query	2	32	32	C244A066-2271-...	Confidential
	7.125	Microsoft SQL Server Management Studio - Query	41	32	32	A7088FD4-759E-...	Confidential, Confidential - GDPR

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview>

## NEW QUESTION 15

- (Exam Topic 3)

You are planning the deployment of Azure Data Lake Storage Gen2. You have the following two reports that will access the data lake:

- > Report1: Reads three columns from a file that contains 50 columns.
- > Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

Report1: 

	▼
Avro	
CSV	
Parquet	
TSV	

Report2: 

	▼
Avro	
CSV	
Parquet	
TSV	

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Report1: CSV

CSV: The destination writes records as delimited data. Report2: AVRO

AVRO supports timestamps.

Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2. Reference:

<https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2>

**NEW QUESTION 17**

- (Exam Topic 3)

You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.

You create five clones of PL1. You configure each clone pipeline to use a different data source.

You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1. What should you do?

- A. Add a new trigger to each cloned pipeline
- B. Associate each cloned pipeline to an existing trigger.
- C. Create a tumbling window trigger dependency for the trigger of PL1.
- D. Modify the Concurrency setting of each pipeline.

**Answer:** B

**NEW QUESTION 18**

- (Exam Topic 3)

You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

- EventDate: 1 million per day
- EventTypeID: 10 million per event type
- WarehouseID: 100 million per warehouse
- ProductCategoryTypeID: 25 million per product category type

You identify the following usage patterns:  
 Analyst will most commonly analyze transactions for a warehouse.

Queries will summarize by product category type, date, and/or inventory event type. You need to recommend a partition strategy for the table to minimize query times. On which column should you recommend partitioning the table?

- A. ProductCategoryTypeID
- B. EventDate
- C. WarehouseID
- D. EventTypeID

**Answer:** D

**NEW QUESTION 20**

- (Exam Topic 3)

You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.

You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information.

What should you include in the recommendation?

- A. data masking
- B. Always Encrypted
- C. column-level security
- D. row-level security

**Answer:** A

**Explanation:**

SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users. The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card.

Example: XXXX-XXXX-XXXX-1234

Reference:

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started>

**NEW QUESTION 24**

- (Exam Topic 3)

You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool. You create a table by using the Transact-SQL statement shown in the following exhibit.

```
CREATE TABLE [DBO].[DimProduct] (
    [ProductKey] [int] IDENTITY(1,1) NOT NULL,
    [ProductSourceID] [int] NOT NULL,
    [ProductName] [nvarchar](100) NOT NULL,
    [ProductNumber] [nvarchar](25) NOT NULL,
    [Color] [nvarchar](15) NULL,
    [Size] [nvarchar](5) NULL,
    [Weight] [decimal](8, 2) NULL,
    [ProductCategory] [nvarchar](100) NULL,
    [SellStartDate] [date] NOT NULL,
    [SellEndDate] [date] NULL,
    [RowInsertedDateTime] [datetime] NOT NULL,
    [RowUpdatedDateTime] [datetime] NOT NULL,
    [ETLAuditID] [int] NOT NULL
)
```

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.

DimProduct is a **[answer choice]** slowly changing dimension (SCD).

	▼
Type 0	
Type 1	
Type 2	

The ProductKey column is **[answer choice]**.

	▼
a surrogate key	
a business key	
an audit column	

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: Type 2

A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.

Reference:

<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics>

**NEW QUESTION 29**



- (Exam Topic 3)

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.

**Values**

CASE
ELSE
OVER
PARTITION
ROW\_NUMBER

**Answer Area**

SELECT
\*,
Value
WHEN hire\_date >= '2019-01-01' THEN
'New' Value 'Standard'
END AS employee\_type
FROM
employees;

- A. Mastered
- B. Not Mastered

**Answer: A**

**Explanation:**

**Values**

CASE
ELSE
OVER
PARTITION
ROW\_NUMBER

**Answer Area**

SELECT
\*,
CASE
WHEN hire\_date >= '2019-01-01' THEN
'New' PARTITION 'Standard'
END AS employee\_type
FROM
employees;

#### NEW QUESTION 34

- (Exam Topic 3)

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.
- D. From Azure DevOps, update the main branch.

**Answer: C**

**Explanation:**

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.

- > In Azure DevOps, open the project that's configured with your data factory.
  - > On the left side of the page, select Pipelines, and then select Releases.
  - > Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
  - > In the Stage name box, enter the name of your environment.
  - > 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.
- > Select the Empty job template. Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment>

### NEW QUESTION 37

- (Exam Topic 3)

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 into Table1 and azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files 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.

A. Yes

B. No

**Answer: B**

### NEW QUESTION 39

- (Exam Topic 3)

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4.KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing. What should you do?

A. Compress the files.

B. Merge the files.

C. Convert the files to JSON

D. Convert the files to Avro.

**Answer: D**

### NEW QUESTION 40

- (Exam Topic 3)

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 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.

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

LAG
LAST
LEAD

LIMIT DURATION
OFFSET
WHEN

A. Mastered

B. Not Mastered

**Answer: A**

#### Explanation:

Box 1: LAG

The LAG analytic operator allows one to look up a “previous” 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>

### NEW QUESTION 42

- (Exam Topic 3)

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

**Answer: AB**

**Explanation:**

B: Example:

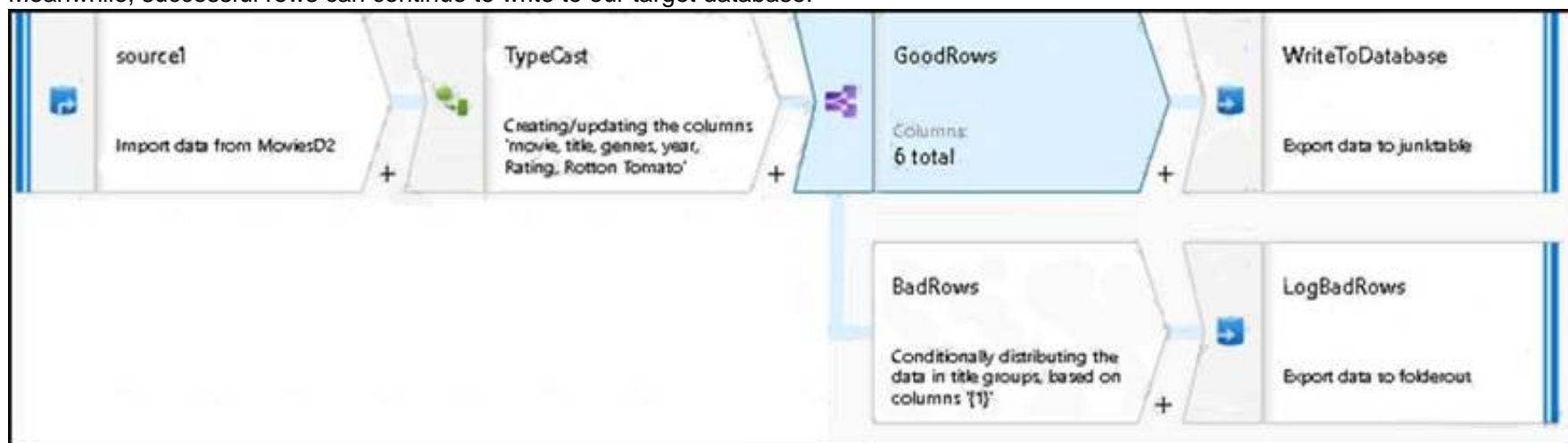
- \* 1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

- \* 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

**NEW QUESTION 47**

- (Exam Topic 3)

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics.

You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest

possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained. What should you recommend?

- A. Parquet
- B. Avro
- C. CSV
- D. JSON

**Answer:** B

**Explanation:**

The Avro format is great for data and message preservation. Avro schema with its support for evolution is essential for making the data robust for streaming architectures like Kafka, and with the metadata that schema provides, you can reason on the data. Having a schema provides robustness in providing meta-data about the data stored in Avro records which are self- documenting the data. References: <http://cloudurable.com/blog/avro/index.html>

**NEW QUESTION 50**

- (Exam Topic 3)

You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account. You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once.

Which windowing function should you use?

- A. a five-minute Session window
- B. a five-minute Sliding window
- C. a five-minute Tumbling window
- D. a five-minute Hopping window that has one-minute hop

**Answer:** C

**Explanation:**

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.

References:

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

**NEW QUESTION 51**

- (Exam Topic 3)

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 into Table1 and azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files 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 Get Metadata activity that retrieves the DateTime of the files.

Does this meet the goal?

- A. Yes
- B. No

**Answer:** B

**NEW QUESTION 53**

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.

You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.

What should you include in the solution?

- A. a default value
- B. dynamic data masking
- C. row-level security (RLS)
- D. column encryption
- E. table partitions

**Answer:** C

**NEW QUESTION 58**

.....



## THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual DP-203 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the DP-203 Product From:

<https://www.2passeasy.com/dumps/DP-203/>

## Money Back Guarantee

### DP-203 Practice Exam Features:

- \* DP-203 Questions and Answers Updated Frequently
- \* DP-203 Practice Questions Verified by Expert Senior Certified Staff
- \* DP-203 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* DP-203 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year