PassLeader®

Leader of IT Certifications

Vendor: Microsoft

Exam Code: DP-203

Exam Name: Data Engineering on Microsoft Azure

Version: DEMO

QUESTION 1
Case Study 1 - Contoso, Ltd

Overview
Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest it integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment
Transactional Data
Contoso has three years of customer, transactional, operational, 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 periods. 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.

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 server-level firewall IP rule
D. a database-level firewall IP rule

Answer: C
Explanation:
There is no VPN between on-premises machines and Azure SQL server, communications use a public endpoint. You can limit the public access to databases through a Server Level IP Firewall rules.
https://docs.microsoft.com/en-us/azure/azure-sql/database/network-access-controls-overview

QUESTION 2
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?

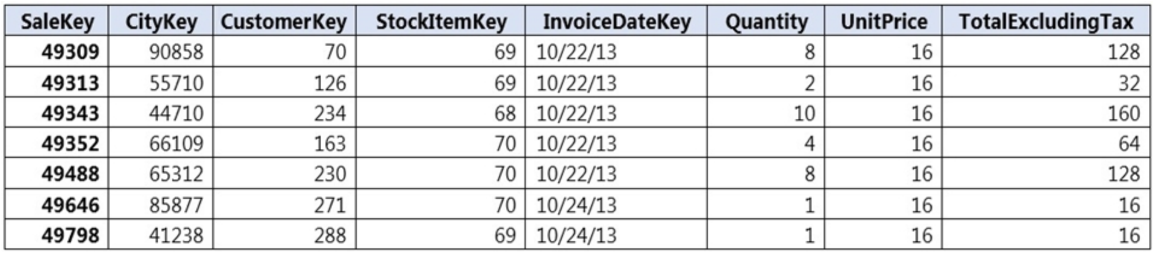
A. data masking
B. Always Encrypted
C. column-level security
D. row-level security

Answer: C
Explanation:
Column-level security simplifies the design and coding of security in your application, allowing you to restrict column access to protect sensitive data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security

QUESTION 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

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.
Incorrect Answers:
C, D: Round-robin tables are useful for improving loading speed.
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

QUESTION 4
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

QUESTION 5
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 in 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: B
Explanation:
Azure Databricks provides three kinds of logging of cluster-related activity:
Cluster event logs, which capture cluster lifecycle events, like creation, termination, configuration edits, and so on.
Apache Spark driver and worker logs, which you can use for debugging.
Cluster init-script logs, valuable for debugging init scripts.
https://docs.microsoft.com/en-us/azure/databricks/clusters/clusters-manage#event-log

QUESTION 6

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

QUESTION 7
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

QUESTION 8
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.dm_pdw_nodes_tran_database_transactions
B. sys.dm_pdw_waits
C. sys.dm_pdw_request_steps
D. sys.dm_pdw_exec_sessions

Answer: A
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),
E. 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-monitor#monitor-transaction-log-rollback

QUESTION 9
You plan to create an Azure Synapse Analytics dedicated SQL pool.

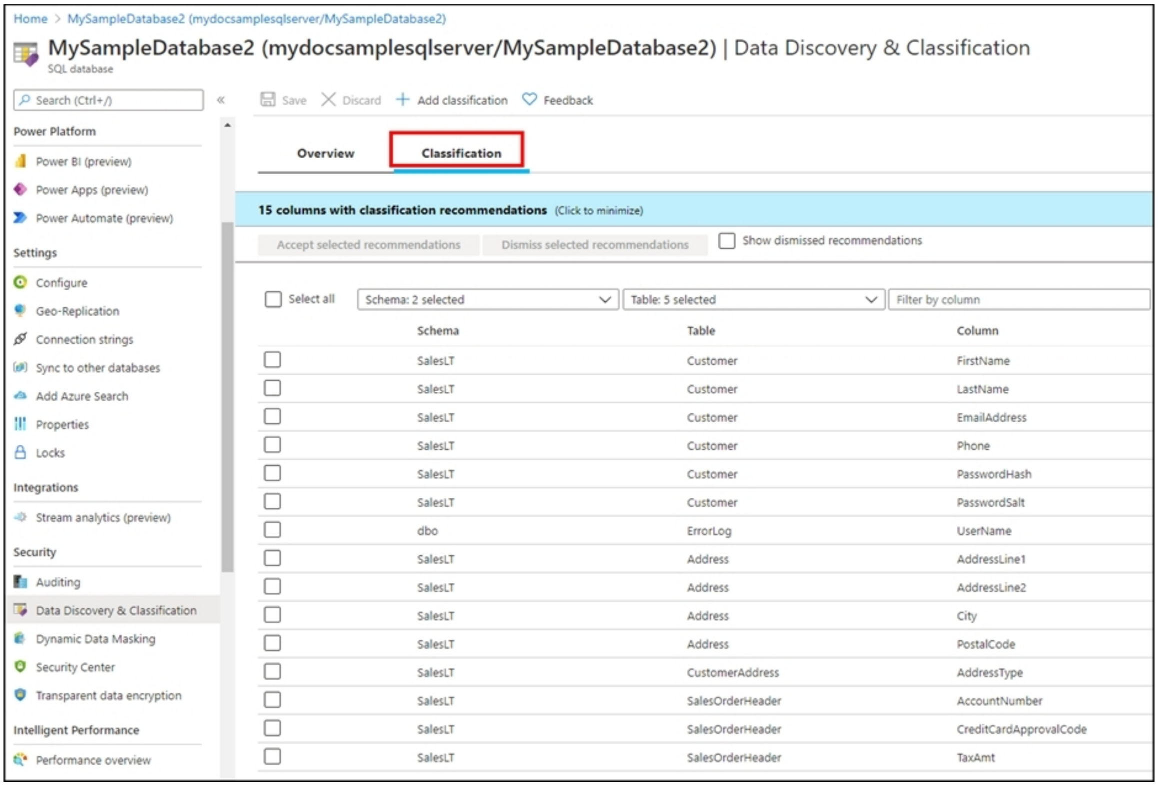
You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.

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

NOTE: Each correct selection is worth one point.

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

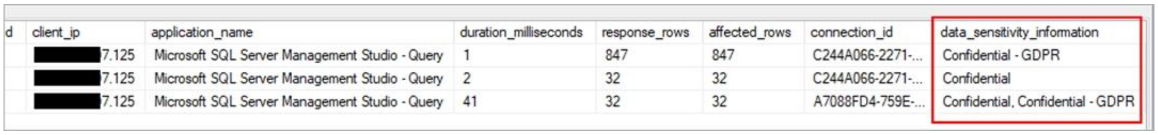
Answer: AC
Explanation:
A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:



15 columns with classification recommendations (Click to minimize)

	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

1. Select Add classification in the top menu of the pane.
2. In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
3. 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

QUESTION 10
Hotspot Question

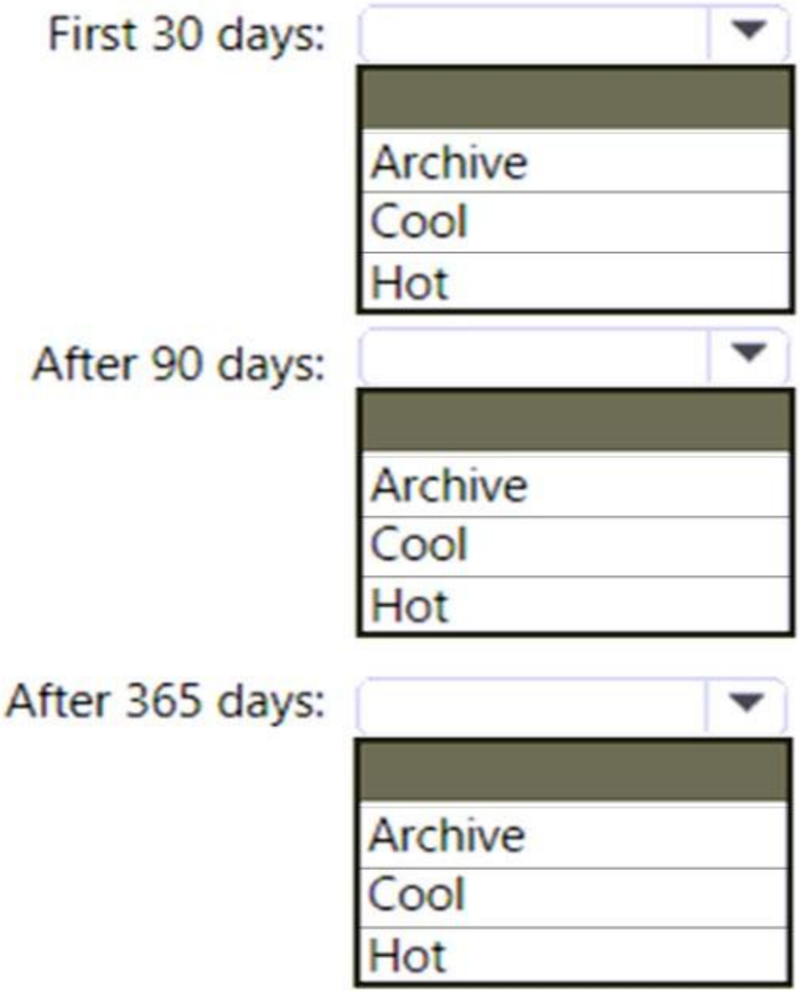
You are designing an application that will use an Azure Data Lake Storage Gen 2 account to store petabytes of license plate photos from toll booths. The account will use zone-redundant storage (ZRS).
You identify the following usage patterns:

```
- The data will be accessed several times a day during the first 30
```

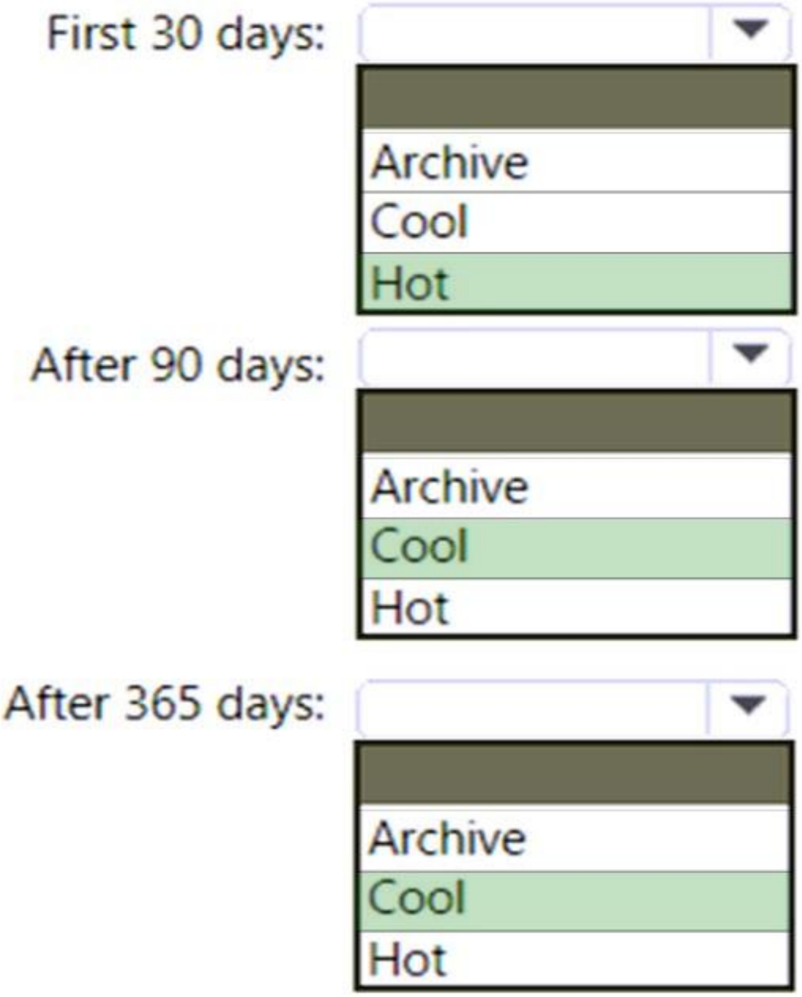
days after the data is created. The data must meet an availability SLA of 99.9%.
- After 90 days, the data will be accessed infrequently but must be available within 30 seconds.
- After 365 days, the data will be accessed infrequently but must be available within five minutes.

You need to recommend a data retention solution. The solution must minimize costs.

Which access tier should you recommend for each time frame? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



First 30 days:

- Archive
- Cool
- [✓ Hot]

After 90 days:

- Archive
- [✓ Cool]
- Hot

After 365 days:

- Archive
- [✓ Cool]
- Hot

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.

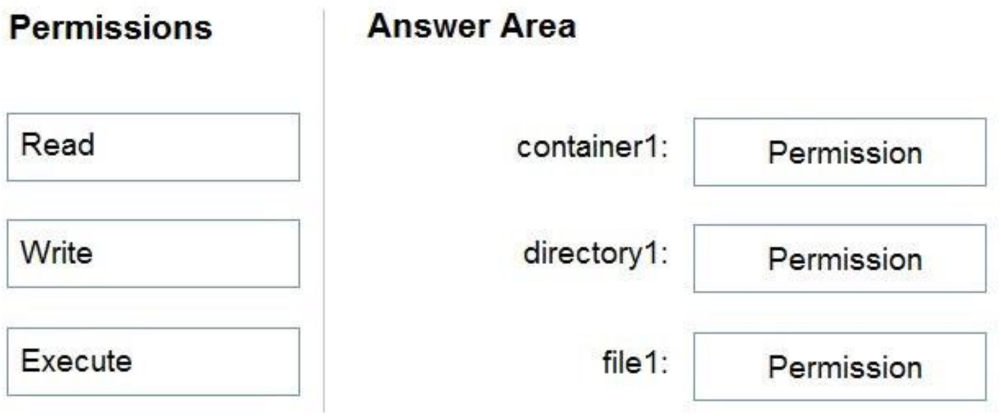
Box 3: Cool
After 365 days, the data will be accessed infrequently but must be available within five minutes.

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

QUESTION 11
Drag and Drop Question

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

Read

Write

Execute

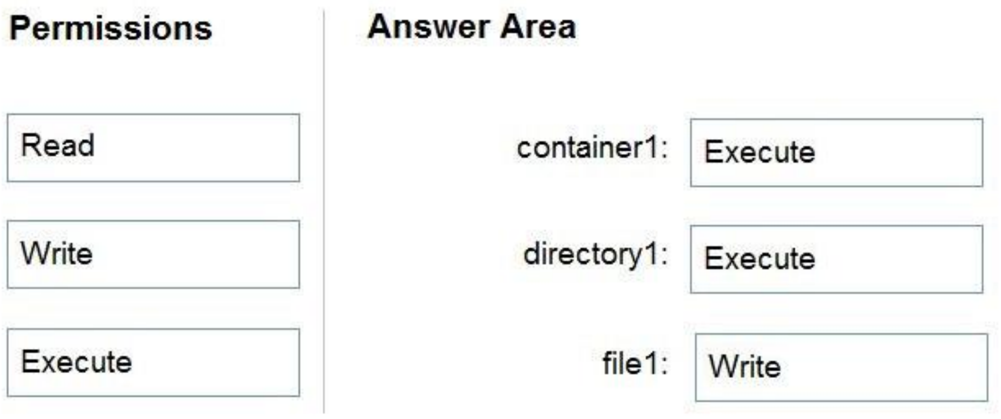
Answer Area

container1: Permission

directory1: Permission

file1: Permission

Answer:



Permissions

Read

Write

Execute

Answer Area

container1: Execute

directory1: Execute

file1: 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

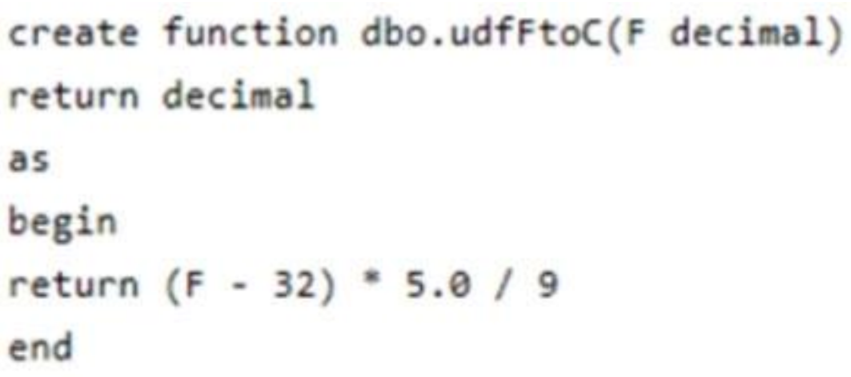
lead to the file.

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.

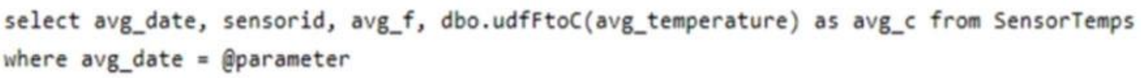
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

QUESTION 12
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: BD
Explanation:
A calculated column is a column that uses an expression to calculate its value based on other columns in the same table. In this case, the udfFtoC function can be used to calculate the avg_c value based on the avg_temperature column, eliminating the need to call the UDF in the SELECT statement.

Enabling result set caching can improve query performance by caching the result set of the query, so subsequent queries that use the same parameters can be retrieved from the cache instead of executing the query again.
Creating an index on the avg_f column or the sensorid column is not useful because there are no join or filter conditions on these columns in the WHERE clause. Changing the table distribution to replicate is also not necessary because it does not affect the query performance in this scenario

QUESTION 13
You have an Azure Data Factory pipeline named Pipeline1. Pipeline1 contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.
Pipeline1 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 Pipeline1 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. Create a pull request.
C. Modify the schedule trigger.
D. Configure the change feed of the new storage account.

Answer: A
Explanation:
CI/CD lifecycle
1. 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.
2. A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes
3. 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.
4. 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

QUESTION 14
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: C
Explanation:
You cannot create user-defined restore points when the Azure Synapse Analytics dedicated SQL pool is currently paused. In order to create a user-defined restore point, the SQL pool must be running.

https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-restore-points

QUESTION 15
You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.
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. ALTER
D. MERGE

Answer: D
Explanation:
When applying updates to a Type 2 slowly changing dimension (SCD) table in Azure Databricks, the best option is to use the MERGE operation in Apache Spark SQL. This operation allows you to combine the data from the source table with the data in the destination table, and then update or insert the appropriate records. The MERGE operation provides a powerful and flexible way to handle updates for SCD tables, as it can handle both updates and inserts in a single operation. Additionally, this operation can be performed on Delta Lake tables, which can easily handle the ACID transactions needed for handling SCD updates.

QUESTION 16
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

QUESTION 17
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: A
Explanation:
Considering the high volume of data, for faster queries its recommended to create fewer partitions.
" If a table contains fewer than the recommended minimum number of rows per partition(i.e. 60 million rows per month for 60 distributed partitions, consider using fewer partitions in order to increase the number of rows per partition."

QUESTION 18
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.