



Exam DP-203

Data Engineering on Microsoft Azure

Version: 20.0

[Total Questions: 303]



Topic 1, Contoso Case StudyTransactional Data

Contoso has three years of customer, transactional, operation, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL server instances contain data from various operational systems. The data is loaded into the instances by using SQL server integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time period. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops and Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes

Contoso plans to implement the following changes:

- * Load the sales transaction dataset to Azure Synapse Analytics.
- * Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.
- * Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

- Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
- Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
- Implement a surrogate key to account for changes to the retail store addresses.
- Ensure that data storage costs and performance are predictable.



- Minimize how long it takes to remove old records. Customer Sentiment Analytics Requirement

Contoso identifies the following requirements for customer sentiment analytics:

- Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own AzureAD credentials.
- Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.
- Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.
- Ensure that the data store supports Azure AD-based access control down to the object level.
- Minimize administrative effort to maintain the Twitter feed data records.
- Purge Twitter feed data records; if they are older than two years.

Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version controlled and developed independently by multiple data engineers.

1. - (Exam Topic 1)

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

- A. change feed
- B. soft delete
- C. time-based retention
- D. lifecycle management

Answer: D

Explanation:

Scenario: Purge Twitter feed data records that are older than two years.

Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

2. - (Exam Topic 1)

You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.

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.

Answer Area

To increase the throughput of ingesting the Twitter feeds:

- Configure Event Hubs partitions.
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

To store the Twitter feed data, use:

- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Answer:

Answer Area

To increase the throughput of ingesting the Twitter feeds:

- Configure Event Hubs partitions.
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

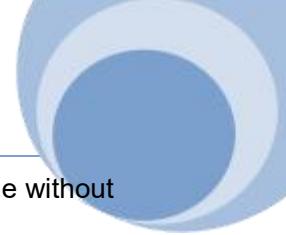
To store the Twitter feed data, use:

- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Explanation:

Graphical user interface, text Description automatically generated

Box 1: Configure Event Hubs partitions



Scenario: Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.

Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose. The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.

Box 2: An Azure Data Lake Storage Gen2 account

Scenario: Ensure that the data store supports Azure AD-based access control down to the object level.

Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features>

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

3. - (Exam Topic 1)

You need to integrate the on-premises data sources and Azure Synapse Analytics. The solution must meet the data integration requirements.

Which type of integration runtime should you use?

- A. Azure-SSIS integration runtime
- B. self-hosted integration runtime
- C. Azure integration runtime

Answer: C

4. - (Exam Topic 1)

You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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.

**Answer Area:**

Table type to store the product sales transactions:

Hash
Round-robin
Replicated

When creating the table for sales transactions:

Configure a clustered index.
Set the distribution column to product ID.
Set the distribution column to the sales date.

Answer:**Answer Area:**

Table type to store the product sales transactions:

Hash
Round-robin
Replicated

When creating the table for sales transactions:

Configure a clustered index.
Set the distribution column to product ID.
Set the distribution column to the sales date.

Explanation:

Graphical user interface, text, application, chat or text message Description automatically generated

Box 1: Hash Scenario:

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

A hash distributed table can deliver the highest query performance for joins and aggregations on large tables. Box 2: Set the distribution column to the sales date.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Reference:

<https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/>

5. - (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>

6. - (Exam Topic 1)

You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.

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

Actions	Answer Area
Publish changes.	
Create a feature branch.	
Merge changes.	>
Create a repository and a main branch.	<
Create a pull request.	

Answer:

Actions	Answer Area
Publish changes.	
Create a feature branch.	
Merge changes.	>
Create a repository and a main branch.	<
Create a pull request.	

Explanation:



Graphical user interface, application Description automatically generated

Scenario: Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app. Step 2: Create a feature branch

Step 3: Create a pull request Step 4: Merge changes

Merge feature branches into the main branch using pull requests. Step 5: Publish changes

Reference:

<https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git>

7. - (Exam Topic 1)

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

- A. time-based retention
- B. change feed
- C. soft delete
- D. lifecycle management

Answer: D

8. - (Exam Topic 1)

You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.

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.



Table type to store retail store data:

Hash
Replicated
Round-robin

Table type to store promotional data:

Hash
Replicated
Round-robin

Answer:

Table type to store retail store data:

Hash
Replicated
Round-robin

Table type to store promotional data:

Hash
Replicated
Round-robin

Explanation:

Graphical user interface, text, application, table Description automatically generated

Box 1: Round-robin

Round-robin tables are useful for improving loading speed.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.

Box 2: Hash

Hash-distributed tables improve query performance on large fact tables. Reference:

[https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribution)

9. - (Exam Topic 1)

You need to implement an Azure Synapse Analytics database object for storing the sales transactions data.

The solution must meet the sales transaction dataset requirements.

What solution must meet the sales transaction dataset requirements.

What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct



selection is worth one point.

Transact-SQL DDL command to use:

CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW

Partitioning option to use in the WITH clause of the DDL statement:

FORMAT_OPTIONS
FORMAT_TYPE
RANGE LEFT FOR VALUES
RANGE RIGHT FOR VALUES

Answer:

Transact-SQL DDL command to use:

CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW

Partitioning option to use in the WITH clause of the DDL statement:

FORMAT_OPTIONS
FORMAT_TYPE
RANGE LEFT FOR VALUES
RANGE RIGHT FOR VALUES

Explanation:

Graphical user interface, text, application, table Description automatically generated

Box 1: Create table

Scenario: Load the sales transaction dataset to Azure Synapse Analytics Box 2: RANGE RIGHT FOR VALUES

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values). FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.

Scenario: Load the sales transaction dataset to Azure Synapse Analytics. Contoso identifies the following requirements for the sales transaction dataset:

- Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
- Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.





- Implement a surrogate key to account for changes to the retail store addresses.
- Ensure that data storage costs and performance are predictable.
- Minimize how long it takes to remove old records. Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse>

10. - (Exam Topic 1)

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements.

Which three Transaction-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.

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

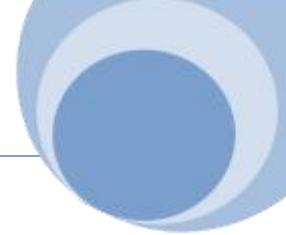
Commands	Answer Area
CREATE EXTERNAL DATA SOURCE	
CREATE EXTERNAL FILE FORMAT	
CREATE EXTERNAL TABLE	
CREATE EXTERNAL TABLE AS SELECT	
CREATE DATABASE SCOPED CREDENTIAL	

Answer:

Commands	Answer Area
CREATE EXTERNAL DATA SOURCE	
CREATE EXTERNAL FILE FORMAT	
CREATE EXTERNAL TABLE	
CREATE EXTERNAL TABLE AS SELECT	
CREATE DATABASE SCOPED CREDENTIAL	

Explanation:

Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level



security (RLS). The users must be authenticated by using their own Azure AD credentials.

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts. Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage. Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

11. - (Exam Topic 1)

You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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.

Partition product sales transactions data by:

Sales date
Product ID
Promotion ID

Store product sales transactions data in:

An Azure Synapse Analytics dedicated SQL pool
An Azure Synapse Analytics serverless SQL pool
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

Answer:



Partition product sales transactions data by:

Sales date
Product ID
Promotion ID

Store product sales transactions data in:

An Azure Synapse Analytics dedicated SQL pool
An Azure Synapse Analytics serverless SQL pool
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

Explanation:

Box 1: Sales date

Scenario: Contoso requirements for data integration include:

- Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Box 2: An Azure Synapse Analytics Dedicated SQL pool Scenario: Contoso requirements for data integration include:

- Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).

Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces the data storage costs, and improves query performance.

Synapse analytics dedicated sql pool Reference:

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

Topic 2, Litware, inc.

Case study

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.



To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

- See inventory levels across the stores. Data must be updated as close to real time as possible.
- Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.
- Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:



- Minimize the number of different Azure services needed to achieve the business goals.
- Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.
- Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
- Use Azure Active Directory (Azure AD) authentication whenever possible.
- Use the principle of least privilege when designing security.
- Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.
- Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.
- Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

- The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.
- Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Daily inventory data comes from a Microsoft SQL server located on a private network.
- Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.
- Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.
- Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network





and Azure.

1. - (Exam Topic 2)

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 serverlevel, not at the database-level.

References:

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

2. - (Exam Topic 2)

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Integration runtime type:	Azure integration runtime Azure-SSIS integration runtime Self-hosted integration runtime
Trigger type:	Event-based trigger Schedule trigger Tumbling window trigger
Activity type:	Copy activity Lookup activity Stored procedure activity

Answer:

Integration runtime type:	Azure integration runtime Azure-SSIS integration runtime Self-hosted integration runtime
Trigger type:	Event-based trigger Schedule trigger Tumbling window trigger
Activity type:	Copy activity Lookup activity Stored procedure activity

Explanation:

Box 1: Self-hosted integration runtime

A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.

Box 2: Schedule trigger Schedule every 8 hours Box 3: Copy activity Scenario:

- Customer data, including name, contact information, and loyalty number, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

3. - (Exam Topic 2)



What should you do to improve high availability of the real-time data processing solution?

- A. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.
- B. Deploy a High Concurrency Databricks cluster.
- C. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- D. Set Data Lake Storage to use geo-redundant storage (GRS).

Answer: A

Explanation:

Guarantee Stream Analytics job reliability during service updates

Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability>

4. - (Exam Topic 2)

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>

Topic 3, Mix Questions

1. - (Exam Topic 3)

You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data. You need to convert a nested JSON string into a DataFrame that will contain multiple rows. Which Spark SQL function should you use?

- A. explode
- B. filter
- C. coalesce
- D. extract

Answer: A

Explanation:

Convert nested JSON to a flattened DataFrame

You can flatten nested JSON, using only `$"column.*"` and `explode` methods. Note: Extract and flatten Use `$"column.*"` and `explode` methods to flatten the struct and array types before displaying the flattened DataFrame.

Scala

```
display(DF.select($"id" as "main_id",$"name",$"batters",$"ppu",explode($"topping")) // Exploding the
topping column using explode as it is an array type
withColumn("topping_id",$"col.id") // Extracting topping_id from col using DOT form
withColumn("topping_type",$"col.type") // Extracting topping_tytype from col using DOT form drop($"col")
select($"**",$"batters.*") // Flattened the struct type batters tto array type which is batter drop($"batters")
select($"**",explode($"batter")) drop($"batter")
```



```
withColumn("batter_id","$col.id") // Extracting batter_id from col using DOT form  
withColumn("batter_type","$col.type") // Extracting batter_type from col using DOT form drop($"col")  
)
```

Reference: <https://learn.microsoft.com/en-us/azure/databricks/kb-scala/flatten-nested-columns-dynamically>

2. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named dbo.Users.

You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?

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

Answer: B

3. - (Exam Topic 3)

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 queues on the cluster at the same time « Reduce overall costs without compromising other requirements

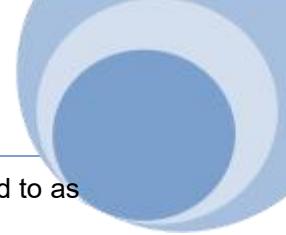
Which cluster type should you recommend?

- A. Standard with Auto termination
- B. Standard with Autoscaling
- C. High Concurrency with Autoscaling
- D. High Concurrency with Auto Termination

Answer: C

Explanation:

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.

Reference:

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

4. - (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: C E

5. - (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

Explanation:

Instead use a serverless SQL pool to create an external table with the extra column. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

6. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains a table named Sales. Sales has row-level security (RLS) applied. RLS uses the following predicate filter.

```
CREATE FUNCTION Security.fn_securiypredicate(@SalesRep AS sysname)
    RETURNS TABLE
WITH SCHEMABINDING
AS
    RETURN SELECT 1 AS fn_securiypredicate_result
WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';
```

A user named SalesUser1 is assigned the db_datareader role for Pool1.

A user named SalesUser1 is assigned the db_datareader role for Pool1. Which rows in the Sales table are returned when SalesUser1 queries the table?

- A. only the rows for which the value in the User_Name column is SalesUser1
- B. all the rows
- C. only the rows for which the value in the SalesRep column is Manager
- D. only the rows for which the value in the SalesRep column is SalesUser1

Answer: A

7. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1. Container1 contains a directory named directory1. Directory1 contains a file named file1.



You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.

You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources.

Each permission 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.

Permissions	Answer Area
Read	container1: <input type="text"/>
Write	directory1: <input type="text"/>
Execute	file1: <input type="text"/>

Answer:

Permissions	Answer Area
Read	container1: <input type="text"/> Execute
Write	directory1: <input type="text"/> Execute
Execute	file1: <input type="text"/> Write

Explanation:

Box 1: Execute

If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file.

Box 2: Execute

On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write

On file: Write (W): Can write or append to a file. Reference:



<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

8. - (Exam Topic 3)

You have an Azure Data Lake Storage account that contains a staging zone.

You need to design a dairy 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

Answer: A

Explanation:

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>

9. - (Exam Topic 3)

You manage an enterprise data warehouse in Azure Synapse Analytics.

Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.

You need to monitor resource utilization to determine the source of the performance issues. Which metric should you monitor?

A. Data IO percentage

B. Local tempdb percentage

C. Cache used percentage





D. DWU percentage

Answer: C

Explanation:

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monit>

10. - (Exam Topic 3)

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

You plan to create a database named D61 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pod.

Which format should you use for the tables in DB1?

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

Answer: A

Explanation:

Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.

For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-storage-files-spark-tables>

11. - (Exam Topic 3)





A company plans to use Apache Spark analytics to analyze intrusion detection data.

You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.

What should you recommend?

- A. Azure Data Lake Storage
- B. Azure Databricks
- C. Azure HDInsight
- D. Azure Data Factory

Answer: B

Explanation:

Three common analytics use cases with Microsoft Azure Databricks

Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

Note: Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

Reference:

<https://azure.microsoft.com/es-es/blog/three-critical-analytics-use-cases-with-microsoft-azure-databricks/>

12. - (Exam Topic 3)

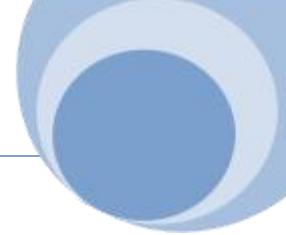
You plan to create an Azure Data Factory pipeline that will include a mapping data flow. You have JSON data containing objects that have nested arrays.

You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays.

Which transformation method should you use in the mapping data flow?

- A. unpivot





- B. flatten
- C. new branch
- D. alter row

Answer: B

Explanation:

Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.

Reference:

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

13. - (Exam Topic 3)

You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables.

Which distribution type should you recommend to minimize data movement?

- A. HASH
- B. REPLICATE
- C. ROUND ROBIN

Answer: B

Explanation:

A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables.

Reference:

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

14. - (Exam Topic 3)

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.

P1:

	▼
Set the Copy method to Bulk insert	
Set the Copy method to PolyBase	
Set the Isolation level to Repeatable read	
Set the Partition option to Dynamic range	

P2:

	▼
Set the Copy method to Bulk insert	
Set the Copy method to PolyBase	
Set the Isolation level to Repeatable read	
Set the Partition option to Dynamic range	

Answer:

P1:

	▼
Set the Copy method to Bulk insert	
<input checked="" type="checkbox"/> Set the Copy method to PolyBase	
Set the Isolation level to Repeatable read	
Set the Partition option to Dynamic range	

P2:

	▼
<input checked="" type="checkbox"/> Set the Copy method to Bulk insert	
<input checked="" type="checkbox"/> Set the Copy method to PolyBase	
Set the Isolation level to Repeatable read	
Set the Partition option to Dynamic range	

Explanation:

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>

15. - (Exam Topic 3)

You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

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

- A. Assign Azure AD security groups to Azure Data Lake Storage.
- B. Configure end-user authentication for the Azure Data Lake Storage account.
- C. Configure service-to-service authentication for the Azure Data Lake Storage account.
- D. Create security groups in Azure Active Directory (Azure AD) and add project members.
- E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Answer: A D E

Explanation: References:

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

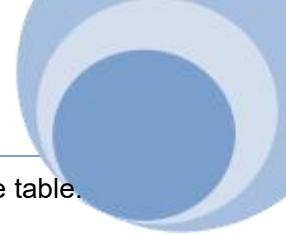
16. - (Exam Topic 3)

You implement an enterprise data warehouse in Azure Synapse Analytics. You have a large fact table that is 10 terabytes (TB) in size.

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

SaleKey	CityKey	CustomerKey	StockItemKey	InvoiceDateKey	Quantity	UnitPrice	TotalExcludingTax
49309	90858	70	69	10/22/13	8	16	128
49313	55710	126	69	10/22/13	2	16	32
49343	44710	234	68	10/22/13	10	16	160
49352	66109	163	70	10/22/13	4	16	64
49488	65312	230	70	10/22/13	8	16	128
49646	85877	271	70	10/24/13	1	16	16
49798	41238	288	69	10/24/13	1	16	16

The No.1 IT Certification Dumps



You need to distribute the large fact table across multiple nodes to optimize performance of the table.

Which technology should you use?

- A. hash distributed table with clustered index
- B. hash distributed table with clustered Columnstore index
- C. round robin distributed table with clustered index
- D. round robin distributed table with clustered Columnstore index
- E. heap table with distribution replicate

Answer: B

Explanation:

Hash-distributed tables improve query performance on large fact tables.

Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.

Reference:

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

<https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance>

17. - (Exam Topic 3)

You have an Azure Data Factory pipeline shown the following exhibit.



The execution log for the first pipeline run is shown in the following exhibit.

Activity runs

Pipeline run ID 87f89922-14fa-468f-b13f-2f867606f4ff

All status

Showing 1 - 2 items

Activity name	Activity type	Run start	Duration	Status
Web_GetIP	Web	Nov 10, 2022, 11:11:36 a	00:00:02	✖ Failed
Exec_COPY_BLOB	Execute Pipeline	Nov 10, 2022, 11:11:25 a	00:00:11	✓ Succeeded





The execution log for the second pipeline run is shown in the following exhibit.

Activity runs

Pipeline run ID a7b5b522-cfaf-4c09-b3a9-f842986be984

All status ▾

Showing 1 - 3 items

Activity name ↑↓	Activity type ↑↓	Run start ↑↓	Duration ↑↓	Status ↑↓
Set status	Set variable	Nov 10, 2022, 11:13:17 a	00:00:01	✓ Succeeded
Web_GetIP	Web	Nov 10, 2022, 11:12:59 a	00:00:16	✓ Succeeded
Exec_COPY_BLOB	Execute Pipeline	Nov 10, 2022, 11:12:48 a	00:00:11	⌚ Skipped

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

Answer Area

Statements	Yes	No
The Retry property of the Web_GetIP activity is set to 1.	<input type="radio"/>	<input type="radio"/>
The waitOnCompletion property of the Exec_COPY_BLOB activity is set to true.	<input type="radio"/>	<input type="radio"/>
The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.	<input type="radio"/>	<input type="radio"/>

Answer:

Answer Area

Statements	Yes	No
The Retry property of the Web_GetIP activity is set to 1.	<input type="radio"/>	<input checked="" type="radio"/>
The waitOnCompletion property of the Exec_COPY_BLOB activity is set to true.	<input type="radio"/>	<input checked="" type="radio"/>
The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.	<input type="radio"/>	<input checked="" type="radio"/>

18. - (Exam Topic 3)

You are implementing a batch dataset in the Parquet format.

Data tiles will be produced by using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.

You need to minimize storage costs for the solution. What should you do?

- A. Store all the data as strings in the Parquet tiles.
- B. Use OPENROWSET to query the Parquet files.



- C. Create an external table that contains a subset of columns from the Parquet files.
- D. Use Snappy compression for the files.

Answer: C

Explanation:

An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage.

External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

19. - (Exam Topic 3)

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.
- D. Scale the SU count for the job up.
- E. Scale the SU count for the job down.
- F. Implement query parallelization by partitioning the data input.

Answer: D F

Explanation:

Reference:

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

20. - (Exam Topic 3)

You are designing a data mart for the human resources (HR) department at your company. The data mart



will contain information and employee transactions. From a source system you have a flat extract that has the following fields:

- EmployeeID
- FirstName
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate

You need to design a star schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each Correct answer present part of the solution.

- A. a dimension table for employee
- B. a fabric for Employee
- C. a dimension table for EmployeeTransaction
- D. a dimension table for Transaction
- E. a fact table for Transaction

Answer: A E

Explanation:

Reference:

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

21. - (Exam Topic 3)

You have two Azure Blob Storage accounts named account1 and account2?

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account2?

You need to recommend a solution to implement the pipeline. The solution must meet the following



requirements:

- Ensure that the pipeline only copies blobs that were created or modified since the most recent replication event.
 - Minimize the effort to create the pipeline. What should you recommend?
- A. Create a pipeline that contains a flowlet.
 - B. Create a pipeline that contains a Data Flow activity.
 - C. Run the Copy Data tool and select Metadata-driven copy task.
 - D. Run the Copy Data tool and select Built-in copy task.

Answer: A

22. - (Exam Topic 3)

You have an Azure subscription.

You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model Pool1 will contain the following tables.

Name	Number of rows	Update frequency	Description
Common.Date	7,300	New rows inserted yearly	<ul style="list-style-type: none"> Contains one row per date for the last 20 years

Table distribution types

- Hash
- Replicated
- Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

Answer:

Table distribution types

- Hash
- Replicated
- Round-robin

Answer Area

Common.Data: Replicated

Marketing.Web.Sessions: Round-robin

Staging. Web.Sessions: Hash



23. - (Exam Topic 3)

The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.) You need to configure the Stream Analytics job to pick up the new reference data. What should you configure? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.

Answer:

See the answer below in explanation

Explanation:

Answer as below

Answer Area

Path pattern:

Date format:

24. - (Exam Topic 3)

You are monitoring an Azure Stream Analytics job.

You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero. You need to ensure that the job can handle all the events.

What should you do?

- A. Change the compatibility level of the Stream Analytics job.
- B. Increase the number of streaming units (SUs).
- C. Remove any named consumer groups from the connection and use \$default.
- D. Create an additional output stream for the existing input stream.

Answer: B

Explanation:

Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units.

Note: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream



Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

Reference:

<https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring>

25. - (Exam Topic 3)

You plan to use an Apache Spark pool in Azure Synapse Analytics to load data to an Azure Data Lake Storage Gen2 account.

You need to recommend which file format to use to store the data in the Data Lake Storage account. The solution must meet the following requirements:

- Column names and data types must be defined within the files loaded to the Data Lake Storage account.
- Data must be accessible by using queries from an Azure Synapse Analytics serverless SQL pool.
- Partition elimination must be supported without having to specify a specific partition. What should you recommend?

- A. Delta Lake
- B. JSON
- C. CSV
- D. ORC

Answer: D

26. - (Exam Topic 3)

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
- D. sliding

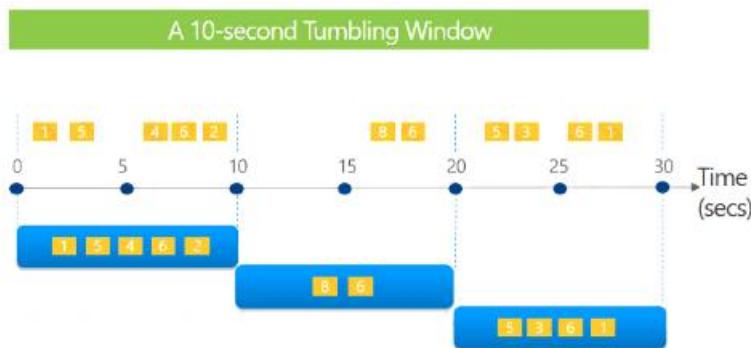


Answer: B

Explanation:

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



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

Reference:

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

27. - (Exam Topic 3)

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

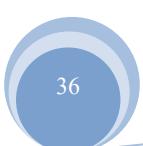
The job receives events based on user actions on the webpage. Each row of data represents an event.

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.



```

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

```

Answer:

```

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

```

Explanation:**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>

28. - (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 SOL.
- 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

Answer: B

Explanation:

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>

29. - (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:

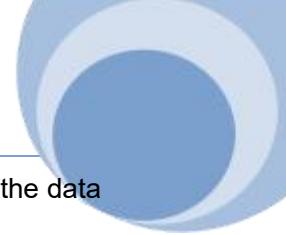
- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SOL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

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>

30. - (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 SOL.
- 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

Answer: B

Explanation:

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>

31. - (Exam Topic 3)

You are designing 2 solution that will use tables in Delta Lake on Azure Databricks. You need to minimize how long it takes to perform the following:

*Queries against non-partitioned tables

* Joins on non-partitioned columns

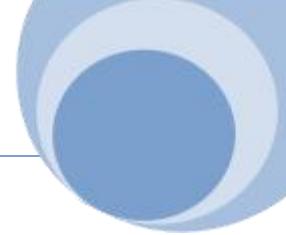
Which two options should you include in the solution? Each correct answer presents part of the solution.

(Choose Correct Answer and Give explanation and References to Support the answers based from Data Engineering on Microsoft Azure)

A. Z-Ordering

B. Apache Spark caching

C. dynamic file pruning (DFP)



D. the clone command

Answer: A C

Explanation:

According to the information I found on the web, two options that you should include in the solution to minimize how long it takes to perform queries and joins on non-partitioned tables are:

- Z-Ordering: This is a technique to colocate related information in the same set of files. This co-locality is automatically used by Delta Lake in data-skipping algorithms. This behavior dramatically reduces the amount of data that Delta Lake on Azure Databricks needs to read.
- Apache Spark caching: This is a feature that allows you to cache data in memory or on disk for faster access. Caching can improve the performance of repeated queries and joins on the same data. You can cache Delta tables using the CACHE TABLE or CACHE LAZY commands.

To minimize the time it takes to perform queries against non-partitioned tables and joins on non-partitioned columns in Delta Lake on Azure Databricks, the following options should be included in the solution:

- * A. Z-Ordering: Z-Ordering improves query performance by co-locating data that share the same column values in the same physical partitions. This reduces the need for shuffling data across nodes during query execution. By using Z-Ordering, you can avoid full table scans and reduce the amount of data processed.
- * B. Apache Spark caching: Caching data in memory can improve query performance by reducing the amount of data read from disk. This helps to speed up subsequent queries that need to access the same data. When you cache a table, the data is read from the data source and stored in memory. Subsequent queries can then read the data from memory, which is much faster than reading it from disk.

References:

- Delta Lake on Databricks: <https://docs.databricks.com/delta/index.html>
- Best Practices for Delta Lake on Databricks: <https://databricks.com/blog/2020/05/14/best-practices-for-delta-lake-on-databricks.html>

32. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly. At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data.



How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Partition the data:

- Partition by date with one partition per day.
- Partition by date with one partition per day.**
- Partition by date with one partition per month.
- Partition by product.

Remove the data:

- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a WHERE clause.**
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

Answer:

Answer Area

Partition the data:

- Partition by date with one partition per day.
- Partition by date with one partition per day.**
- Partition by date with one partition per month.
- Partition by product.

Remove the data:

- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a WHERE clause.**
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

33. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours.

You have the following function.

```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
return (F - 32) * 5.0 / 9
end
```

You have the following query.

```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date. You need to minimize the time it takes for the query to return results.

Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each



correct selection is worth one point.

- A. Create an index on the avg_f column.
- B. Convert the avg_c column into a calculated column.
- C. Create an index on the sensorid column.
- D. Enable result set caching.
- E. Change the table distribution to replicate.

Answer: B D

Explanation:

[https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-s
et-cac](https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cac)

34. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers. You need to create an external data source in Pool1 that will be used to read .orc files in storage1. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 = 'abfs://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE = BLOB_STORAGE
);
```

Answer:





Answer Area

```

CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 '://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE - );

```

Location1

abfs
abfss
wasb
wasbs

TYPE

BLOB_STORAGE
HADOOP
RDBMS
SHARP MAP MANAGER

Explanation:

Graphical user interface, text, application, email Description automatically generated

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw>

35. - (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 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

B. No

**Answer: B**

Explanation:

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>

36. - (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**Answer Area**

```

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

Answer:


Values Answer Area

```

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

```

The diagram shows a SQL query with several parts highlighted in boxes. The 'CASE' keyword is in a green box, 'ELSE' is in a red box, and 'OVER' and 'PARTITION BY' are in dashed green boxes. The 'ROW_NUMBER()' function is also in a dashed green box.

Explanation:

Graphical user interface, text, application Description automatically generated

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>

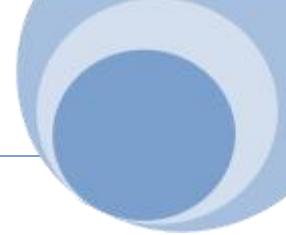
37. - (Exam Topic 3)

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: A

Explanation:

Each partition should have around 1 millions records. Dedicated SQL pools already have 60 partitions. We have the formula: Records/(Partitions*60)= 1 million

Partitions= Records/(1 million * 60)

Partitions= $2.4 \times 1,000,000,000 / (1,000,000 * 60) = 40$

Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

38. - (Exam Topic 3)

You have an Azure Stream Analytics job.

You need to ensure that the job has enough streaming units provisioned. You configure monitoring of the SU % Utilization metric.

Which two additional metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Backlogged Input Events

B. Watermark Delay

C. Function Events

D. Out of order Events

E. Late Input Events

Answer: A B

Explanation:

To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU



Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job, by increasing the SUs.

Reference:

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

39. - (Exam Topic 3)

You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:

- SensorTypeID
- GeographyRegionID
- Year
- Month
- Day
- Hour
- Minute
- Temperature
- WindSpeed
- Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.

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

NOTE: Each correct selection is worth one point.



```
df.write
```

.bucketBy	(*)
.format	("GeographyRegionID")
.partitionBy	("GeographyRegionID", "Year", "Month", "Day")
.sortBy	("Year", "Month", "Day", "GeographyRegionID")

```
.mode ("append")
```

.csv("/DBTBL1")	(*)
.json("/DBTBL1")	
.parquet("/DBTBL1")	
.saveAsTable("/DBTBL1")	

Answer:

```
df.write
```

.bucketBy	(*)
.format	("GeographyRegionID")
.partitionBy	("GeographyRegionID", "Year", "Month", "Day")
.sortBy	("Year", "Month", "Day", "GeographyRegionID")

```
.mode ("append")
```

.csv("/DBTBL1")	(*)
.json("/DBTBL1")	
.parquet("/DBTBL1")	
.saveAsTable("/DBTBL1")	

Explanation:

Graphical user interface, text, application Description automatically generated

40. - (Exam Topic 3)

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.

Answer Area

Architecture requirement	Technology
Ingest	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> Logic Apps <input type="checkbox"/> Azure Data Factory <input type="checkbox"/> Azure Automation</div>
Store	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> Azure Data Lake Storage <input type="checkbox"/> Azure Blob storage <input type="checkbox"/> Azure files</div>
Prepare and Train	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> HDInsight Apache Spark cluster <input type="checkbox"/> Azure Databricks <input type="checkbox"/> HDInsight Apache Storm cluster</div>
Model and Serve	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> HDInsight Apache Kafka cluster <input type="checkbox"/> Azure Synapse Analytics <input type="checkbox"/> Azure Data Lake Storage</div>

Answer:





Answer Area

Architecture requirement	Technology
Ingest	<ul style="list-style-type: none"> Logic Apps Azure Data Factory Azure Automation
Store	<ul style="list-style-type: none"> Azure Data Lake Storage Azure Blob storage Azure files
Prepare and Train	<ul style="list-style-type: none"> HDInsight Apache Spark cluster Azure Databricks HDInsight Apache Storm cluster
Model and Serve	<ul style="list-style-type: none"> HDInsight Apache Kafka cluster Azure Synapse Analytics Azure Data Lake Storage

Explanation:

Graphical user interface, application, table, email Description automatically generated

41. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains the users shown in the following table.

Name	Role
User1	Server admin
User2	db_datereader

User1 executes a query on the database, and the query returns the results shown in the following exhibit.





```

1  SELECT c.name,
2    |   tbl.name as table_name,
3    |   typ.name as datatype,
4    |   c.is_masked,
5    |   c.masking_function
6  FROM sys.masked_columns AS c
7  INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
8  INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
9  WHERE is_masked = 1;
10

```

Results Messages

	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

User1 is the only user who has access to the unmasked data.

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.

When User2 queries the YearlyIncome column,

the values returned will be [answer choice].

a random number
the values stored in the database
XXXX
0

When User1 queries the BirthDate column, the

values returned will be [answer choice].

a random date
the values stored in the database
XXXX
1900-01-01

Answer:





When User2 queries the YearlyIncome column,
the values returned will be [answer choice].

a random number
the values stored in the database
XXXX
0

When User1 queries the BirthDate column, the
values returned will be [answer choice].

a random date
the values stored in the database
XXXX
1900-01-01

Explanation:

Graphical user interface, text, application, email Description automatically generated

Box 1: 0

The YearlyIncome column is of the money data type.

The Default masking function: Full masking according to the data types of the designated fields

- Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).

Box 2: the values stored in the database

Users with administrator privileges are always excluded from masking, and see the original data without any mask.

Reference:

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

42. - (Exam Topic 3)

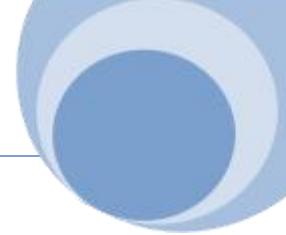
You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.

You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.

Which type of data redundancy should you use?

- A. zone-redundant storage (ZRS)
- B. read-access geo-redundant storage (RA-GRS)





- C. locally-redundant storage (LRS)
- D. geo-redundant storage (GRS)

Answer: B

Explanation:

Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

43. - (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 from the last five minutes every minute. Which windowing function should you use?

- A. Sliding
- B. Session
- C. Tumbling
- D. Hopping

Answer: D

Explanation:

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>





44. - (Exam Topic 3)

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster1. You need to reduce the time it takes for cluster1 to start and scale up. The solution must minimize costs. What should you do first?

- A. Upgrade workspace1 to the Premium pricing tier.
- B. Create a cluster policy in workspace1.
- C. Create a pool in workspace1.
- D. Configure a global init script for workspace1.

Answer: C

Explanation:

You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.

Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.

Reference:

<https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html>

45. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark_df.

You need to write the contents of pyspark_df to a table in SQLPool by using a PySpark notebook. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area



Answer:





Answer Area



46. - (Exam Topic 3)

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.

```
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] [varchar](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)
```

Answer:



```

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] [varchar](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
)
  
```

Explanation:

Text Description automatically generated

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-ver1>

47. - (Exam Topic 3)

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:

- * The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.
- * Line total sales amount and line total tax amount will be aggregated in Databricks.
- * Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

You need to recommend an output mode for the dataset that will be processed by using Structured





Streaming. The solution must minimize duplicate data.

What should you recommend?

A. Append

B. Update

C. Complete

Answer: B

Explanation:

By default, streams run in append mode, which adds new records to the table.

<https://docs.databricks.com/delta/delta-streaming.html>

48. - (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 enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly. Solution: You convert the files to compressed delimited text files. Does this meet the goal?

A. Yes

B. No

Answer: A

Explanation:

All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.

Reference:

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





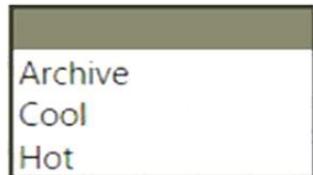
49. - (Exam Topic 3)

You are designing an application that will use an Azure Data Lake Storage Gen 2 account to store petabytes of license plate photos from toll booths. The account will use zone-redundant storage (ZRS).

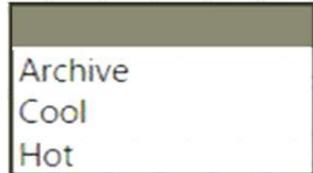
You identify the following usage patterns:

- The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SU of 99.9%.
- After 90 days, the data will be accessed infrequently but must be available within 30 seconds.
- After 365 days, the data will be accessed infrequently but must be available within five minutes.

First 30 days:



After 90 days:



After 365 days:

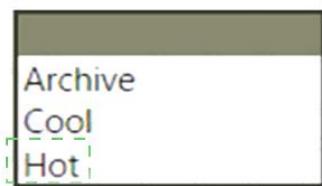


Answer:





First 30 days:



After 90 days:



After 365 days:



Explanation:

Box 1: Hot

The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SLA of 99.9%.

Box 2: Cool

After 90 days, the data will be accessed infrequently but must be available within 30 seconds. Data in the Cool tier should be stored for a minimum of 30 days.

When your data is stored in an online access tier (either Hot or Cool), users can access it immediately. The Hot tier is the best choice for data that is in active use, while the Cool tier is ideal for data that is accessed less frequently, but that still must be available for reading and writing.

Box 3: Cool

After 365 days, the data will be accessed infrequently but must be available within five minutes. Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

<https://docs.microsoft.com/en-us/azure/storage/blobs/archive-rehydrate-overview>

50. - (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to determine the size of the transaction log file for each distribution of DW1.

What should you do?





A. On DW1, execute a query against the sys.database_files dynamic management view.

B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.

C. Execute a query against the logs of DW1 by using the

Get-AzOperationalInsightsSearchResult PowerShell cmdlet.

D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Answer: A

Explanation:

For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/logs/manage-the-size-of-the-transaction-log-file>

51. - (Exam Topic 3)

You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.

You need to recommend an authentication mechanism to ensure that the solution can access the source data.

What should you recommend?

A. a managed identity

B. anonymous public read access

C. a shared key

Answer: A

Explanation:

Managed Identity authentication is required when your storage account is attached to a VNet. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples>

52. - (Exam Topic 3)

You have the following Azure Stream Analytics query.





WITH

```

step1 AS (SELECT *
    FROM input1
    PARTITION BY StateID
    INTO 10),
step2 AS (SELECT *
    FROM input2
    PARTITION BY StateID
    INTO 10)

SELECT *
INTO output
FROM step1
PARTITION BY StateID
UNION
SELECT * INTO output
    FROM step2
    PARTITION BY StateID

```

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

NOTE: Each correct selection is worth one point.

Statements	Yes	No
The query combines two streams of partitioned data.	<input type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input type="radio"/>	<input type="radio"/>

Answer:

Statements	Yes	No
The query combines two streams of partitioned data.	<input type="radio"/>	<input checked="" type="radio"/>
The stream scheme key and count must match the output scheme.	<input checked="" type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input checked="" type="radio"/>	<input type="radio"/>

Explanation:

Box 1: No

Note: You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.

The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10),





step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)

SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID

Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.

Box 2: Yes

When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count. Box 3: Yes

Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY. Here there are 10 partitions, so $6 \times 10 = 60$ SUs is good.

Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.

Reference:

<https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/>

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption>

53. - (Exam Topic 3)

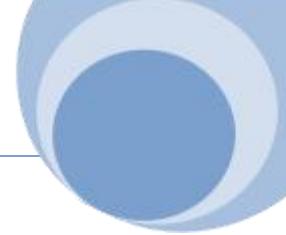
A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.

The company must be able to monitor the devices in real-time. You need to design the solution.

What should you recommend?

- A. Azure Stream Analytics cloud job using Azure PowerShell
- B. Azure Analysis Services using Azure Portal
- C. Azure Data Factory instance using Azure Portal
- D. Azure Analysis Services using Azure PowerShell





Answer: C

Explanation:

Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.

Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks.

Reference:

[https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-a](https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-product-ion-a)

54. - (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	<ul style="list-style-type: none"> Azure SQL Database Azure Blob Storage Azure Cosmos DB Azure Data Lake Store
--------------	--

Batch processing	<ul style="list-style-type: none"> HDInsight Spark HDInsight Hadoop Azure Databricks HDInsight Interactive Query
------------------	--

Analytical data store	<ul style="list-style-type: none"> HDInsight HBase Azure SQL Data Warehouse Azure Analysis Services Azure Cosmos DB
-----------------------	---

Answer:

Architecture requirement	Technology
--------------------------	------------

Data storage	<ul style="list-style-type: none"> Azure SQL Database Azure Blob Storage Azure Cosmos DB Azure Data Lake Store
--------------	--

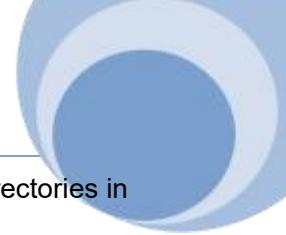
Batch processing	<ul style="list-style-type: none"> HDInsight Spark HDInsight Hadoop Azure Databricks HDInsight Interactive Query
------------------	--

Analytical data store	<ul style="list-style-type: none"> HDInsight HBase Azure SQL Data Warehouse Azure Analysis Services Azure Cosmos DB
-----------------------	---

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

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

HDIInsight is a managed Hadoop service. Use it 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-namespace>

<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>

55. - (Exam Topic 3)

You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.

You execute the Transact-SQL query shown in the following exhibit.

```
SELECT
    payment_type,
    SUM(fare_amount) AS fare_total
FROM OPENROWSET(
    BULK 'csv/busfare/tripdata_2020*.csv',
    DATA_SOURCE = 'BusData',
    FORMAT = 'CSV', PARSER_VERSION = '2.0',
    FIRSTROW = 2
)
WITH (
    payment_type INT 10,
    fare_amount FLOAT 11
) AS nyc
GROUP BY payment_type
ORDER BY payment_type;
```

What do the query results include?



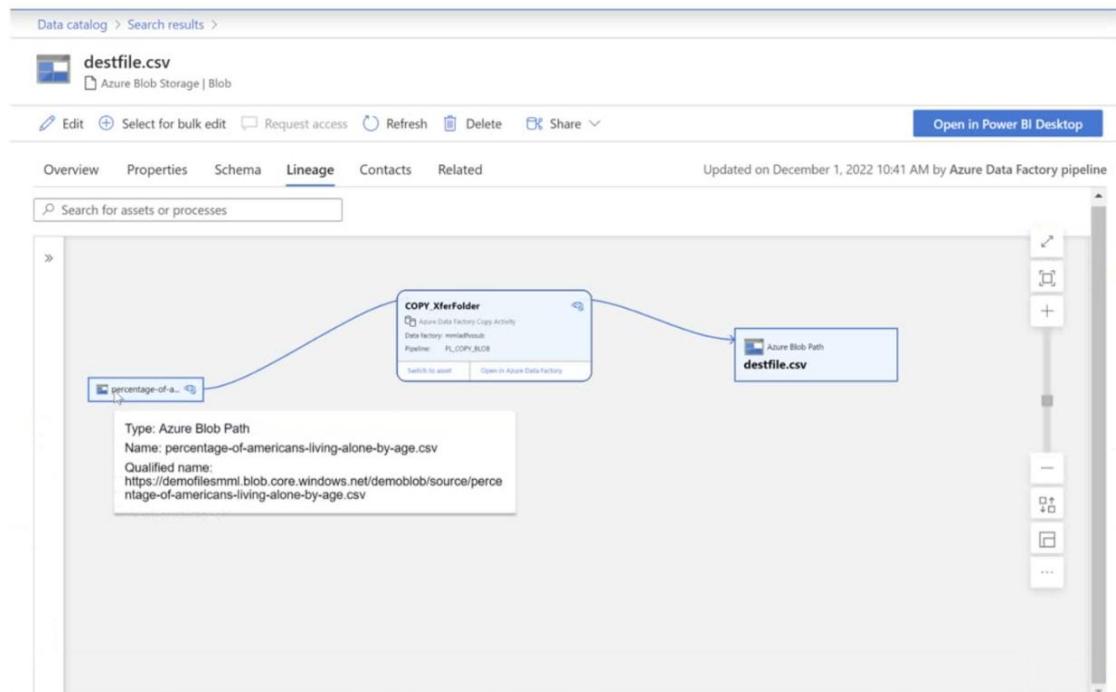


- A. Only CSV files in the tripdata_2020 subfolder.
- B. All files that have file names that begin with "tripdata_2020".
- C. All CSV files that have file names that contain "tripdata_2020".
- D. Only CSV that have file names that begin with "tripdata_2020".

Answer: D

56. - (Exam Topic 3)

You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.



How is the data for the lineage populated?

- A. manually
- B. by scanning data stores
- C. by executing a Data Factory pipeline

Answer: B

Explanation:

According to Microsoft Purview Data Catalog lineage user guide¹, data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems². Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source².





57. - (Exam Topic 3)

You are designing the folder structure for an Azure Data Lake Storage Gen2 account. You identify the following usage patterns:

- Users will query data by using Azure Synapse Analytics serverless SQL pools and Azure Synapse Analytics serverless Apache Spark pods.
- Most queries will include a filter on the current year or week.
- Data will be secured by data source.

You need to recommend a folder structure that meets the following requirements:

- Supports the usage patterns
- Simplifies folder security
- Minimizes query times

Which folder structure should you recommend?

A)

\YYYY\WW\DataSource\SubjectArea\FileData_YYYY_MM_DD.parquet

B)

DataSource\SubjectArea\WW\YYYY\FileData_YYYY_MM_DD.parquet

C)

\DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet

D)

\DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet

E)

WW\YYYY\SubjectArea\DataSource\FileData_YYYY_MM_DD.parquet

A. Option A

B. Option B

C. Option C

D. Option D

E. Option E

Answer: C

Explanation:





Data will be secured by data source. -> Use DataSource as top folder.

Most queries will include a filter on the current year or week -> Use \YYYY\WW\ as subfolders. Common Use Cases

A common use case is to filter data stored in a date (and possibly time) folder structure such as /YYYY/MM/DD/ or /YYYY/MM/YYYY-MM-DD/. As new data is generated/sent/copied/moved to the storage account, a new folder is created for each specific time period. This strategy organises data into a maintainable folder structure.

Reference: <https://www.serverlesssql.com/optimisation/azurestoragefilteringusingfilepath/>

58. - (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 SOL.
- 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 High Concurrency cluster for the jobs.





Does this meet the goal?

- A. Yes
- B. No

Answer: A

Explanation:

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>

59. - (Exam Topic 3)

You have an Azure Storage account that generates 200.000 new files daily. The file names have a format of (YYY)/(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.

Answer:

See the answer below in

Explanation:

Answer Area

Load methodology: Incremental load

Trigger: Tumbling window

60. - (Exam Topic 3)

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline.



From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns.

You need to add an activity to pipeline to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables.

Which two types of activities can you use to execute SP1? (Refer to Data Engineering on Microsoft Azure documents or guide for Answers explanation available at Microsoft.com)

- A. Stored Procedure
- B. Lookup
- C. Script
- D. Copy

Answer: A B

Explanation:

the two types of activities that you can use to execute SP1 are Stored Procedure and Lookup.

A Stored Procedure activity executes a stored procedure on an Azure SQL Database or Azure Synapse Analytics or SQL Server1. You can specify the stored procedure name and parameters in the activity setting1s.

A Lookup activity retrieves a dataset from any data source that returns a single row of data with four columns2. You can use a query to execute a stored procedure as the source of the Lookup activit2y. You can then store the values in the columns as pipeline variables by using expressions2.

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure>

61. - (Exam Topic 3)

You have an Azure Synapse Analytics Apache Spark pool named Pool1.

You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.

You need to load the files into the tables. The solution must maintain the source data types. What should you do?

- A. Use a Get Metadata activity in Azure Data Factory.
- B. Use a Conditional Split transformation in an Azure Synapse data flow.
- C. Load the data by using the OPEHROwset Transact-SQL command in an Azure Synapse Anarytics



serverless SQL pool.

D. Load the data by using PySpark.

Answer: A

Explanation:

Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.

Serverless SQL pool enables you to query data in your data lake. It offers a T-SQL query surface area that accommodates semi-structured and unstructured data queries.

To support a smooth experience for in place querying of data that's located in Azure Storage files, serverless SQL pool uses the OPENROWSET function with additional capabilities.

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

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage>

62. - (Exam Topic 3)

DRAG DROP

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

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	Answer Area
CLUSTERED INDEX	CREATE TABLE table1
COLLATE	(
DISTRIBUTION	ID INTEGER,
PARTITION	col1 VARCHAR(10),
PARTITION FUNCTION	col2 VARCHAR(10)
PARTITION SCHEME) WITH
	(
	[] = HASH(ID),
	[] (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
) ;

Answer:





Values
CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
    ID INTEGER,
    col1 VARCHAR(10),
    col2 VARCHAR(10)
) WITH
(
    DISTRIBUTION = HASH(ID),
    PARTITION (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

Explanation:

Box 1: DISTRIBUTION

Table distribution options include DISTRIBUTION = HASH (distribution_column_name), assigns each row to one distribution by hashing the value stored in distribution_column_name. Box 2: PARTITION

Table partition options. Syntax:

```
PARTITION ( partition_column_name RANGE [ LEFT | RIGHT ] FOR VALUES ( [ boundary_value [,...n] ] ) )
```

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?>

63. - (Exam Topic 3)

You need to design a solution that will process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage. The solution must ensure that analysts can interactively query the streaming data.

What should you use?

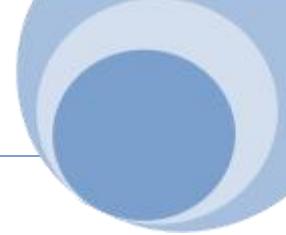
- A. event triggers in Azure Data Factory
- B. Azure Stream Analytics and Azure Synapse notebooks
- C. Structured Streaming in Azure Databricks
- D. Azure Queue storage and read-access geo-redundant storage (RA-GRS)

Answer: C

Explanation:

Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real-time.

With Structured Streaming, you can use SQL queries to process streaming data in the same way that you



would process static data.

Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.

Azure Event Hubs can be integrated with Spark Structured Streaming to perform the processing of messages in near real-time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.

Reference:

<https://k21academy.com/microsoft-azure/data-engineer/structured-streaming-with-azure-event-hubs/>

64. - (Exam Topic 3)

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/{YYYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYYY}{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.

Parameter:

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

Naming pattern:

/{{deviceID}}/out/{{YYYYY}}/{{MM}}/{{DD}}/{{HH}}.json
/{{YYYYY}}/{{MM}}/{{DD}}/{{deviceType}}.json
/{{YYYYY}}/{{MM}}/{{DD}}/{{HH}}.json
/{{YYYYY}}/{{MM}}/{{DD}}/{{HH}}_{{deviceType}}.json

Copy behavior:

Add dynamic content
Flatten hierarchy
Merge files

**Answer:**

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

Explanation:

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>

65. - (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQL pool, an Azure Synapse Analytics dedicated SQL pool, an Apache Spark pool, and an Azure Data Lake Storage Gen2 account.

You need to create a table in a lake database. The table must be available to both the serverless SQL pool and the Spark pool.

Where should you create the table, and Which file format should you use for data in the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.





Create the table in:

File format:

Answer:

Create the table in:

File format:

Explanation:

The dedicated SQL pool Apache Parquet

66. - (Exam Topic 3)

You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.

You need to verify the duration of the activity when it ran last. What should you use?

- A. activity runs in Azure Monitor
- B. Activity log in Azure Synapse Analytics
- C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics
- D. an Azure Resource Manager template

Answer: A**Explanation:****Reference:**

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually>

67. - (Exam Topic 3)

You are designing a folder structure for the files in an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.

You need to recommend a folder structure that meets the following requirements:

- Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pool
- Supports fast data retrieval for data from the current month
- Simplifies data security management by department Which folder structure should you recommend?

A. \YYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet

B. \Depdftment\DataSource\YYY\MM\DataFile_YYYYMMDD.parquet

C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet

D. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet

Answer: B

Explanation:

Department top level in the hierarchy to simplify security management.

Month (MM) at the leaf/bottom level to support fast data retrieval for data from the current month.

68. - (Exam Topic 3)

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 executes mapping data Flow, and then inserts the data info the data warehouse.

Does this meet the goal?

A. Yes

B. No

Answer: B

Explanation:

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow,5 with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.



Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

69. - (Exam Topic 3)

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.

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input type="radio"/>	<input type="radio"/>

Answer:





Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="checkbox"/>	<input checked="" type="checkbox"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Explanation:

Graphical user interface, text, application Description automatically generated

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>

70. - (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

71. - (Exam Topic 3)

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.

You update a Data Factory pipeline.

You need to ensure that the updated lineage is available in Microsoft Purview.

What You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.

You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible.

Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

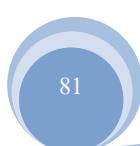
- A. If Condition
- B. ForEach
- C. Lookup
- D. Get Metadata

Answer: C D

72. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:

- One billion rows
- A clustered columnstore index
- A hash-distributed column named Product Key



- A column named Sales Date that is of the date data type and cannot be null. Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

- A. once per month
- B. once per year
- C. once per day
- D. once per week

Answer: B

Explanation:

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition.

Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions.

Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

Reference:

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

73. - (Exam Topic 3)

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. storage event

Answer: D

Explanation:

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>

74. - (Exam Topic 3)

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.

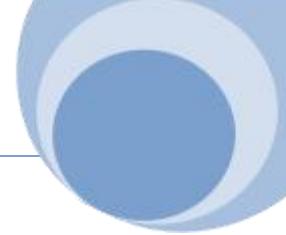
You create the following components:

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

In which order should you perform the actions? 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
Mount the Data Lake Storage onto DBFS.	
Write the results to a table in Azure Synapse.	
Specify a temporary folder to stage the data.	
Read the file into a data frame.	
Perform transformations on the data frame.	



**Answer:**

Actions	Answer Area
Mount the Data Lake Storage onto DBFS.	Mount the Data Lake Storage onto DBFS.
Write the results to a table in Azure Synapse.	Read the file into a data frame.
Specify a temporary folder to stage the data.	Perform transformations on the data frame.
Read the file into a data frame.	Specify a temporary folder to stage the data.
Perform transformations on the data frame.	Write the results to a table in Azure Synapse.

Explanation:

Table Description automatically generated

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>

75. - (Exam Topic 3)

You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.

You monitor the Stream Analytics job and discover high latency. You need to reduce the latency.

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

- A. Add a pass-through query.
- B. Add a temporal analytic function.



- C. Scale out the query by using PARTITION BY.
- D. Convert the query to a reference query.
- E. Increase the number of streaming units.

Answer: C E

Explanation:

C: Scaling a Stream Analytics job takes advantage of partitions in the input or output. Partitioning lets you divide data into subsets based on a partition key. A process that consumes the data (such as a Stream Analytics job) can consume and write different partitions in parallel, which increases throughput.

E: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.

References:

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

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption>

76. - (Exam Topic 3)

You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

```
source(output(
    License_plate as string,
    Make as string,
    Time as string
),
allowSchemaDrift: true,
```

Answer:

See below answer.

Answer Area

Number of columns: 22

Number of rows: 4





77. - (Exam Topic 3)

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.

Select TimeZone, count (*) AS MessageCount

FROM MessageStream	▼	CreatedAt
		LAST
		OVER
		SYSTEM.TIMESTAMP()
		TIMESTAMP BY

GROUP BY TimeZone,	▼	(second,15)
		HOPPINGWINDOW
		SESSIONWINDOW
		SLIDINGWINDOW
		TUMBLINGWINDOW

Answer:

Select TimeZone, count (*) AS MessageCount

FROM MessageStream	▼	CreatedAt
		LAST
		OVER
		SYSTEM.TIMESTAMP()
		TIMESTAMP BY

GROUP BY TimeZone,	▼	(second,15)
		HOPPINGWINDOW
		SESSIONWINDOW
		SLIDINGWINDOW
		TUMBLINGWINDOW

Explanation:

Table Description automatically generated

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.

Timeline Description automatically generated

Reference:

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

78. - (Exam Topic 3)

You have a SQL pool in Azure Synapse.

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

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

NOTE: Each correct selection is worth one point.

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Answer:



Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Explanation:

Graphical user interface, application, table Description automatically generated

Box 1: Hash

Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance.

Box 2: Clustered columnstore

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed.

Box 3: Date

Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

Partition switching can be used to quickly remove or replace a section of a table. Reference:

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

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

79. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named pool1.

You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.





```

CREATE TABLE dbo.[DimCustomer](
    [CustomerKey] int NOT NULL,
    [CustomerSourceID] [int] NOT NULL,
    [Title] [nvarchar](8) NULL,
    [FirstName] [nvarchar](50) NOT NULL,
    [MiddleName] [nvarchar](50) NULL,
    [LastName] [nvarchar](50) NOT NULL,
    [Suffix] [nvarchar](10) NULL,
    [CompanyName] [nvarchar](128) NULL,
    [SalesPerson] [nvarchar](256) NULL,
    [EmailAddress] [nvarchar](50) NULL,
    [Phone] [nvarchar](25) NULL,
    [InsertedDate] [datetime] NOT NULL,
    [ModifiedDate] [datetime] NOT NULL,
    [HashKey] [varchar](100) NOT NULL,
    [IsCurrentRow] [bit] NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
GO

```

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL
- B. [EffectiveEndDate] [datetime] NOT NULL
- C. [PreviousModifiedDate] [datetime] NOT NULL
- D. [RowID] [bigint] NOT NULL
- E. [EffectiveStartDate] [datetime] NOT NULL

Answer: A B

80. - (Exam Topic 3)

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.





```

SELECT [▼]
    Collect(Score)
    CollectTop(1) OVER(ORDER BY Score Desc)
    Game, MAX(Score)
    TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)
] as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY [▼]
    Game
    Hopping(minute,5)
    Tumbling(minute,5)
    Windows(TumblingWindow(minute,5),Hopping(minute,5))
]

```

Answer:

```

SELECT [▼]
    Collect(Score)
    CollectTop(1) OVER(ORDER BY Score Desc)
    Game, MAX(Score)
    TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) [▼]
] as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY [▼]
    Game
    Hopping(minute,5) [▼]
    Tumbling(minute,5)
    Windows(TumblingWindow(minute,5),Hopping(minute,5))
]

```

Explanation:

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.

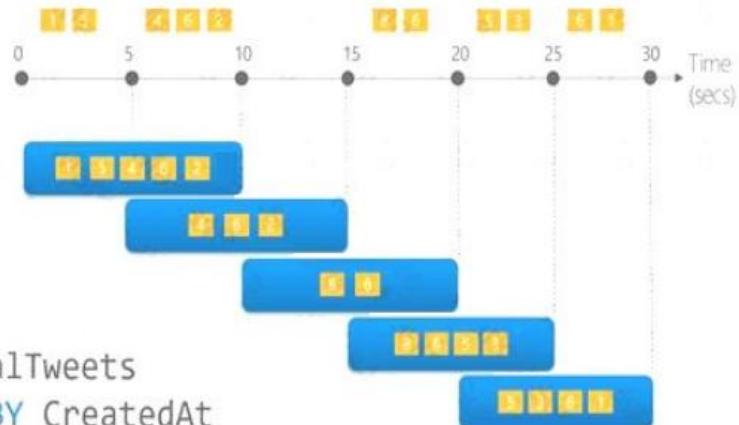
Box 2: Hopping(minute,5)

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.

A picture containing timeline Description automatically generated

Every 5 seconds give me the count of Tweets over the last 10 seconds

A 10-second Hopping Window with a 5-second "Hop"



```
SELECT Topic, COUNT(*) AS TotalTweets
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, HoppingWindow(second, 10 , 5)
```

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>

81. - (Exam Topic 3)

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files.

Each file contains the same data attributes and data from a subsidiary of your company.

You need to move the files to a different folder and transform the data to meet the following requirements:

- Provide the fastest possible query times.
- Automatically infer the schema from the underlying files.

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

NOTE: Each correct selection is worth one point.



Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

Answer:

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

Explanation:

Box 1: Preserver herarchy

Compared to the flat namespace on Blob storage, the hierarchical namespace greatly improves the performance of directory management operations, which improves overall job performance.

Box 2: Parquet

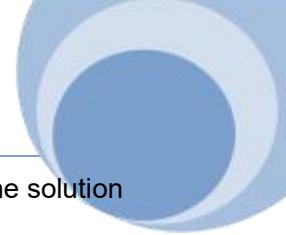
Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2. Parquet supports the schema property.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction>

<https://docs.microsoft.com/en-us/azure/data-factory/format-parquet>

82. - (Exam Topic 3)



You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

- Send the output to Azure Synapse.
- Identify spikes and dips in time series data.
- Minimize development and configuration effort. Which should you include in the solution?
 - A. Azure Databricks
 - B. Azure Stream Analytics
 - C. Azure SQL Database

Answer: B

Explanation:

You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics.

Reference:

<https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/>

83. - (Exam Topic 3)

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Create a date dimension table that has a DateTime key.
- B. Use built-in SQL functions to extract date attributes.
- C. Create a date dimension table that has an integer key in the format of yyyyymmdd.
- D. In the fact table, use integer columns for the date fields.
- E. Use DateTime columns for the date fields.

Answer: B D

84. - (Exam Topic 3)

You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics





named DW1.

Several users execute ad hoc queries to DW1 concurrently. You regularly perform automated data loads to DW1.

You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run. What should you do?

- A. Hash distribute the large fact tables in DW1 before performing the automated data loads.
- B. Assign a smaller resource class to the automated data load queries.
- C. Assign a larger resource class to the automated data load queries.
- D. Create sampled statistics for every column in each table of DW1.

Answer: C

Explanation:

The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution.

Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources assigned to each query. There's a trade-off between memory and concurrency.

Smaller resource classes reduce the maximum memory per query, but increase concurrency. Larger resource classes increase the maximum memory per query, but reduce concurrency. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management>

85. - (Exam Topic 3)

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID
	ArrivalDateTime
Weather	AirportID
	ReportDateTime

You need to recommend a solution that maximum query performance. What should you include in the recommendation?





- A. In each table, create a column as a composite of the other two columns in the table.
- B. In each table, create an IDENTITY column.
- C. In the tables, use a hash distribution of ArriveDateTime and ReportDateTime.
- D. In the tables, use a hash distribution of ArriveAirPortID and AirportID.

Answer: D

86. - (Exam Topic 3)

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. complete
- B. update
- C. append

Answer: C

Explanation:

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.

<https://docs.databricks.com/getting-started/spark/streaming.html>

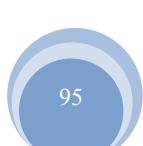
87. - (Exam Topic 3)

You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.

You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

- Track the usage of encryption keys.
- Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

What should you include in the recommendation? To answer, select the appropriate options in the answer





area.

NOTE: Each correct selection is worth one point.

To track encryption key usage:

Always Encrypted
TDE with customer-managed keys
TDE with platform-managed keys

To maintain client app access in
the event of a datacenter outage:

Create and configure Azure key vaults in two Azure regions.
Enable Advanced Data Security on Server1.
Implement the client apps by using a Microsoft .NET Framework data provider.

Answer:

To track encryption key usage:

Always Encrypted
TDE with customer-managed keys
TDE with platform-managed keys

To maintain client app access in
the event of a datacenter outage:

Create and configure Azure key vaults in two Azure regions.
Enable Advanced Data Security on Server1.
Implement the client apps by using a Microsoft .NET Framework data provider.

Explanation:

Box 1: TDE with customer-managed keys

Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.

Box 2: Create and configure Azure key vaults in two Azure regions

The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

<https://docs.microsoft.com/en-us/azure/key-vault/general/logging>





88. - (Exam Topic 3)

You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.

Which type of slowly changing dimension (SCD) should use?

- A. Type 0
- B. Type 1
- C. Type 2
- D. Type 3

Answer: C

Explanation:

Type 2 - Creating a new additional record. In this methodology all history of dimension changes is kept in the database. You capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key(or other durable identifier). Also 'effective date' and 'current indicator' columns are used in this method. There could be only one record with current indicator set to 'Y'. For 'effective date' columns, i.e. start_date and end_date, the end_date for current record usually is set to value 9999-12-31. Introducing changes to the dimensional model in type 2 could be very expensive database operation so it is not recommended to use it in dimensions where a new attribute could be added in the future.

<https://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html>

89. - (Exam Topic 3)

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.

HubA:	<input type="checkbox"/> Stream <input type="checkbox"/> Reference
HubB:	<input type="checkbox"/> Stream <input type="checkbox"/> Reference
Database1:	<input type="checkbox"/> Stream <input type="checkbox"/> Reference

Answer:

HubA:	<input checked="" type="checkbox"/> Stream <input type="checkbox"/> Reference
HubB:	<input checked="" type="checkbox"/> Stream <input type="checkbox"/> Reference
Database1:	<input type="checkbox"/> Stream <input checked="" type="checkbox"/> Reference

Explanation:

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:



90. - (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

Type	Designated retention period
Application	360 days
Infrastructure	60 days

You do not expect that the logs will be accessed during the retention periods.

You need to recommend a solution for account1 that meets the following requirements:

- Automatically deletes the logs at the end of each retention period
- Minimizes storage costs

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

NOTE: Each correct selection is worth one point.

To minimize storage costs:

Store the infrastructure logs and the application logs in the Archive access tier	▼
Store the infrastructure logs and the application logs in the Cool access tier	
Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier	

To delete logs automatically:

Azure Data Factory pipelines	▼
Azure Blob storage lifecycle management rules	
Immutable Azure Blob storage time-based retention policies	

Answer:

To minimize storage costs:

Store the infrastructure logs and the application logs in the Archive access tier	▼
Store the infrastructure logs and the application logs in the Cool access tier	
Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier	

To delete logs automatically:

Azure Data Factory pipelines	▼
Azure Blob storage lifecycle management rules	
Immutable Azure Blob storage time-based retention policies	



Explanation:

Table Description automatically generated

Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

For infrastructure logs: Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the cool tier should be stored for a minimum of 30 days. The cool tier has lower storage costs and higher access costs compared to the hot tier.

For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the archive tier should be stored for a minimum of 180 days.

Box 2: Azure Blob storage lifecycle management rules

Blob storage lifecycle management offers a rule-based policy that you can use to transition your data to the desired access tier when your specified conditions are met. You can also use lifecycle management to expire data at the end of its life.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

91. - (Exam Topic 3)

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.





Input type:

Stream
Reference

Function:

Aggregate
Geospatial
Windowing

Answer:

Input type:

Stream
Reference

Function:

Aggregate
Geospatial
Windowing

Explanation:

Diagram, table Description automatically generated

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> <https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios>

92. - (Exam Topic 3)

You have the following Azure Data Factory pipelines





- ingest Data from System 1
- Ingest Data from System2
- Populate Dimensions
- Populate facts

ingest Data from System1 and Ingest Data from System1 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System*. Populate Facts must execute after the 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. Create a parent pipeline that contains the four pipelines and use an event trigger.
- C. Create a parent pipeline that contains the four pipelines and use a schedule trigger.
- D. Add a schedule trigger to all four pipelines.

Answer: C

Explanation:

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>

93. - (Exam Topic 3)

You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).

You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.

You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.

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

NOTE: Each correct selection is worth one point





Actions	Answer Area
Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key	
Create an external data source that uses the abfs location	▶
Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages	◀
Create an external file format and set the First_Row option	

**Answer:**

Actions	Answer Area
Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key	
Create an external data source that uses the abfs location	
Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages	
Create an external file format and set the First_Row option	

Explanation:

A picture containing timeline Description automatically generated

Step 1: Create an external data source that uses the abfs location

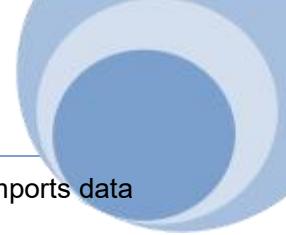
Create External Data Source to reference Azure Data Lake Store Gen 1 or 2 Step 2: Create an external file format and set the First_Row option.

Create External File Format.

Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

To use PolyBase, you must create external tables to reference your external data. Use reject options.

Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is



run. Instead, they're specified here so that the database can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects>

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql>

94. - (Exam Topic 3)

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.

Values

Answer Area

{deviceID}	/	Value	/	Value	/	Value	.json
{mm}/{HH}/{DD}/{MM}/{YYYY}							
{regionID}/{deviceID}							
{regionID}/raw							
{YYYY}/{MM}/{DD}/{HH}							
{YYYY}/{MM}/{DD}/{HH}/{mn}							
raw/{deviceID}							
raw/{regionID}							

Answer:





Values	Answer Area
{deviceID}	
{mm}/{HH}/{DD}/{MM}/{YYYY}	/ {YYYY}/{MM}/{DD}/{HH} /
{regionID}/{deviceID}	{regionID}/raw
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	{deviceID}
raw/{regionID}	.json

Explanation:

Box 1: {YYYY}/{MM}/{DD}/{HH}

Date Format [optional]: if the date token is used in the prefix path, you can select the date format in which your files are organized. Example: YYYY/MM/DD

Time Format [optional]: if the time token is used in the prefix path, specify the time format in which your files are organized. Currently the only supported value is HH.

Box 2: {regionID}/raw

Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

Box 3: {deviceID} Reference:

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

95. - (Exam Topic 3)

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies. You need to ensure that users from each company can view only the data of their respective company. Which two objects should you include in the solution? Each correct answer presents part of the solution

NOTE: Each correct selection it worth one point.

- A. a custom role-based access control (RBAC) role.
- B. asymmetric keys
- C. a predicate function
- D. a column encryption key





E. a security policy

Answer: A E

Explanation:

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview>

96. - (Exam Topic 3)

You are designing an Azure Synapse Analytics dedicated SQL pool.

Groups will have access to sensitive data in the pool as shown in the following table.

Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

You have policies for the sensitive data. The policies vary by region as shown in the following table.

Region	Data considered sensitive
RegionA	Financial, Personally Identifiable Information (PII)
RegionB	Financial, Personally Identifiable Information (PII), medical
RegionC	Financial, medical

You have a table of patients for each region. The tables contain the following potentially sensitive columns.

Name	Sensitive data	Description
CardOnFile	Financial	Debit/credit card number for charges
Height	Medical	Patient's height in cm
ContactEmail	PII	Email address for secure communications

You are designing dynamic data masking to maintain compliance.

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

NOTE: Each correct selection is worth one point.



Statements	Yes	No
------------	-----	----

Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].

Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]

Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]

Answer:

Statements	Yes	No
------------	-----	----

Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].

Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]

Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]

Explanation:

Text Description automatically generated

Reference:

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

97. - (Exam Topic 3)

You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.

You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.

Which authentication solution or solutions should you include in the recommendation? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.





MFA:

- Azure AD authentication
- Microsoft SQL Server authentication
- Passwordless authentication
- Windows authentication

Database-level authentication:

- Application roles
- Contained database users
- Database roles
- Microsoft SQL Server logins

Answer:

MFA:

- Azure AD authentication**
- Microsoft SQL Server authentication
- Passwordless authentication
- Windows authentication

Database-level authentication:

- Application roles**
- Contained database users**
- Database roles
- Microsoft SQL Server logins

Explanation:

Graphical user interface, text, application, chat or text message Description automatically generated

Box 1: Azure AD authentication

Azure Active Directory authentication supports Multi-Factor authentication through Active Directory Universal Authentication.

Box 2: Contained database users

Azure Active Directory Uses contained database users to authenticate identities at the database level.

Reference:

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

98. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.

You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and



land the data in another folder.

You need to ensure that the data factory can read and write data from any folder in the DataLake1 file system. The solution must meet the following requirements:

- Minimize the risk of unauthorized user access.
- Use the principle of least privilege.
- Minimize maintenance effort.

How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Use	to authenticate by using
<input type="checkbox"/> Azure Active Directory (Azure AD) <input type="checkbox"/> a shared access signature (SAS) <input type="checkbox"/> a shared key	<input type="checkbox"/> a managed identity <input type="checkbox"/> a stored access policy <input type="checkbox"/> an Authorization header

Answer:

Use	to authenticate by using
<input checked="" type="checkbox"/> Azure Active Directory (Azure AD) <input type="checkbox"/> a shared access signature (SAS) <input type="checkbox"/> a shared key	<input checked="" type="checkbox"/> a managed identity <input type="checkbox"/> a stored access policy <input type="checkbox"/> an Authorization header

Explanation:

Text Description automatically generated with low confidence

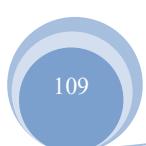
Box 1: Azure Active Directory (Azure AD)

On Azure, managed identities eliminate the need for developers having to manage credentials by providing an identity for the Azure resource in Azure AD and using it to obtain Azure Active Directory (Azure AD) tokens.

Box 2: a managed identity

A data factory can be associated with a managed identity for Azure resources, which represents this specific data factory. You can directly use this managed identity for Data Lake Storage Gen2 authentication, similar to using your own service principal. It allows this designated factory to access and copy data to or from your Data Lake Storage Gen2.

Note: The Azure Data Lake Storage Gen2 connector supports the following authentication types.





- Account key authentication
- Service principal authentication
- Managed identities for Azure resources authentication Reference:

<https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview>

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

99. - (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: You use a dedicated SQL pool to create an external table that has a additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Answer: A

100. - (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics.

Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.

The external table has three columns.

You discover that the Parquet files have a fourth column named ItemID.

Which command should you run to add the ItemID column to the external table?



```

A. ALTER EXTERNAL TABLE [Ext].[Items]
    ADD [ItemID] int;

B. DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
    FORMAT_TYPE = PARQUET,
    DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);

C. DROP EXTERNAL TABLE [Ext].[Items]
CREATE EXTERNAL TABLE [Ext].[Items]
([ItemID] [int] NULL,
[ItemName] nvarchar(50) NULL,
[ItemType] nvarchar(20) NULL,
[ItemDescription] nvarchar(250))
WITH
(
    LOCATION= '/Items/',
    DATA_SOURCE = AzureDataLakeStore,
    FILE_FORMAT = PARQUET,
    REJECT_TYPE = VALUE,
    REJECT_VALUE = 0
);

D. ALTER TABLE [Ext].[Items]
ADD [ItemID] int;

```

A. Option A

B. Option B

C. Option C

D. Option D

Answer: C

Explanation:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql>

101. - (Exam Topic 3)

You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool.

Each batch of incoming data is staged before being loaded into the fact tables. |

You need to ensure that the incoming data is staged as quickly as possible. |

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





Table distribution:

HASH
REPLICATE
ROUND_ROBIN

Table structure:

Clustered index
Columnstore index
Heap

Answer:

Table distribution:

HASH
REPLICATE
ROUND_ROBIN

Table structure:

Clustered index
Columnstore index
Heap

Explanation:

Round-robin distribution is recommended for staging tables because it distributes data evenly across all the distributions without requiring a hash column. This can improve the speed of data loading and avoid data skew. Heap tables are recommended for staging tables because they do not have any indexes or partitions that can slow down the data loading process. Heap tables are also easier to truncate and reload than clustered index or columnstore index tables.

102. - (Exam Topic 3)

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Tablet.

Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. MERGE
- D. ALTER

Answer: C

Explanation:



The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.

Example:

```
// Implementing SCD Type 2 operation using merge function customersTable  
as("customers") merge(  
stagedUpdates.as("staged_updates"), "customers.customerId = mergeKey")  
whenMatched("customers.current = true AND customers.address <> staged_updates.address")  
updateExpr(Map(  
"current" -> "false",  
"endDate" -> "staged_updates.effectiveDate")) whenNotMatched()  
insertExpr(Map(  
"customerId" -> "staged_updates.customerId", "address" -> "staged_updates.address", "current" -> "true",  
"effectiveDate" -> "staged_updates.effectiveDate",  
"endDate" -> "null")) execute()  
}
```

Reference:

<https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks>

103. - (Exam Topic 3)

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.
- C. Set Cluster Mode to High Concurrency.
- D. Create a cluster policy in workspace1.

Answer: B

Explanation:

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>

104. - (Exam Topic 3)

You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.

Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted.

You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

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	Answer Area
CustomerKey	
HASH	
ROUND_ROBIN	
REPLICATE	
OrderDateKey	
SalesOrderNumber	

```

CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]      int      NOT NULL
    , [OrderDateKey]   int      NOT NULL
    , [CustomerKey]   int      NOT NULL
    , [SalesOrderNumber] nvarchar ( 20 ) NOT NULL
    , [OrderQuantity] smallint NOT NULL
    , [UnitPrice]     money    NOT NULL
)
WITH
(   CLUSTERED           COLUMNSTORE      INDEX
    , DISTRIBUTION =  Value ([ProductKey])
    , PARTITION ( [  Value ] ) RANGE RIGHT FOR VALUES
        (20170101,20180101,20190101,20200101,20210101)
)

```

Answer:

Values	Answer Area
CustomerKey	
HASH	
ROUND_ROBIN	
REPLICATE	
OrderDateKey	
SalesOrderNumber	

```

CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]      int      NOT NULL
    , [OrderDateKey]   int      NOT NULL
    , [CustomerKey]   int      NOT NULL
    , [SalesOrderNumber] nvarchar ( 20 ) NOT NULL
    , [OrderQuantity] smallint NOT NULL
    , [UnitPrice]     money    NOT NULL
)
WITH
(   CLUSTERED           COLUMNSTORE      INDEX
    , DISTRIBUTION = HASH ([ProductKey])
    , PARTITION ( [ OrderDateKey ] ) RANGE RIGHT FOR VALUES
        (20170101,20180101,20190101,20200101,20210101)
)

```

Explanation:

Box 1: HASH

Box 2: OrderDateKey

In most cases, table partitions are created on a date column.

A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the





partition with data for an empty partition from another table.

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

105. - (Exam Topic 3)

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
- D. Microsoft.EventHub

Answer: C

Explanation:

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>

106. - (Exam Topic 3)

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.

Components

- a database scoped credential
- an asymmetric key
- an external data source
- a database encryption key
- an external file format

Answer Area



Answer:

Components

- a database scoped credential
- an asymmetric key
- an external data source
- a database encryption key
- an external file format

Answer Area

- a database scoped credential
- an external data source
- an external file format



107. - (Exam Topic 3)

You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of <https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/>.

ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.

Resource	Permission
container1	Access – Execute
Folder1	Access – Execute
Folder2	Access – Read





You need to ensure that ServicePrincipal1 can perform the following actions:

- Traverse child items that are created in Folder2.
- Read files that are created in Folder2.

The solution must use the principle of least privilege.

Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Access - Read
- B. Access - Write
- C. Access - Execute
- D. Default-Read
- E. Default - Write
- F. Default - Execute

Answer: D F

Explanation:

Execute (X) permission is required to traverse the child items of a folder.

There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs. Access ACLs: These control access to an object. Files and folders both have Access ACLs.

Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs.

Reference:

<https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control>

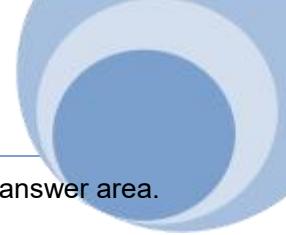
108. - (Exam Topic 3)

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

Answer:

Report1:

Avro
CSV
Parquet
TSV

Report2:

Avro
CSV
Parquet
TSV

Explanation:

Report1: CSV

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

AVRO supports timestamps.

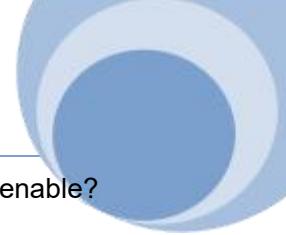
Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2. Reference:

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

109. - (Exam Topic 3)

You have a data warehouse in Azure Synapse Analytics.





You need to ensure that the data in the data warehouse is encrypted at rest. What should you enable?

- A. Advanced Data Security for this database
- B. Transparent Data Encryption (TDE)
- C. Secure transfer required
- D. Dynamic Data Masking

Answer: B

Explanation:

Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.

- Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.
- Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature.

Reference:

<https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest>

110. - (Exam Topic 3)

You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java. Which service should you recommend using to process the streaming data?

- A. Azure Data Factory
- B. Azure Stream Analytics
- C. Azure Databricks
- D. Azure Event Hubs

Answer: C

Explanation:

<https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing>

111. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.



You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.

What should you do?

- A. Switch the first partition from dbo.Sales to stg.Sales.
- B. Switch the first partition from stg.Sales to dbo. Sales.
- C. Update dbo.Sales from stg.Sales.
- D. Insert the data from stg.Sales into dbo.Sales.

Answer: A

112. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pod.

You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input (or a downstream activity). The solution must minimize development effort.

Which Type of activity should you use in the pipeline?

- A. Notebook
- B. U-SQL
- C. Script
- D. Stored Procedure

Answer: D

113. - (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: A B

Explanation: B: Example:

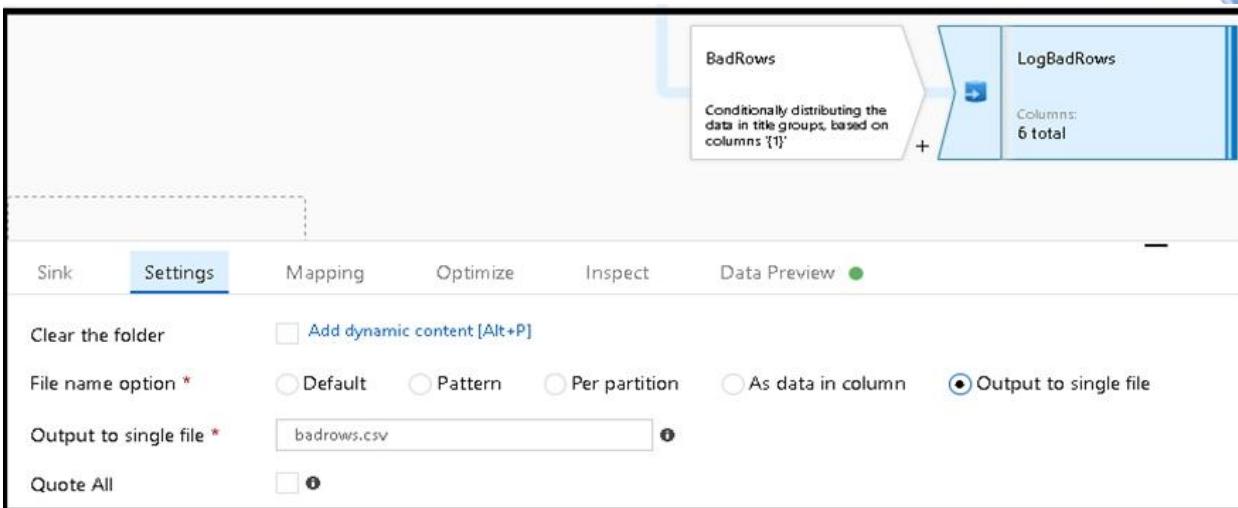
- * 1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

STREAM NAMES	CONDITION
GoodRows	length(title) <= 5
BadRows	Rows that do not meet any condition will use this output stream

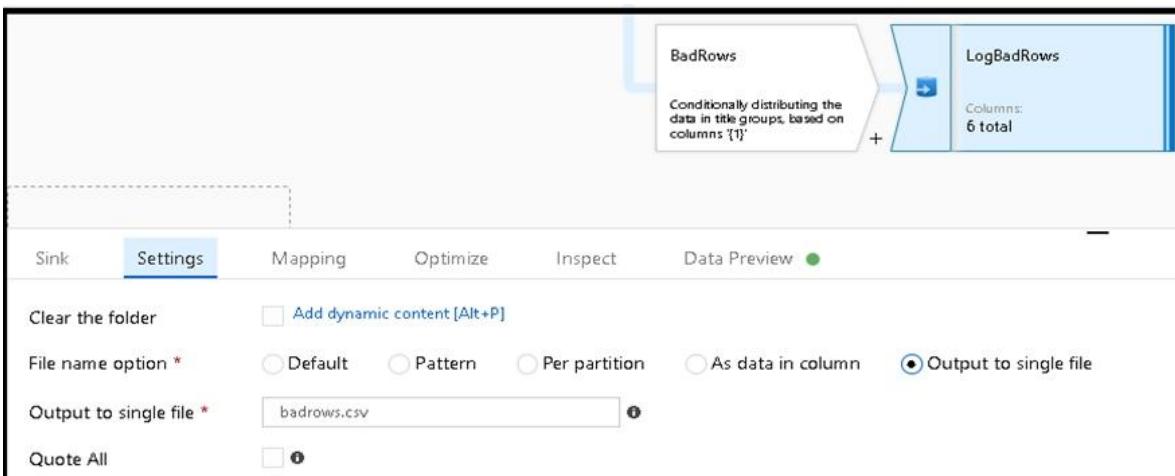
- * 2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream. A:

- * 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>

114. - (Exam Topic 3)

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





B. Azure-SSIS integration runtime

C. Self-hosted integration runtime

Answer: A

Explanation:

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

Reference:

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

115. - (Exam Topic 3)

You have an Azure Synapse serverless SQL pool.

You need to read JSON documents from a file by using the OPENROWSET function.

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

Each correct selection is worth one point.

Answer Area

```

SELECT *
FROM OPENROWSET(
    BULK
        'https://sourcedatalake.blob.core.windows.net/public/docs.json',
        FORMAT = 'JSON',
        FIELDTERMINATOR = '0x0b',
        FIELDQUOTE = '0x0b',
        ROWTERMINATOR = '0x09'
    )
    WITH (jsondoc nvarchar '0x0c') AS (SELECT Documents
    
```

Answer:





Answer Area

```

SELECT *
FROM OPENROWSET
(
    BULK
        'https://sourcedatalake.blob.core.windows.net/public/docs.json',
        FORMAT = 'JSON',
        'CSV',
        'DELTA',
        'JSON' ----- JSON ----- PARQUET
        FIELDTERMINATOR = '0x0b',
        FIELDQUOTE = '0x0b' ----- 0x0b -----
        ROWTERMINATOR = . '0x09'
        . '0x0a'
        . '0x0b' ----- 0x0b -----
        WITH (jsondoc nvarc '0x0c') onDocuments
)

```

116. - (Exam Topic 3)

You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.

You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1.

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

Actions	Answer Area
Enable TDE on Pool1.	
Assign a managed identity to Server1.	
Configure key1 as the TDE protector for Server1.	< >
Add key1 to the Azure key vault.	↑ ↓
Create an Azure key vault and grant the managed identity permissions to the key vault.	

Answer:



Actions	Answer Area
Enable TDE on Pool1.	Assign a managed identity to Server1.
Assign a managed identity to Server1.	Create an Azure key vault and grant the managed identity permissions to the key vault.
Configure key1 as the TDE protector for Server1.	Add key1 to the Azure key vault.
Add key1 to the Azure key vault.	Configure key1 as the TDE protector for Server1.
Create an Azure key vault and grant the managed identity permissions to the key vault.	Enable TDE on Pool1.

Explanation:

Graphical user interface, text, application Description automatically generated

Step 1: Assign a managed identity to Server1

You will need an existing Managed Instance as a prerequisite.

Step 2: Create an Azure key vault and grant the managed identity permissions to the vault Create Resource and setup Azure Key Vault.

Step 3: Add key1 to the Azure key vault

The recommended way is to import an existing key from a .pfx file or get an existing key from the vault.

Alternatively, generate a new key directly in Azure Key Vault.

Step 4: Configure key1 as the TDE protector for Server1 Provide TDE Protector key

Step 5: Enable TDE on Pool1 Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-best-practices>

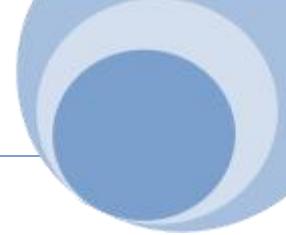
117. - (Exam Topic 3)

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.

You need to recommend a format for the transformed files. The solution must meet the following requirements:

- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.





What should you recommend?

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

Answer: D

Explanation:

Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format.

Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance.

It is especially good for queries that read particular columns from a “wide” (with many columns) table since only needed columns are read, and IO is minimized.

Reference: <https://www.clairvoyant.ai/blog/big-data-file-formats>

118. - (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder1 and Folder2.

You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

```
Operation on target Copy_sks failed: Failure happened on 'Sink' side.  
ErrorCode=DelimitedTextMoreColumnsThanDefined,  
'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,  
Message>Error found when processing 'Csv/Tsv Format Text' source  
'0_2020_11_09_11_43_32.avro' with row number 53: found more columns  
than expected column count 27., Source=Microsoft.DataTransfer.Common,'
```

You receive the following error.

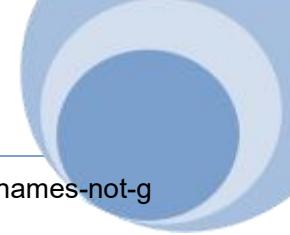
- What should you do to resolve the error.
- A. Add an explicit mapping.
 - B. Enable fault tolerance to skip incompatible rows.
 - C. Lower the degree of copy parallelism
 - D. Change the Copy activity setting to Binary Copy

Answer: A

Explanation:

Reference:





<https://knowledge.informatica.com/s/article/Microsoft-Azure-Data-Lake-Store-Gen2-target-file-names-not-gene>

119. - (Exam Topic 3)

You use PySpark in Azure Databricks to parse the following JSON input.

```
{
  "persons": [
    {
      "name": "Keith",
      "age": 30,
      "dogs": ["Fida", "Fluffy"]
    },
    {
      "name": "Donna",
      "age": 46,
      "dogs": ["Spot"]
    }
  ]
}
```

You need to output the data in the following tabular format.

owner	age	dog
Keith	30	Fida
Keith	30	Fluffy
Donna	46	Spot

How should you complete the PySpark code? 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	Answer Area
alias	dbutils.fs.put("/tmp/source.json", source_json, True)
array_union	source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
createDataFrame	persons = source_df. <input type="text"/> Value Value ("persons").alias("persons")
explode	persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode ("persons.dogs"). <input type="text"/> Value ("dog"))
select	display(persons_dogs)
translate	

Answer:





Values	Answer Area
	<pre> dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json") persons = source_df. select explode ("persons").alias("persons") persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode ("persons.dogs"). alias ("dog")) display(persons_dogs) </pre>

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: select

Box 2: explode

Box 3: alias

pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode).

Reference: <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html>

<https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode>

120. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and query sysdm_pdw_sys_info.
- B. Connect to Pool1 and run DBCC CHECKALLOC.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Answer: D

Explanation:

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

Reference:

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



121. - (Exam Topic 3)

You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service. You have an Azure Key vault named vault1 that contains an encryption key named key1.

You need to encrypt Df1 by using key1. What should you do first?

- A. Add a private endpoint connection to vault 1.
- B. Enable Azure role-based access control on vault 1.
- C. Remove the linked service from Df1.
- D. Create a self-hosted integration runtime.

Answer: C

Explanation:

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key>

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services>

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

122. - (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
        ORDER BY time
         LAG
         LAST
         LEAD)
         (hour,1))
         LIMIT DURATION
         OFFSET
         WHEN
FROM input
```

Answer:



```

SELECT sensorId,
       growth = reading -
                  (reading) OVER (PARTITION BY sensorId
                                ORDER BY time
                                ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING)
FROM input
  
```

The diagram illustrates the LAG analytic operator. It shows two boxes: Box 1 (LAG) containing 'LAG', 'LAST', and 'LEAD' with a downward arrow; and Box 2 (LIMIT DURATION) containing 'LIMIT DURATION', 'OFFSET', and 'WHEN' with a downward arrow. The output of Box 1 is connected to the input of Box 2.

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>

123. - (Exam Topic 3)

You have an Azure subscription.

You need to deploy an Azure Data Lake Storage Gen2 Premium account. The solution must meet the following requirements:

- Blobs that are older than 365 days must be deleted.
- Administrator efforts must be minimized.
- Costs must be minimized

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

Answer Area

The screenshot shows two sections of a Microsoft Azure interface. The top section, titled 'To minimize costs:', lists four storage tiers: Locally-redundant storage (LRS), The Archive access tier, The Cool access tier, and Zone-redundant storage (ZRS). The bottom section, titled 'To delete blobs:', lists three methods: Azure Automation runbooks, Azure Storage lifecycle management, and Soft delete. A mouse cursor is visible over the 'Soft delete' option.

Answer:

Answer Area

This screenshot is identical to the one above, but with several items highlighted by red boxes. In the 'To minimize costs:' section, 'Locally-redundant storage (LRS)' is highlighted. In the 'To delete blobs:' section, both 'Azure Automation runbooks' and 'Azure Storage lifecycle management' are highlighted.

Explanation:

<https://learn.microsoft.com/en-us/azure/storage/blobs/premium-tier-for-data-lake-storage>

124. - (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: A

Explanation:





125. - (Exam Topic 3)

You have an Azure data factory named ADM that contains a pipeline named Pipeline1. Pipeline1 must execute every 30 minutes with a 15-minute offset.

You need to create a trigger for Pipeline1. The trigger must meet the following requirements:

- Backfill data from the beginning of the day to the current time.
- If Pipeline1 fails, ensure that the pipeline can re-execute within the same 30-minute period.
- Ensure that only one concurrent pipeline execution can occur.
- Minimize development and configuration effort. Which type of trigger should you create?

A. schedule

B. event-based

C. manual

D. tumbling window

Answer: A

126. - (Exam Topic 3)

You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCD) when loading the data into a table named DimCustomer1 in an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that Dataflow1 can perform the following tasks:

- * Detect whether the data of a given customer has changed in the DimCustomer table.
- Perform an upsert to the DimCustomer table.

Which type of transformation should you use for each task? To answer, select the appropriate options in the answer area

NOTE: Each correct selection is worth one point.





Answer Area

Detect whether the data of a given customer has changed in the DimCustomer table:

Aggregate
Derived column
Surrogate key

Perform an upsert to the DimCustomer table:

Alter row
Assert
Cast

Answer:

Answer Area

Detect whether the data of a given customer has changed in the DimCustomer table:

Aggregate
Derived column
Surrogate key

Perform an upsert to the DimCustomer table:

Alter row
Assert
Cast

127. - (Exam Topic 3)

You have an Azure Databricks resource.

You need to log actions that relate to changes in compute for the Databricks resource. Which Databricks services should you log?

- A. clusters
- B. workspace
- C. DBFS
- D. SSH
- E. jobs

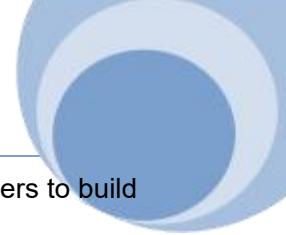
Answer: B

Explanation:

Cloud Provider Infrastructure Logs. Databricks logging allows security and admin teams to demonstrate conformance to data governance standards within or from a Databricks workspace. Customers, especially in the regulated industries, also need records on activities like:- User access control to cloud data storage- Cloud Identity and Access Management roles- User access to cloud network and compute

Azure Databricks offers three distinct workloads on several VM Instances tailored for your data analytics





workflow—the Jobs Compute and Jobs Light Compute workloads make it easy for data engineers to build and execute jobs, and the All-Purpose Compute workload makes it easy for data scientists to explore, visualize, manipulate, and share data and insights interactively.

128. - (Exam Topic 3)

You are developing an application that uses Azure Data Lake Storage Gen 2.

You need to recommend a solution to grant permissions to a specific application for a limited time period.

What should you include in the recommendation?

- A. Azure Active Directory (Azure AD) identities
- B. shared access signatures (SAS)
- C. account keys
- D. role assignments

Answer: B

Explanation:

A shared access signature (SAS) provides secure delegated access to resources in your storage account.

With a SAS, you have granular control over how a client can access your data. For example:

What resources the client may access.

What permissions they have to those resources. How long the SAS is valid.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview>

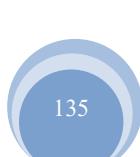
129. - (Exam Topic 3)

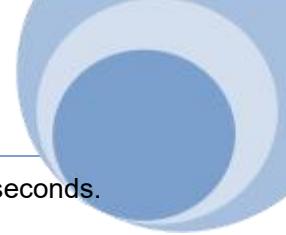
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

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 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

Answer: B

Explanation:

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>

130. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.

You plan to access the files in Account1 by using an external table.

You need to create a data source in Pool1 that you can reference when you create the external table. How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

```
CREATE EXTERNAL DATA SOURCE source1
WITH
    ( LOCATION = 'https://account1.  .core.windows.net',
      
      
      
```

Answer:





```
CREATE EXTERNAL DATA SOURCE source1
WITH
( LOCATION = 'https://account1.blob.core.windows.net',
  
)
)
```

Explanation:

Graphical user interface, diagram Description automatically generated

Box 1: blob

The following example creates an external data source for Azure Data Lake Gen2 CREATE EXTERNAL DATA SOURCE YellowTaxi

```
WITH ( LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE =
HADOOP)
```

Box 2: HADOOP

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

131. - (Exam Topic 3)

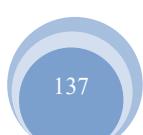
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

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 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 an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Answer: B

Explanation:

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>

132. - (Exam Topic 3)

You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool.

You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone.

How should you complete the Transact-SQL code? 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

- BEGIN DISTRIBUTED TRANSACTION
- BEGIN TRAN
- COMMIT TRAN
- ROLLBACK TRAN
- SET RESULT_SET_CACHING ON

Answer Area

```
BEGIN TRY
    INSERT INTO dbo.Table1 (col1, col2, col3)
    SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
    IF @@TRANCOUNT > 0
        BEGIN
            ROLLBACK TRAN;
        END
    END CATCH;
    IF @@TRANCOUNT > 0
        BEGIN
            COMMIT TRAN;
        END

```

**Answer:****Values**

BEGIN DISTRIBUTED TRANSACTION
BEGIN TRAN
COMMIT TRAN
ROLLBACK TRAN
SET RESULT_SET_CACHING ON

Answer Area

```
BEGIN TRAN
BEGIN TRY
    INSERT INTO dbo.Table1 (col1, col2, col3)
    SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
    IF @@TRANCOUNT > 0
        BEGIN
            ROLLBACK TRAN
        END
    END CATCH;
    IF @@TRANCOUNT >0
        BEGIN
            COMMIT TRAN;
        END

```

133. - (Exam Topic 3)

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.

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

Answer:



Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

Explanation:

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>

134. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1. You load 5 TB of data into table1.



You need to ensure that columnstore compression is maximized for table1. Which statement should you execute?

- A. ALTER INDEX ALL on table1 REORGANIZE
- B. ALTER INDEX ALL on table1 REBUILD
- C. DBCC DBREINDEX (table1)
- D. DBCC INDEXDEFRAG (pool1,table1)

Answer: B

Explanation:

Columnstore and columnstore archive compression

Columnstore tables and indexes are always stored with columnstore compression. You can further reduce the size of columnstore data by configuring an additional compression called archival compression. To perform archival compression, SQL Server runs the Microsoft XPRESS compression algorithm on the data.

Add or remove archival compression by using the following data compression types:

Use COLUMNSTORE_ARCHIVE data compression to compress columnstore data with archival compression.

Use COLUMNSTORE data compression to decompress archival compression. The resulting data continue to be compressed with columnstore compression.

To add archival compression, use ALTER TABLE (Transact-SQL) or ALTER INDEX (Transact-SQL) with the REBUILD option and DATA COMPRESSION = COLUMNSTORE_ARCHIVE.

Reference: <https://learn.microsoft.com/en-us/sql/relational-databases/data-compression/data-compression>

135. - (Exam Topic 3)

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.



Load methodology:

	▼
Full Load	
Incremental Load	
Load individual files as they arrive	

Trigger:

	▼
Fixed schedule	
New file	
Tumbling window	

Answer:

Load methodology:

	▼
Full Load	
Incremental Load	
Load individual files as they arrive	

Trigger:

	▼
Fixed schedule	
New file	
Tumbling window	

Explanation:

Table Description automatically generated

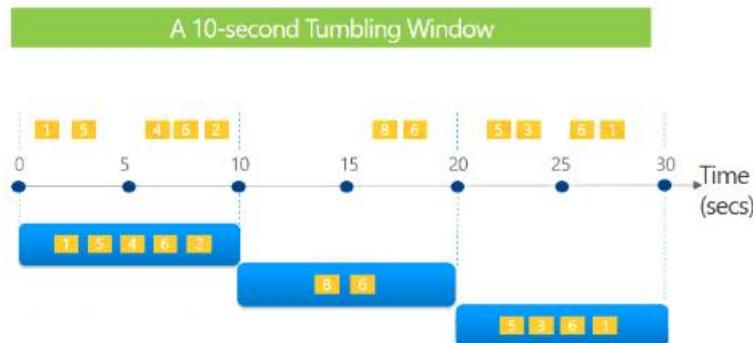
Box 1: Incremental load Box 2: Tumbling window

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.

Timeline Description automatically generated



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



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

Reference:

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

136. - (Exam Topic 3)

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.

Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.

Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128.

You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1. What should you configure? To answer, select the appropriate options in the answer area.

To Pipeline1, add:

- A custom activity
- A Get Metadata activity
- An If Condition activity

For Dataflow1, set the core count by using:

- Dynamic content
- Parameters
- User properties

Answer:

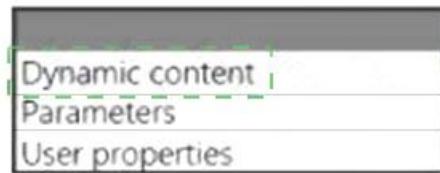




To Pipeline1, add:



For Dataflow1, set the core count by using:



Explanation:

Box 1: A Get Metadata activity

Dynamically size data flow compute at runtime

The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties.

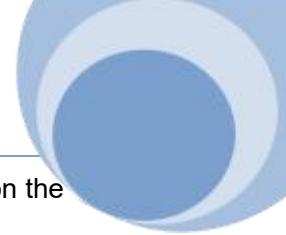
Box 2: Dynamic content

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

137. - (Exam Topic 3)

You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.

```
{
  "rules": [
    {
      "enabled": true,
      "name": "contosorule",
      "type": "lifecycle",
      "definition": {
        "actions": {
          "version": {
            "delete": {
              "daysAfterCreationGreaterThan": 60
            }
          },
          "baseBlob": {
            "tierToCool": {
              "daysAfterModificationGreaterThan": 30
            }
          }
        }
      },
      "filters": {
        "blobTypes": [
          "blockBlob"
        ],
        "prefixMatch": [
          "container1/contoso"
        ]
      }
    }
  ]
}
```



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.

The files are [answer choice] after 30 days:

▼
deleted from the container
moved to archive storage
moved to cool storage
moved to hot storage

The storage policy applies to [answer choice]:

▼
container1/contoso.csv
container1/docs/contoso.json
container1/mycontoso/contoso.csv

Answer:

The files are [answer choice] after 30 days:

▼
deleted from the container
moved to archive storage
<u>moved to cool storage</u>
moved to hot storage

The storage policy applies to [answer choice]:

▼
<u>container1/contoso.csv</u>
container1/docs/contoso.json
container1/mycontoso/contoso.csv

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: moved to cool storage

The ManagementPolicyBaseBlob.TierToCool property gets or sets the function to tier blobs to cool storage.

Support blobs currently at Hot tier.

Box 2: container1/contoso.csv As defined by prefixMatch.

prefixMatch: An array of strings for prefixes to be matched. Each rule can define up to 10 case-sensitive prefixes. A prefix string must start with a container name.

Reference:

<https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.management.storage.fluent.models.managementpoli>

138. - (Exam Topic 3)

You have an Azure subscription.



You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model. Pool1 will contain the following tables.

Name	Number of rows	Update frequency	Description
Common.Date	7,300	New rows inserted yearly	<ul style="list-style-type: none"> Contains one row per date for the last 20 years Contains columns named Year, Month, Quarter, and IsWeekend
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

You need to design the table storage for pool1. The solution must meet the following requirements:

- Maximize the performance of data loading operations to Staging.WebSessions.
- Minimize query times for reporting queries against the dimensional model.

Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables. Each table distribution type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Table distribution types	Answer Area
Hash	Common.Data: <input type="text"/>
Replicated	Marketing.Web.Sessions: <input type="text"/>
Round-robin	Staging. Web.Sessions: <input type="text"/>

Answer:



Table distribution types	Answer Area
Hash	Common.Data: <input type="checkbox"/>
Replicated	Marketing.Web.Sessions: <input type="checkbox"/>
Round-robin	Staging. Web.Sessions: <input type="checkbox"/>

Explanation:

Box 1: Replicated

The best table storage option for a small table is to replicate it across all the Compute nodes.

Box 2: Hash
Hash-distribution improves query performance on large fact tables.

Box 3: Round-robin
Round-robin distribution is useful for improving loading speed.

Reference:

[https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribution)

139. - (Exam Topic 3)

You have an Azure Data Lake Storage Gen 2 account named storage1.

You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

- List and read permissions must be granted at the storage account level.
- Additional permissions can be applied to individual objects in storage1.
- Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

What should you use? To answer, drag the appropriate components to the correct requirements. Each component 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.



Components
Access control lists (ACLs)
Role-based access control (RBAC) roles
Shared access signatures (SAS)
Shared account keys

Answer Area

To grant permissions at the storage account level:

To grant permissions at the object level:

Answer:

Components
Access control lists (ACLs)
Role-based access control (RBAC) roles
Shared access signatures (SAS)
Shared account keys

Answer Area

To grant permissions at the storage account level:

To grant permissions at the object level:

Explanation:**Box 1: Role-based access control (RBAC) roles**

List and read permissions must be granted at the storage account level.

Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

Role-based access control (Azure RBAC)

Azure RBAC uses role assignments to apply sets of permissions to security principals. A security principal is an object that represents a user, group, service principal, or managed identity that is defined in Azure Active Directory (AD). A permission set can give a security principal a "coarse-grain" level of access such as read or write access to all of the data in a storage account or all of the data in a container.

Box 2: Access control lists (ACLs)

Additional permissions can be applied to individual objects in storage1. Access control lists (ACLs)

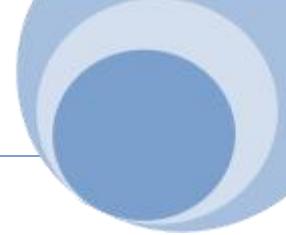
ACLs give you the ability to apply "finer grain" level of access to directories and files. An ACL is a permission construct that contains a series of ACL entries. Each ACL entry associates security principal with an access level.

Reference: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model>

140. - (Exam Topic 3)

You are designing a streaming data solution that will ingest variable volumes of data. You need to ensure that you can change the partition count after creation.





Which service should you use to ingest the data?

- A. Azure Event Hubs Dedicated
- B. Azure Stream Analytics
- C. Azure Data Factory
- D. Azure Synapse Analytics

Answer: B

Explanation:

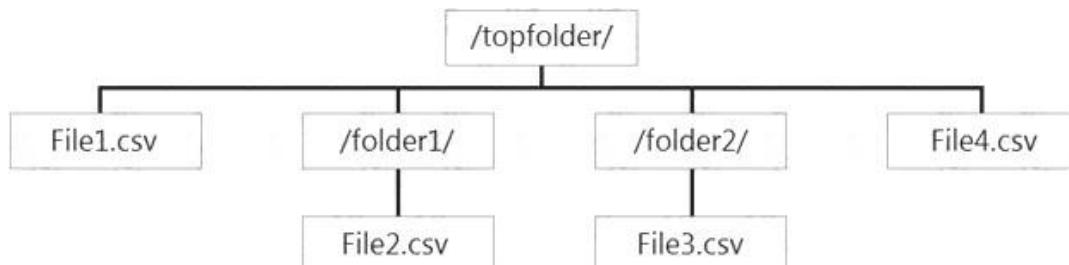
You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features>

141. - (Exam Topic 3)

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.



You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

- A. File2.csv and File3.csv only
- B. File1.csv and File4.csv only
- C. File1.csv, File2.csv, File3.csv, and File4.csv
- D. File1.csv only

Answer: B

Explanation:

To run a T-SQL query over a set of files within a folder or set of folders while treating them as a single entity or rowset, provide a path to a folder or a pattern (using wildcards) over a set of files or folders. Reference:



<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders>

142. - (Exam Topic 3)

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool.

You need to create a surrogate key for the table. The solution must provide the fastest query performance.

What should you use for the surrogate key?

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column

Answer: C

Explanation:

Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.

Note: 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>

143. - (Exam Topic 3)

You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool. You need to design queries to maximize the benefits of partition elimination. What should you include in the Transact-SQL queries?

- A. JOIN
- B. WHERE
- C. DISTINCT
- D. GROUP BY

Answer: B



144. - (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: C

Explanation:

Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

145. - (Exam Topic 3)

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage1.

New files are uploaded daily to storage1.

- Incrementally process new files as they are uploaded to storage1 as a structured streaming source. The solution must meet the following requirements:
 - Minimize implementation and maintenance effort.
 - Minimize the cost of processing millions of files.
 - Support schema inference and schema drift. Which should you include in the recommendation?

- A. Auto Loader
- B. Apache Spark FileStreamSource
- C. COPY INTO
- D. Azure Data Factory

Answer: D



146. - (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 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

Answer: A

Explanation:

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>

147. - (Exam Topic 3)

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?

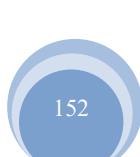
A. Azure Stream Analytics

B. Azure SQL Database

C. Azure Databricks

D. Azure Synapse Analytics

Answer: A





Explanation:

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics>

148. - (Exam Topic 3)

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. [EffectiveScarcDate] [datetime] NOT NULL,
- B. [CurrentProduccCacegory] [nvarchar] (100) NOT NULL,
- C. [EffectiveEndDace] [dacecime] NULL,
- D. [ProductCategory] [nvarchar] (100) NOT NULL,
- E. [OriginalProduccCacegory] [nvarchar] (100) NOT NULL,

Answer: B E

Explanation:

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.





Graphical user interface, application, email Description automatically generated

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/>

149. - (Exam Topic 3)

You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated
- C. interactive

Answer: C

Explanation:

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.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-schedule-d-bat>



150. - (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>

151. - (Exam Topic 3)

You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.



You plan to keep a record of changes to the available fields. The supplier data contains the following columns.

Name	Description
SupplierSystemID	Unique supplier ID in an enterprise resource planning (ERP) system
SupplierName	Name of the supplier company
SupplierAddress1	Address of the supplier company
SupplierAddress2	Second address line of the supplier company
SupplierCity	City of the supplier company
SupplierStateProvince	State or province of the supplier company
SupplierCountry	Country of the supplier company
SupplierPostalCode	Postal code of the supplier company
SupplierDescription	Free-text description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. surrogate primary key
- B. foreign key
- C. effective start date
- D. effective end date
- E. last modified date
- F. business key

Answer: C D F

Explanation:

Reference:

<https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension>

152. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named pool1.

You need to perform a monthly audit of SQL statements that affect sensitive data. The solution must minimize administrative effort.



What should you include in the solution?

- A. Microsoft Defender for SQL
- B. dynamic data masking
- C. sensitivity labels
- D. workload management

Answer: B

153. - (Exam Topic 3)

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

Table	Comments
Sales	The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Severity-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DateKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping.

Queries that use the data warehouse take a long time to complete.

You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.

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

NOTE: Each correct selection is worth one point

Table	Distribution type	Distribution column
-------	-------------------	---------------------

Sales:

Hash-distributed
Round-robin

DateKey
ProductKey
RegionKey

Invoices:

Hash-distributed
Round-robin

DateKey
ProductKey
RegionKey



**Answer:**

Table	Distribution type	Distribution column
Sales:	Hash-distributed Round-robin	DateKey ProductKey RegionKey
Invoices:	Hash-distributed Round-robin	DateKey ProductKey RegionKey

Explanation:

Box 1: Hash-distributed

Box 2: ProductKey

ProductKey is used extensively in joins.

Hash-distributed tables improve query performance on large fact tables. Box 3: Round-robin

Box 4: RegionKey

Round-robin tables are useful for improving loading speed.

Consider using the round-robin distribution for your table in the following scenarios:

- When getting started as a simple starting point since it is the default
- If there is no obvious joining key
- If there is not good candidate column for hash distributing the table
- If the table does not share a common join key with other tables
- If the join is less significant than other joins in the query
- When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions.

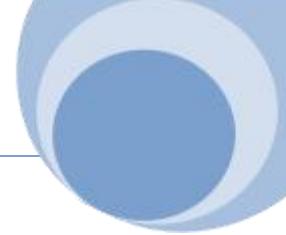
The rows are distributed with a hash or round-robin algorithm.

Reference:

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

154. - (Exam Topic 3)

From a website analytics system, you receive data extracts about user interactions such as downloads, link



clicks, form submissions, and video plays.

The data contains the following columns.

Name	Sample value
Date	15 Jan 2021
EventCategory	Videos
EventAction	Play
EventLabel	Contoso Promotional
ChannelGrouping	Social
TotalEvents	150
UniqueEvents	120
SessionWithEvents	99

You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.

To which table should you add each column? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

EventCategory:

▼

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

▼

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

▼

DimChannel
DimDate
DimEvent
FactEvents

Answer:





EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

Explanation:

Table Description automatically generated

Box 1: DimEvent

Box 2: DimChannel

Box 3: FactEvents

Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc

Reference:

<https://docs.microsoft.com/en-us/power-bi/guidance/star-schema>

155. - (Exam Topic 3)

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

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month
- AccountType: 500 million per account type

You have the following query requirements:





- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account 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. CustomerSegment
- B. AccountType
- C. TransactionType
- D. TransactionMonth

Answer: C

Explanation:

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

Example: Any partitioning added to a table is in addition to the distributions created behind the scenes.

Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

156. - (Exam Topic 3)

You are deploying a lake database by using an Azure Synapse database template.

You need to add additional tables to the database. The solution must use the same grouping method as the template tables.

'Which grouping method should you use?

- A. business area
- B. size
- C. facts and dimensions
- D. partition style

Answer: A





Explanation:

- Business area: This is how the Azure Synapse database templates group tables by default. Each template consists of one or more enterprise templates that contain tables grouped by business areas. For example, the Retail template has business areas such as Customer, Product, Sales, and Store123. Using the same grouping method as the template tables can help you maintain consistency and compatibility with the industry-specific data model.

<https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/database-templates-in-azure-synapse-anal>

157. - (Exam Topic 3)

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.

You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.

What should you do?

- A. Clone the cluster after it is terminated.
- B. Terminate the cluster manually when processing completes.
- C. Create an Azure runbook that starts the cluster every 90 days.
- D. Pin the cluster.

Answer: D

Explanation:

To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

References:

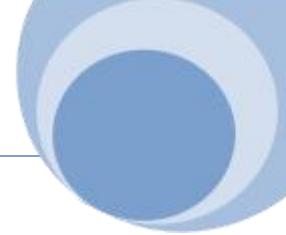
<https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination>

158. - (Exam Topic 3)

You are designing a highly available Azure Data Lake Storage solution that will induce geo-zone-redundant storage (GZRS).

You need to monitor for replication delays that can affect the recovery point objective (RPO). What should





you include in the monitoring solution?

- A. Last Sync Time
- B. Average Success Latency
- C. Error errors
- D. availability

Answer: A

Explanation:

Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get>

159. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

- Customer
- Salesperson
- Region
- Amount

You need to identify the three top performing salespersons by amount for a region named HQ.

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.





Values	Answer Area
agg(col('SalesPerson'))	df_sales.filter(col('Region')=='HQ').
filter(col('SalesPerson'))	.agg(sum('Amount').alias
groupBy(col('SalesPerson'))	('TotalAmount')).
groupBy(col('TotalAmount'))	
orderBy(col('TotalAmount'))	
orderBy(desc('TotalAmount'))	.limit(3)

Answer:

Values	Answer Area
agg(col('SalesPerson'))	df_sales.filter(col('Region')=='HQ').
filter(col('SalesPerson'))	filter(col('SalesPerson'))
groupBy(col('SalesPerson'))	.agg(sum('Amount').alias
groupBy(col('TotalAmount'))	('TotalAmount')).
orderBy(col('TotalAmount'))	
orderBy(desc('TotalAmount'))	.limit(3)

160. - (Exam Topic 3)

You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS).

You need to monitor for replication delays that can affect the recovery point objective (RPO). What should you include in the monitoring solution?

- A. availability
- B. Average Success E2E Latency
- C. 5xx: Server Error errors
- D. Last Sync Time

Answer: D

Explanation:

Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location.



Writes made to the primary region after the last sync time property may or may not be available for reads yet.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get>

161. - (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: B

Explanation:

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.

Reference:

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

162. - (Exam Topic 3)

You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.



```
CREATE TABLE [dbo].[DimEmployee] (
[EmployeeKey] [int] IDENTITY(1,1) NOT NULL,
[EmployeeID] [int] NOT NULL,
[FirstName] [varchar](100) NOT NULL,
[LastName] [varchar](100) NOT NULL,
[JobTitle] [varchar](100) NULL,
[LastHireDate] [date] NULL,
[StreetAddress] [varchar](500) NOT NULL,
[City] [varchar](200) NOT NULL,
[StateProvince] [varchar](50) NOT NULL,
[Portalcode] [varchar](10) NOT NULL
)
```

You need to alter the table to meet the following requirements:

- Ensure that users can identify the current manager of employees.
- Support creating an employee reporting hierarchy for your entire company.
- Provide fast lookup of the managers' attributes such as name and job title.

Which column should you add to the table?

- A. [ManagerEmployeeID] [int] NULL
- B. [ManagerEmployeeID] [smallint] NULL
- C. [ManagerEmployeeKey] [int] NULL
- D. [ManagerName] [varchar](200) NULL

Answer: A

Explanation:

Use the same definition as the EmployeeID column. Reference:

<https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular>

163. - (Exam Topic 3)

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.

Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Library:

Azure Databricks Monitoring Library
Microsoft Azure Management Monitoring Library
PyTorch
TensorFlow

Workspace:

Azure Databricks
Azure Log Analytics
Azure Machine Learning

Answer:**Library:**

Azure Databricks Monitoring Library
Microsoft Azure Management Monitoring Library
PyTorch
TensorFlow

Workspace:

Azure Databricks
Azure Log Analytics
Azure Machine Learning

Explanation:

You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.

References:

<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

164. - (Exam Topic 3)

You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.

You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

- No transformations must be performed.



- The original folder structure must be retained.
- Minimize time required to perform the copy activity.

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

NOTE: Each correct selection is worth one point.

Source dataset type:

Binary
Parquet
Delimited text

Copy activity copy behavior:

FlattenHierarchy
MergeFiles
PreserveHierarchy

Answer:

Source dataset type:

Binary
Parquet
Delimited text

Copy activity copy behavior:

FlattenHierarchy
MergeFiles
PreserveHierarchy

Explanation:

Graphical user interface, text, application, chat or text message Description automatically generated

Box 1: Parquet

For Parquet datasets, the type property of the copy activity source must be set to ParquetSource. Box 2:

PreserveHierarchy

PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/format-parquet>

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

165. - (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 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

Answer: B

Explanation:

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>

166. - (Exam Topic 3)

You are designing an Azure Synapse Analytics dedicated SQL pool.

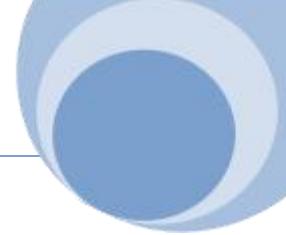
You need to ensure that you can audit access to Personally Identifiable information (PII). What should you include in the solution?

A. dynamic data masking

B. row-level security (RLS)

C. sensitivity classifications

D. column-level security



Answer: C

Explanation:

Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.

Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:

- Helping to meet standards for data privacy and requirements for regulatory compliance.
- Various security scenarios, such as monitoring (auditing) access to sensitive data.
- Controlling access to and hardening the security of databases that contain highly sensitive data.

Reference:

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

167. - (Exam Topic 3)

You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account. Pipeline 1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipelinel executes, you discover that data is NOT copied to the new storage account. You need to ensure that the data is copied to the new storage account.

What should you do?

- A. Publish from the collaboration branch.
- B. Configure the change feed of the new storage account.
- C. Create a pull request.
- D. Modify the schedule trigger.

Answer: A

Explanation: CI/CD lifecycle

- A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.

- A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes
- After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.
- After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

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

168. - (Exam Topic 3)

You plan to monitor an Azure data factory by using the Monitor & Manage app.

You need to identify the status and duration of activities that reference a table in a source database.

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

Actions	Answer Area
From the Data Factory monitoring app, add the Source user property to the Activity Runs table.	
From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.	
From the Data Factory authoring UI, publish the pipelines.	
From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.	
From the Data Factory authoring UI, generate a user property for Source on all activities.	
From the Data Factory authoring UI, generate a user property for Source on all datasets.	

Answer:



Actions	Answer Area
From the Data Factory monitoring app, add the Source user property to the Activity Runs table.	From the Data Factory authoring UI, generate a user property for Source on all activities.
From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.	From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.
From the Data Factory authoring UI, publish the pipelines.	From the Data Factory authoring UI, publish the pipelines.
From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.	From the Data Factory authoring UI, publish the pipelines.
From the Data Factory authoring UI, generate a user property for Source on all activities.	From the Data Factory authoring UI, generate a user property for Source on all datasets.

Explanation:

Step 1: From the Data Factory authoring UI, generate a user property for Source on all activities. Step 2:

From the Data Factory monitoring app, add the Source user property to Activity Runs table.

You can promote any pipeline activity property as a user property so that it becomes an entity that you can monitor. For example, you can promote the Source and Destination properties of the copy activity in your pipeline as user properties. You can also select Auto Generate to generate the Source and Destination user properties for a copy activity.

Step 3: From the Data Factory authoring UI, publish the pipelines

Publish output data to data stores such as Azure SQL Data Warehouse for business intelligence (BI) applications to consume.

References:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually>

169. - (Exam Topic 3)

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.

Values	Answer Area
all, ecommerce, retail, wholesale	CleanData
dept=='ecommerce', dept=='retail', dept=='wholesale'	split(
dept=='ecommerce', dept== 'wholesale', dept=='retail'	[]
disjoint: false	[]
disjoint: true) ~> SplitByDept@([])
ecommerce, retail, wholesale, all	

Answer:

Values	Answer Area
all, ecommerce, retail, wholesale	CleanData
dept=='ecommerce', dept=='retail', dept=='wholesale'	split(
dept=='ecommerce', dept== 'wholesale', dept=='retail'	dept=='ecommerce', dept=='retail', dept=='wholesale'
disjoint: false	disjoint: false
disjoint: true) ~> SplitByDept@([])
ecommerce, retail, wholesale, all	ecommerce, retail, wholesale, all)

Explanation:

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>

170. - (Exam Topic 3)

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. Delete the files in the destination before loading new data.
- B. Filter by the last modified date of the source files.
- C. Delete the source files after they are copied.
- D. Specify a file naming pattern for the destination.

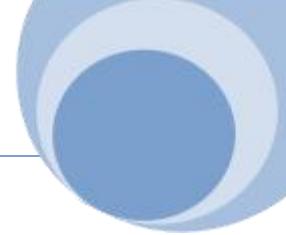
Answer: B D

Explanation:

Copy data from one place to another. The requirements are : 1- need to minimize transfert and 2- need to adapte data to the destination folder structure. Filter on LastModifiedDate will copy everything that have changed since the latest load while minimizing the data transfert. Specifying the file naming pattern allows to copy data at the right place to the destination Data Lake.

171. - (Exam Topic 3)

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



Connections

- Linked services
- Integration runtimes
- Source control
- Git configuration**
- ARM template
- Parameterization template
- Author
- Triggers
- Global parameters
- Security
- Customer managed key
- Managed private endpoints

Git repositoryGit repository information associated with your data factory. CI/CD best practices
 Setting
 Disconnect

Repository type	Azure DevOps Git
Azure DevOps Account	CONTOSO
Project name	Data
Repository name	dwh_batchetl
Collaboration branch	main
Publish branch	adf_publish
Root folder	/

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.

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

Answer:





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	

Explanation:

Letter Description automatically generated

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

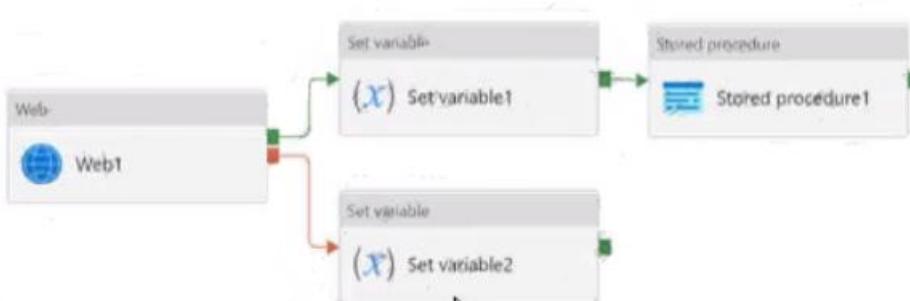
Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.

Reference:

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

172. - (Exam Topic 3)

You have an Azure Data Factory pipeline that has the activity 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.



Answer Area

Stored procedure1 will execute if Web1 and Set variable1 [answer choice]

complete
fail
succeed

These are the selections for the statement Stored procedure1 will execute if Web1 and Set variable1 [answer choice]

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

Canceled
Failed
Succeeded

These are the selections for the statement If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

Answer:

Answer Area

Stored procedure1 will execute if Web1 and Set variable1 [answer choice] succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice] Failed

173. - (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics.

You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads

Which is the best metric to monitor?

More than one answer choice may achieve the goal. Select the BEST answer.

- A. Data 10 percentage
- B. CPU percentage
- C. DWU used
- D. DWU percentage

Answer: C

174. - (Exam Topic 3)

You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contain approximately 60 million rows per month and will be partitioned by month. The table will use a clustered column store index and round-robin distribution.

Approximately how many rows will there be for each combination of distribution and partition?

- A. 1 million





- B. 5 million
- C. 20 million
- D. 60 million

Answer: D

Explanation:

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

175. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1.

You need to identify tables that have a high percentage of deleted rows. What should you run?

A)

`sys.pdw_nodes_column_store_segments`

B)

`sys.dm_db_column_store_row_group_operational_stats`

C)

`sys.pdw_nodes_column_store_row_groups`

D)

`sys.dm_db_column_store_row_group_physical_stats`

A. Option

B. Option

C. Option

D. Option

Answer: B

176. - (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowedBlobpublicAccess property is disabled for storage1.

You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to





access storage1 from Pool1.

What should you create first?

- A. an external resource pool
- B. a remote service binding
- C. database scoped credentials
- D. an external library

Answer: C

Explanation: Security

User must have SELECT permission on an external table to read the data. External tables access underlying Azure storage using the database scoped credential defined in data source.

Note: A database scoped credential is a record that contains the authentication information that is required to connect to a resource outside SQL Server. Most credentials include a Windows user and password.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql>

177. - (Exam Topic 3)

You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.

You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.

Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.





Databricks:

Azure Active Directory credential passthrough
Azure Key Vault secrets
Personal access tokens

Data Lake Storage:

Azure Active Directory credential passthrough
Shared access keys
Shared access signatures

Answer:

Databricks:

Azure Active Directory credential passthrough
Azure Key Vault secrets
Personal access tokens

Data Lake Storage:

Azure Active Directory credential passthrough
Shared access keys
Shared access signatures

Explanation:

Table Description automatically generated

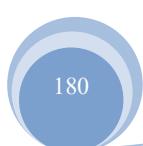
Box 1: Personal access tokens

You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.

You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.

Box 2: Azure Active Directory credential passthrough

You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS Gen2) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.



After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfss:// path:

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-ac>

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

178. - (Exam Topic 3)

You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company. You have the following data access requirements:

- After initial processing, the HR department data will be retained for seven years.
- The operations department data will be accessed frequently for the first six months, and then accessed once per month.

You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.

Answer:

See the answer in explanation:..

Explanation:

Answer Area

HR: Archive storage after one day and delete storage after 2.555 days.

Operations: Cool storage after 180 days.

179. - (Exam Topic 3)

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.

```

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:



```

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	

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: WHERE EventType='HeartBeat' Box 2: ,TumblingWindow(Second, 5)

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

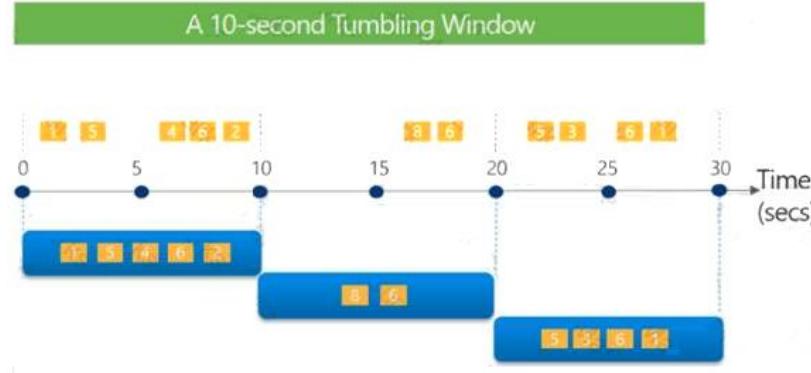
The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Timeline Description automatically generated





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



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

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>

180. - (Exam Topic 3)

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.

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	▼

Answer:

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	▼

Explanation:

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>

181. - (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 have an Azure Storage account that contains 100 GB of files. The files contain rows of 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 enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly. Solution: You copy the files to a table that has a columnstore index. Does this meet the goal?

A. Yes

B. No

Answer: B

Explanation:

Instead convert the files to compressed delimited text files. Reference:

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

182. - (Exam Topic 3)

You have an Azure data factory that has the Git repository settings shown in the following exhibit.

Git repository

Git repository information associated with your data factory. CI/CD best practices 

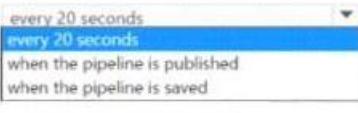
 Edit	 Overwrite live mode	 Disconnect	 Import resources
Repository type	Azure DevOps Git		
Azure DevOps Account			
Project name	ADFDemo		
Repository name	ADFDemo		
Collaboration branch	main		
Publish branch	adf_publish		
Root folder	/		
Last published commit	23b144ac4aa7daf16f2fe7c2ab0eb303a8e4ed65		
Publish (from ADF Studio)	Enabled		

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

NOTE: Each correct answer is worth one point.

Answer Area

Changes to pipelines will be saved in Azure DevOps [answer choice].



To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].


Answer:**Answer Area**

Changes to pipelines will be saved in Azure DevOps [answer choice].



To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].


183. - (Exam Topic 3)

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
- C. Azure IoT Hub
- D. Azure Event Hubs

Answer: B

Explanation:

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>

184. - (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 Analytics dedicated SQL pool.

The CSV file contains 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. Add a select transformation that selects only the rows which will cause truncation errors.
- B. Add a sink transformation that writes the rows to a file in blob storage.
- C. Add a filter transformation that filters out rows which will cause truncation errors.
- D. Add a Conditional Split transformation that separates the rows which will cause truncation errors.

Answer: B D





185. - (Exam Topic 3)

You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1.

You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.

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

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

Actions	Answer Area
Create a database role named Role1 and grant Role1 SELECT permissions to schema1.	
Create a database role named Role1 and grant Role1 SELECT permissions to dw1.	
Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.	
Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.	
Assign Role1 to the Group1 database user.	

Answer:

Actions	Answer Area
Create a database role named Role1 and grant Role1 SELECT permissions to schema1.	Create a database role named Role1 and grant Role1 SELECT permissions to schema1.
Create a database role named Role1 and grant Role1 SELECT permissions to dw1.	
Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.	Assign Role1 to the Group1 database user.
Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.	Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.
Assign Role1 to the Group1 database user.	

Explanation:

Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema You need to grant Group1 read-only permissions to all the tables and views in schema1.

Place one or more database users into a database role and then assign permissions to the database role.

Step 2: Assign Rol1 to the Group database user

Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1

Reference:

<https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql>

186. - (Exam Topic 3)

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage. MP1 is configured

10 scan storage1. DF1 is connected to MP1 and contains 3 dataset named DS1. DS1 references 2 file in storage.

In DF1, you plan to create a pipeline that will process data from DS1.

You need to review the schema and lineage information in MP1 for the data referenced by DS1.

Which two features can you use to locate the information? Each correct answer presents a complete solution. NOTE: Each correct answer is worth one point.

- A. the Storage browser of storage1 in the Azure portal
- B. the search bar in the Azure portal
- C. the search bar in Azure Data Factory Studio
- D. the search bar in the Microsoft Purview governance portal

Answer: C D

Explanation:

- The search bar in the Microsoft Purview governance portal: This is a feature that allows you to search for assets in your data estate using keywords, filters, and facets. You can use the search bar to find the files in storage1 that are referenced by DS1, and then view their schema and lineage information in the asset details page12.
- The search bar in Azure Data Factory Studio: This is a feature that allows you to search for datasets, linked services, pipelines, and other resources in your data factory. You can use the search bar to find DS1 in DF1, and then view its schema and lineage information in the dataset details page. You can also click on



the Open in Purview button to open the corresponding asset in MP13.

The two features that can be used to locate the schema and lineage information for the data referenced by DS1 are the search bar in Azure Data Factory Studio and the search bar in the Microsoft Purview governance portal.

The search bar in Azure Data Factory Studio allows you to search for the dataset DS1 and view its properties and lineage. This can help you locate information about the source and destination data stores, as well as the transformations that were applied to the data.

The search bar in the Microsoft Purview governance portal allows you to search for the storage account and view its metadata, including schema and lineage information. This can help you understand the different data assets that are stored in the storage account and how they are related to each other.

The Storage browser of storage1 in the Azure portal may allow you to view the files that are stored in the storage account, but it does not provide lineage or schema information for those files. Similarly, the search bar in the Azure portal may allow you to search for resources in the Azure subscription, but it does not provide detailed information about the data assets themselves.

References:

- What is Azure Purview?
- Use Azure Data Factory Studio

187. - (Exam Topic 3)

You use Azure Data Lake Storage Gen2.

You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.

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

- A. Reregister the Microsoft Data Lake Store resource provider.
- B. Reregister the Azure Storage resource provider.
- C. Create a storage policy that is scoped to a container.
- D. Register the query acceleration feature.
- E. Create a storage policy that is scoped to a container prefix filter.

Answer: B D





188. - (Exam Topic 3)

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.

Answer: D

Explanation:

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:

- Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
- Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
- 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>

189. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contains data on a single sale, including the name of the salesperson.

You need to implement row-level security (RLS). The solution must ensure that the salespeople can access





only their respective sales.

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

Create:

- A materialized view in Pool1
- A security policy for Sales
- Database scoped credentials in Pool1

Restrict row access by using:

- A masking rule
- A table-valued function
- The CONTAINS predicate

Answer:

Create:

- A materialized view in Pool1
- A security policy for Sales
- Database scoped credentials in Pool1

Restrict row access by using:

- A masking rule
- A table-valued function
- The CONTAINS predicate

Explanation:

Box 1: A security policy for sale

Here are the steps to create a security policy for Sales:

- Create a user-defined function that returns the name of the current user:
- CREATE FUNCTION dbo.GetCurrentUser()
- RETURNS NVARCHAR(128)
- AS
- BEGIN
- RETURN SUSER_SNAME();
- END;
- Create a security predicate function that filters the Sales table based on the current user:
- CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128))
- RETURNS TABLE





- WITH SCHEMABINDING
- AS
- RETURN SELECT 1 AS access_result
- WHERE @salesperson = SalespersonName;
- Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:
- CREATE SECURITY POLICY SalesFilter
- ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales
- WITH (STATE = ON);

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user.

Box 2: table-value function

to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security policy that applies a predicate on the table.

190. - (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 less than 1 MB. Does this meet the goal?

A. Yes

B. No

Answer: A



Explanation:

When exporting data into an ORC File Format, you might get Java out-of-memory errors when there are large text columns. To work around this limitation, export only a subset of the columns.

References:

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

191. - (Exam Topic 3)

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.

You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.

What should you implement?

- A. a materialized view
- B. a replicated table
- C. in ordered clustered columnstore index
- D. result set caching

Answer: A

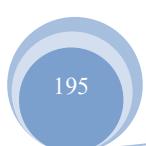
Explanation:

Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change.

Note: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached results set do not use any concurrency slots and thus do not count against existing concurrency limits.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-views>



[https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-s
et-cac](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cac)

192. - (Exam Topic 3)

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>
- B. %<Language>
- C. \<Language>
- D. \<Language>

Answer: B

Explanation:

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-a-zur>

193. - (Exam Topic 3)

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.

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

Answer:

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

Explanation:

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>

194. - (Exam Topic 3)



You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool.

The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

FactPurchase will have 1 million rows of data added daily and will contain three years of data.

Transact-SQL queries similar to the following query will be executed daily.

SELECT

SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase

WHERE DateKey >= 20210101 AND DateKey <= 20210131

GROUP By SupplierKey, StockItemKey

Which table distribution will minimize query times?

- A. round-robin
- B. replicated
- C. hash-distributed on DateKey
- D. hash-distributed on PurchaseKey

Answer: D

Explanation:

Hash-distributed tables improve query performance on large fact tables, and are the focus of this article.

Round-robin tables are useful for improving loading speed.

Reference:

[https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribution)



195. - (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQ1 pool.

You have an Azure Data Lake Storage account named adls1 that contains a public container named container1. The container 1 container contains a folder named folder 1.

You need to query the top 100 rows of all the CSV files in folder 1.

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

NOTE Each correct selection is worth one point.

Values	Answer Area
BULK	
DATA_SOURCE	
LOCATION	
OPENROWSET	

```

SELECT TOP 100 *
FROM [OPENROWSET
BULK 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
  
```

Answer:

Values	Answer Area
BULK	
DATA_SOURCE	
LOCATION	
OPENROWSET	

```

SELECT TOP 100 *
FROM [OPENROWSET
BULK 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
  
```

196. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1. The synapse1 workspace contains an Apache Spark pool named pool1.

You need to share an Apache Hive catalog of pool1 with databricks1.

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

NOTE: Each correct selection is worth one point.

From synapse1, create a linked service to:

- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:

- An Azure Purview account
- A Hive metastore
- A managed Hive metastore service

Answer:

From synapse1, create a linked service to:

- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:

- An Azure Purview account
- A Hive metastore
- A managed Hive metastore service

Explanation:

Box 1: Azure SQL Database

Use external Hive Metastore for Synapse Spark Pool

Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace.

- Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service.
- Set up Hive Metastore linked service
- Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.

- Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.
- You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually.
- Provide User name and Password to set up the connection.
- Test connection to verify the username and password.



- Click Create to create the linked service.

Box 2: A Hive Metastore

nce: <https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

197. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

You need to Create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:

Show order counts by week.

- Calculate sales totals by region.
- Calculate sales totals by product.
- Find all the orders from a given month. Which data should you use to partition Table1?

A. region

B. product

C. week

D. month

Answer: D

Explanation:

Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

Benefits to queries

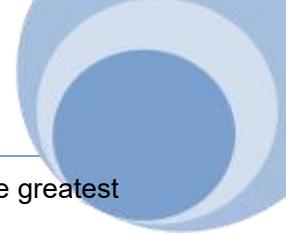
Partitioning can also be used to improve query performance. A query that applies a filter to partitioned data can limit the scan to only the qualifying partitions. This method of filtering can avoid a full table scan and only scan a smaller subset of data. With the introduction of clustered columnstore indexes, the predicate elimination performance benefits are less beneficial, but in some cases there can be a benefit to queries.

For example, if the sales fact table is partitioned into 36 months using the sales date field, then queries that filter on the sale date can skip searching in partitions that don't match the filter.

Note: Benefits to loads

The primary benefit of partitioning in dedicated SQL pool is to improve the efficiency and performance of loading data by use of partition deletion, switching and merging. In most cases data is partitioned on a date





column that is closely tied to the order in which the data is loaded into the SQL pool. One of the greatest benefits of using partitions to maintain data is the avoidance of transaction logging. While simply inserting, updating, or deleting data can be the most straightforward approach, with a little thought and effort, using partitioning during your load process can substantially improve performance.

Reference:

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

198. - (Exam Topic 3)

You have a C# application that processes data from an Azure IoT hub and performs complex transformations. You need to replace the application with a real-time solution. The solution must reuse as much code as possible from the existing application.

- A. Azure Databricks
- B. Azure Event Grid
- C. Azure Stream Analytics
- D. Azure Data Factory

Answer: C

Explanation:

Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data. UDFs are available in C# for IoT Edge jobs.

Azure Stream Analytics on IoT Edge runs within the Azure IoT Edge framework. Once the job is created in Stream Analytics, you can deploy and manage it using IoT Hub.

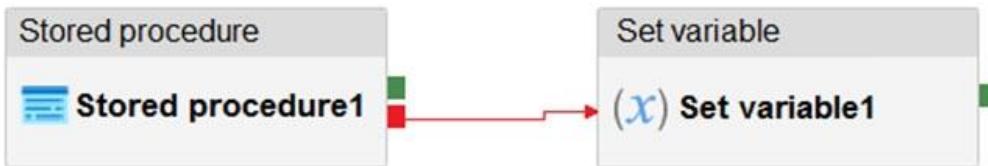
References:

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

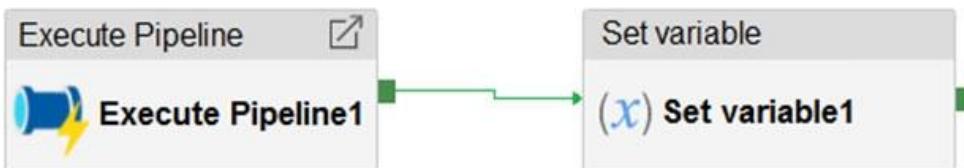
199. - (Exam Topic 3)

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.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

Answer: A

Explanation:

Activities are linked together via dependencies. A dependency has a condition of one of the following:

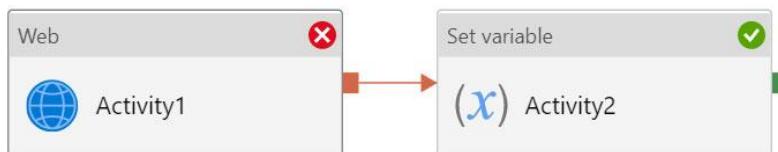
Succeeded, Failed, Skipped, or Completed.

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.

Waterfall chart Description automatically generated with medium confidence

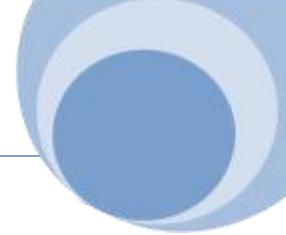


The failure dependency means this pipeline reports success. Note:

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/>



200. - (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>

201. - (Exam Topic 3)

You are designing a 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



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:

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:

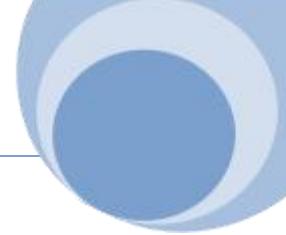
Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Explanation:

Reference:

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

202. - (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: A C

Explanation:

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

Schema	Table	Column
SalesLT	Customer	FirstName
SalesLT	Customer	LastName
SalesLT	Customer	EmailAddress
SalesLT	Customer	Phone
SalesLT	Customer	PasswordHash
SalesLT	Customer	PasswordSalt
dbo	ErrorLog	UserName
SalesLT	Address	AddressLine1
SalesLT	Address	AddressLine2
SalesLT	Address	City
SalesLT	Address	PostalCode
SalesLT	CustomerAddress	AddressType
SalesLT	SalesOrderHeader	AccountNumber
SalesLT	SalesOrderHeader	CreditCardApprovalCode
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
	[REDACTED] 7.125	Microsoft SQL Server Management Studio - Query	1	847	847	C244A066-2271...	Confidential - GDPR
	[REDACTED] 7.125	Microsoft SQL Server Management Studio - Query	2	32	32	C244A066-2271...	Confidential
	[REDACTED] 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>

203. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap.

Most queries against the table aggregate values from approximately 100 million rows and return only two columns.

You discover that the queries against the fact table are very slow. Which type of index should you add to provide the fastest query times?

- A. nonclustered columnstore
- B. clustered columnstore
- C. nonclustered
- D. clustered

Answer: B

Explanation:

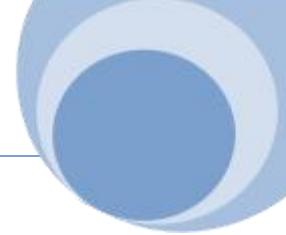
Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool.

Columnstore tables won't benefit a query unless the table has more than 60 million rows. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

204. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1



contains a fact table named Table1.

You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and run dbcc pdw_showspaceused.
- B. Connect to the built-in pool and run dbcc checkalloc.
- C. Connect to Pool1 and query sys.dm_pdw_node_scacus.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_scacs.

Answer: A

Explanation:

A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.

DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales'); Reference:

[https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribution)

205. - (Exam Topic 3)

Note: The 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 dairy 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 executes a mapping data flow and then inserts the data into the data warehouse.

Does this meet the goal?

- A. Yes
- B. No





Answer: B

206. - (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB.

What should you do?

- A. On the master database, execute a query against the sys.dmv_nodes_os_performance_counters dynamic management view.
- B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
- C. On DW1, execute a query against the sys.database_files dynamic management view.
- D. Execute a query against the logs of DW1 by using the

Get-AzOperationalInsightSearchResult PowerShell cmdlet.

Answer: A

Explanation:

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

```
-- Transaction log size
SELECT
    instance_name AS distribution_db, cntr_value * 1.0 / 1048576 AS log_file_size_used_Gb, pdw_node_id
FROM sys.dmv_nodes_os_performance_counters
WHERE
    instance_name LIKE 'Distribution_%'
    AND counter_name = 'Log File(s) Used Size (KB)'
```

References:

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

207. - (Exam Topic 3)

You are designing the folder structure for an Azure Data Lake Storage Gen2 container.

Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.





Which folder structure should you recommend to support fast queries and simplified folder security?

- A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}_{YYYY}_{MM}_{DD}.csv
- B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
- C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
- D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}_{YYYY}_{MM}_{DD}.csv

Answer: D

Explanation:

There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise, if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on.

Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout:

{Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/

208. - (Exam Topic 3)

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. Storage event
- B. on-demand





- C. schedule
- D. tumbling window

Answer: D

Explanation:

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>

209. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Sales.Orders.

Sales.Orders contains a column named SalesRep.

You plan to implement row-level security (RLS) for Sales.Orders.

You need to create the security policy that will be used to implement RLS. The solution must ensure that sales representatives only see rows for which the value of the SalesRep column matches their username.

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

Answer Area

```

CREATE SCHEMA Security;
GO
CREATE FUNCTION Security.tvf_securitypredicate(@SalesRep AS nvarchar(50))
RETURNS TABLE
WITH SCHEMABINDING
ENCRYPTION
RETURNS NULL ON NULL INPUT
SCHEMABINDING
AS
RETURN SELECT 1 AS tvf_securitypredicate_result
WHERE @SalesRep = USER_NAME();
GO
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)
ADD BLOCK PREDICATE tvf_securitypredicate_result
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)

```



**Answer:****Answer Area**

```

CREATE SCHEMA Security;
GO

CREATE FUNCTION Security.tvf_securitypredicate(@SalesRep AS nvarchar(50))
RETURNS TABLE
WITH SCHEMABINDING
ENCRYPTION
RETURNS NULL ON NULL INPUT
SCHEMABINDING
AS
RETURN SELECT 1 AS tvf_securitypredicate_result
WHERE @SalesRep = USER_NAME();
GO

CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)
ADD BLOCK PREDICATE tvf_securitypredicate_result
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)

```

210. - (Exam Topic 3)

You plan to develop a dataset named Purchases by using Azure databricks Purchases will contain the following columns:

- ProductID
- ItemPrice
- lineTotal
- Quantity
- StoreID
- Minute
- Month
- Hour
- Year
- Day

You need to store the data to support hourly incremental load pipelines that will vary for each StoreID. the solution must minimize storage costs. How should you complete the rode? To answer, select the appropriate options In the answer area.





NOTE: Each correct selection is worth one point.

df.write

.bucketBy
.partitionBy
.range
.sortBy

{"*"} ("StoreID", "Hour") ("StoreID", "Year", "Month", "Day", "Hour")

.mode("append")

.csv("/Purchases") .json("/Purchases") .parquet("/Purchases") .saveAsTable("/Purchases")

Answer:

df.write

.bucketBy
.partitionBy
.range
.sortBy

{"*"} ("StoreID", "Hour") ("StoreID", "Year", "Month", "Day", "Hour")

.mode("append")

.csv("/Purchases") .json("/Purchases") .parquet("/Purchases") .saveAsTable("/Purchases")

Explanation:

Box 1: partitionBy

We should overwrite at the partition level. Example: df.write.partitionBy("y","m","d")

mode(SaveMode.Append)

parquet("/data/hive/warehouse/db_name.db/" + tableName) Box 2: ("StoreID", "Year", "Month", "Day", "Hour", "StoreID") Box 3: parquet("/Purchases")

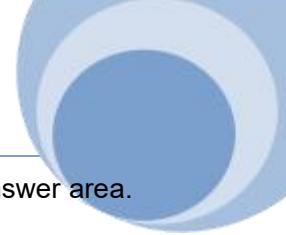
Reference:

<https://intellipaat.com/community/11744/how-to-partition-and-write-dataframe-in-spark-without-deleting-partitions>

211. - (Exam Topic 3)

You need to implement an Azure Databricks cluster that automatically connects to Azure Data Lake Storage Gen2 by using Azure Active Directory (Azure AD) integration.





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

NOTE: Each correct selection is worth one point.

Cluster Mode:

High Concurrency
Premium
Standard

Advanced option to enable:

Azure Data Lake Storage Gen1 Credential Passthrough
Table Access Control

Answer:

Cluster Mode:

High Concurrency
Premium
Standard

Advanced option to enable:

Azure Data Lake Storage Gen1 Credential Passthrough
Table Access Control

Explanation:

Box 1: High Concurrency

Enable Azure Data Lake Storage credential passthrough for a high-concurrency cluster. Incorrect:

Support for Azure Data Lake Storage credential passthrough on standard clusters is in Public Preview.

Standard clusters with credential passthrough are supported on Databricks Runtime 5.5 and above and are limited to a single user.

Box 2: Azure Data Lake Storage Gen1 Credential Passthrough

You can authenticate automatically to Azure Data Lake Storage Gen1 and Azure Data Lake Storage Gen2 from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.

References:



212. - (Exam Topic 3)

You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.

You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible.

Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. If Condition
- B. ForEach
- C. Lookup
- D. Get Metadata

Answer: B C

Explanation:

Lookup: This is a control activity that retrieves a dataset from any of the supported data sources and makes it available for use by subsequent activities in the pipeline. You can use a Lookup activity to read File1.txt from storage1 and store its content as an array variable1.

ForEach: This is a control activity that iterates over a collection and executes specified activities in a loop. You can use a ForEach activity to loop over the array variable from the Lookup activity and pass each table name as a parameter to a Copy activity that copies data from DB1 to storage11.

213. - (Exam Topic 3)

You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table.



Files with an invalid schema cause errors to occur. You need to monitor for an invalid schema error. For which error should you monitor?

- A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external files.'
- B. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'
- C. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11": for linked server "(null)", Query aborted- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.
- D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external files.'

Answer: C

Explanation: Customer Scenario:

SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.

SSMS Error:

Select query on the external table gives the following error: Msg 7320, Level 16, State 110, Line 14
Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)".
Query aborted-- the maximum reject threshold (0 rows) was reached while reading from an external source:
1 rows rejected out of total 1 rows processed.

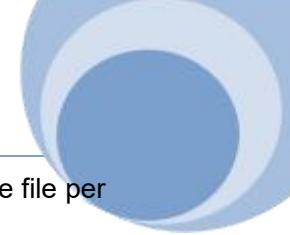
Possible Reason:

The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.

Possible Solution:

If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular





file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

214. - (Exam Topic 3)

You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days. What should you use?

- A. the Activity log blade for the Data Factory resource
- B. the Monitor & Manage app in Data Factory
- C. the Resource health blade for the Data Factory resource
- D. Azure Monitor

Answer: D

Explanation:

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

215. - (Exam Topic 3)

You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.

You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.

Which type of transformation should you add to the data flow?

- A. join
- B. select
- C. surrogate key
- D. alter row

Answer: D

Explanation:

The alter row transformation allows you to specify insert, update, delete, and upsert policies on rows based



on expressions. You can use the alter row transformation to perform upserts on a sink table by matching on a key column and setting the appropriate row policy

216. - (Exam Topic 3)

You plan to create an Azure Data Lake Storage Gen2 account

You need to recommend a storage solution that meets the following requirements:

- Provides the highest degree of data resiliency
- Ensures that content remains available for writes if a primary data center fails

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

Answer Area

Replication mechanism:

- | |
|---|
| Change feed |
| Zone-redundant storage (ZRS) |
| Read-access geo-redundant storage (RA-GRS) |
| Read-access geo-zone-redundant storage (RA-GRS) |

Failover process:

- | |
|---|
| Failover initiated by Microsoft |
| Failover manually initiated by the customer |
| Failover automatically initiated by an Azure Automation job |

Answer:

Answer Area

Replication mechanism:

- | |
|---|
| Change feed |
| Zone-redundant storage (ZRS) |
| Read-access geo-redundant storage (RA-GRS) |
| Read-access geo-zone-redundant storage (RA-GRS) |

Failover process:

- | |
|---|
| Failover initiated by Microsoft |
| Failover manually initiated by the customer |
| Failover automatically initiated by an Azure Automation job |

Explanation:

Graphical user interface, text, application Description automatically generated

Availability : "Microsoft recommends RA-GZRS for maximum availability and durability for your applications."

Failover: "The customer initiates the account failover to the secondary endpoint. "

<https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/>

<https://docs.microsoft.com/en-us/answers/questions/32583/azure-data-lake-gen2-disaster-recoverystorage-acco.h>

217. - (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network.

You are designing a SQL pool in Azure Synapse that will use adls2 as a source.

What should you use to authenticate to adls2?

- A. a shared access signature (SAS)
- B. a managed identity
- C. a shared key
- D. an Azure Active Directory (Azure AD) user

Answer: B

Explanation:

Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD. You can use the Managed Identity capability to authenticate to any service that supports Azure AD authentication.

Managed Identity authentication is required when your storage account is attached to a VNet. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-exa>

218. - (Exam Topic 3)

You have an Azure Data Lake Storage account that has a virtual network service endpoint configured.

You plan to use Azure Data Factory to extract data from the Data Lake Storage account. The data will then be loaded to a data warehouse in Azure Synapse Analytics by using PolyBase.

Which authentication method should you use to access Data Lake Storage?



- A. shared access key authentication
- B. managed identity authentication
- C. account key authentication
- D. service principal authentication

Answer: B

Explanation:

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-d>

219. - (Exam Topic 3)

You have an Azure data factory.

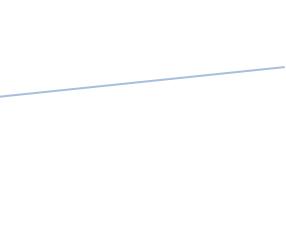
You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.

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

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

Actions	Answer Area
Select the PipelineRuns category.	
Create a Log Analytics workspace that has Data Retention set to 120 days.	
Stream to an Azure event hub.	
Create an Azure Storage account that has a lifecycle policy.	
From the Azure portal, add a diagnostic setting.	
Send the data to a Log Analytics workspace.	
Select the TriggerRuns category.	

Answer:





Actions	Answer Area
Select the PipelineRuns category.	Create an Azure Storage account that has a lifecycle policy.
Create a Log Analytics workspace that has Data Retention set to 120 days.	Create a Log Analytics workspace that has Data Retention set to 120 days.
Stream to an Azure event hub.	From the Azure portal, add a diagnostic setting.
Create an Azure Storage account that has a lifecycle policy.	Send the data to a Log Analytics workspace.
From the Azure portal, add a diagnostic setting.	
Send the data to a Log Analytics workspace.	
Select the TriggerRuns category.	

Explanation:

Step 1: Create an Azure Storage account that has a lifecycle policy

To automate common data management tasks, Microsoft created a solution based on Azure Data Factory.

The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.

Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.

Step 3: From Azure Portal, add a diagnostic setting. Step 4: Send the data to a log Analytics workspace, Event Hub: A pipeline that transfers events from services to Azure Data Explorer. Keeping Azure Data Factory metrics and pipeline-run data.

Configure diagnostic settings and workspace.

Create or add diagnostic settings for your data factory.

- In the portal, go to Monitor. Select Settings > Diagnostic settings.
- Select the data factory for which you want to set a diagnostic setting.



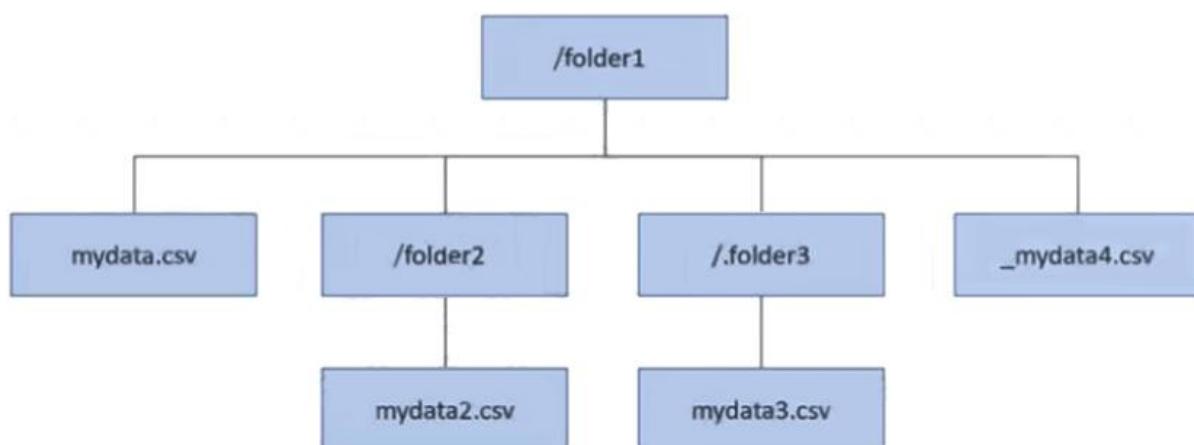


- If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.
- Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.
- Select Save. Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

220. - (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account that contains a container named container1. You have an Azure Synapse Analytics serverless SQL pool that contains a native external table named dbo.Table1. The source data for dbo.Table1 is stored in container1. The folder structure of container1 is shown in the following exhibit.



The external data source is defined by using the following statement.

```

CREATE EXTERNAL DATA SOURCE DataLake
WITH
(
    LOCATION      = 'https://mydatalake.dfs.core.windows.net/container1/folder1/**'
    , CREDENTIAL = DataLakeCred
);
  
```

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

NOTE: Each correct selection is worth one point.



Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="radio"/>	<input type="radio"/>

Answer:

Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Explanation:

Box 1: Yes

In the serverless SQL pool you can also use recursive wildcards `/logs/**` to reference Parquet or CSV files in any sub-folder beneath the referenced folder.

Box 2: Yes

Box 3: No

Reference: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

221. - (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 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 schedule an Azure Databricks job that executes an R notebook, and then inserts the data into



the data warehouse.

Does this meet the goal?

A. Yes

B. No

Answer: B

Explanation:

Must use an Azure Data Factory, not an Azure Databricks job. Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

222. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named SQLPool1.

SQLPool1 is currently paused.

You need to restore the current state of SQLPool1 to a new SQL pool. What should you do first?

A. Create a workspace.

B. Create a user-defined restore point.

C. Resume SQLPool1.

D. Create a new SQL pool.

Answer: B

Explanation:

Reference:

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

223. - (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: A

Explanation:

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>

224. - (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 container.

Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.

You need to design a data archiving solution that meets the following requirements:

- New data is accessed frequently and must be available as quickly as possible.
- Data that is older than five years is accessed infrequently but must be available within one second when requested.
- Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible.
- Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area. NOTE:

Each correct selection is worth one point

Five-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Seven-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Answer:

Five-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Seven-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Explanation:

Box 1: Move to cool storage Box 2: Move to archive storage

Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.

The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.



	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum storage duration	N/A	N/A	30 days ¹	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours ²

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers>

225. - (Exam Topic 3)

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

Answer: A B C

226. - (Exam Topic 3)

You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes.

You need to ensure that pipeline1 will execute only if the previous execution completes successfully. How should you configure the self-dependency for Trigger1?

- A. offset: "-00:01:00" size: "00:01:00"
- B. offset: "01:00:00" size: "-01:00:00"
- C. offset: "01:00:00" size: "01:00:00"





D. offset: "-01:00:00" size: "01:00:00"

Answer: D

Explanation:

Tumbling window self-dependency properties

In scenarios where the trigger shouldn't proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that's dependent on the success of earlier runs of itself within the preceding hour will have the properties indicated in the following code.

Example code:

```
"name": "DemoSelfDependency",
"properties": { "runtimeState": "Started", "pipeline": { "pipelineReference": { "referenceName": "Demo", "type": "PipelineReference"
},
"type": "TumblingWindowTrigger", "typeProperties": {
"frequency": "Hour", "interval": 1,
"startTime": "2018-10-04T00:00:00Z", "delay": "00:01:00",
"maxConcurrency": 50, "retryPolicy": { "intervalInSeconds": 30
},
"dependsOn": [
{
"type": "SelfDependencyTumblingWindowTriggerReference", "size": "01:00:00",
"offset": "-01:00:00"
}
]
}
}
```

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency>





227. - (Exam Topic 3)

You need to implement an Azure Databricks cluster that automatically connects to Azure Data lake Storage Gen2 by using Azure Active Directory (Azure AD) integration. How should you configure the new clutter?

To answer, select the appropriate options in the answers area. NOTE: Each correct selection is worth one point.

Answer Area

Tier: Premium Standard

Advanced option to enable: Azure Data Lake Storage Credential Passthrough

Answer:

Answer Area

Tier: Premium

Advanced option to enable: Azure Data Lake Storage Credential Passthrough

Explanation:

<https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html>

228. - (Exam Topic 3)

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.



Values	Answer Area
	<pre> select* FROM [] (BULK 'https://contoso.blob.core.windows.net/contosodw', FORMAT= 'CSV', fieldterminator = '0x0b', fieldquote = '0x0b', rowterminator = '0x0b') with (id varchar(50), contextdateeventTime varchar(50) '\$.context.data.eventTime', contextdatasamplingRate varchar(50) '\$.context.data.samplingRate', contextdataisSynthetic varchar(50) '\$.context.data.isSynthetic', contextsessionisFirst varchar(50) '\$.context.session.isFirst', contextsession varchar(50) '\$.context.session.id', contextcustomdimensions varchar(max) '\$.context.custom.dimensions') as q cross apply [] (contextcustomdimensions) with (ProfileType varchar(50) '\$.customerInfo.ProfileType', RoomName varchar(50) '\$.customerInfo.RoomName', CustomerName varchar(50) '\$.customerInfo.CustomerName', UserName varchar(50) '\$.customerInfo.UserName') </pre>

Answer:

Values	Answer Area
	<pre> select* FROM [] (BULK 'https://contoso.blob.core.windows.net/contosodw', FORMAT= 'CSV', fieldterminator = '0x0b', fieldquote = '0x0b', rowterminator = '0x0b') with (id varchar(50), contextdateeventTime varchar(50) '\$.context.data.eventTime', contextdatasamplingRate varchar(50) '\$.context.data.samplingRate', contextdataisSynthetic varchar(50) '\$.context.data.isSynthetic', contextsessionisFirst varchar(50) '\$.context.session.isFirst', contextsession varchar(50) '\$.context.session.id', contextcustomdimensions varchar(max) '\$.context.custom.dimensions') as q cross apply [] openjson (contextcustomdimensions) with (ProfileType varchar(50) '\$.customerInfo.ProfileType', RoomName varchar(50) '\$.customerInfo.RoomName', CustomerName varchar(50) '\$.customerInfo.CustomerName', UserName varchar(50) '\$.customerInfo.UserName') </pre>

Explanation:

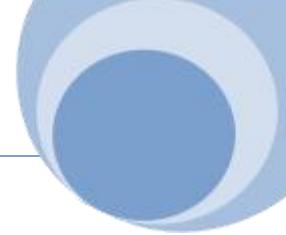
Graphical user interface, text, application, email Description automatically generated

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>

229. - (Exam Topic 3)

You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud.

Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers.

You need to recommend a solution that meets the following requirements:

- Users must be able to identify potentially fraudulent transactions.
- Users must be able to use credit cards as a potential feature in models.
- Users must NOT be able to access the actual credit card numbers.

What should you include in the recommendation?

- A. Transparent Data Encryption (TDE)
- B. row-level security (RLS)
- C. column-level encryption
- D. Azure Active Directory (Azure AD) pass-through authentication

Answer: C

Explanation:



Use Always Encrypted to secure the required columns. You can configure Always Encrypted for individual database columns containing your sensitive data. Always Encrypted is a feature designed to protect sensitive data, such as credit card numbers or national identification numbers (for example, U.S. social security numbers), stored in Azure SQL Database or SQL Server databases.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine>

230. - (Exam Topic 3)

You are building a database in an Azure Synapse Analytics serverless SQL pool. You have data stored in Parquet files in an Azure Data Lake Storage Gen2 container. Records are structured as shown in the following sample.

```
{  
  "id": 123,  
  "address_housenumber": "19c", "address_line": "Memory Lane", "applicant1_name": "Jane",  
  "applicant2_name": "Dev"  
}
```

The records contain two applicants at most.

You need to build a table that includes only the address fields.

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

NOTE: Each correct selection is worth one point.

**Answer Area**

```

▼ applications
CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW
WITH (
    LOCATION = 'applications/',
    DATA_SOURCE = applications_ds,
    FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
    ▼ (BULK 'https://contosol.dfs.core.windows.net/applications/year=/*/*.parquet',
CROSS APPLY
OPENJSON
OPENROWSET
FORMAT='PARQUET') AS [r]
GO

```

Answer:**Answer Area**

```

▼ applications
CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW
WITH (
    LOCATION = 'applications/',
    DATA_SOURCE = applications_ds,
    FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
    ▼ (BULK 'https://contosol.dfs.core.windows.net/applications/year=/*/*.parquet',
CROSS APPLY
OPENJSON
OPENROWSET
FORMAT='PARQUET') AS [r]
GO

```

Explanation:**Box 1: CREATE EXTERNAL TABLE**

An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage.

External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

Syntax:

```

CREATE EXTERNAL TABLE { database_name.schema_name.table_name | schema_name.table_name |
table_name }

( <column_definition> [ ,...n ] ) WITH (

```



LOCATION = 'folder_or_filepath', DATA_SOURCE = external_data_source_name, FILE_FORMAT = external_file_format_name

Box 2. OPENROWSET

When using serverless SQL pool, CETAS is used to create an external table and export query results to Azure Storage Blob or Azure Data Lake Storage Gen2.

Example: AS

```
SELECT decennialTime, stateName, SUM(population) AS population FROM  
OPENROWSET(BULK  
'https://azureopendatastorage.blob.core.windows.net/censusdatacontainer/release/us_population_county/  
year=/*'  
FORMAT='PARQUET') AS [r]  
GROUP BY decennialTime, stateName GO
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

231. - (Exam Topic 3)

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.



Actions	Answer Area
Mount the Data Lake Storage onto DBFS.	
Write the results to a table in Azure Synapse.	
Perform transformations on the file.	
Specify a temporary folder to stage the data.	
Write the results to Data Lake Storage.	
Read the file into a data frame.	
Drop the data frame.	
Perform transformations on the data frame.	

Answer:

Actions	Answer Area
Mount the Data Lake Storage onto DBFS.	Mount the Data Lake Storage onto DBFS.
Write the results to a table in Azure Synapse.	
Perform transformations on the file.	Read the file into a data frame.
Specify a temporary folder to stage the data.	
Write the results to Data Lake Storage.	Perform transformations on the data frame.
Read the file into a data frame.	Specify a temporary folder to stage the data.
Drop the data frame.	
Perform transformations on the data frame.	Write the results to a table in Azure Synapse.

Explanation:

- 1) mount onto DBFS
- 2) read into data frame
- 3) transform data frame
- 4) specify temporary folder
- 5) write the results to table in in Azure Synapse

<https://docs.databricks.com/data/data-sources/azure/azure-datalake-gen2.html>

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

232. - (Exam Topic 3)

You have an Azure Synapse Analytics job that uses Scala. You need to view the status of the job.





What should you do?

- A. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
- B. From Azure Monitor, run a Kusto query against the SparkLogyng1 Event.CL table.
- C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.
- D. From Synapse Studio, select the workspace. From Monitor, select SQL requests.

Answer: C

Explanation:

Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. If the Apache Spark application is still running, you can monitor the progress.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications>

233. - (Exam Topic 3)

You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.

You need to ensure that the table meets the following requirements:

- Minimizes the processing time to delete data that is older than 10 years
- Minimizes the I/O for queries that use year-to-date values

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

NOTE: Each correct selection is worth one point.



**Answer Area**

```

CREATE TABLE [dbo].[FactTransaction]
(
    [TransactionTypeID] int NOT NULL,
    [TransactionDateID] int NOT NULL,
    [CustomerID] int NOT NULL,
    [RecipientID] int NOT NULL,
    [Amount] money NOT NU::,
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
    DISTRIBUTION
    PARTITION
    TRUNCATE_TARGET
)
    (
        [TransactionDateID]
        [TransactionDateID], [TransactionTypeID]
        HASH([TransactionTypeID])
        ROUND_ROBIN
    )
    RANGE RIGHT FOR VALUES
(20200101,20200201,20200301,20200401,20200501,20200601)

```

Answer:**Answer Area**

```

CREATE TABLE [dbo].[FactTransaction]
(
    [TransactionTypeID] int NOT NULL,
    [TransactionDateID] int NOT NULL,
    [CustomerID] int NOT NULL,
    [RecipientID] int NOT NULL,
    [Amount] money NOT NU::,
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
    DISTRIBUTION
    PARTITION
    TRUNCATE_TARGET
)
    (
        [TransactionDateID]
        [TransactionDateID], [TransactionTypeID]
        HASH([TransactionTypeID])
        ROUND_ROBIN
    )
    RANGE RIGHT FOR VALUES
(20200101,20200201,20200301,20200401,20200501,20200601)

```

Explanation:

Table Description automatically generated





Box 1: PARTITION

RANGE RIGHT FOR VALUES is used with PARTITION.

Part 2: [TransactionDateID] Partition on the date column.

Example: Creating a RANGE RIGHT partition function on a datetime column

The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.

```
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
```

```
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401',
'20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');
```

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql>

234. - (Exam Topic 3)

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 executes an Azure Databricks notebook, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. No

Answer: B

Explanation:

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not an Azure Databricks notebook, with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.

Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>





235. - (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

Explanation:

Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache's Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.

Reference:

<https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/>

236. - (Exam Topic 3)

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

Answer: D





Explanation:

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/>

237. - (Exam Topic 3)

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.

Distribution:

- Hash
- Round robin
- Replicated

Index:

- Clustered columnstore
- Clustered
- Nonclustered

Answer:



Distribution:

Hash
Round robin
Replicated

Index:

Clustered columnstore
Clustered
Nonclustered

Explanation:

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.

Reference:

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

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

238. - (Exam Topic 3)

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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





- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the snared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.

Answer: A D F

Explanation:

The managed identity grants permissions to the dedicated SQL pools in the workspace.

Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity>

239. - (Exam Topic 3)

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID
	ArrivalDateTime
Weather	AirportID
	ReportDateTime

You need to recommend a solution that maximizes query performance. What should you include in the recommendation?

- A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
- B. In the tables use a hash distribution of ArrivalAirportID and AirportID.
- C. In each table, create an identity column.
- D. In each table, create a column as a composite of the other two columns in the table.

Answer: B

Explanation:

Hash-distribution improves query performance on large fact tables.





240. - (Exam Topic 3)

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.

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.	

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 .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.	

Explanation:

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.





Create a new project

The screenshot shows the Azure Stream Analytics project creation interface. At the top, there are dropdown menus for 'Language' (set to C#) and 'Platform' (set to Cloud). Below that is a 'Project type' dropdown. A sidebar on the left lists 'Recent project templates' with a note: 'A list of your recently accessed templates will be displayed here.' Below this are five template options:

- Empty Azure Stream Analytics Edge Application**: An empty project for an Azure Stream Analytics Edge application.
- Azure Stream Analytics Edge Application**: A project for creating Azure Stream Analytics Edge application.
- Empty Azure Stream Analytics Application**: An empty project for an Azure Stream Analytics application.
- Azure Stream Analytics Application**: A project for creating an Azure Stream Analytics application.
- Azure Stream Analytics Custom Deserializer Project (.NET)**: A .NET Standard project for Azure Stream Analytics Custom Deserializer.

The last template, 'Azure Stream Analytics Custom Deserializer Project (.NET)', is highlighted with a red border.

- * 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

- 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.
- 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>

241. - (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

a surrogate key
a business key
an audit column

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

Answer:

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

Type 0
Type 1
Type 2

a surrogate key
a business key
an audit column

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

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>

242. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a table named FactInternetSales that will be a large fact table in a dimensional model. FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity. Queries executed on FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time.

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

NOTE: Each correct selection is worth one point.

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
(
    [ProductKey] int NOT NULL
    , [OrderDateKey] int NOT NULL
    , [CustomerKey] int NOT NULL
    , [PromotionKey] int NOT NULL
    , [SalesOrderNumber] nvarchar(20) NOT NULL
    , [OrderQuantity] smallint NOT NULL
    , [UnitPrice] money NOT NULL
    , [SalesAmount] money NOT NULL
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
    ( CLUSTERED INDEX ([OrderDateKey])
    ( HEAP
    ( INDEX on [ProductKey]
    )
    , DISTRIBUTION =
    );
    Hash([OrderDateKey])
    Hash([ProductKey])
    REPLICATE
    ROUND_ROBIN
)
```

Answer:



Answer Area

```

CREATE TABLE [dbo].[FactInternetSales]
(
    [ProductKey] int NOT NULL
    , [OrderDateKey] int NOT NULL
    , [CustomerKey] int NOT NULL
    , [PromotionKey] int NOT NULL
    , [SalesOrderNumber] nvarchar(20) NOT NULL
    , [OrderQuantity] smallint NOT NULL
    , [UnitPrice] money NOT NULL
    , [SalesAmount] money NOT NULL
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
    CLUSTERED INDEX ([OrderDateKey])
    HEAP
    INDEX on [ProductKey]
    , DISTRIBUTION =
);
    Hash([OrderDateKey])
    Hash([ProductKey])
    REPLICATE
    ROUND_ROBIN

```

Explanation:

Box 1: (CLUSTERED COLUMNSTORE INDEX CLUSTERED COLUMNSTORE INDEX)

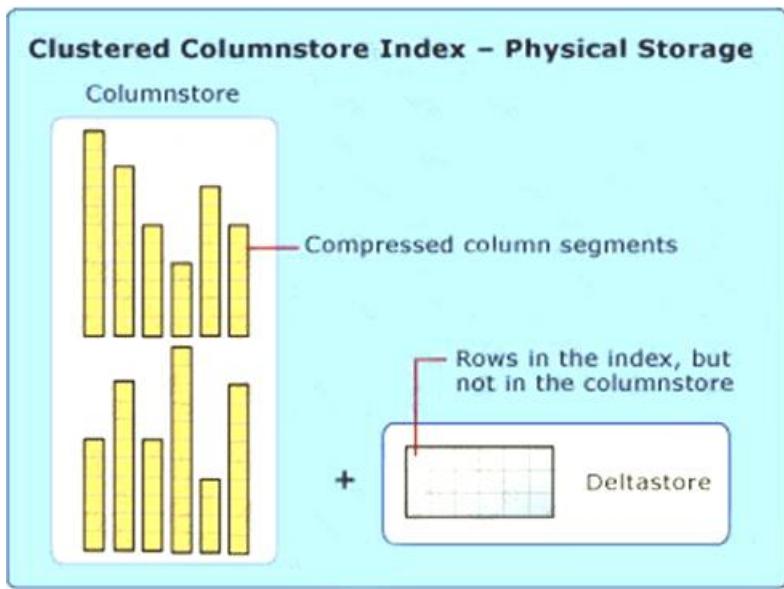
Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage. You can also achieve gains up to 10 times the data compression over the uncompressed data size. Beginning with SQL Server 2016 (13.x) SP1, columnstore

indexes enable operational analytics: the ability to run performant real-time analytics on a transactional workload.

Note: Clustered columnstore index

A clustered columnstore index is the physical storage for the entire table. Diagram Description automatically generated





To reduce fragmentation of the column segments and improve performance, the columnstore index might store some data temporarily into a clustered index called a deltastore and a B-tree list of IDs for deleted rows. The deltastore operations are handled behind the scenes. To return the correct query results, the clustered columnstore index combines query results from both the columnstore and the deltastore.

Box 2: HASH([ProductKey])

A hash distributed table distributes rows based on the value in the distribution column. A hash distributed table is designed to achieve high performance for queries on large tables.

Choose a distribution column with data that distributes evenly

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-overview>

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

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

243. - (Exam Topic 3)

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)

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)

Answer: B

Explanation:

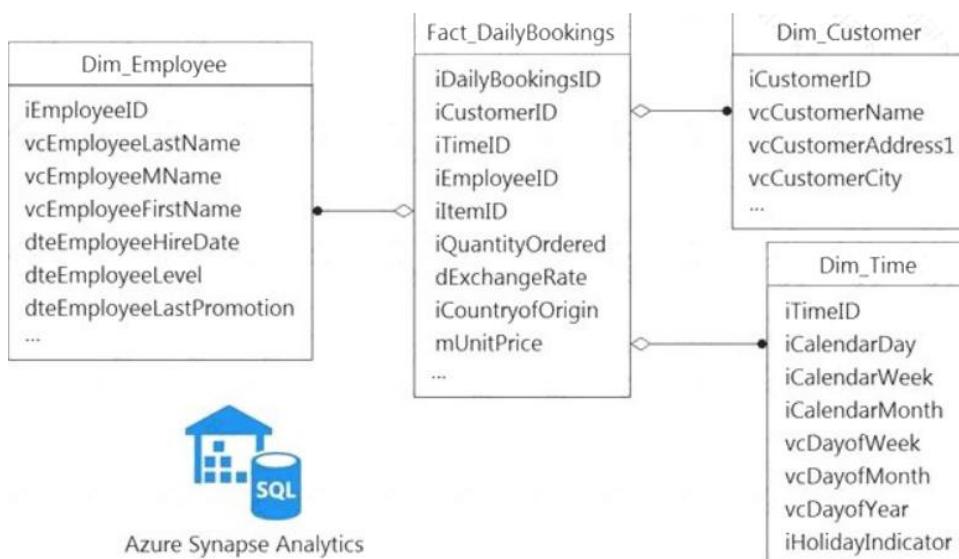
Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Example: Reference:

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

244. - (Exam Topic 3)

You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.



All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB.



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

NOTE: Each correct selection is worth one point.

Answer Area

Dim_Customer:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated
Dim_Employee:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated
Dim_Time:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated
Fact_DailyBookings:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated

Answer:**Answer Area**

Dim_Customer:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated
Dim_Employee:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated
Dim_Time:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated
Fact_DailyBookings:	<ul style="list-style-type: none">Hash distributedRound-robinReplicated





245. - (Exam Topic 3)

You have an Azure data factory.

You execute a pipeline that contains an activity named Activity1. Activity1 produces the following output.

```
{
  ...
  "dataRead": 1208,
  "dataWritten": 1208,
  "filesRead": 1,
  "filesWritten": 1,
  "sourcePeakConnections": 3,
  "sinkPeakConnections": 2,
  "copyDuration": 13,
  "throughput": 0.147,
  "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (West Central US)",
  "usedDataIntegrationUnits": 4,
  "reportLineageToPurview": {
    "status": "Succeeded",
    "durationInSecond": "4"
  }
}
}
```

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

Answer Area

Statements	Yes	No
Activity1 is a Copy activity.	<input type="radio"/>	<input type="radio"/>
Activity1 is executed by using a self-hosted integration runtime.	<input type="radio"/>	<input type="radio"/>
The data factory that executed the pipeline is connected to Microsoft Purview.	<input type="radio"/>	<input type="radio"/>

Answer:

Answer Area

Statements	Yes	No
Activity1 is a Copy activity.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Activity1 is executed by using a self-hosted integration runtime.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
The data factory that executed the pipeline is connected to Microsoft Purview.	<input type="checkbox"/>	<input checked="" type="checkbox"/>





246. - (Exam Topic 3)

You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace. CREATE

TABLE mytestdb.myParquetTable(

EmployeeID int, EmployeeName string, EmployeeStartDate date) USING Parquet

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace. SELECT

EmployeeID

FROM mytestdb.dbo.myParquetTable WHERE name = 'Alice';

What will be returned by the query?

A. 24

B. an error

C. a null value

Answer: B

Explanation:

Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet as the storage format. Table names will be converted to lower case and need to be queried using the lower case name. These tables will immediately become available for querying by any of the Azure Synapse workspace Spark pools. They can also be used from any of the Spark jobs subject to permissions.

Note: For external tables, since they are synchronized to serverless SQL pool asynchronously, there will be a delay until they appear.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/table>

247. - (Exam Topic 3)

You have an Azure subscription that contains the resources shown in the following table.





Name	Type	Description
ws1	Azure Synapse Analytics workspace	None
kv1	Azure Key Vault	None
UAMI1	User-assigned managed identity	Associated with ws1
sp1	Apache Spark pool in Azure Synapse Analytics	Associated with ws1

You need to ensure that you can Spark notebooks in ws1. The solution must ensure secrets from kv1 by using UAMI1. What should you do? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

In the Azure portal: Add a role-based access control (RBAC) role to kv1.

Add a role-based access control (RBAC) role to kv1.

Add a role-based access control (RBAC) role to ws1.

Create a linked service to kv1.

In Synapse Studio: Create a linked service to kv1.

Add a role-based access control (RBAC) role to kv1.

Add a role-based access control (RBAC) role to ws1.

Create a linked service to kv1.

Answer:

Answer Area

In the Azure portal: Add a role-based access control (RBAC) role to kv1.

Add a role-based access control (RBAC) role to kv1.

Add a role-based access control (RBAC) role to ws1.

Create a linked service to kv1.

In Synapse Studio: Create a linked service to kv1.

Add a role-based access control (RBAC) role to kv1.

Add a role-based access control (RBAC) role to ws1.

Create a linked service to kv1.

248. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that data in the pool is encrypted at rest. The solution must NOT require modifying applications that query the data.

What should you do?

- A. Enable encryption at rest for the Azure Data Lake Storage Gen2 account.
- B. Enable Transparent Data Encryption (TDE) for the pool.
- C. Use a customer-managed key to enable double encryption for the Azure Synapse workspace.
- D. Create an Azure key vault in the Azure subscription grant access to the pool.

Answer: B





Explanation:

Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest. When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.

Reference:

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

249. - (Exam Topic 3)

You are building an Azure Stream Analytics job that queries reference data from a product catalog file. The file is updated daily.

The reference data input details for the file are shown in the Input exhibit. (Click the Input tab.)

Input Details ×

products

Test Delete

Container

Create new Use existing

refdata

Path pattern

product.csv

Date format

YYYY/MM/DD

Time format

HH

Event serialization format *

CSV

Delimiter

comma (,)

Encoding

UTF-8

Save • If the chosen resource and the stream analytics job are located in different regions, you will be billed to move data between regions.

The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)




refdata
Container

 Search (Ctrl + /)

 Upload

 + Add Directory

 Refresh

 Rename

 Delete

 Overview

 Access Control (IAM)
Settings
 Access policy

 Properties

 Metadata

Authentication method: Access key ([Switch to Azure AD User Account](#))

Location: [refdata](#) / 2020-03-20

 Search blobs by prefix (case-sensitive)

Name

 [..]

 product.csv

You need to configure the Stream Analytics job to pick up the new reference data.

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

NOTE: Each correct selection is worth one point.

Path pattern:

{date}/product.csv
{date}/{time}/product.csv
product.csv
*/product.csv

Date format:

MM/DD/YYYY
YYYY/MM/DD
YYYY-DD-MM
YYYY-MM-DD

Answer:

Path pattern:

{date}/product.csv
{date}/{time}/product.csv
product.csv
*/product.csv

Date format:

MM/DD/YYYY
YYYY/MM/DD
YYYY-DD-MM
YYYY-MM-DD

Explanation:

Graphical user interface, application, table Description automatically generated

Box 1: {date}/product.csv

In the 2nd exhibit we see: Location: refdata / 2020-03-20

Note: Path Pattern: This is a required property that is used to locate your blobs within the specified container. Within the path, you may choose to specify one or more instances of the following 2 variables: {date}, {time}

Example 1: products/{date}/{time}/product-list.csv

Example 2: products/{date}/product-list.csv

Example 3: product-list.csv

Box 2: YYYY-MM-DD

Note: Date Format [optional]: If you have used {date} within the Path Pattern that you specified, then you can select the date format in which your blobs are organized from the drop-down of supported formats.

Example: YYYY/MM/DD, MM/DD/YYYY, etc. Reference:

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

250. - (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:

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

- Analysts 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 partition the table?

A. ProductCategoryTypeID

B. EventDate

C. WarehouseID

D. EventTypeID

Answer: C

Explanation:



The number of records for each warehouse is big enough for a good partitioning.

Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

251. - (Exam Topic 3)

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.

```
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
```



**Answer:**

```

WITH LastInWindow AS
(
    SELECT
        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

```

Explanation:

Graphical user interface, text, application Description automatically generated

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>

252. - (Exam Topic 3)

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Number of partitions:

1
8
16
32

Partition key:

Fraud indicator
Fraud score
Individual line items
Payment details
Transaction ID

Answer:

Number of partitions:

1
8
16
32

Partition key:

Fraud indicator
Fraud score
Individual line items
Payment details
Transaction ID

Explanation:

Box 1: 16

For Event Hubs you need to set the partition key explicitly.

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output. Box 2: Transaction ID

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions>

253. - (Exam Topic 3)

You are monitoring an Azure Stream Analytics job.

The Backlogged Input Events count has been 20 for the last hour. You need to reduce the Backlogged



Input Events count.

What should you do?

- A. Drop late arriving events from the job.
- B. Add an Azure Storage account to the job.
- C. Increase the streaming units for the job.
- D. Stop the job.

Answer: C

Explanation:

General symptoms of the job hitting system resource limits include:

- If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs>

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

254. - (Exam Topic 3)

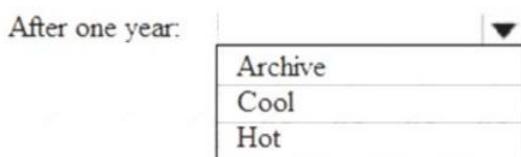
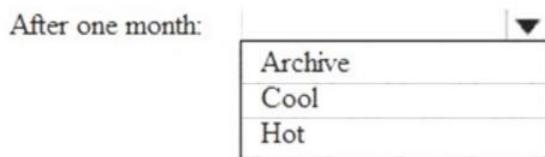
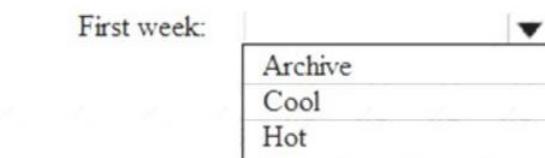
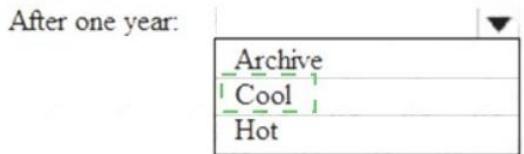
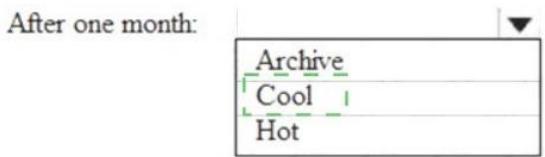
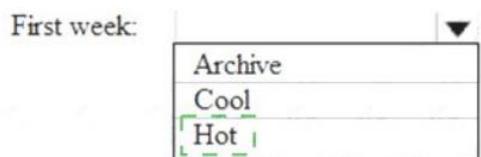
You are designing an application that will store petabytes of medical imaging data

When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.

You need to select a storage strategy for the data. The solution must minimize costs.

Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Answer:****Explanation:**

First week: Hot

Hot - Optimized for storing data that is accessed frequently. After one month: Cool

Cool - Optimized for storing data that is infrequently accessed and stored for at least 30 days.

After one year: Cool

255. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Blob Storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool named Pool1.

You need to store data in storage1. The data will be read by Pool1. The solution must meet the following requirements:

- Enable Pool1 to skip columns and rows that are unnecessary in a query.





- Automatically create column statistics.
- Minimize the size of files. Which type of file should you use?

A. JSON

B. Parquet

C. Avro

D. CSV

Answer: B

Explanation:

Automatic creation of statistics is turned on for Parquet files. For CSV files, you need to create statistics manually until automatic creation of CSV files statistics is supported.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-statistics>

256. - (Exam Topic 3)

You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns.

Each day, 1,500 new files are added to the folder.

You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace.

You need to minimize how long it takes to perform the incremental loads.

What should you use to store the files and format?

Storage: These are the se

- Multiple blob storage accounts
- Multiple containers in the blob storage account
- Timeslice partitioning in the folders

Format:

- Apache Parquet
- CSV
- JSON

Answer:



Storage: These are the se

- Multiple blob storage accounts
- Multiple containers in the blob storage account
- Timeslice partitioning in the folders

Format:

- Apache Parquet
- CSV
- JSON

Explanation:

Box 1 = timeslice partitioning in the folders This means that you should organize your files into folders based on a time attribute, such as year, month, day, or hour. For example, you can have a folder structure like /yyyy/mm/dd/file.csv. This way, you can easily identify and load only the new files that are added each day by using a time filter in your Azure Synapse pipeline¹². Timeslice partitioning can also improve the performance of data loading and querying by reducing the number of files that need to be scanned

Box 2 = Apache Parquet This is because Parquet is a columnar file format that can efficiently store and compress data with many columns. Parquet files can also be partitioned by a time attribute, which can improve the performance of incremental loading and querying by reducing the number of files that need to be scanned^{1 23}. Parquet files are supported by both dedicated SQL pool and serverless SQL pool in Azure Synapse Analytics².

257. - (Exam Topic 3)

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.
- C. Include a watermark column.
- D. Use a JSON format for physical data storage.

Answer: A

Explanation:



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>

258. - (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 <code>DELETE</code> statement where the value in the Date column is more than 36 months ago.	
Copy the data to a new table by using <code>CREATE TABLE AS SELECT (CTAS)</code> .	

Answer:

**Actions**

- 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).

Answer Area

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

Switch the partition containing the stale data from SalesFact to SalesFact_Work.

Drop the SalesFact_Work table.

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>

259. - (Exam Topic 3)

You need to build a solution to ensure that users can query specific files in an Azure Data Lake Storage Gen2 account from an Azure Synapse Analytics serverless SQL pool.

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

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



**Actions****Answer Area**

Create an external file format object

Create an external data source

Create a query that uses Create Table as Select

Create a table

Create an external table

**Answer:****Actions****Answer Area**

Create an external file format object

Create an external data source

Create an external data source

Create an external file format object

Create a query that uses Create Table as Select

Create an external table

Create a table

Create an external table

Explanation:

Graphical user interface, text, application, email Description automatically generated

Step 1: Create an external data source

You can create external tables in Synapse SQL pools via the following steps:

- CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.
- CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.
- CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

Step 2: Create an external file format object

Creating an external file format is a prerequisite for creating an external table. Step 3: Create an external table

Reference:

260. - (Exam Topic 3)

You have an Azure subscription that contains the following resources:

- An Azure Active Directory (Azure AD) tenant that contains a security group named Group1
- An Azure Synapse Analytics SQL pool named Pool1

You need to control the access of Group1 to specific columns and rows in a table in Pool1.

Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area.

To control access to the columns:

CREATE CRYPTOGRAPHIC PROVIDER
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

To control access to the rows:

CREATE CRYPTOGRAPHIC PROVIDER
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

Answer:

To control access to the columns:

CREATE CRYPTOGRAPHIC PROVIDER
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

To control access to the rows:

CREATE CRYPTOGRAPHIC PROVIDER
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

Explanation:

Text Description automatically generated

Box 1: GRANT

You can implement column-level security with the GRANT T-SQL statement. Box 2: CREATE SECURITY POLICY

Implement Row Level Security by using the CREATE SECURITY POLICY Transact-SQL statement



Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security>

261. - (Exam Topic 3)

You have an Azure Data Factory pipeline that is triggered hourly. The pipeline has had 100% success for the past seven days.

The pipeline execution fails, and two retries that occur 15 minutes apart also fail. The third failure returns the following error.

```
ErrorCode=UserErrorFileNotFound, 'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=ADLS Gen2 operation failed for:  
Operation returned an invalid status code 'NotFound'. Account: 'contosoproduksouth'. FileSystem: wwi. Path:  
'BIKES/CARBON/year=2021/month=01/day=10/hour=06'. ErrorCode: 'PathNotFound'. Message: 'The specified path does not exist.'. RequestId: '6d269b78-  
901f-001b-4924-e7a7bc000000'.TimeStamp: 'Sun, 10 Jan 2021 07:45:05'
```

What is a possible cause of the error?

- A. The parameter used to generate year=2021/month=01/day=10/hour=06 was incorrect.
- B. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON.
- C. From 06:00 to 07:00 on January 10, 2021, the file format of data in wwi/BIKES/CARBON was incorrect.
- D. The pipeline was triggered too early.

Answer: C

262. - (Exam Topic 3)

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- Can return an employee record from a given point in time.
- Maintains the latest employee information.
- Minimizes query complexity.

How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table
- C. as a degenerate dimension table
- D. as a Type 2 slowly changing dimension (SCD) table

Answer: D

Explanation:



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>

263. - (Exam Topic 3)

You have the following Azure Stream Analytics query.

```
WITH
    step1 AS (SELECT *
               FROM input1
              PARTITION BY StateID
              INTO 10),
    step1 AS (SELECT *
               FROM input2
              PARTITION BY StateID
              INTO 10)

    SELECT *
    INTO output
    FROM step1
    PARTITION BY StateID
    UNION step2
    BY StateID
```

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

NOTE: Each correct selection is worth one point.

Statements	Yes	No
The query joins two streams of partitioned data.	<input type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input type="radio"/>	<input type="radio"/>

Answer:

Statements	Yes	No
The query joins two streams of partitioned data.	<input checked="" type="checkbox"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input checked="" type="checkbox"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input checked="" type="checkbox"/>	<input type="radio"/>

Explanation:

Box 1: Yes

You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.

The outcome is a stream that has the same partition scheme. Please see below for an example:

```
WITH
step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10),
step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)
```

```
SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID
```

Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.

Box 2: Yes

When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.

Box 3: Yes

10 partitions x six SUs = 60 SUs is fine.

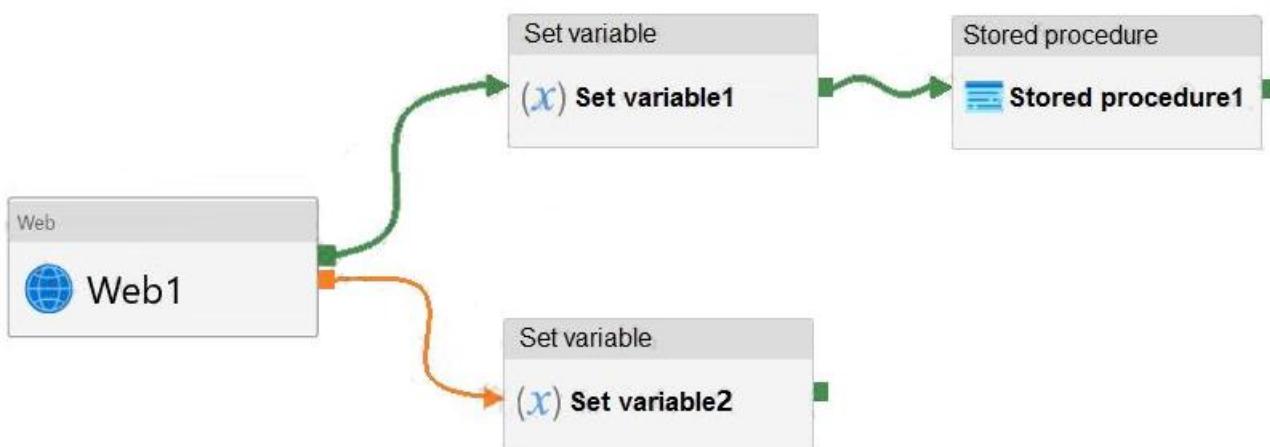
Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.

Reference:

<https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/>

264. - (Exam Topic 3)

You have an Azure Data Factory pipeline that has the activities 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.

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

complete
fail
succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

Canceled
Failed
Succeeded

Answer:

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

complete
fail
succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

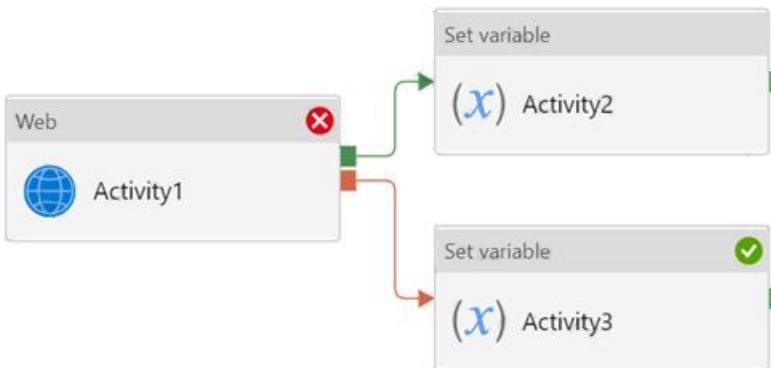
Canceled
Failed
Succeeded

Explanation:

Box 1: succeed

Box 2: failed Example:

Now let's say we have a pipeline with 3 activities, where Activity1 has a success path to Activity2 and a failure path to Activity3. If Activity1 fails and Activity3 succeeds, the pipeline will fail. The presence of the success path alongside the failure path changes the outcome reported by the pipeline, even though the activity executions from the pipeline are the same as the previous scenario.



Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure. Reference:

<https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/>

265. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos D6 database account named Cosmos1. Cosmos1 contains a container named container 1 and ws1 contains a serverless1 SQL pool.

you need to ensure that you can Query the data in container by using the serverless1 SQL pool. Which three actions should you perform? Each correct answer presents part of the solution NOTE: Each correct



selection is worth one point.

- A. Enable Azure Synapse Link for Cosmos1
- B. Disable the analytical store for container1.
- C. In ws1. create a linked service that references Cosmos1
- D. Enable the analytical store for container1
- E. Disable indexing for container1

Answer: A C D

266. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

You need to monitor the database for long-running queries and identify which queries are waiting on resources Which dynamic management view should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE; Each correct answer is worth one point.

Answer Area

Monitor the database for long-running queries:

sys.dm_pdw_exec_requests	▼
sys.dm_pdw_sql_requests	
sys.dm_pdw_exec_sessions	

Identify which queries are waiting on resources:

sys.dm_pdw_waits	▼
sys.dm_pdw_lock_waits	
sys.resource_governor_workload_groups	

Answer:

Answer Area

Monitor the database for long-running queries:

sys.dm_pdw_exec_requests	▼
sys.dm_pdw_sql_requests	
sys.dm_pdw_exec_sessions	

Identify which queries are waiting on resources:

sys.dm_pdw_waits	▼
sys.dm_pdw_lock_waits	
sys.resource_governor_workload_groups	

267. - (Exam Topic 3)





You have a Microsoft SQL Server database that uses a third normal form schema.

You plan to migrate the data in the database to a star schema in an Azure Synapse Analytics dedicated SQL pool.

You need to design the dimension tables. The solution must optimize read operations.

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.

Transform data for the dimension tables by:

Maintaining to a third normal form
Normalizing to a fourth normal form
Denormalizing to a second normal form

For the primary key columns in the dimension tables, use:

New IDENTITY columns
A new computed column
The business key column from the source sys

Answer:

Transform data for the dimension tables by:

Maintaining to a third normal form
Normalizing to a fourth normal form
Denormalizing to a second normal form

For the primary key columns in the dimension tables, use:

New IDENTITY columns
A new computed column
The business key column from the source sys

Explanation:

Text, table Description automatically generated

Box 1: Denormalize to a second normal form

Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation. Denormalization increases the performance in data retrieval at cost of bringing update anomalies to a database.

Box 2: New identity columns

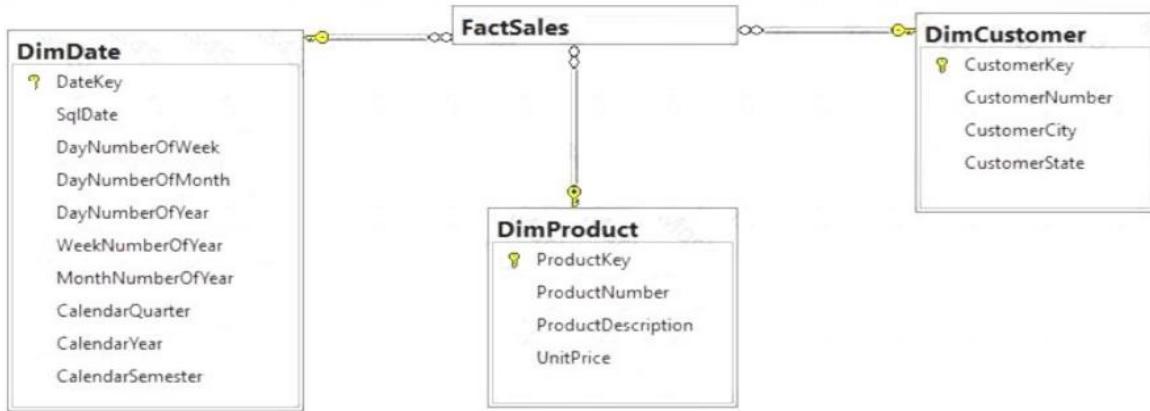
The collapsing relations strategy can be used in this step to collapse classification entities into component entities to obtain at dimension tables with single-part keys that connect directly to the fact table. The



single-part key is a surrogate key generated to ensure it remains unique over time.

Example:

Diagram Description automatically generated



Note: 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://www.mssqltips.com/sqlservertip/5614/explore-the-role-of-normal-forms-in-dimensional-modeling/>

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

268. - (Exam Topic 3)

You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.

You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com instead.

What should you do?

- A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.
- B. From the Azure portal, set a mask on the Email column.
- C. From Microsoft SQL Server Management studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.
- D. From the Azure portal, set a sensitivity classification of Confidential for the Email column.





Answer: D

Explanation:

From Microsoft SQL Server Management Studio, set an email mask on the Email column. This is because "This feature cannot be set using portal for Azure Synapse (use PowerShell or REST API) or SQL Managed Instance." So use Create table statement with Masking e.g. CREATE TABLE Membership
(MemberID int IDENTITY PRIMARY KEY, FirstName varchar(100) MASKED WITH (FUNCTION =
'partial(1,"XXXXXXX",0)'), LastName varchar(100), Email varchar(100) MASKED WITH (FUNCTION =
'partial(1,"XXXXXXX",0)'), . . .

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

upvoted 24 times

269. - (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: A

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>

270. - (Exam Topic 3)

You have a SQL pool in Azure Synapse.



A user reports that queries against the pool take longer than expected to complete. You need to add monitoring to the underlying storage to help diagnose the issue.

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

- A. Cache used percentage
- B. DWU Limit
- C. Snapshot Storage Size
- D. Active queries
- E. Cache hit percentage

Answer: A E

Explanation:

A: Cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes.

E: Cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes

Reference:

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

271. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.

In workspace1, you complete testing of pipeline1. You need to schedule pipeline1 to run daily at 6 AM.

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

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



Actions	Answer Area
Create a new branch in Repo1.	
Merge the changes from branch1 into main.	
Associate the schedule trigger with pipeline1.	▶
Switch to Synapse live mode.	◀
Create a schedule trigger.	
Publish the contents of main.	

**Answer:**

Actions	Answer Area
Create a new branch in Repo1.	Create a schedule trigger.
Merge the changes from branch1 into main.	Associate the schedule trigger with pipeline1.
Associate the schedule trigger with pipeline1.	Merge the changes from branch1 into main.
Switch to Synapse live mode.	Publish the contents of main.
Create a schedule trigger.	
Publish the contents of main.	

Explanation:**Timeline Description automatically generated****272. - (Exam Topic 3)**

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table 1. Table1 contains a timestamp column named LastModifiedOn.

LastModifiedOn contains the timestamp of the most recent update for each individual row.

DB2 contains a table named Watermark. Watermark contains a single timestamp column named WatermarkValue.

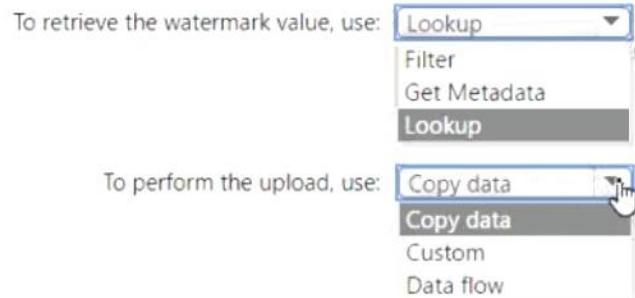
You plan to create an Azure Data Factory pipeline that will incrementally upload into Azure Blob Storage all

the rows in Table1 for which the LastModifiedOn column contains a timestamp newer than the most recent value of the WatermarkValue column in Watermark.

You need to identify which activities to include in the pipeline. The solution must meet the following requirements:

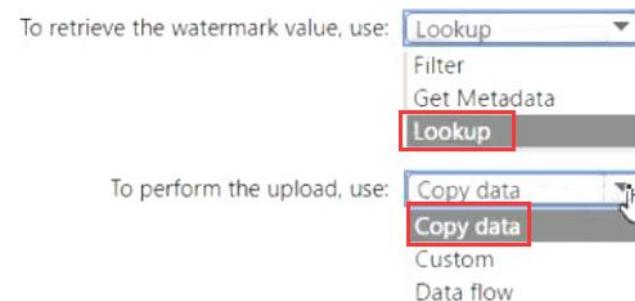
- Minimize the effort to author the pipeline.
- Ensure that the number of data integration units allocated to the upload operation can be controlled. What should you identify? To answer, select the appropriate options in the answer area.

Answer Area



Answer:

Answer Area



273. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool. You plan to deploy a solution that will analyze sales data and include the following:

- A table named Country that will contain 195 rows
- A table named Sales that will contain 100 million rows
- A query to identify total sales by country and customer from the past 30 days

You need to create the tables. The solution must maximize query performance.

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

**Answer Area**

```

CREATE TABLE [dbo].[Sales]
(
    [OrderDate]         date        NOT NULL
,   [CustomerId]      int NOT NULL
,   [CountryId]       int NOT NULL
,   [Total]            money NOT NULL
)

WITH
(
    DISTRIBUTION = HASH([CustomerId])
,   CLUSTERED COLUMNSTORE INDEX HASH([CustomerId])
,   HASH([OrderDate])
)
REPLICATE
ROUND_ROBIN
CREATE TABLE [dbo].[Country]
(
    [CountryName]      nvarchar(50) NOT NULL
,   [CountryCode]     char(2) NOT NULL
,   [Region]          nvarchar(50) NOT NULL
,   [Population]      int NOT NULL
,   [Area]            float NOT NULL
,   [GDP]             money NOT NULL
)
REPLICATE
ROUND_ROBIN
    
```

Answer:**Answer Area**

```

CREATE TABLE [dbo].[Sales]
(
    [OrderDate]         date        NOT NULL
,   [CustomerId]      int NOT NULL
,   [CountryId]       int NOT NULL
,   [Total]            money NOT NULL
)

WITH
(
    DISTRIBUTION = HASH([CustomerId])
,   CLUSTERED COLUMNSTORE INDEX HASH([CustomerId])
,   HASH([OrderDate])
)
REPLICATE
ROUND_ROBIN
CREATE TABLE [dbo].[Country]
(
    [CountryName]      nvarchar(50) NOT NULL
,   [CountryCode]     char(2) NOT NULL
,   [Region]          nvarchar(50) NOT NULL
,   [Population]      int NOT NULL
,   [Area]            float NOT NULL
,   [GDP]             money NOT NULL
)
REPLICATE
ROUND_ROBIN
    
```

274. - (Exam Topic 3)

You create an Azure Databricks cluster and specify an additional library to install. When you attempt to load the library to a notebook, the library is not found.

You need to identify the cause of the issue. What should you review?

- A. notebook logs
- B. cluster event logs
- C. global init scripts logs



D. workspace logs

Answer: C

Explanation:

Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.

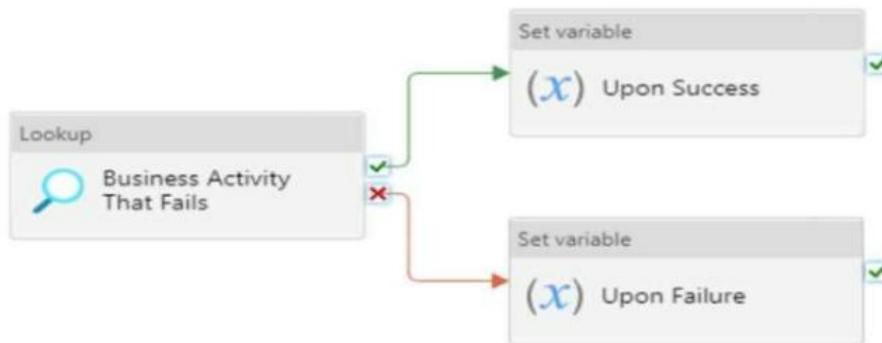
Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference:

<https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>

275. - (Exam Topic 3)

You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful.

What should you configure for the set variable activity?

- A. a success dependency on the Business Activity That Fails activity
- B. a failure dependency on the Upon Failure activity
- C. a skipped dependency on the Upon Success activity
- D. a skipped dependency on the Upon Failure activity

Answer: A

Explanation:

A failure dependency means that the activity will run only if the previous activity fails. In this case, setting a

failure dependency on the Upon Failure activity will ensure that the set variable activity will run after the pipeline fails and set the status of the pipeline to successful.

276. - (Exam Topic 3)

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.

Service:	<ul style="list-style-type: none">An Azure Synapse Analytics Apache Spark poolAn Azure Synapse Analytics serverless SQL poolAzure Data FactoryAzure Stream Analytics
Window:	<ul style="list-style-type: none">HoppingNo windowSessionTumbling
Analysis type:	<ul style="list-style-type: none">Event pattern matchingLagged record comparisonPoint within polygonPolygon overlap

Answer:



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

Explanation:

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>

277. - (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

You run PDW_SHOWSPACEUSED(dbo,FactInternetSales'); and get the results shown in the following table.





BINS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PW_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
168	2632	528	32	2072	1	5
195	2696	516	32	2128	1	6
595	3464	1424	32	2088	1	7
0	2232	496	0	1736	1	8
264	2576	544	48	1992	1	9
3008	3016	960	32	2024	1	10
-	-	-	-	-	-	-
1550	2832	752	48	2032	1	50
1238	2832	696	48	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	784	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1417	2832	728	48	2064	1	57
0	2232	496	0	1736	1	58
384	2632	560	32	2040	1	59
225	2768	544	48	2184	1	60

Which statement accurately describes the dbo.FactInternetSales table?

- A. The table contains less than 1,000 rows.
- B. All distributions contain data.
- C. The table is skewed.
- D. The table uses round-robin distribution.

Answer: C

Explanation:

Data skew means the data is not distributed evenly across the distributions. Reference:

[https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribution)

278. - (Exam Topic 3)

You have an Azure subscription that contains an Azure Data Lake Storage account named myaccount1.

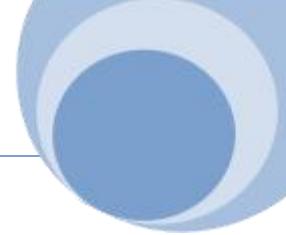
The myaccount1 account contains two containers named container1 and contained. The subscription is linked to an Azure Active Directory (Azure AD) tenant that contains a security group named Group1.

You need to grant Group1 read access to container1. The solution must use the principle of least privilege.

Which role should you assign to Group1?

- A. Storage Blob Data Reader for container1
- B. Storage Table Data Reader for container1





- C. Storage Blob Data Reader for myaccount1
- D. Storage Table Data Reader for myaccount1

Answer: A

279. - (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 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

B. No

Answer: A

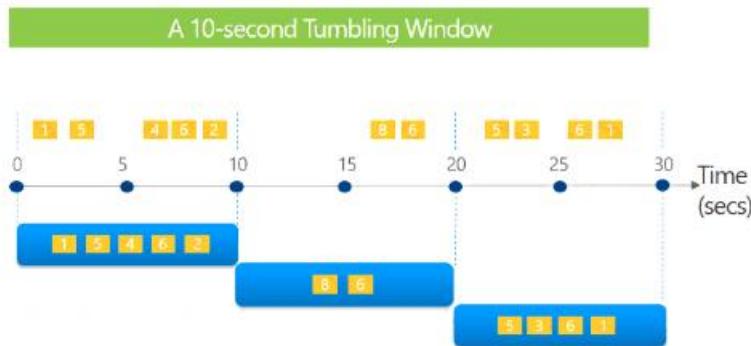
Explanation:

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



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

Reference:

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

280. - (Exam Topic 3)

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the UX Authoring canvas for ADF1.

Which two actions should you perform? Each correct answer presents part of the solution

NOTE: Each correct selection is worth one point.

- A. Create an Azure Data Factory trigger
- B. From the UX Authoring canvas, select Set up code repository
- C. Create a GitHub action
- D. From the Azure Data Factory Studio, run Publish All.
- E. Create a Git repository
- F. From the UX Authoring canvas, select Publish

Answer: D E

Explanation:

Reference:

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





281. - (Exam Topic 3)

You are implementing Azure Stream Analytics windowing functions.

Which windowing function should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap:

Hopping
Sliding
Tumbling

Segment the data stream into distinct time segments that repeat and can overlap:

Hopping
Sliding
Tumbling

Segment the data stream to produce an output only when an event occurs:

Hopping
Sliding
Tumbling

Answer:

Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap:

Hopping
Sliding
Tumbling

Segment the data stream into distinct time segments that repeat and can overlap:

Hopping
Sliding
Tumbling

Segment the data stream to produce an output only when an event occurs:

Hopping
Sliding
Tumbling

282. - (Exam Topic 3)

A company uses Azure Stream Analytics to monitor devices.

The company plans to double the number of devices that are monitored.

You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load.

Which metric should you monitor?

- A. Early Input Events
- B. Late Input Events



C. Watermark delay

D. Input Deserialization Errors

Answer: A

Explanation:

There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

- Not enough processing resources in Stream Analytics to handle the volume of input events.
- Not enough throughput within the input event brokers, so they are throttled.
- 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>

283. - (Exam Topic 3)

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.


Values 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

```

Answer:
Values 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'
)
[ ](
    PIVOT (
        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

```

Explanation:

Text Description automatically generated

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.

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>





284. - (Exam Topic 3)

You have an Azure data factory.

You need to examine the pipeline failures from the last 180 days. What should you use?

- A. the Activity log blade for the Data Factory resource
- B. Azure Data Factory activity runs in Azure Monitor
- C. Pipeline runs in the Azure Data Factory user experience
- D. the Resource health blade for the Data Factory resource

Answer: B

Explanation:

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

285. - (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>

286. - (Exam Topic 3)

You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB. You need to create the table to meet the following requirements:

- Provide the fastest Query time.
- Minimize data movement during queries. Which type of table should you use?

A. hash distributed

B. heap

C. replicated

D. round-robin

Answer: C

Explanation:

A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables>

287. - (Exam Topic 3)

You have a SQL pool in Azure Synapse.

You discover that some queries fail or take a long time to complete. You need to monitor for transactions that have rolled back.

Which dynamic management view should you query?

A. sys.dmv_request_steps



- B. sys.dm_pdw_nodes_tran_database_transactions
- C. sys.dm_pdw_waits
- D. sys.dm_pdw_exec_sessions

Answer: B

Explanation:

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.

If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.

Example:

```
-- Monitor rollback SELECT  
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END),  
t.pdw_node_id,  
nod.[type]  
FROM sys.dm_pdw_nodes_tran_database_transactions t  
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id GROUP BY t.pdw_node_id,  
nod.[type]
```

Reference:

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

288. - (Exam Topic 3)

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest.

Which two components should you include in the recommendation? Each coned answer presents part of the solution

NOTE: Each correct selection is worth one point.

- A. an X509 certificate
- B. an RSA key
- C. an Azure key vault that has purge protection enabled





- D. an Azure virtual network that has a network security group (NSG)
- E. an Azure Policy initiative

Answer: B C

Explanation:

Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace. Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM keys.

The Key Vault itself needs to have purge protection enabled. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

