

Microsoft DP-203 Exam

Certification Provider: Microsoft

Exam: Data Engineering on Microsoft Azure

Duration: 2 Hours

Number of questions in the database: 320

Exam Version: Sept. 2, 2023

Exam Topics:

Topic 1: Question Set 1

Topic 2: Question Set 2

Topic 3: Question Set 3

Topic 4: Question Set 4

Topic 5: Testlet 1

Topic 6: Testlet 2

Topic 7: Testlet 3

Topic 8: Testlet 4

Topic 9: Testlet 5

Topic 10: Testlet 6

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.

Transactional 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 an 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.

Question 1

DRAG DROP -

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool.

The solution must meet the customer sentiment analytic requirements.

Which three Transact-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.

Select and Place:

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

ET User:

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

mbravo Highly Voted 2 years, 2 months ago

DATA SOURCE

FILE FORMAT

TABLE (WITHOUT AS SELECT)

EXTERNAL TABLE AS SELECT is nonsense in this context. It is used to unload data from DW to external data store. We are trying to load data into DW here.

upvoted 14 times

Alekx42 2 years, 2 months ago

I agree. Documentation on this:

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

upvoted 3 times

bsa_2021 Highly Voted 2 years, 2 months ago

Pls read the text - "Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units" which means we should not use more resource to create a path and folder while creating and dumping the output files.

Now read the below text from below link which says that TABLE (WITHOUT AS SELECT) will not create the path and folder. In contrast, when we use CREATE TABLE AS SELECT statement, it will automatically create the path and folder.

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated>

In my personal view, this is the reason why CREATE TABLE AS SELECT is more appropriate. Hence my answer would be

DATA SOURCE

FILE FORMAT

CREATE TABLE AS SELECT

upvoted 10 times

<p>✉️👤 csd Most Recent 1 year ago</p> <p>CREATE EXTERNAL DATA SOURCE CREATE EXTERNAL FILE FORMAT CREATE EXTERNAL TABLE above 3 steps needed.</p> <p>CREATE DATABASE SCOPED CREDENTIAL , this is not needed cause, CREDENTIAL = <database scoped credential> is optional credential that will be used to authenticate on Azure storage. External data source without credential can access public storage account or use the caller's Azure AD identity to access files on storage.</p> <p> upvoted 3 times</p>
<p>✉️👤 Deeksha1234 1 year ago</p> <p>the answer is wrong, I agree with the explanation provided by Namit If it is public data, then no need for SCOPE Credential. Answer should be:- CREATE EXTERNAL DATA SOURCE CREATE EXTERNAL FILE FORMAT CREATE EXTERNAL TABLE</p> <p> upvoted 2 times</p>

ET Answer:

Commands

```
CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE
CREATE EXTERNAL TABLE AS SELECT
CREATE DATABASE SCOPED CREDENTIAL
```

Answer Area

```
CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE AS SELECT
```

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.

Incorrect Answers:

CREATE EXTERNAL TABLE -

The CREATE EXTERNAL TABLE command creates an external table for Synapse SQL to access data stored in Azure Blob Storage or Azure Data Lake Storage.

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

Question 2

HOTSPOT -

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.

Hot Area:

Answer Area

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
An Azure Synapse Analytics workspace

Answer:

Answer Area

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
An Azure Synapse Analytics workspace

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

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

✉ erssiws 2 years, 2 months ago

agree with Alekx42

◀️ upvoted 2 times

✉ Yatoom 2 years, 2 months ago

The text provides the following statements:

- 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.

It seems that the first points towards using Sales Date as partition key, while the second one hints towards using the product ID as partition key. Which is the right answer and why?

◀️ upvoted 2 times

✉ Alekx42 2 years, 2 months ago

The provided answer is correct. About your doubt, this is my reasoning:

- PARTITION by Sales Date, in order to take advantage of partition switching to delete the last month and query data only on a subset of partitions.
- DISTRIBUTE the table (which will have to be hash-distributed, of course) by Product ID to improve the join performance. Filters on product ID won't be as quick, but for that I guess you could create a non-clustered index on Product ID.

◀️ upvoted 10 times

Question 3

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

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.

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

 **dragos_dragos62000** 2 years, 2 months ago

Correct, surrogate key is created with IDENTITY statement.

   upvoted 3 times

Question 4

HOTSPOT -

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.

Hot Area:

Answer Area

Table type to store retail store data:

Hash
Replicated
Round-robin

Table type to store promotional data:

Hash
Replicated
Round-robin

ET User:

Replicated

Hash

alexleonvalencia Highly Voted 1 year, 8 months ago

REPLICATED & HASH, La primera serviría como dimensión y la segunda como una entidad FACT.

upvoted 38 times

PallaviPatel Highly Voted 1 year, 7 months ago

Replicated and Hash are the answers

upvoted 7 times

wwdba Most Recent 1 year, 5 months ago

It should be REPLICATED and HASH

upvoted 2 times

ovokpus 1 year, 6 months ago

Replicated and Hash makes the most intuitive sense here

upvoted 1 times

ET Answer:

Answer Area

Table type to store retail store data:

Hash
Replicated
Round-robin

Table type to store promotional data:

Hash
Replicated
Round-robin

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.

Scenario:

☞ 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.

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

Reference:

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

Question 5

HOTSPOT -

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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:

Answer Area

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

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.

<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-data-warehouse/sql-data-warehouse-tables-partition>

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

✉️ **Canary_2021** Highly Voted 1 year, 8 months ago

Answers are correct.

- Sales data are loaded from on-premises SQL server to Synapse Dedicated SQL Pool using SSIS.
- 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.

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

👍 ↗️ 🎉 upvoted 8 times

✉️ **Canary_2021** 1 year, 8 months ago

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

Boundary values must belong to the partition on the right.

Based on this doc online, Answer of Box 2 should be 'AS RANGE LEFT FOR VALUES.'

👍 ↗️ 🎉 upvoted 1 times

✉️ **Canary_2021** 1 year, 8 months ago

Correct my answer Box 2 should be 'AS RANGE RIGHT FOR VALUES.'

👍 ↗️ 🎉 upvoted 2 times

✉️ **[Removed]** Most Recent 7 months, 3 weeks ago

AS RANGE RIGHT FOR VALUES

👍 ↗️ 🎉 upvoted 1 times

Question 6

HOTSPOT -

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.

Hot Area:

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.

ET User:

Hash

Set the distribution column to Product ID

Jerrie86 Highly Voted 7 months, 1 week ago

This case study was in my exam and I scored 970. I chose productid.

upvoted 31 times

RoyP654 2 months, 3 weeks ago

Good Job, Congrats!

upvoted 2 times

Julia01 Highly Voted 11 months, 4 weeks ago

Id choose product id as well since it will be used in joins "Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible."

upvoted 18 times

mokrani 10 months ago

Why not sales date for distribution column ?
Partition date 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...

upvoted 1 times

k18585 9 months, 2 weeks ago

because it's asking about distribution, not partition. The requirements say "ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible". The best way to do so is hash distributing on product ID, this way all rows with the same product id will be on the same node and there will be no data shuffling, hence fast queries

upvoted 11 times

ET 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.

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/>

Question 7

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

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.

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

✉ **Kondzio** 1 year, 6 months ago

Selected Answer: D

There is a blade on storage account called "Lifecycle management" which have the functionality to purge old blobs. So I think the answer is correct

Like 1 Upvote 3 times

✉ **Mad_001** 1 year, 6 months ago

Following requirement is placed in the text: "Purge Twitter feed data records that are older than two years."

In my opinion, the solution should be time-based-retention

Like 1 Upvote 1 times

✉ **dev2dev** 1 year, 7 months ago

why not time-based retention

Like 1 Upvote 4 times

✉ **Adsk323** 1 year, 7 months ago

sounds correct

Like 1 Upvote 4 times

Question 8

DRAG DROP -

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 the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

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



Answer:

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

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 -

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

✉ ItHYMeRish **Highly Voted** 1 year, 8 months ago

The answer provided is correct

👍 ↗️ 📈 upvoted 34 times

✉ SameerL 1 year, 1 month ago

The provided sequence is correct per below link:

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

👍 ↗️ 📈 upvoted 9 times

Question 9

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

Question 10

HOTSPOT -

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.

Hot Area:

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

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.

Incorrect Answers:

- ☞ Event Hubs Dedicated: Event Hubs clusters offer single-tenant deployments for customers with the most demanding streaming needs. This single-tenant offering has a guaranteed 99.99% SLA and is available only on our Dedicated pricing tier.
- ☞ Auto-Inflate: The Auto-inflate feature of Event Hubs automatically scales up by increasing the number of TUs, to meet usage needs.

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).

Incorrect Answers:

- ☞ Azure Databricks: An Azure administrator with the proper permissions can configure Azure Active Directory conditional access to control where and when users are permitted to sign in to Azure Databricks.
- ☞ Azure Storage supports using Azure Active Directory (Azure AD) to authorize requests to blob data.

You can scope access to Azure blob resources at the following levels, beginning with the narrowest scope:

- An individual container. At this scope, a role assignment applies to all of the blobs in the container, as well as container properties and metadata.
- The storage account. At this scope, a role assignment applies to all containers and their blobs.
- The resource group. At this scope, a role assignment applies to all of the containers in all of the storage accounts in the resource group.
- The subscription. At this scope, a role assignment applies to all of the containers in all of the storage accounts in all of the resource groups in the subscription.
- A management group.

<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>

✉️ noobprogrammer Highly Voted 1 year, 4 months ago

Answer looks correct to me:

1) Configure Event Hubs partition - The description says: "Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units."

2) An Azure Data Lake Storage Gen2 account.

Databricks cluster has nothing to do with storage, and a Data lake fits the needs

👍👎🚩 upvoted 12 times

Case Study 2 - Litware, inc.

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.

Question 1

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

Scenario:

- ⇒ Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
- ⇒ Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Since Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure, they will have to create firewall IP rules to allow connection from the IP ranges of the on-premise network. They can also use the firewall rule 0.0.0.0 to allow access from Azure services.

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

✉️  **Kyle1**  1 year, 11 months ago

I think it should be C. The company doesn't want any virtual network stuff and server-level is more comprehensive, thus safer than just database-level rule.

   upvoted 33 times

✉️  **Marcus1612**  1 year, 11 months ago

The answer is C. Since 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/sql-database/network-access-controls-overview>

   upvoted 12 times

Question 2

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

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

ET User: C

✉️  **Alekx42**  2 years, 2 months ago

But can you limit access to data with sensitivity labels? It seems to me that they are used more for auditing and monitoring. I would have answered C - column level security.

   upvoted 23 times

✉️  **escoins**  2 years, 2 months ago

The answer should be C

   upvoted 2 times

✉️  **bsa_2021** 2 years, 2 months ago

The given answer is correct. Please read the below link which says below -
"Finally, as an admin, when you configure a sensitivity label to apply encryption, you can choose either to:

Assign permissions now, so that you determine exactly which users get which permissions to content with that label.

Let users assign permissions when they apply the label to content. This way, you can allow people in your organization some flexibility that they might need to collaborate and get their work done.

The encryption settings are available when you create a sensitivity label in the Microsoft 365 compliance center.

<https://docs.microsoft.com/en-us/microsoft-365/compliance/encryption-sensitivity-labels?view=o365-worldwide>

   upvoted 2 times

✉️  **Marcello83** 2 years, 2 months ago

Your citation is about sensitivity labels in Office 365. For SQL, this is the link : <https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview>

"Labeling: You can apply sensitivity-classification labels persistently to columns by using new metadata attributes that have been added to the SQL Server database engine. This metadata can then be used for sensitivity-based auditing and protection scenarios."

So, in my opinion, the correct answer is C

   upvoted 1 times

✉️  **erssiws** 2 years, 2 months ago

Agree with Alekx42. The answer should be C

   upvoted 3 times

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.

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

ET Answer: D

Scenario: Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Labeling: You can apply sensitivity-classification labels persistently to columns by using new metadata attributes that have been added to the SQL Server database engine. This metadata can then be used for advanced, sensitivity-based auditing and protection scenarios.

Incorrect Answers:

A: Transparent Data Encryption (TDE) encrypts SQL Server, Azure SQL Database, and Azure Synapse Analytics data files, known as encrypting data at rest. TDE does not provide encryption across communication channels.

Reference:

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

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

Question 3

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

- A. Deploy a High Concurrency Databricks cluster.
- B. 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.
- C. Set Data Lake Storage to use geo-redundant storage (GRS).
- D. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.

Answer: D

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

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

✉️ petulda Highly Voted 2 years ago

There is a request 'Minimize number of Azure services'. With <https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-capture-overview> Event capture, data can be stored in DL without using Stream Analytics. In this case just Regional redundancy for DL would be needed.

👍 ↗️ 🎉 upvoted 12 times

✉️ sachabess79 1 year, 11 months ago

NB : it's an asynchronous copy.

👍 ↗️ 🎉 upvoted 1 times

✉️ ian_viana 1 year, 11 months ago

Agree, they also want a stage on data lake 2.

"Stage Inventory data in Azure Data Lake Storage Gen2"

we don't need Stream Analytics to do that. Event Hub enables you to automatically capture the streaming data in Event Hubs in an Azure Blob storage or Azure Data Lake Storage Gen 1 or Gen 2 account of your choice, with the added flexibility of specifying a time or size interval.

👍 ↗️ 🎉 upvoted 1 times

✉️ ian_viana 1 year, 11 months ago

Please consider my answer!

Event Hub can capture data to Data Lake and Blob. But I think the key word in the question is: real-time data PROCESSING solution azure. Event hub is just for capture. Stream Analytics do the processing so I'm going with answer D

👍 ↗️ 🎉 upvoted 9 times

Question 4

HOTSPOT -

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.

Hot Area:

Answer Area

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:

Answer Area

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

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.

✉️ **evericka** 2 years ago

Shouldn't the first box be azure integration runtime?

1 upvoted 2 times

✉️ **BlackMal** 2 years ago

Why would it be Self Hosted Integration Runtime if it explicitly says "Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure."

1 upvoted 1 times

✉️ **bsa_2021** 2 years, 2 months ago

You can use tumbling but watch out for the 2nd line which says "Minimize the number of different Azure services needed to achieve the business goals.". Schedule window is a simple one while tumbling window is a higher powered/capacity one with lot more flexibility. But we want to minimize and hence go for the basic one.

1 upvoted 1 times

✉️ **Alekx42** 2 years, 3 months ago

I guess if data must be copied daily there is no need for a tumbling window trigger (there are no non-overlapping trigger requirements, no dependencies, no retries to be set, etc)

1 upvoted 2 times

✉️ **rumosgf** 2 years, 3 months ago

Why is schedule instead tumbling? Someone can explain?

1 upvoted 1 times

1 upvoted 3 times

✉️ **yoblip** 2 years, 2 months ago

coz requirement said need to be schedule every 8 hours

1 upvoted 3 times

Solution Challenge

Question 1 2 3 4

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.

- A. Solution: You convert the files to compressed delimited text files.
- B. Solution: You copy the files to a table that has a columnstore index.
- C. Solution: You modify the files to ensure that each row is more than 1 MB.
- D. Solution: You modify the files to ensure that each row is less than 1 MB.

Answer: A, D 有小争议, Bing 给的单 A

To copy data from an Azure Storage account to an enterprise data warehouse in Azure Synapse Analytics, you can use the Copy Activity in Azure Data Factory or Synapse pipelines ¹. Azure Storage is used as the staging area in the copy operation ². To achieve the best performance, it is recommended to use PolyBase or COPY statement to load data into Azure Synapse Analytics ¹.

Based on the information provided, the best solution would be **A. Solution: You convert the files to compressed delimited text files.** This will reduce the size of the files and make the data copy faster. The other solutions may not be as effective in ensuring that the data copies quickly. For example, modifying the files to ensure that each row is more or less than 1 MB (solutions C and D) may not have a significant impact on the speed of the data copy. Copying the files to a table that has a columnstore index (solution B) may improve query performance, but it may not necessarily speed up the data copy process.

A 的评论, ET 给的 Y, 用户 80%以上 Y

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

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

- ✉️ **Fahd92** Highly Voted 1 year, 11 months ago
They said you need to prepare the files to copy, maybe the mean we should make them less than 1MB ? so it will be A else would be B !!!!
 upvoted 16 times
- ✉️ **ANath** 1 year, 7 months ago
The answer should be A.
<https://azure.microsoft.com/en-gb/blog/increasing-polybase-row-width-limitation-in-azure-sql-data-warehouse/>
 upvoted 4 times
- ✉️ **Thij** 1 year, 11 months ago
After reading the other questions oh this topic I go with A because the relevant part seems to be the compression.
 upvoted 4 times

B 的评论, ET 给的 N

- ✉️ **bhanuprasad9331** Highly Voted 1 year, 6 months ago
From the documentation, loads to heap table are faster than indexed tables. So, better to use heap table than columnstore index table in this case.
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index#heap-tables>
 upvoted 9 times
- ✉️ **Odoxtoom** Highly Voted 1 year, 10 months ago
Consider this sets one question:
What should you do to improve loading times?
What | Yes | No |
compressed | O | O |
columnstore | O | O |
> 1MB | O | O |

So now answers should be clear
 upvoted 7 times

C 的评论, ET 给的 N

- ✉️ **Gilvan** Highly Voted 1 year, 11 months ago
No, rows need to have less than 1 MB. A batch size between 100 K to 1M rows is the recommended baseline for determining optimal batch size capacity.
 upvoted 12 times
- ✉️ **vigilante89** Most Recent 9 months ago
Selected Answer: B
B is correct!!!
 upvoted 2 times

D 的评论, ET 给的 Y, 70% 的用户选的 Y

Polybase loads rows that are smaller than 1 MB.

Note on Polybase Load: PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.

Extract, Load, and Transform (ELT)

Extract, Load, and Transform (ELT) is a process by which data is extracted from a source system, loaded into a data warehouse, and then transformed.

The basic steps for implementing a PolyBase ELT for dedicated SQL pool are:

Extract the source data into text files.

Land the data into Azure Blob storage or Azure Data Lake Store.

Prepare the data for loading.

Load the data into dedicated SQL pool staging tables using PolyBase.

Transform the data.

Insert the data into production tables.

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

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

✉  **Tj87**  1 year ago

I think we had this question in the previous pages and the correct answer was set as " compress the files"

   upvoted 21 times

✉  **semauni** 1 month, 1 week ago

More than 1 solution might be right. The question here is: if row size is reduced to 1MB, will loading go faster? The answer then is yes: whether compression is better or not, is not relevant.

   upvoted 2 times

✉  **kim32** 3 months, 4 weeks ago

The question before was more than that 1 MB but here is less than 1 MB, since, it is less, then answer is Yes.

   upvoted 3 times

✉  **dom271219** 12 months ago

Exactly

compress because a lot of row have more than 1MB length

   upvoted 4 times

✉  **Phund**  12 months ago

Selected Answer: A

"ensure that each row is less than 1 MB" and the condition for polybase is <1M, whatever method you used

   upvoted 15 times

Question 5 6 7 8

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.

- A. Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds.
- B. Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.
- C. Solution: You use a tumbling window, and you set the window size to 10 seconds.
- D. Solution: You use a session window that uses a timeout size of 10 seconds.

Answer: C, A 有小争议, Bing 给的单 C

To count the tweets in each 10-second window and ensure that each tweet is counted only once, the best solution would be C. Solution: You use a tumbling window, and you set the window size to 10 seconds. A tumbling window function is used to segment a data stream into distinct time segments and perform a function against them ¹. 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 ¹. This means that each tweet will be counted only once in each 10-second window.

The other solutions may not be as effective in ensuring that each tweet is counted only once. For example, using a hopping window (solutions A and B) may result in tweets being counted more than once if the hop size is smaller than the window size ¹. Using a session window (solution D) may not be appropriate for this scenario, as session windows are typically used to group events based on periods of activity and inactivity ¹.

A 的评论, ET 给的 N, 70%以上的用户选的 Y

👤 steeee [Highly Voted] 2 years ago

The answer should be "Yes". Hopping window with hop size equals window size should be the same as Tumbling window.

👍👎💡 upvoted 113 times

👤 DataEngineer7331 4 months, 2 weeks ago

A Tumbling Window would be correct. But as stated in the following, a hopping window can be the same as a tumbling window: "To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size."

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

👍👎💡 upvoted 6 times

B 的评论, ET 给的 N

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

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

Ati1362 Highly Voted 2 years, 2 months ago

answer is correct

upvoted 20 times

allagowf Highly Voted 11 months, 1 week ago

Selected Answer: B

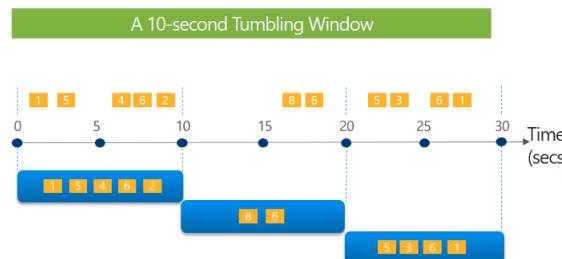
if the hop size is equivalent to the window size then it can be true, but because the hop size is smaller, then each tweet can be count more than one and the windows will overlap with each others.

upvoted 5 times

C 的评论，ET 给的 Y

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)
```

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

Prabagar Highly Voted 2 years, 2 months ago

correct answer

upvoted 37 times

Deeksha1234 Most Recent 1 year, 1 month ago

correct

upvoted 1 times

practicewizards 1 year, 1 month ago

this question appears at topic 2 question 18 and it said the correct answer was hopping window with 10" window... so, what's the right correct answer?

upvoted 2 times

kmmrhc 10 months, 3 weeks ago

Both are correct. A Hopping window with hop-size = window-size is identical to a Tumbling window.

upvoted 4 times

D 的评论，ET 给的 N

Ati1362 Highly Voted 2 years, 2 months ago

answer correct

upvoted 22 times

MoDar Highly Voted 2 years ago

False as we need to count tweets in EACH 10 sec. Session windows can have gaps if there is no event happening during the window size

upvoted 11 times

mamahani Most Recent 3 months, 3 weeks ago

Selected Answer: B

B is correct answer

upvoted 1 times

Question 9 10 11 12

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

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

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

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

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

You need to create the Databricks clusters for the workloads.

- A. 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.
- B. 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.
- C. 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.
- D. 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.

Answer: C

A 的评论，ET 给的 N

□  **Nieswurz**  3 years, 1 month ago

As job notebooks include scala and high concurrency clusters do not support scala, the answer should be no.

   upvoted 26 times

□  **TashaP** 11 months, 4 weeks ago

this is 100% correct, it's as simple as this, job requires scala, high concurrency does not support scala. The answer is no.

   upvoted 1 times

□  **Equalizer** 3 years ago

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

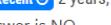
   upvoted 3 times

B 的评论, ET 给的 N

□  **ACSC**  2 years, 8 months ago

A workload that data scientists will use to perform ad hoc analysis in Scala and R. High Concurrency clusters don't support Scala. Answer is "No".

   upvoted 9 times

□  **cadio30**  2 years, 4 months ago

Definitely the answer is NO

   upvoted 1 times

C 的评论, ET 给的 N, 86% 的用户选的 Y

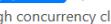
We would need a High Concurrency cluster for the jobs.

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.

<https://docs.azuredatabricks.net/clusters/configure.html>

□  **Amalberrebal**  2 years ago

- data engineers: high concurrency cluster
- jobs: Standard cluster
- data scientists: Standard cluster

   upvoted 86 times

□  **gogosgh** 4 months ago

The issue is the jobs are going to be ran by multiple users i.e. engineers and scientists? So it needs to be high concurrency cluster?

   upvoted 1 times

□  **auwia** 2 months, 1 week ago

If you enable high concurrency then all scale scripts doesn't work, so scientists will stop to work). Standard cluster is scalable, will support all jobs and users! :-)

   upvoted 2 times

□  **supriyako** 11 months, 2 weeks ago

Correct. Because jobs could be for Scala notebook, which is supported by Standard cluster mode

   upvoted 2 times

□  **Egocentric** 1 year, 4 months ago

agreed

   upvoted 1 times

□  **Julius7000** 1 year, 11 months ago

Tell me one thing: is this answer 9jobs) based on the text:

"A Single Node cluster has no workers and runs Spark jobs on the driver node.

In contrast, a Standard cluster requires at least one Spark worker node in addition to the driver node to execute Spark jobs."?

I dont understand the connection between worker nodes and the requirements given in the question about jobs workspace.

   upvoted 1 times

□  **Aditya0891** 1 year, 2 months ago

single node cluster and standard cluster are different. In single node cluster you only have 1 node which act as driver and worker node while in standard cluster you can have separate driver and worker node and for jobs you can use standard or high concurrency cluster as well. So the requirements are satisfied here

   upvoted 1 times

□  **gangstfear**  2 years ago

The answer must be A!

   upvoted 34 times

D 的评论, ET 给的 Y, 用户全部选的 N

✉️👤 **dfdsfdsfsd** Highly Voted 2 years, 3 months ago

High-concurrency clusters do not support Scala. So the answer is still 'No' but the reasoning is wrong.
<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

👍👎FLAG upvoted 45 times

✉️👤 **Preben** 2 years, 2 months ago

I agree that High concurrency does not support Scala. But they specified using a Standard cluster for the jobs, which does support Scala. Why is the answer 'No'?
👍👎FLAG upvoted 3 times

✉️👤 **eng1** 2 years, 2 months ago

Because the High Concurrency cluster for each data scientist is not correct, it should be standard for a single user!
👍👎FLAG upvoted 6 times

✉️👤 **FRAN_CO_HO** Highly Voted 2 years, 2 months ago

Answer should be NO, which
Data scientist: STANDARD as need to run scala
Jobs: STANDARD as need to run scala
Data Engineers: High-concurrency clusters as better resource sharing
👍👎FLAG upvoted 14 times

Question 13 14 15 16

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.

- A. Solution: In an Azure Synapse Analytics pipeline, you use a Get Metadata activity that retrieves the DateTime of the files.
- B. Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.
- C. Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column.
- D. Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.

Answer: 争议好大啊，看投票率像是 B, Bing 给的单 A, 另外一个题库给的是 B, C 在那个题库里出现了两次一次 Y 一次 N

To ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1, the best solution would be **A. Solution: In an Azure Synapse Analytics pipeline, you use a Get Metadata activity that retrieves the DateTime of the files.** This solution will allow you to retrieve the DateTime of the files and store it as an additional column in Table1.

The other solutions may not be as effective in ensuring that the DateTime is stored as an additional column in Table1. For example, using a data flow that contains a Derived Column transformation (solution B) may allow you to add a new column to the data, but it may not necessarily retrieve the DateTime of the files. Creating an external table with an additional DateTime column (solutions C and D) may allow you to store the DateTime information, but it may not necessarily retrieve the DateTime of the files when they are loaded to container1.

A 的评论，第一个同样的题 ET 给的 Y，第二个同样的题 ET 给的 N，用户 NY 七三开
You can use the Get Metadata activity to retrieve the metadata of any data in Azure Data Factory or a Synapse pipeline. You can use the output from the Get Metadata activity in conditional expressions to perform validation, or consume the metadata in subsequent activities.

<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-get-metadata-activity>

tezen_tolga Highly Voted 1 year, 6 months ago

This is a DP-203 question.

upvoted 7 times

CaptainJameson Highly Voted 1 year, 6 months ago

Selected Answer: A

Yes, Get Metadata can be used to retrieve the DateTime of the files and allow you to use this data. The question is to add it to Table1, not to an external table.

upvoted 6 times

U_C Most Recent 4 months, 3 weeks ago

Using a Get Metadata activity in an Azure Synapse Analytics pipeline to retrieve the DateTime of the files will not directly ensure that the DateTime is stored as an additional column in Table1. The Get Metadata activity only retrieves metadata information about the files, such as their names, size, and date created or modified.

To achieve the goal of storing the DateTime as an additional column in Table1, you would need to use other pipeline activities, such as a Data Flow or a Copy activity, to extract data from the files, transform it as necessary, and load it into Table1. During this process, you could use derived columns or mappings to add the DateTime column and populate it with the appropriate values.

Therefore, B is the correct answer.

upvoted 1 times

Ciupaz 10 months, 3 weeks ago

Exam DP-203: Data Engineering on Microsoft Azure

upvoted 1 times

Backy 1 year, 2 months ago

Get Metadata activity retrieves the DateTime of the files but it does not create a column in Table1, so answer is B

upvoted 4 times

RehanRajput 1 year, 3 months ago

Not sure if the answer is Yes. However, the explanation makes absolutely no sense.

We want to load data into our SQL Pools, why would we load the data in our SQL serverless pools? :-/

Using metadata activity might be PART of a solution but in itself does not give a complete indication of what the solution should be.

upvoted 1 times

o2091 1 year, 9 months ago

Answer looks correct

upvoted 3 times

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

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

juanlu46 Highly Voted 1 year, 4 months ago

Selected Answer: B

Is part of a possible solution, but it isn't sufficient to meet the goal, yo need to pass the "Get metadata"'s output as a parameter to the ingest process, processing each file inside a "for" loop, for example.

<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-get-metadata-activity>

upvoted 14 times

oldpony Highly Voted 1 year, 3 months ago

Selected Answer: A

https://docs.microsoft.com/en-us/azure/data-factory/control-flow-get-metadata-activity
points that Get Metadata activity can retrieve the corresponding Metadata type of: Created datetime of the file or folder.

upvoted 8 times

Ram9198 Most Recent 3 weeks, 1 day ago

Selected Answer: B

Does not meet the goal

upvoted 1 times

vctrhuggo 2 months, 2 weeks ago

According to ChatGPT:

B. No

The proposed solution of using a Get Metadata activity in an Azure Synapse Analytics pipeline will retrieve the DateTime of the files, but it does not address the requirement of storing the DateTime as an additional column in Table1.

upvoted 2 times

esaade 6 months ago

No, using a Get Metadata activity in an Azure Synapse Analytics pipeline to retrieve the DateTime of the files does not meet the goal of storing the DateTime as an additional column in Table1. The Get Metadata activity retrieves metadata about the files, such as file size, file name, or last modified date, but it does not provide the file content needed to extract the DateTime value and store it as an additional column in Table1. To achieve the goal, you need to use a data flow in the pipeline that loads the data from container1, extracts the DateTime value, and transforms the data by adding the DateTime column to Table1.

upvoted 5 times

OldSchool 9 months, 2 weeks ago

If DateTime is part of data in files in container1 than answer is A, but if it is not part of data in files but only Meta data of files then B. Wording in question is really strange but I think it is A because it says "data from files in container1"

upvoted 2 times

Deeksha1234 1 year ago

Its confusing, if we need to insert the dateTime of insertion then answer should be No, but if we need to insert the datetime of file modified then answer should be yes.

To me looks like the question is about 1st case so the answer should be No

upvoted 2 times

Dusica 7 months, 3 weeks ago

AGREED

upvoted 1 times

Strix 1 year, 1 month ago

Selected Answer: B

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

upvoted 2 times

Davico93 1 year, 2 months ago

I'm confusing more every time I read the solution, I don't know if it says that you have to do it in two steps, that changes everything

upvoted 1 times

MvanG 1 year, 2 months ago

It seems rather odd that in the same two previous questions "Use the derived column transformation to generate new columns in your data flow or to modify existing fields." was the answer. This is very confusing.

Reference:

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

upvoted 3 times

g2000 1 year, 4 months ago

Get Metadata seems possible

<https://www.mssqltips.com/sqlservertip/6246/azure-data-factory-get-metadata-example/>

upvoted 2 times

B 的评论, ET 给 T, 用户 YN 八二开

corebit Highly Voted 1 year, 8 months ago

Selected Answer: A

Data flows are available both in Azure Data Factory and Azure Synapse Pipelines"

"Use the derived column transformation to generate new columns in your data flow or to modify existing fields."

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

upvoted 21 times

Canary_2021 Highly Voted 1 year, 7 months ago

Selected Answer: B

Derived Column cannot get DateTime (created or lastmodified datetime) of the files.

Get Metadata activity can retrieve the DateTime of the files.

so answer should be B.

upvoted 6 times

Canary_2021 1 year, 7 months ago

If it is a real time process and pipeline is triggered to load data to table1 when file drop to container immediately, the created datetime of the file is similar as the pipeline process datetime. In this way Derived Column works.

The question is not clear.

upvoted 8 times

Jerrie86 7 months, 1 week ago

Can we just use the current datetime when the data is loaded. It doesn't say that we need to get data from the files. Just datetime which is kind of confusing. I will say, use derived column

upvoted 3 times

kkk5566 Most Recent 5 days, 8 hours ago

A. Use this transformation to add any new columns to existing data.

upvoted 1 times

Deeksha1234 1 year ago

correct

upvoted 2 times

Anandtr 1 year ago

Selected Answer: A

Correct

upvoted 2 times

mkthoma3 1 year, 2 months ago

What is the DateTime measuring? The DML transaction time or a file property?

If the measurement gives respect to the DML transaction time, you can use this: <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-expressions-usage#currentTimestamp>

upvoted 1 times

C 的评论, ET 给的 N

- ④ Canary_2021 **Highly Voted** 1 year, 8 months ago
Selected Answer: B
Answer should be B.
An external table is based on a source flat file structure. It seems to make no sense to add additional date time columns to such a table.
upvoted 16 times
- ④ Deeksha1234 **Most Recent** 1 year ago
Selected Answer: B
B is correct
upvoted 1 times
- ④ youngbug 1 year ago
From the words in the Solution part, it seems to use PolyBase to read external tables. PolyBase can't change the schemas of external tables(files). You can only transform the data after loading data in the staging directory. And then load the data into tables
upvoted 3 times
- ④ sdokmak 1 year, 3 months ago
Selected Answer: B
serverless works for data lake
dedicated doesn't
upvoted 2 times
- ④ GDJ2022 1 year, 7 months ago
It's clearly mentioned "You plan to insert data from the files in container1 into Table1". External tables don't get the data inserted into themselves, but instead refer outside data.
upvoted 4 times
- ④ edba 1 year, 7 months ago
If using dedicated SQL pool, after creating an external table, need a further CTAS for adding derived columns.
upvoted 3 times

D 的评论, ET 给的 N

A (53%) B (47%)

- ④ rainbowyu **Highly Voted** 1 year, 7 months ago
You can't use serverless pool to create table in dedicated pool
upvoted 19 times
- ④ Knoushore1 **Highly Voted** 10 months ago
Selected Answer: B
Table1 is in dedicated SQL pool
upvoted 5 times
- ④ Ram9198 **Most Recent** 3 weeks, 1 day ago
Selected Answer: B
You can't use serverless pool to create table in dedicated pool
upvoted 1 times
- ④ AliakseiM 3 weeks, 1 day ago
Selected Answer: B
B since Table1 is in dedicated pool
upvoted 2 times
- ④ g2000 1 month ago
Selected Answer: A
You can use external tables to read external data using dedicated SQL pool or serverless SQL pool.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>
upvoted 2 times
- ④ OldSchool 9 months ago
Selected Answer: A
Q: 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.
Park for a while Table1 and dedicated SQL pool, that is where the transformation will happen AFTER loading from container1 to Table1.
Here is about loading data to ADLSG2 container1 and adding a column which can be done with serverless SQL as an external table.
upvoted 3 times
- ④ berend1 10 months, 1 week ago
if table 1 would be serverless, yes, now no
upvoted 1 times
- ④ emna2022 11 months, 3 weeks ago
The job is to insert data from the files in container1 into Table1 (in the dedicated SQL pool) and transform the data after that and we need to add a new additional column.
External table are just references to the data, only metadata is really stored in the SQL pool.
Hence anything including external table will not be a solution.
If you follow the different proposed solutions from previous questions, the most efficient solution is to use derived column transformation.
upvoted 3 times

- ✉  **Deeksha1234** 1 year ago
Selected Answer: A
yes, with serverless pool we can add a new column while creating an external table
   upvoted 1 times
- ✉  **youngbug** 1 year ago
The aim of the solution is to load data from Data Lake's files to dedicated SQL pool's tables. There are three ways: DF's Copy Activity, PolyBase and Bulk insert. It's not serverless SQL pool's business...
   upvoted 1 times
- ✉  **StudentFromAus** 1 year, 2 months ago
The answer should be yes as we can create an additional column using CETAS in a serverless SQL pool though it is not a complete solution but a step closer to the required result.
   upvoted 1 times
- ✉  **sdokmak** 1 year, 3 months ago
Serverless pool works for data lake
Dedicated doesn't
   upvoted 1 times
- ✉  **nefarious_smalls** 1 year, 3 months ago
Apparently when dealing with dedicated sql pools you can only create an external table by importing the data from source using ctas. However, when using serverless using cetas will actually export a new file to your data source as well as create an external table. With that being said I think the answer is A.
   upvoted 3 times
- ✉  **Andushi** 1 year, 4 months ago
Selected Answer: A
Answer should be Yes
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas#examples>
   upvoted 3 times
- ✉  **Billybob0604** 9 months ago
it doesn't say in the link you can add a column using external table, so no.
   upvoted 3 times
- ✉  **ranjisi01** 1 year, 5 months ago
answer is Yes
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas>
   upvoted 1 times
- ✉  **g2000** 1 year, 4 months ago
Table1 is not an external table
   upvoted 1 times
- ✉  **edba** 1 year, 8 months ago
correct to me.
   upvoted 4 times
- ✉  **edba** 1 year, 7 months ago
after further looking into it, I think the answer should be YES. pls refer to <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas#examples>
   upvoted 6 times
- ✉  **Aditya0891** 1 year, 2 months ago
edba can you please suggest where in the link is it mentioned that you can use extra columns ?
   upvoted 1 times
- ✉  **Aditya0891** 1 year, 2 months ago
Ignore my comments, I got your point thanks :)
   upvoted 1 times
- ✉  **alex623** 1 year, 7 months ago
I think it's possible modify the files using cetas, but you have to create very much cetas to modify the files, so I think thw answer is no
   upvoted 1 times

Question 17 18 19 20

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.

- A. Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse.
- B. Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse.
- C. 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.
- D. 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.

Answer: B C, 从 DP-300 里扒的，DP-203 里的答案也是 C

A 的评论，ET 给的 N

Correct 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.

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

✉️  **Ciupaz** 10 months ago
Not related to DP-300 exam.
   upvoted 2 times

✉️  **cusman** 1 year, 4 months ago
DP-203
   upvoted 4 times

✉️  **o2091** 1 year, 9 months ago
looks correct
   upvoted 1 times

B 的评论，ET 给的 N, DP-300 中部分用户选 Y, DP-203 中用户全选的 Y

[Exam DP-203 topic 2 question 65 discussion - ExamTopics](#)

Must use an Azure Data Factory, not an Azure Databricks job.

Correct 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.

✉️  **U.C** 4 months, 1 week ago
Yes, this solution meets the goal. By scheduling an Azure Databricks job that executes an R notebook, you can transform the data from the staging zone in your Azure Data Lake Storage account. Then, by inserting the data into the data warehouse in Azure Synapse Analytics, you can complete the daily process of ingesting incremental data. So, the answer is A. Yes.
   upvoted 2 times

✉️  **KIET2131** 5 months, 3 weeks ago
A. Yes, this solution meets the goal of ingesting incremental data from the staging zone, transforming the data by executing an R script, and inserting the transformed data into a data warehouse in Azure Synapse Analytics using Azure Databricks. The scheduled Azure Databricks job can be used to execute the R notebook and insert the transformed data into the data warehouse.
   upvoted 2 times

✉️  **Ciupaz** 10 months ago
Not related to DP-300 exam.
   upvoted 1 times

✉️  **cusman** 1 year, 4 months ago
DP-203
   upvoted 3 times

✉️  **o2091** 1 year, 9 months ago
looks good, what do you think?
   upvoted 1 times

C 的评论，ET 给的 Y

An Azure Data Factory can trigger a Databricks notebook.

<https://docs.microsoft.com/en-us/azure/data-factory/transform-data-using-databricks-notebook> <https://docs.databricks.com/spark/latest/sparkr/index.html>

✉️  **U.C** 4 months, 1 week ago
Yes, this solution meets the goal. By using an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, you can transform the data from the staging zone in your Azure Data Lake Storage account. Then, by inserting the data into the data warehouse in Azure Synapse Analytics, you can complete the daily process of ingesting incremental data. So, the answer is A. Yes.
   upvoted 1 times

✉️  **Ciupaz** 10 months ago
Not related to DP-300 exam.
   upvoted 1 times

✉️  **Icyb3r** 11 months, 2 weeks ago
In the Question mentioned "Execute R script" not "executes an Azure Databricks notebook"
The correct answer should be - B (No)
   upvoted 1 times

✉️  **Icyb3r** 11 months, 2 weeks ago
Sorry ignore my comment, it should be A. Yes.
Databricks can run R script Notebook.
<https://docs.databricks.com/spark/latest/sparkr/index.html>
   upvoted 1 times

✉️  **cusman** 1 year, 4 months ago
DP-203
   upvoted 2 times

✉️  **o2091** 1 year, 9 months ago
looks correct
   upvoted 2 times

D 的评论，ET 给的 N, NY 都有人选

Azure Synapse Analytics does not support R script.

Correct 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.

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

<https://docs.microsoft.com/en-us/sql/machine-learning/tutorials/quickstart-r-create-script?view=sql-server-ver15>

✉ **U_C** 4 months, 1 week ago
B. No.

The solution described does not fully meet the stated goal. While it includes a pipeline to copy data to a staging table in the data warehouse, it does not account for the incremental nature of the data. Additionally, using a stored procedure to execute the R script may not be the most efficient approach for transforming the data.

👍👎🚩 upvoted 1 times

✉ **KIET2131** 5 months, 2 weeks ago
Selected Answer: A
A. Yes.

This solution meets the goal of ingesting incremental data from the staging zone, transforming the data using an R script, and inserting the transformed data into a data warehouse in Azure Synapse Analytics. By using Azure Data Factory to copy the data to a staging table in the data warehouse, and then using a stored procedure to execute the R script, you can ensure that the data is transformed correctly before it is inserted into the data warehouse. Additionally, using a stored procedure can help simplify the pipeline and reduce maintenance efforts, since the R script can be updated in a single location

👍👎🚩 upvoted 1 times

✉ **Ciupaz** 9 months, 1 week ago
This question is not for DBA (DP-300 exam).
👍👎🚩 upvoted 1 times

✉ **Icyb3r** 11 months, 2 weeks ago
Selected Answer: B
Insert to DWH comes after the Execution of R script, so the correct Answer should be (B. No)

👍👎🚩 upvoted 3 times

✉ **cusman** 1 year, 4 months ago
DP-203
👍👎🚩 upvoted 4 times

✉ **CaptainJameson** 1 year, 6 months ago
Selected Answer: A
Seems right
<https://docs.microsoft.com/en-us/sql/machine-learning/tutorials/quickstart-r-create-script?view=sql-server-ver15>
By default, sp_execute_external_script accepts a single dataset as input, which typically you supply in the form of a valid SQL query. It then returns a single R data frame as output.

👍👎🚩 upvoted 1 times

✉ **o2091** 1 year, 9 months ago
looks correct
👍👎🚩 upvoted 1 times

Common Question 1

Question 1

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,
    [Postalcode] [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] [smallint] NULL
- B. [ManagerEmployeeKey] [smallint] NULL
- C. [ManagerEmployeeKey] [int] NULL
- D. [ManagerName] [varchar](200) NULL

Answer: C

We need an extra column to identify the Manager. Use the data type as the EmployeeKey column, an int column.

<https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular>

Question 2

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 EmployeeName = 'Alice';
```

What will be returned by the query?

- A. 24
- B. an error
- C. a null value

ET User: B

 **dmitriypo**  10 months, 2 weeks ago

Answer is B, but not because of the lowercase. The case has nothing to do with the error.

If you look attentively, you will notice that we create table mytestdb.myParquetTable, but the select statement contains the reference to table mytestdb.dbo.myParquetTable (!! - dbo).

Here is the error message I got:

Error: spark_catalog requires a single-part namespace, but got [mytestdb, dbo].

   upvoted 37 times

ET Answer: A

✉️  **gerrie1979**  10 months, 1 week ago

I did a test, waited for one minute and tried the query in a serverless sql pool and received 24 as the result, so I don't understand that B has been voted so much because the answer is A) 24 without a doubt

   upvoted 32 times

Bing: B

The query will return **an error**. This is because the table `myParquetTable` was created in the `mytestdb` database, but the query is trying to access it from the `dbo` schema of the `mytestdb` database. The correct schema name should be used in the query to access the table. So, the correct answer is **B**. Is there anything else you would like to know? 😊

Question 3

DRAG DROP -

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 index

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.

Select and Place:

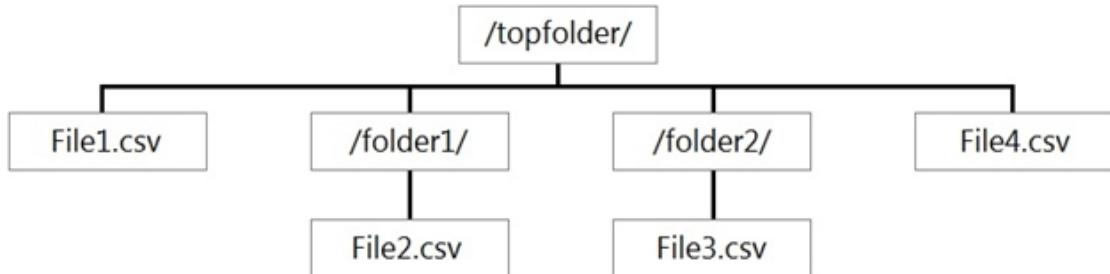
Actions	Answer Area
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	
Truncate the partition containing the stale data.	
Drop the SalesFact_Work table.	
Create an empty table named SalesFact_Work that has the same schema as SalesFact.	
Execute a <code>DELETE</code> statement 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	Answer Area
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Truncate the partition containing the stale data.	Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Drop the SalesFact_Work table.	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> .	

Question 4

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

Chillem1900 2 years, 4 months ago

I believe the answer should be B.

In case of a serverless pool a wildcard should be added to the location.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table>

upvoted 109 times

captainpike 1 year, 10 months ago

I tested and prove you right, the answer is B. Remind the question is referring to serverless SQL and not dedicated SQL pool. "Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Hadoop tables will return all files within any subfolder."

upvoted 32 times

Question 5

HOTSPOT -

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.

Hot Area:

Answer Area

Report1:	<input type="checkbox"/> Avro
	<input type="checkbox"/> CSV
	<input type="checkbox"/> Parquet
	<input type="checkbox"/> TSV
Report2:	<input type="checkbox"/> Avro
	<input type="checkbox"/> CSV
	<input type="checkbox"/> Parquet
	<input type="checkbox"/> TSV

ET User:

- 1: Parquet - column-oriented binary file format
- 2: AVRO - Row based format, and has logical type timestamp

  **alain2**  2 years, 3 months ago

- 1: Parquet - column-oriented binary file format
 - 2: AVRO - Row based format, and has logical type timestamp
- <https://youtu.be/UrWthx8T3UY>

   upvoted 176 times

ET Answer:

Answer Area

Report1:	<input checked="" type="checkbox"/> Avro
	<input checked="" type="checkbox"/> CSV
	<input type="checkbox"/> Parquet
	<input type="checkbox"/> TSV
Report2:	<input checked="" type="checkbox"/> Avro
	<input type="checkbox"/> CSV
	<input type="checkbox"/> Parquet
	<input type="checkbox"/> TSV

Question 6

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

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}/

Question 7

HOTSPOT -

You need to output files from Azure Data Factory.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

Answer:

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

Question 8

HOTSPOT -

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.

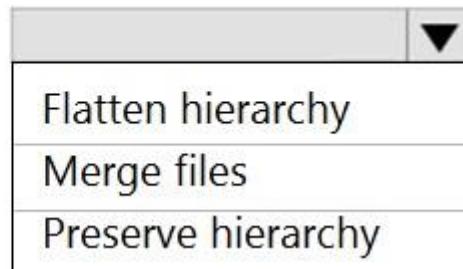
How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Copy behavior:



Sink file type:



ET User:

1. Merge Files
2. Parquet

  alain2  2 years, 3 months ago

1. Merge Files

2. Parquet

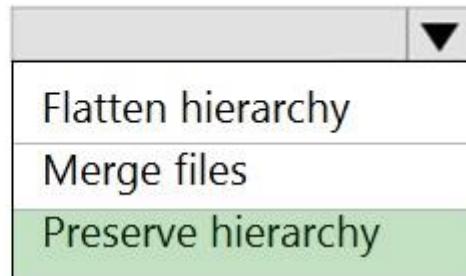
<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance>

   upvoted 146 times

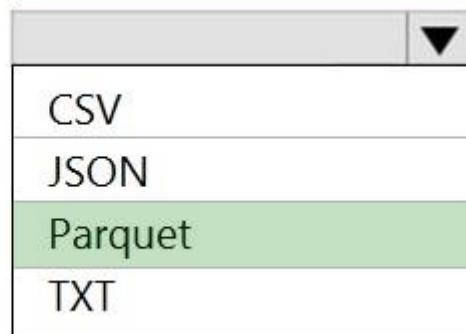
ET Answer:

Answer Area

Copy behavior:



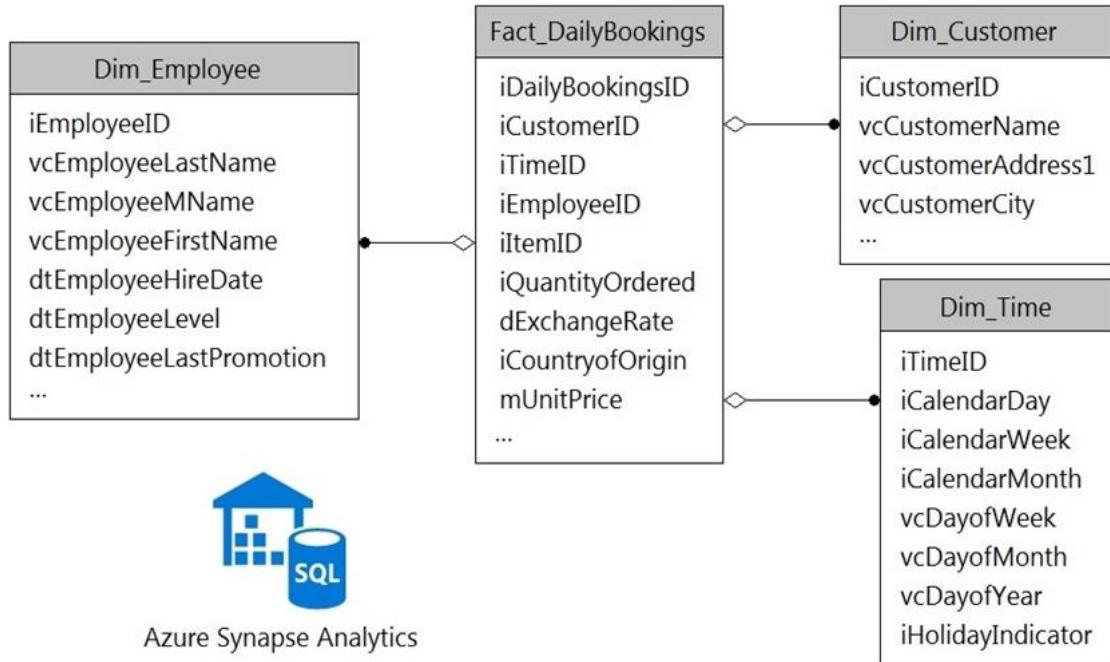
Sink file type:



Question 9

HOTSPOT -

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. The dimension tables will be relatively static with very few data inserts and updates.

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.

Hot Area:

Answer Area

Dim_Customer:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				
Dim_Employee:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				
Dim_Time:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				
Fact_DailyBookings:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				

Answer:

Answer Area

Dim_Customer:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				
Dim_Employee:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				
Dim_Time:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				
Fact_DailyBookings:	<table border="1"><tr><td>Hash distributed</td></tr><tr><td>Round-robin</td></tr><tr><td>Replicated</td></tr></table>	Hash distributed	Round-robin	Replicated
Hash distributed				
Round-robin				
Replicated				

Question 10

HOTSPOT -

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

Hot Area:

Answer Area

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:

Answer Area

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.

Question 11

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.

Select and Place:

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	Answer Area
CLUSTERED INDEX	CREATE TABLE table1
COLLATE	(
DISTRIBUTION	ID INTEGER,
PARTITION	col1 VARCHAR(10),
PARTITION FUNCTION	col2 VARCHAR(10)
PARTITION SCHEME) WITH

(

DISTRIBUTION	= HASH(ID),
PARTITION	(ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))

);

 **Sasha_in_San_Francisco** Highly Voted  1 year, 10 months ago

Correct answer by how to remember? Distribution option before the Partition option because... 'D' comes before 'P' or because the system needs to know the algorithm (hash, round-robin, replicate) before it can start to Partition or segment the data. (seem reasonable?)

   upvoted 84 times

Question 12

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

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.

<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types>

Question 13

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 shared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.

Answer: ABF

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 -

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity>

Question 14

HOTSPOT -

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.

Hot Area:

Answer Area

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:

Answer Area

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

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>

Question 15

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;
```

Answer: C

Incorrect Answers:

A, D: Only these Data Definition Language (DDL) statements are allowed on external tables:

- ☞ CREATE TABLE and DROP TABLE
- ☞ CREATE STATISTICS and DROP STATISTICS
- ☞ CREATE VIEW and DROP VIEW

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

Question 16

HOTSPOT -

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.

Hot Area:

Answer Area

Source dataset type:

Binary
Parquet
Delimited text

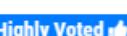
Copy activity copy behavior:

FlattenHierarchy
MergeFiles
PreserveHierarchy

ET User:

Binary / Parquet 争议

PreserveHierarchy

  **EddyRoboto**  2 years ago

This could be binary as source and sink, since there are no transformations on files. I tend to believe that would be binary the correct answer.

   upvoted 70 times

  **GameLift** 1 year, 10 months ago

But the doc says "When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset." So I guess it's parquet then?

   upvoted 8 times

  **captainpike** 1 year, 10 months ago

This note is referring to the fact that, in the template, you have to specify "BinarySink" as the type for the target Sink; and that exactly what the Copy data tool does. (you can check this by editing the created copy pipeline and see the code). Choosing Binary and PreserveHierarchy copy all file as they are perfectly.

   upvoted 4 times

✉  **AbhiGola**  1 year, 12 months ago

Answer seems correct as data is store is parquet already and requirement is to do no transformation so answer is right

   upvoted 57 times

✉  **NintyFour** 1 year, 3 months ago

As question has mentioned, Minimize time required to perform the copy activity. And binary is faster than Parquet. Hence, Binary is answer

   upvoted 6 times

✉  **anto69** 8 months, 3 weeks ago

No: req1 "no transformation", req2 "Minimize time required to perform the copy activity". Both must be met hence it's Parquet cause it's the second fastest choice and it requires no transformations.

   upvoted 6 times

✉  **mhi** 3 months, 4 weeks ago

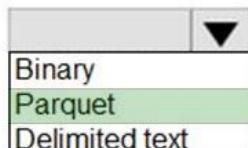
when doing a binary copy, you're not doing any transformation!

   upvoted 2 times

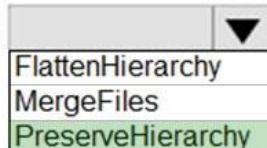
ET Answer:

Answer Area

Source dataset type:



Copy activity copy behavior:



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.

Incorrect Answers:

- ☞ FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.
- ☞ MergeFiles: Merges all files from the source folder to one file. If the file name is specified, the merged file name is the specified name. Otherwise, it's an autogenerated file name.

<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>

Question 17

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

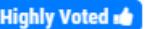
You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.

Which type of data redundancy should you use?

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

ET User: 统计的不准，推荐 B

A (73%) **B (27%)**

✉  **meetj**  2 years ago

B is right

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.

   upvoted 97 times

✉  **dev2dev** 1 year, 7 months ago

A looks correct answer. RA-GRS is always available because its auto failover. Since this is not asked in the question but more importantly the question is about reducing cost which GRS.

   upvoted 31 times

✉  **BK10** 1 year, 6 months ago

It should be A because of two reasons:

1. Minimize cost
2. When primary is unavailable.

Hence No need for RA_GRS

ET Answer: B

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.

Incorrect Answers:

A: While Geo-redundant storage (GRS) is cheaper than Read-Access Geo-Redundant Storage (RA-GRS), GRS does NOT initiate automatic failover.

C, D: Locally redundant storage (LRS) and Zone-redundant storage (ZRS) provides redundancy within a single region.

Reference:

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

Bing: B

The correct answer is **B. read-access geo-redundant storage (RA-GRS)**. This type of data redundancy ensures that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region, while minimizing costs. RA-GRS replicates your data to another physical location in the secondary region, but also provides read access to the data in the secondary location. This means that you can continue to read data even if there is an outage in the primary region. Is there anything else you would like to know? 😊

Question 18

You plan to implement an Azure Data Lake Gen 2 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. geo-zone-redundant storage (GZRS)
- C. locally-redundant storage (LRS)
- D. zone-redundant storage (ZRS)

Answer: D

Zone-redundant storage (ZRS) copies your data synchronously across three Azure availability zones in the primary region.

Incorrect Answers:

C: 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, but is not recommended for applications requiring high availability or durability

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

Question 19

HOTSPOT -

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.

Hot Area:

Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

ET User:

Round-Robin

Heap

None

 **A1000** Highly Voted 2 years ago

Round-Robin

Heap

None

   upvoted 335 times

 **laszek** Highly Voted 2 years ago

Round-robin - this is the simplest distribution model, not great for querying but fast to process

Heap - no brainer when creating staging tables

No partitions - this is a staging table, why add effort to partition, when truncated daily?

   upvoted 56 times

ET Answer:

Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Question 20

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, IsOrderFinalized, COUNT(*)  
FROM FactPurchase -  
WHERE DateKey >= 20210101 -  
AND DateKey <= 20210131 -  
GROUP By SupplierKey, StockItemKey, IsOrderFinalized  
Which table distribution will minimize query times?  
A. replicated  
B. hash-distributed on PurchaseKey  
C. round-robin  
D. hash-distributed on IsOrderFinalized
```

Answer: B

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

To balance the parallel processing, select a distribution column that:

- ☞ Has many unique values. The column can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions, some distributions can have > 1 unique values while others may end with zero values.
- ☞ Does not have NULLs, or has only a few NULLs.
- ☞ Is not a date column.

Incorrect Answers:

C: 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-distribute>

Question 21

HOTSPOT -

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.

Hot Area:

Answer Area

EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

Answer:

Answer Area

EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

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>

Question 22

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. an ordered clustered columnstore index
- B. a materialized view
- C. result set caching
- D. a replicated table

Answer: B

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.

Incorrect Answers:

C: One daily execution does not make use of result cache caching.

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.

<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-set-caching>

Question 23

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

You plan to create a database named DB1 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 pool.

Which format should you use for the tables in DB1?

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

Answer: D

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.

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

Question 24

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

Answer: D

The following tables summarize the key differences in capabilities for stream processing technologies in Azure.

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Programmability	Stream analytics query language, JavaScript	C#/F# ↗, Java, Python, Scala	C#/F# ↗, Java, Python, R, Scala	C#, Java

General capabilities -

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure	HDInsight with Storm
Inputs	Azure Event Hubs, IoT Hubs, Hub, Kafka, HDFS, Storage Blob Storage, Blobs, Azure Data Lake Store, Data Lake Store	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store
Sinks	Azure Data Lake Store, Azure SQL Database, Storage Blobs, Event	HDFS, Kafka, Storage Blobs, Azure Data Lake	HDFS, Kafka, Storage Blobs, Azure Data Lake Store, Cosmos DB	Event Hubs, Service Bus, Kafka, Azure Data Lake Store, Cosmos DB

Integration capabilities -

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

Question 25

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. Convert the files to JSON
- B. Convert the files to Avro
- C. Compress the files
- D. Merge the files

ET User: 点赞 B 的没怎么投票，选 D 的投了好多票，还有几个选 C 的

✉️  **VeroDon**  1 year, 8 months ago

You can not merge the files if u don't know how many files exist in ADLS2. In this case, you could easily create a file larger than 100 GB in size and decrease performance. so B is the correct answer. Convert to AVRO

   upvoted 48 times

✉️  **auwia** 2 months, 2 weeks ago

Option B: Convert the files to Avro (WRONG FOR ME)

While converting the files to Avro is a valid option for optimizing data storage and processing, it may not be the most suitable choice in this specific scenario. Avro is a binary serialization format that is efficient for compact storage and fast data processing. It provides schema evolution support and is widely used in big data processing frameworks like Apache Hadoop and Apache Spark.

However, in the given scenario, the files are already in CSV format. Converting them to Avro would require additional processing and potentially introduce complexity. Avro is better suited for scenarios where data is generated or consumed by systems that natively support Avro or for cases where schema evolution is a critical requirement.

On the other hand, merging the files (Option D) is a more straightforward and common approach to optimize batch processing. It helps reduce the overhead associated with managing a large number of small files, improves data scanning efficiency, and enhances overall processing performance. Merging files is a recommended practice to achieve better performance and cost efficiency in scenarios where file sizes vary.

   upvoted 4 times

✉️ Canary_2021 **Highly Voted** 1 year, 8 months ago

Selected Answer: D

If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256 MB to 100 GB in size).
<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#optimize-for-data-ingest>

👍 ↪️ 🏴 upvoted 20 times

✉️ kkk5566 **Most Recent** 13 hours, 8 minutes ago

Selected Answer: D

Should be D

👍 ↪️ 🏴 upvoted 1 times

✉️ sidh_r 1 week, 6 days ago

Selected Answer: B

Merging might not be the best one since the size of files are varying, we wouldn't be able to come with a simple approach to merge. If we have to merge then again we would cause some overhead, hence converting to avro seems to be a best choice

👍 ↪️ 🏴 upvoted 1 times

✉️ rajaanto 5 months, 1 week ago

As per ChatGPT "By converting CSV files to the Avro format, you can take advantage of these benefits and optimize your data for batch processing in Azure Data Lake Storage Gen2."

👍 ↪️ 🏴 upvoted 1 times

✉️ esaade 5 months, 3 weeks ago

Selected Answer: C

The best option for optimizing CSV files stored in an Azure Data Lake Storage Gen2 container for batch processing would be to compress the files, option C.

Batch processing typically involves processing large amounts of data at once, and reducing the size of the data can significantly improve processing times and reduce costs.

Compressing the files can help reduce their size and optimize them for batch processing. The compression method used will depend on the specific requirements of the batch processing workload, but options such as gzip, bzip2, or Snappy can be effective for compressing CSV files.

👍 ↪️ 🏴 upvoted 1 times

ET Answer: B

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.

<https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/>

Bing: B

To optimize the files stored in an Azure Data Lake Storage Gen2 container for batch processing, you should **convert the files to a binary format** that is more performant for querying because it stores data in a columnar format, and often provides indexes and inline statistics about the data ¹. One example of such a binary format is **Avro**. Therefore, the correct answer is **B. Convert the files to Avro**. Converting the files to JSON (option A) or compressing the files (option C) would not necessarily optimize them for batch processing. Merging the files (option D) could potentially improve performance, but it would not be as effective as converting the files to a binary format like Avro.

Question 26

HOTSPOT -

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": {  
                            "daysAfterCreationGreaterThanOrEqual": 60  
                        }  
                    },  
                    "baseBlob": {  
                        "tierToCool": {  
                            "daysAfterModificationGreaterThanOrEqual":  
                                30  
                        }  
                    },  
                    "blobType": {  
                        "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.

Hot Area:

Answer Area

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:

Answer Area

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

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.managementpolicybaseblob.tiertocool>

Question 27

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: D

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.

Question 28

HOTSPOT -

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.

Hot Area:

Answer Area

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:

Answer Area

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

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.

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

✉️  **gf2tw**  1 year, 8 months ago

"Data must remain in the Archive tier for at least 180 days or be subject to an early deletion charge. For example, if a blob is moved to the Archive tier and then deleted or moved to the Hot tier after 45 days, you'll be charged an early deletion fee equivalent to 135 (180 minus 45) days of storing that blob in the Archive tier." <- from the sourced link.

This explains why we have to use two different access tiers rather than both as archive.

   upvoted 77 times

✉️  **RoyP654** 2 months, 3 weeks ago

"You do not expect that the logs will be accessed during the retention periods." - including deletes, i suppose. you just let lifecycle management rule do the deletes after the retention period ... Archiving for cost-reduction?

   upvoted 2 times

✉️  **Anshuman_B** 7 months ago

Thanks for sharing this info.
   upvoted 1 times

✉️  **dsp17** 1 year, 1 month ago

Thanks a ton for explaining.
   upvoted 2 times

✉️  **ANath**  1 year, 8 months ago

The answers are correct.

Data must remain in the Archive tier for at least 180 days or be subject to an early deletion charge. For example, if a blob is moved to the Archive tier and then deleted or moved to the Hot tier after 45 days, you'll be charged an early deletion fee equivalent to 135 (180 minus 45) days of storing that blob in the Archive tier.

A blob in the Cool tier in a general-purpose v2 accounts is subject to an early deletion penalty if it is deleted or moved to a different tier before 30 days has elapsed. This charge is prorated. For example, if a blob is moved to the Cool tier and then deleted after 21 days, you'll be charged an early deletion fee equivalent to 9 (30 minus 21) days of storing that blob in the Cool tier.

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

   upvoted 16 times

Question 29

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 files can be queried quickly and that the data type information is retained.

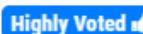
What should you recommend?

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

Answer: B

Need Parquet to support both Databricks and PolyBase.

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

✉️  **demirsamuel**  1 year, 3 months ago

Selected Answer: B

Avro schema definitions are JSON records. Polybase does not support JSON so why supporting Avro then. A CSV does not contain the schema as it is everything marked as string. so only parquet is left to choose.

   upvoted 25 times

✉️  **hrastogi7**  1 year, 8 months ago

Parquet can be quickly retrieved and maintain metadata in itself. Hence Parquet is correct answer.

   upvoted 22 times

Question 30

You have an Azure Synapse Analytics dedicated SQL pool named 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. Insert the data from stg.Sales into dbo.Sales.
- B. Switch the first partition from dbo.Sales to stg.Sales.
- C. Switch the first partition from stg.Sales to dbo.Sales.
- D. Update dbo.Sales from stg.Sales.

ET User: C

 **Aslam208** Highly Voted 1 year, 8 months ago

Selected Answer: C

The correct answer is C

   upvoted 62 times

 **Nifl91** Highly Voted 1 year, 8 months ago

this must be C. since the need is to overwrite dbo.Sales with the content of stg.Sales.
SWITCH source TO target

   upvoted 32 times

ET Answer: B

Question 31

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 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. effective start date
- C. business key
- D. last modified date
- E. effective end date
- F. foreign key

ET User: ABE

 **ItHYMeRlsh**  1 year, 8 months ago

Selected Answer: ABE

The answer is ABE. A type 2 SCD requires a surrogate key to uniquely identify each record when versioning.

See <https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types> under SCD Type 2 " the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member."

A business key is already part of this table - SupplierSystemID. The column is derived from the source data.

   upvoted 101 times

ET Answer: BCE

C: The Slowly Changing Dimension transformation requires at least one business key column.

BE: Historical attribute changes create new records instead of updating existing ones. The only change that is permitted in an existing record is an update to a column that indicates whether the record is current or expired. This kind of change is equivalent to a Type 2 change. The Slowly Changing Dimension transformation directs these rows to two outputs: Historical Attribute Inserts Output and New Output.

Reference:

<https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension-transformation>

Question 32

HOTSPOT -

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.

Hot Area:

Answer Area

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:

Answer Area

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

Box 1: Denormalize to a second normal form

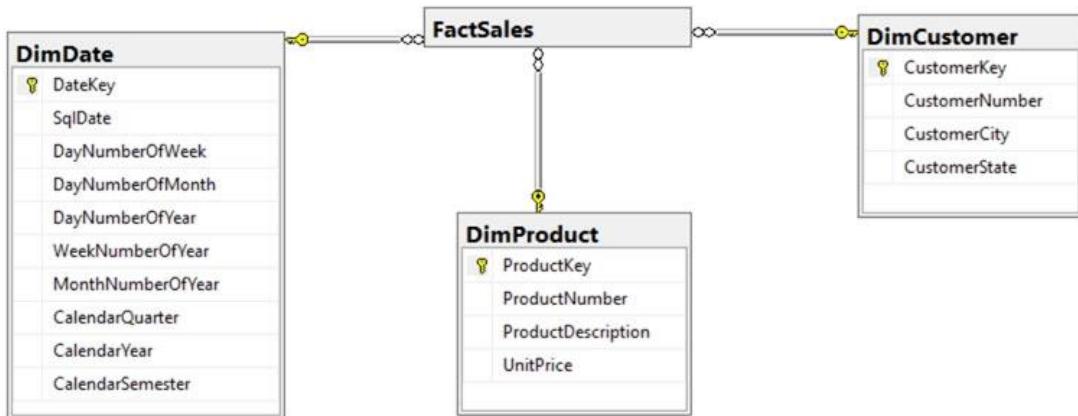
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation.

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 flat 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:



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.

<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>

Question 33

HOTSPOT -

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 Store ID. 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.

Hot Area:

Answer Area

df.write

.bucketBy	▼
.partitionBy	▼
.range	▼
.sortBy	▼

.mode("append")

.csv("/Purchases")	▼
.json("/Purchases")	▼
.parquet("/Purchases")	▼
.saveAsTable("/Purchases")	▼

("*")	▼
("StoreID", "Hour")	▼
("StoreID", "Year", "Month", "Day", "Hour")	▼

Answer:

Answer Area

df.write

.bucketBy	▼
.partitionBy	▼
.range	▼
.sortBy	▼

.mode("append")

.csv("/Purchases")	▼
.json("/Purchases")	▼
.parquet("/Purchases")	▼
.saveAsTable("/Purchases")	▼

("*")	▼
("StoreID", "Hour")	▼
("StoreID", "Year", "Month", "Day", "Hour")	▼

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-with-no-new-data>

Question 34

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 for the clustered columnstore index?

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

Answer: A

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

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.

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

Question 35

HOTSPOT -

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.

Hot Area:

Answer Area

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

Type 0
Type 1
Type 2

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

a surrogate key
a business key
an audit column

ET User:

Type 1

a surrogate key

✉  **nkav**  2 years, 4 months ago

product key is a surrogate key as it is an identity column

   upvoted 185 times

✉  **sagga**  2 years, 3 months ago

Type2 because there are start and end columns and ProductKey is a surrogate key.
ProductNumber seems a business key.

   upvoted 44 times

✉  **DrC** 2 years, 3 months ago

The start and end columns are for when to when the product was being sold, not for metadata purposes. That makes it:

Type 1 – No History

Update record directly, there is no record of historical values, only current state

   upvoted 97 times

ET Answer:

Answer Area

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]**.

Question 36

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. replicated
- B. hash-distributed on PurchaseKey
- C. round-robin
- D. hash-distributed on DateKey

Answer: B / D 争议。

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.

Incorrect:

Not D: Do not use a date column. . All data for the same date lands in the same distribution.

If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

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

✉️  **AugustineUba**  2 years ago

From the documentation the answer is clear enough. B is the right answer.

When choosing a distribution column, select a distribution column that: "Is not a date column. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work."

   upvoted 49 times

✉️  **YipingRuan** 1 year, 10 months ago

To minimize data movement, select a distribution column that:

Is used in JOIN, GROUP BY, DISTINCT, OVER, and HAVING clauses.

"PurchaseKey" is not used in the group by

   upvoted 8 times

✉️  **cem_kalender** 10 months ago

A distribution column should have high cardinality to ensure even distribution over nodes.

   upvoted 1 times

✉️  **YipingRuan** 1 year, 10 months ago

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 no 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

   upvoted 7 times

✉️  **waterbender19**  2 years ago

I think the answer should be D for that specific query. If you look at the datatypes, DateKey is an INT datatype not a DATE datatype.

   upvoted 17 times

✉️  **AnandEMani** 1 year, 12 months ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute> this link says date filed , NOT a date Data type. B is correct

   upvoted 4 times

✉️  **waterbender19** 2 years ago

and the statement that Fact table will be added 1 million rows daily means that each datekey value has an equal amount of rows associated with that value.

   upvoted 5 times

Question 37

You are implementing a batch dataset in the Parquet format.

Data files 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. Use Snappy compression for the files.
- B. Use OPENROWSET to query the Parquet files.
- C. Create an external table that contains a subset of columns from the Parquet files.
- D. Store all data as string in the Parquet files.

ET User:



✉️ **m2shines** 1 year, 8 months ago

Answer should be A, because this talks about minimizing storage costs, not querying costs

upvoted 60 times

✉️ **assU2** 1 year, 7 months ago

Isn't snappy a default compressionCodec for parquet in azure?

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

upvoted 21 times

✉️ **Aslam208** 1 year, 8 months ago

C is the correct answer, as an external table with a subset of columns with parquet files would be cost-effective.

upvoted 20 times

✉️ **RehanRajput** 1 year, 3 months ago

This is not correct.

1. External tables are not saved in the database. (This is why they're external)
2. You're assuming that the SQL Serverless pools have a local storage. They don't --> <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-serverless-sql-pool>

upvoted 4 times

ET Answer: C

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.

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

Bing: A

To minimize storage costs for a batch dataset in the Parquet format, you should use a **compression method** that provides a good balance between compression ratio and speed for reading and writing ¹. One such compression method is **Snappy**, which is known for its fast compression and decompression speeds, as well as its reasonable compression ratio. Therefore, the correct answer is **A. Use Snappy compression for the files**. Using OPENROWSET to query the Parquet files (option B) or creating an external table that contains a subset of columns from the Parquet files (option C) would not necessarily minimize storage costs. Storing all data as string in the Parquet files (option D) would likely increase storage costs, as it would prevent the use of more efficient data types and encodings.

Question 38

DRAG DROP -

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.

Select and Place:

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 data source
	Create an external file format object
Create a query that uses Create Table as Select	
Create a table	
	Create an external table

Step 1: Create an external data source

You can create external tables in Synapse SQL pools via the following steps:

1. CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.
2. CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.
3. 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

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

Question 39

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

The data mart will contain employee 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 presents part of the solution.

NOTE: Each correct selection is worth one point.

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

Answer: CE

C: Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table.

Instead, the fact table and the dimension table can share a customer ID. A query can join the two tables to associate a customer's profile and transactions.

E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool. For example, a retail business generates sales transactions every day, and then loads the data into a dedicated SQL pool fact table for analysis.

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

Question 40

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 you use?

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

Answer: C

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.

Incorrect Answers:

B: A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten.

D: 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.

<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types>

Question 41

DRAG DROP -

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 a dedicated SQL pool 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

Select and Place:

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: 争议

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
	Create an external file format and set the First_Row option
	Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

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.

<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>

✉️  **sunil_smile**  11 months, 2 weeks ago

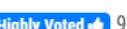
- 1) create database scoped credentials
- 2) create external source
- 3) create file format
- 4) create external table (it not supports CTAS)

   upvoted 26 times

✉️  **auwia** 2 months, 1 week ago

It supports, it is a dedicated SQL pool (means not severless), reading the question:
You are implementing a pattern that batch loads the files daily into a dedicated SQL pool
in Azure Synapse Analytics by using PolyBase.
--> provided answers are correct in my opinion.

   upvoted 1 times

✉️  **OldSchool**  9 months, 2 weeks ago

Because it's saying "You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2" and "You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase" assumption is that we already have database credentials, so the answer is:

- 1) create external source
- 2) create file format
- 3) create external table

   upvoted 12 times

✉️  **Rob77** 3 months, 1 week ago

No, CETAS is not used for loading Azure Synapse Analytics. It's used to export data from and not to!

   upvoted 1 times

✉️  **kkk5566**  16 hours, 8 minutes ago

source ,format ,external

   upvoted 1 times

✉️  **eladioyovera** 1 month, 1 week ago

The answer is correct,
- Create database scoped credential: "This step is required only for Kerberos-secured Hadoop clusters."
In this case, the previous step does not apply.

   upvoted 1 times

✉️  **vctrhugo** 2 months, 1 week ago

1. Create database scoped credential
2. Create external data source
3. Create external file format

<https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16#create-external-tables-for-azure-blob-storage>

   upvoted 4 times

Question 42

HOTSPOT -

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.

Hot Area:

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]          RANGE RIGHT FOR VALUES
    [TransactionDateID], [TransactionTypeID]
    HASH([TransactionTypeID])
    ROUND_ROBIN
)
(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] RANGE RIGHT FOR VALUES
        [TransactionDateID], [TransactionTypeID]
        HASH([TransactionTypeID])
        ROUND ROBIN
    )
    (20200101,20200201,20200301,20200401,20200501,20200601)
)
```

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>

Question 43

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 beginning with "tripdata_2020".
- C. All CSV files that have file names that contain "tripdata_2020".
- D. Only CSV that have file names that beginning with "tripdata_2020".

Answer: D

✉️  **panda_azzurro** 7 months, 1 week ago

Selected Answer: D

Sorry but I don't understand.
File or Directory that start with "tripdata_2020" can selected.
/tripdata_2020/a.csv
/tripdata_2020_a_b.csv
/tripdata_2020/2020/1/1/myfile.csv
So question is very not clear.
D question is partially correct

   upvoted 2 times

Question 44

DRAG DROP -

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

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

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

owner	age	dog
Keith	30	Fido
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 spit bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
alias	dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
array_union	persons = source_df. <input type="button" value="Value"/> <input type="button" value="Value"/> ("persons").alias("persons")
createDataFrame	persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), ("persons-dogs"). <input type="button" value="Value"/> ("dog"))
explode	display(persons_dogs)
select	
translate	

Answer:

Values	Answer Area
array_union	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")
createDataFrame	persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode [] alias [] ("dog")) display(persons_dogs)
[]	
translate	

✉  **kb8bo**  1 year, 2 months ago

The final line with the blank looks incorrect... surely it should be:
explode("persons.dogs").alias("dog"))

(Assuming this, the answer is correct, otherwise I don't think it makes any sense).

   upvoted 19 times

✉  **esaade**  5 months, 3 weeks ago

```
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(col("persons.dog")).alias("dog_name"))  
persons_dogs.display()
```

   upvoted 7 times

✉  **kkk5566**  16 hours, 9 minutes ago

syntax is correct

   upvoted 1 times

✉  **urassi** 6 months ago

ah "persons".alias("persons") what a fun and useful and nice alias

   upvoted 7 times

✉  **Deeksha1234** 1 year, 1 month ago

Correct

   upvoted 3 times

✉  **Dicer** 1 year, 1 month ago

Correct, but last .alias("dog") is quite unnecessary because the column name is already 'dog'. I guess that is for safety measurement.

   upvoted 4 times

✉  **Anton2020** 6 months, 1 week ago

The column name in the json is dogs, not dog

   upvoted 2 times

✉  **galacaw** 1 year, 4 months ago

Correct

   upvoted 3 times

Question 45

HOTSPOT -

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.

Hot Area:

Answer Area

First week:

Archive
Cool
Hot

After one month:

Archive
Cool
Hot

After one year:

Archive
Cool
Hot

Answer:

Answer Area

First week:

Archive
Cool
Hot

After one month:

Archive
Cool
Hot

After one year:

Archive
Cool
Hot

Box 1: Hot -

Hot tier - An online tier optimized for storing data that is accessed or modified frequently.

The Hot tier has the highest storage costs, but the lowest access costs.

Box 2: Cool -

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.

Box 3: Cool -

Not 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.

	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 object size	N/A	N/A	N/A	N/A
Minimum storage duration	N/A	N/A	30 days ¹	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours ²

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

<https://www.altaro.com/hyper-v/azure-archive-storage/>

Question 46

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 Conditional Split transformation in an Azure Synapse data flow.
- B. Use a Get Metadata activity in Azure Data Factory.
- C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.

ET User: D

✉️  **galacaw** Highly Voted  1 year, 4 months ago

Should be D, it's about Apache Spark pool, not serverless SQL pool.

   upvoted 32 times

✉️  **kkk5566** Most Recent  16 hours, 12 minutes ago

Selected Answer: D

Should be D

   upvoted 1 times

✉️  **vctrhugo** 2 months, 1 week ago

Selected Answer: D

PySpark provides a powerful and flexible programming interface for processing and loading data in Azure Synapse Analytics Apache Spark pools. With PySpark, you can leverage its JSON reader capabilities to infer the schema and maintain the source data types during the loading process.

   upvoted 2 times

ET Answer: C

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>

Bing: D

To load JSON files from an Azure Data Lake Storage Gen2 container into the tables in an Azure Synapse Analytics Apache Spark pool named Pool1, while maintaining the source data types, you should **load the data by using PySpark**. PySpark is the Python API for Apache Spark and provides a high-level API for distributed data processing. It can be used to read JSON files from Azure Data Lake Storage Gen2 and load them into tables in an Azure Synapse Analytics Apache Spark pool, while preserving the source data types ¹. Therefore, the correct answer is **D. Load the data by using PySpark**. Using a Conditional Split transformation in an Azure Synapse data flow (option A), using a Get Metadata activity in Azure Data Factory (option B), or loading the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool (option C) would not necessarily maintain the source data types.

Question 47

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

Answer: D

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.

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

  **hkay**  1 year, 3 months ago

Answer D is correct. Azure Databricks pools reduce cluster start and auto-scaling times by maintaining a set of idle, ready-to-use instances.

   upvoted 13 times

  **kim32**  3 months, 4 weeks ago

D is accurate answer and this link show this info explicitly

<https://learn.microsoft.com/en-us/azure/databricks/clusters/cluster-config-best-practices>

   upvoted 9 times

Bing:

To reduce the time it takes for an all-purpose cluster named cluster1 in an Azure Databricks workspace named workspace1 to start and scale up, while minimizing costs, you should **create a pool in workspace1**. A pool can reduce cluster start and scale-up times by maintaining a set of available, ready-to-use instances ¹. Therefore, the correct answer is **D. Create a pool in workspace1**. Configuring a global init script for workspace1 (option A) or creating a cluster policy in workspace1 (option B) would not necessarily reduce the time it takes for cluster1 to start and scale up. Upgrading workspace1 to the Premium pricing tier (option C) would increase costs, rather than minimizing them.

Question 48

HOTSPOT -

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 X

products

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

ⓘ 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.

Hot Area:

Answer Area

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:

Answer Area

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

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>

Question 49

HOTSPOT -

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.

Hot Area:

Answer Area

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"/>

ET User: YYY

  **objecto**  1 year, 2 months ago

I feel its all YES. Since it does use a UNION and UNION combines. No matter it repartitions the result is the combination of two sources, a UNION of two sources. Am I missing something here?

   upvoted 23 times

 **flaviodiasps**  1 year, 2 months ago

No reason for NO in the first box.

Should be all yes

   upvoted 7 times

 **Saintu**  1 week, 2 days ago

Question 1 is false; the question says the union of streams and not data. The union combines 2 streams which are the same and thus, the output is the same stream.

   upvoted 1 times

 **auwia** 2 months, 1 week ago

False, True, False.

<https://learn.microsoft.com/en-us/azure/stream-analytics/repartition>

The first is False, because this:

"The following example query joins two streams of repartitioned data."

It's extracted from the link above, and it's pointing to our query! Repartitioned and not partitioned.

Second is True, it's explicitly written

The output scheme should match the stream scheme key and count so that each substream can be flushed independently.

Third is False,

"In general, six SUs are needed for each partition."

In the example we have 10 positions for step 1 and 10 for step 2, it should be 120 and not 60.

   upvoted 5 times

ET Answer:

Answer Area

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"/>

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>

Question 50

HOTSPOT -

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 Storege 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.

Hot Area:

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://contosol1.dfs.core.windows.net/applications/year=*/*.parquet',
CROSS APPLY
OPENJSON
OPENROWSET
FORMAT='PARQUET') AS [r]
GO
```

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=*/*.parquet',
```

```
FORMAT='PARQUET') AS [r]
```

```
GROUP BY decennialTime, stateName
```

GO -

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

Question 51

HOTSPOT -

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.

Hot Area:

Answer Area

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

|                     |
|---------------------|
| PUSHDOWN = ON       |
| TYPE = BLOB_STORAGE |
| TYPE = HADOOP       |


```

ET User:

TYPE = HADOOP

dfs

  galacaw  1 year, 4 months ago
1. dfs (for Azure Data Lake Storage Gen2)
   upvoted 47 times

ET Answer:

Answer Area

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

|                     |
|---------------------|
| PUSHDOWN = ON       |
| TYPE = BLOB_STORAGE |
| TYPE = HADOOP       |


```

Question 52

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

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.

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

Question 53

DRAG DROP -

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.

Select and Place:

Values	Answer Area
CustomerKey	
HASH	
ROUND_ROBIN	
REPLICATE	
OrderDateKey	
SalesOrderNumber	
	<pre>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 = <input type="text"/> Value ([ProductKey]) , PARTITION ([<input type="text"/> Value] RANGE RIGHT FOR VALUES (20170101,20180101,20190101,20200101,20210101))</pre>

Answer:

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>

Question 54

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

- ☞ 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.
- ☞ 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.

Hot Area:

Answer Area

Data over five years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Answer:

Answer Area

Data over five years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Question 55

HOTSPOT -

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.

NOTE: Each correct selection is worth one point.

Hot 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

ET User:

ZRS

Failover initiated by Microsoft.

 **aniagnesighile1**  9 months, 4 weeks ago

I am surprised you all missed this requirement 'Ensures that content remains available for writes if a primary data center fails'. RA-GRS and RAGZRS provide read access only after failover. The correct answer is ZRS as stated in the link below "Microsoft recommends using ZRS in the primary region for Azure Data Lake Storage Gen2 workloads."
<https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy?toc=%2Fazur...>

   upvoted 48 times

 **vctrhugo** 2 months, 1 week ago

You can still write data to second region if first one fails. RA only allows you to read data in second region even if the first does not fail.

   upvoted 6 times

✉  **chinomoreno**  12 months ago

Failover initiated by Microsoft.

Customer-managed account failover is not yet supported in accounts that have a hierarchical namespace (Azure Data Lake Storage Gen2). To learn more, see Blob storage features available in Azure Data Lake Storage Gen2.

   upvoted 23 times

✉  **Gg2** 11 months, 4 weeks ago

RA-GZRS

Failover initiated by Microsoft.

   upvoted 19 times

ET 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

<https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/blobs/toc.json>

<https://docs.microsoft.com/en-us/answers/questions/32583/azure-data-lake-gen2-disaster-recoverystorage-acco.html>

Question 56

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) NOT 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.

[EffectiveEndDate] [datetime] NULL,

B.

[CurrentProductCategory] [nvarchar](100) NOT NULL,

C.

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

D.

[EffectiveStartDate] [datetime] NOT NULL,

E.

[OriginalProductCategory] [nvarchar](100) NOT NULL,

Answer: BE

A Type 3 SCD supports storing two versions of a dimension member as separate columns.

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

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

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

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

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

Question 57

DRAG DROP -

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 yearsContains 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.

Select and Place:

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: Replicated
Replicated	Marketing.Web.Sessions: Hash
Round-robin	Staging. Web.Sessions: Round-robin

Box 1: Replicated -

The best table storage option for a small table is to replicate it across all the Compute nodes.

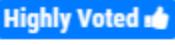
Box 2: Hash -

Hash-distribution improves query performance on large fact tables.

Box 3: Round-robin -

Round-robin distribution is useful for improving loading speed.

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

✉️  **anks84**  12 months ago

Replicated (Because its a Dimension table)

Hash (Fact table with High volume of data)

Round-Robin (Staging table)

   upvoted 24 times

Question 58

HOTSPOT -

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.

Hot Area:

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
```

Box 1: (CLUSTERED COLUMNSTORE INDEX

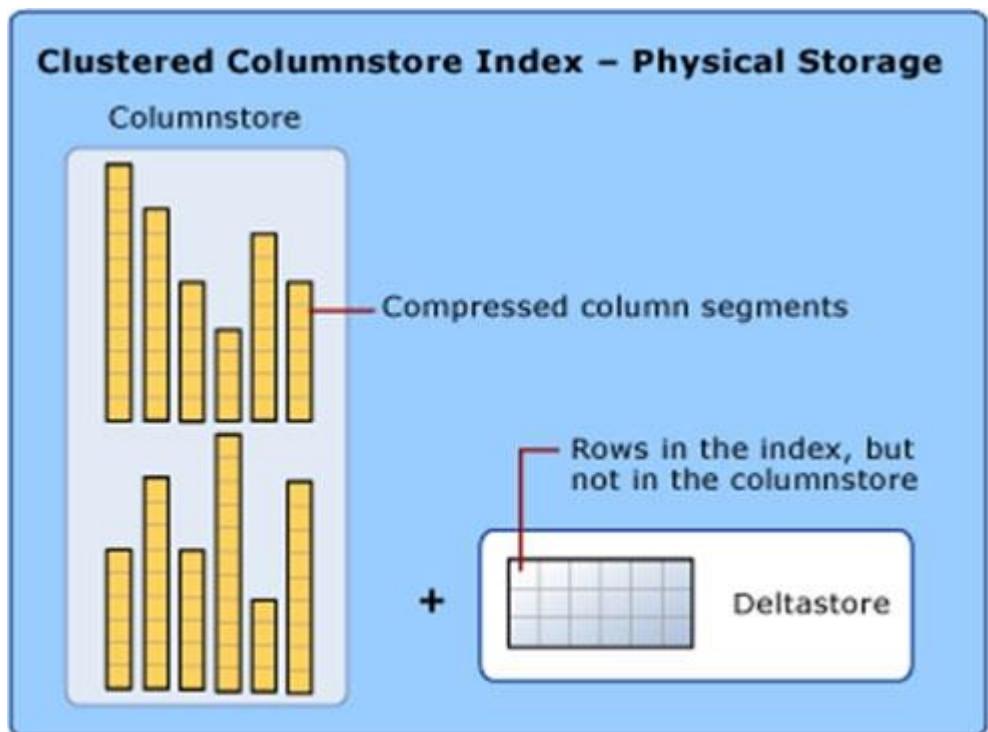
CLUSTERED COLUMNSTORE INDEX -

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.



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

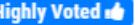
Incorrect:

- * Not HASH([OrderDateKey]). Is not a date column. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work
- * 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.
- * A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods.

<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-distribute>

✉  **ted0809**  10 months, 1 week ago

you don't hash the date.. never..

   upvoted 32 times

✉  **smsme323**  11 months, 1 week ago

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#choose-a-distribution-column-with-data-that-distributes-evenly>

To balance the parallel processing, select a distribution column or set of columns that:

Has many unique values. The distribution column(s) can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions, some distributions can have > 1 unique values while others may end with zero values.

Does not have NULLs, or has only a few NULLs. For an extreme example, if all values in the distribution column(s) are NULL, all the rows are assigned to the same distribution. As a result, query processing is skewed to one distribution, and does not benefit from parallel processing.

Is not a date column. All data for the same date lands in the same distribution, or will cluster records by date. If several users are all filtering on the same date (such as today's date), then only 1 of the 60 distributions do all the processing work.

Ans: Hash(ProductKey)

   upvoted 7 times

✉  **Lestrang**  7 months, 1 week ago

By using the product key as the distribution key, the data for a specific product will be stored on the same node, allowing for faster aggregation of the values in SalesAmount and OrderQuantity for that product.

   upvoted 7 times

✉  **Phund** 12 months ago

must hash on OrderDateKey because that field was not a date and it was used for filter condition "from the last year for a specific product"

   upvoted 7 times

✉  **Lestrang** 11 months, 2 weeks ago

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

the example in the sample-test is in this page and they used productkey for hashing, so yeah, the answer is productkey

   upvoted 14 times

✉  **anks84** 12 months ago

correct

   upvoted 4 times

Question 59

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

ET User:



 **vrodriguesp**  8 months, 2 weeks ago

Remembering that we have data splitted in distribution (60 nodes) and considering that we Need a MINIMUM 1 million rows per distribution, we have:

- A. once per month = $30 \text{ milion} / 60 = 500\text{k}$ record per partition
- B. once per year = $360 \text{ milion} / 60 = 6 \text{ milion}$ record per partition
- C. once per day = $\text{about } 1 \text{ milion} / 60 = 16\text{k}$ record per partition
- D. once per week = $\text{about } 7.5 \text{ milion} / 60 = 125\text{k}$ record per partition

correct should be B



ET Answer: B

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.

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

Bing:

In an Azure Synapse Analytics dedicated SQL pool, the optimal partitioning strategy for a table depends on several factors, including the size of the table, the rate of data growth, and the nature of the queries that will be run against the table. For a large table like Table1, which contains one billion rows and will have thirty million rows added each month, partitioning the table based on the Sales Date column can help optimize query performance and data loading. The frequency at which new partitions should be created depends on the specific requirements of your workload. However, a common approach is to create a new partition for each time period that corresponds to a common query predicate [1](#). For example, if most queries against Table1 filter on a monthly basis, then creating a new partition once per month (option A) would be a reasonable choice. If most queries filter on a daily, weekly or yearly basis, then creating a new partition once per day (option C), once per week (option D), or once per year (option B) respectively would be more appropriate. Ultimately, the optimal partitioning strategy will depend on your specific workload and may require some experimentation to determine.

Question 60

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

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>

Question 61

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

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.

Incorrect:

Not C:

The Avro format is the ideal candidate for storing data in a data lake landing zone because:

1. Data from the landing zone is usually read as a whole for further processing by downstream systems (the row-based format is more efficient in this case).
2. Downstream systems can easily retrieve table schemas from Avro files (there is no need to store the schemas separately in an external meta store).
3. Any source schema change is easily handled (schema evolution).

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

Question 62

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. replicated
- B. hash distributed
- C. heap
- D. round-robin

Answer: A

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.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables>

 **anks84**  12 months ago

Given answer is correct !

Replicated because

- Dimension table
- Less than 2 GB (less than 1 GB in this case)

   upvoted 9 times

Question 63

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

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.

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

✉️  nicky87654 7 months, 3 weeks ago

Selected Answer: C

- C. an IDENTITY column

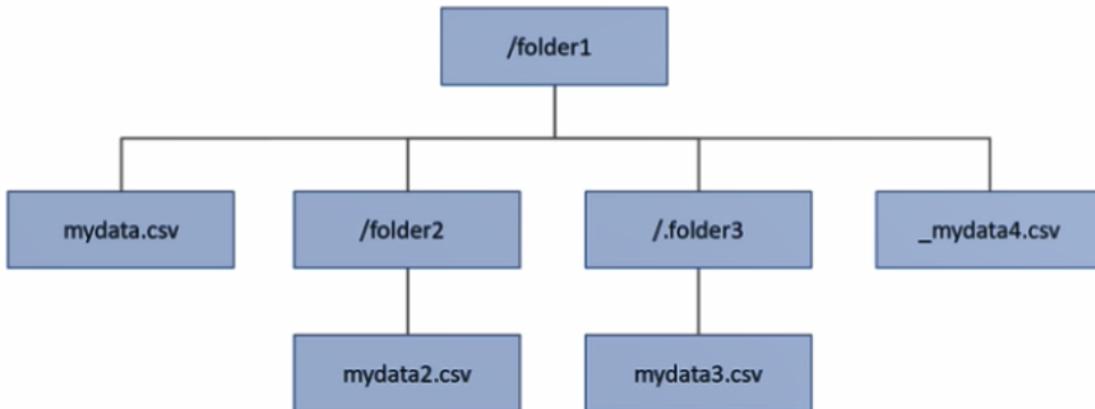
When designing a dimension table in a data warehouse, it's important to consider the types of queries that will be run against it. IDENTITY columns are generally the best option for surrogate keys in dimension tables because they provide the fastest query performance. IDENTITY columns are auto-incremented and indexed by default, which makes them ideal for use as primary keys. They also require less storage space than GUID columns and are less likely to cause fragmentation in indexes.

   upvoted 4 times

Question 64

HOTSPOT -

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.

Answer Area

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"/>

ET User: YNN/YYN

[Exam DP-203 topic 1 question 68 discussion - ExamTopics](#)

✉️ **shoottedduck** Highly Voted 6 months, 2 weeks ago

I have just tested this on Synapse Serverless: ./Folder3 AND _mydata4.csv were ignored. Therefor; Yes, No, No
👍👎FLAG upvoted 49 times

✉️ **MuruAzure** 4 months, 4 weeks ago

its not ./Folder3 . it is ./Folder3 still ignored?
👍👎FLAG upvoted 4 times

✉️ **Vikkic** 2 months, 3 weeks ago

Folder or file that starts with . or ..
Reference documentation: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1
👍👎FLAG upvoted 6 times

✉️  **PGiagkoulas**  7 months, 3 weeks ago

1.Yes, 2.Yes:

"Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /* at the end of path" which is the case here.

3. No:

"Both Hadoop and native external tables will skip the files with the names that begin with an underline (_) or a period (.), refers to files, not directories, so the last file with the underscore will be excluded.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table>

   upvoted 34 times

✉️  **Paulkuzzio** 2 months, 1 week ago

@PGiagkoulas, read this link again : https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

See this part in the link:

native external tables don't return subfolders unless you specify /* at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder.

Based on this, the answer is Yes, No and No

   upvoted 5 times

✉️  **semauni** 1 month, 1 week ago

But /* is specified at the end?

   upvoted 2 times

✉️  **matiandal** 2 months, 3 weeks ago

éva like yia to nickname ;-)

   upvoted 1 times

✉️  **UzairMir**  1 month, 2 weeks ago

The answer is Yes No No.

I found this resource

"COPY ignores hidden folders and doesn't return files that begin with an underline (_) or a period(.) unless explicitly specified in the path. This behavior is the same even when specifying a path with a wildcard."

<https://learn.microsoft.com/en-us/sql/t-sql/statements/copy-into-transact-sql?view=azure-sqldw-latest>.

   upvoted 3 times

✉️  **bakamon** 3 months, 1 week ago

Correct Answers :: 100000% True

Statement 1: Yes. The data from the mydata2.csv file will be returned when selecting all the rows in dbo.Table1 because the file is located in the folder structure specified by the external data source.

Statement 2: No. The data from the mydata3.csv file will not be returned when selecting all the rows in dbo.Table1 because folders that start with a dot are treated as hidden folders and are not included in wildcard searches.

Statement 3: No. The data from the _mydata4.csv file will not be returned when selecting all the rows in dbo.Table1 because files that start with an underscore are treated as hidden files and are not included in wildcard searches.

   upvoted 13 times

✉️  **rocky48** 3 months, 3 weeks ago

1.Yes, 2.Yes, 3. No

   upvoted 1 times

✉️  **Victor_Kings** 4 months, 2 weeks ago

It's definitely Yes-No-No. According to Microsoft documentation, "It won't return mydata3.txt because it's a file in a hidden subfolder. And it won't return _hidden.txt because it's a hidden file.", and as we can see the folder is named "./hiddenfolder/", which means in this case the "./Folder3/" should be ignored too as it is hidden.

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver16&tabs=dedicated>

   upvoted 6 times

✉️  **victorie** 6 months, 1 week ago

Tricky question as Folder3 and mydata4.csv have (.) and (_) in the name. So they should be ignored. IMHO: Y N N

   upvoted 10 times

✉️  **hereiamken** 6 months, 1 week ago

Yes, Yes, No be correct

   upvoted 1 times

✉️  **bubby248** 6 months, 3 weeks ago

All yes.

   upvoted 3 times

✉️  **UGOTCOOKIES** 6 months, 3 weeks ago

Answer should be Yes Yes Yes.

From the learn documentation: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files>
Serverless SQL pool can recursively traverse folders if you specify /* at the end of path. The following query will read all files from all folders and subfolders located in the csv taxi folder.

   upvoted 2 times

✉️  **Karl_Cen** 7 months, 1 week ago

should be yes, no, no, according to following URL

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=native#create-external-data-source>

this external data source is an native data source,

"Both Hadoop and native external tables will skip the files with the names that begin with an underline (_) or a period (.)."

   upvoted 5 times

✉️  **DindaS** 7 months, 2 weeks ago

The main option is missing. Records from all files will be returned. As ** is mentioned hence it will return data from all subfolders.

   upvoted 1 times

✉️  **shakes103** 7 months, 2 weeks ago

Answer is correct. Read explanation posted by @PGiagkoulas.

   upvoted 1 times

✉ **shakes103** 7 months, 2 weeks ago

Answer is correct. Read explanation posted by @PGiagkoulas.

upvoted 1 times

✉ **DAYENKAR** 7 months, 3 weeks ago

1. Yes, 2. Yes:

3. No:

upvoted 1 times

✉ **[Removed]** 7 months, 3 weeks ago

Similar to question 4 topic 1 should be

No

No

Yes

upvoted 2 times

✉ **Lestrang** 7 months, 2 weeks ago

Unlike that question, this has a wildcard in the end, so it will in fact return all subdirectories. so all the answers are yes.

upvoted 1 times

✉ **Lestrang** 7 months, 1 week ago

correction, last one is no due to the underscore.

if file name starts with _ or . it will not be retrieved.

upvoted 1 times

✉ **labrijj** 7 months, 3 weeks ago

Should be all yes!

>>> The Location path contains at the end /** which allows for a recursive search within all folders.

upvoted 4 times

✉ **Amo1** 7 months, 3 weeks ago

It should be all yes!

upvoted 3 times

✉ **ZIMARAKI** 7 months, 3 weeks ago

I think its NO,NO,YES

upvoted 1 times

✉ **Jerrie86** 7 months, 2 weeks ago

You have to see ** after the /. This allows it to search through the folders as well.

upvoted 1 times

ET Answer:

Answer Area

Statements

When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.

Yes

No

When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.

Yes

No

When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.

Yes

No

Question 65

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. product
- B. month
- C. week
- D. region

Answer: B

 nicky87654 Highly Voted 7 months, 3 weeks ago

Selected Answer: B

When designing a fact table in a data warehouse, it is important to consider the types of queries that will be run against it. In this case, the queries that need to be optimized include: show order counts by week, calculate sales totals by region, calculate sales totals by product, and find all the orders from a given month.

Partitioning the table by month would be the best option in this scenario as it would allow for efficient querying of data by month, which is necessary for the query operations described above. For example, it would be easy to find all the orders from a given month by only searching the partition for that specific month.

   upvoted 12 times

Question 66

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 pools.
- 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. \DataSource\SubjectArea\YYYY\WW\FileDialog_YYYY_MM_DD.parquet
- B. \DataSource\SubjectArea\YYYY-WW\FileDialog_YYYY_MM_DD.parquet
- C. DataSource\SubjectArea\WW\YYYY\FileDialog_YYYY_MM_DD.parquet
- D. \YYYY\WW\DataSource\SubjectArea\FileDialog_YYYY_MM_DD.parquet
- E. WW\YYYY\SubjectArea\DataSource\FileDialog_YYYY_MM_DD.parquet

Answer: A

 **nicky87654**  7 months, 3 weeks ago

Selected Answer: A

A. \DataSource\SubjectArea\YYYY\WW\FileDialog_YYYY_MM_DD.parquet

The recommended folder structure that best meets the requirements is option A. It separates data by data source, year and week. It allows for easy filtering of data by year or week, which aligns with the usage pattern where most queries include a filter on the current year or week.

   upvoted 10 times

Bing: A

When designing the folder structure for an Azure Data Lake Storage Gen2 account, it is important to consider the usage patterns, security requirements, and query performance. In this case, the usage patterns indicate that most queries will include a filter on the current year or week, and data will be secured by data source. To support these usage patterns, simplify folder security, and minimize query times, you should organize the data in a hierarchical structure that reflects the most common query filters and security boundaries.

One way to achieve this is to organize the data by data source, subject area, year, and week. This would allow you to easily secure data at the data source level and efficiently filter data by year and week. Therefore, the recommended folder structure is A. \DataSource\SubjectArea\YYYY\WW\FileDialog_YYYY_MM_DD.parquet. This structure supports the usage patterns by organizing data by year and week, simplifies folder security by organizing data by data source, and minimizes query times by allowing efficient filtering on common query predicates.

Question 67

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. DBCC INDEXDEFRAG (pool1, table1)
- B. DBCC DBREINDEX (table1)
- C. ALTER INDEX ALL on table1 REORGANIZE
- D. ALTER INDEX ALL on table1 REBUILD

ET User:



MrWood47 Highly Voted 7 months, 3 weeks ago

Selected Answer: D

D. ALTER INDEX ALL on table1 REBUILD

This statement will rebuild all indexes on table1, which can help to maximize columnstore compression. The other options are not appropriate for this task.

DBCC INDEXDEFRAG (pool1, table1) is for defragmenting the indexes and DBCC DBREINDEX (table1) is for recreating the indexes. ALTER INDEX ALL on table1 REORGANIZE is for reorganizing the indexes.

upvoted 18 times

aemilka Highly Voted 5 months, 1 week ago

Selected Answer: C

Reorganizing an index is less resource intensive than rebuilding an index. For that reason it should be your preferred index maintenance method, unless there is a specific reason to use index rebuild.

<https://learn.microsoft.com/en-us/sql/relational-databases/indexes/reorganize-and-rebuild-indexes?view=sql-server-ver16>

upvoted 5 times

OfficeSaracus 3 months, 3 weeks ago

As far as I can see, your quoted article does not refer to Azure Synapse Analytics dedicated SQL pool. I think rebuild is the only supported option for dedicated SQL as can be found here:
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index>

upvoted 3 times

aemilka 3 months, 3 weeks ago

Yes, I agree with you, I haven't noticed that the article does not apply to Synapse Analytics.

D seems to be only possible answer.

upvoted 2 times

Deeksha1234 **Most Recent** 2 weeks, 5 days ago

Selected Answer: D

D is correct

upvoted 1 times

vctrhugo 2 months, 1 week ago

Selected Answer: D

ALTER INDEX REORGANIZE is used for rebuilding or reorganizing indexes, but it does not maximize columnstore compression.

upvoted 1 times

Matt2000 3 weeks ago

I agree. A rebuild can compress the data more efficiently within each combination of distribution and partition: It can open such existing columnstore segments and shuffle data within them (and the deltastore) to maximize compression for the resulting compressed columnstore segments. That is not possible when reorganizing. That process only changes compressed columnstore segments by physically deleting logically deleted rows and

combining small columnstore segments into larger ones.

upvoted 1 times

Rajan191083 3 months, 2 weeks ago

Reorganize is for row store indexes. The question here clearly mentions column store indexes. Correct answer is D

upvoted 2 times

MuhilMahil 3 months, 4 weeks ago

Selected Answer is C.

reorganizing only help in optimizing compression and performance.

upvoted 1 times

Vedjha 7 months ago

Why not C?

When reorganizing a columnstore index, the Database Engine compresses each closed row group in delta store into columnstore as a compressed row group. Starting with SQL Server 2016 (13.x) and in Azure SQL Database, the REORGANIZE command performs the following additional defragmentation optimizations online:

Physically removes rows from a row group when 10% or more of the rows have been logically deleted. For example, if a compressed row group of 1 million rows has 100,000 rows deleted, the Database Engine will remove the deleted rows and recompress the row group with 900,000 rows, reducing storage footprint.

upvoted 1 times

ET Answer: D

Bing: D

To maximize columnstore compression for a table named table1 in an Azure Synapse Analytics dedicated SQL pool named Pool1, you should execute the **ALTER INDEX ALL on table1 REBUILD** statement. This statement rebuilds all indexes on the table, including the clustered columnstore index, which can improve compression rates and query performance for columnstore indexes ¹. Therefore, the correct answer is D. **ALTER INDEX ALL on table1 REBUILD**. The DBCC INDEXDEFrag (option A) and DBCC DBREINDEX (option B) statements are not applicable to columnstore indexes. The ALTER INDEX ALL on table1 REORGANIZE (option C) statement can be used to defragment a columnstore index, but it would not necessarily maximize columnstore compression.

Question 68

You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You plan to implement a star schema in pool 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

ET User: BE

BE (68%) BD (32%)

ET Answer: BD

✉ **aditya816** Highly Voted 4 months, 4 weeks ago

Selected Answer: BE

Surrogate is already there as customerkey column

Like Reply Flag upvoted 9 times

✉ **laurasscastro** 3 months, 3 weeks ago

that's the business key, not the surrogate key. If a new record is generated, there would be a duplicate key. SK is necessary to identify the record

Like Reply Flag upvoted 3 times

✉ **[Removed]** Highly Voted 4 months, 1 week ago

Selected Answer: BD

I think, there is already a column called InsertedDate, therefore E is not necessary. So we just need another column to track the end date, which is B. And RowID should be a surrogate key in this case.

Like Reply Flag upvoted 7 times

✉ **kkk5566** Most Recent 1 week, 2 days ago

B and D ,its a star schema on which has a fact table include a customerID property.

Like Reply Flag upvoted 1 times

✉ **Deeksha1234** 2 weeks, 5 days ago

B and D makes more sense, since inserted date is there already

Like Reply Flag upvoted 1 times

✉ **YikesYikes2023** 2 months ago

Selected Answer: BE

If RowID was the surrogate, wouldn't it be an IDENTITY column? Therefore, it has to be B and E. Right? Please explain if this doesn't make sense

Like Reply Flag upvoted 1 times

✉ **auwia** 2 months, 1 week ago

Selected Answer: BE

<https://www.sqlshack.com/implementing-slowly-changing-dimensions-scds-in-data-warehouses/>

"For the SCD Type 2, we need to include three more attributes such as StartDate, EndDate and IsCurrent"

IsCurrentRow is already present! ... ;)

CustomerKey (in reality is the RowID that many guys wants to add here), effectiveEndDate will probably set to: 31.12.9999, (to justify the not null).

My final answer wil lbe : B and E.

Like Reply Flag upvoted 1 times

✉ **_ry_** 2 months, 1 week ago

what is the answer ?

Like Reply Flag upvoted 1 times

✉ **ArunMat** 3 months, 1 week ago

Selected Answer: BE

For SCD Type 2 we need record valid from and to date i.e effective date to identify latest row for that id.

Like Reply Flag upvoted 2 times

✉ **jlad26** 4 months ago

I'm confused by the NOT NULL for the EffectiveEndDate. What value is this column going to hold for the row that holds the current information ?

Like Reply Flag upvoted 2 times

Question 69

HOTSPOT -

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])
                  HASH([OrderDate])
                  REPLICATE
                  ROUND_ROBIN
    CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
(
    [CountryId] int NOT NULL
    ,   [CountryCode] varchar(10) NOT NULL
)
WITH
(
    DISTRIBUTION = HASH([CountryCode])
                  HASH([CountryId])
                  REPLICATE
                  ROUND_ROBIN
    CLUSTERED COLUMNSTORE INDEX
)
```

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])
        HASH([OrderDate])
        REPLICATE
        ROUND_ROBIN
)
CLUSTERED COLUMNSTORE INDEX
)

CREATE TABLE [dbo].[Country]
(
    [CountryId] int NOT NULL
,   [CountryCode] varchar(10) NOT NULL
)
WITH
(
    DISTRIBUTION =
        HASH([CountryCode])
        HASH([CountryId])
        REPLICATE
        ROUND_ROBIN
)
CLUSTERED COLUMNSTORE INDEX
)
```

Question 70

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and an Azure Synapse Analytics workspace named workspace1. You need to create an external table in a serverless SQL pool in workspace1. The external table will reference CSV files stored in account1. The solution must maximize performance. How should you configure the external table?

- A. Use a native external table and authenticate by using a shared access signature (SAS).
- B. Use a native external table and authenticate by using a storage account key.
- C. Use an Apache Hadoop external table and authenticate by using a shared access signature (SAS).
- D. Use an Apache Hadoop external table and authenticate by using a service principal in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra.

Answer: A

✉  **FRANCIS_A_M**  5 months ago

Selected Answer: A

Correct! Serverless SQL Pools cannot use Hadoop, Only Native. Access Key Auth is never best practice therefore leaving only A as a viable answer.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>

   upvoted 22 times

✉  **Rob77** 3 months, 1 week ago

It's not about the best practice - there is no option to use storage keys...

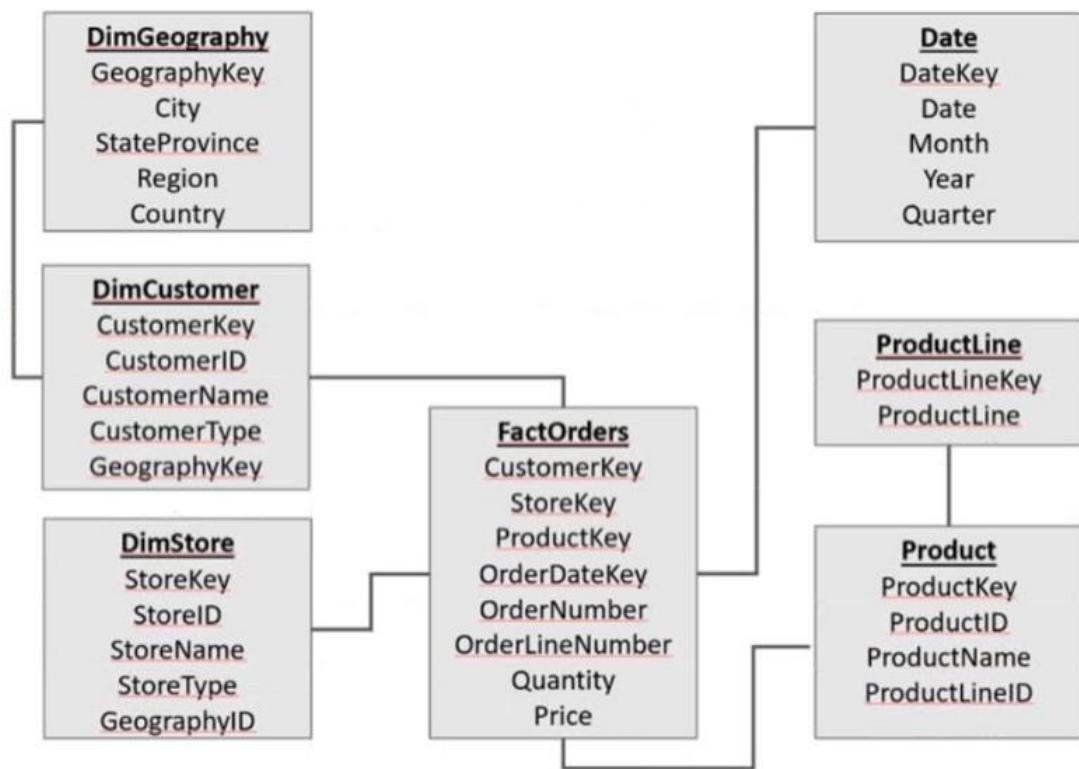
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#security>

   upvoted 6 times

Question 71

HOTSPOT -

You have an Azure Synapse Analytics serverless SQL pool that contains a database named db1. The data model for db1 is 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 exhibit.

NOTE: Each correct selection is worth one point.

Answer Area

To convert the data model to a star schema, [answer choice].

join DimGeography and DimCustomer
join DimGeography and FactOrders
union DimGeography and DimCustomer
union DimGeography and FactOrders

Once the data model is converted into a star schema, there will be [answer choice] tables.

4
5
6
7

ET User:

join DimGeography and DimCustomer and 5 tables

blenau Highly Voted 5 months ago

Correct answer should be join DimGeography and DimCustomer and 5 tables.

You also need to combine ProductLine and Product in order for the schema to be considered a star schema. This would result in 5 remaining tables: DimCustomer (DimCustomer JOIN DimGeography), DimStore, Date, Product (Product JOIN ProductLine) and FactOrders.

upvoted 49 times

Dataminer Highly Voted 4 months, 2 weeks ago

Agree with explanation. It will still be snowflake if Product and ProductLine is not combined
upvoted 11 times

ccesarr Most Recent 2 weeks, 4 days ago

This question is really messy. It doesn't explicitly say that by joining or unioning the tables this means they will be combined into a single table, so it seems like we'll still have 2 tables (DimGeography and DimCustomer) in both options, besides the fact that just fixing DimGeography and DimCustomer won't generate a Star Schema

upvoted 1 times

Deeksha1234 2 weeks, 5 days ago

join DimGeography and DimCustomer and 5 tables
upvoted 1 times

Zak_Zakaria 1 month, 2 weeks ago

ProductLine and Product also should be joined to switch to a star schema, if not we will be still on Snowflake Schema, so the remained tables should be 5, not 6.

upvoted 1 times

DataEngDP 1 month, 3 weeks ago

Customer is already joined with Geography (see the lines), the only thing needed is to combine it with Orders and ProductLine with Orders too, in order to convert this design to a star schema.

In this way we get 6 dimension tables plus the fact tables: Orders.

upvoted 1 times

Reloadedvn 3 months, 3 weeks ago

2. should be 5 tables

upvoted 6 times

rocky48 3 months, 3 weeks ago

DimGeography and DimCustomer and 5 tables.

upvoted 3 times

ET Answer:

Answer Area

To convert the data model to a star schema, [answer choice].

- join DimGeography and DimCustomer
- join DimGeography and FactOrders
- union DimGeography and DimCustomer
- union DimGeography and FactOrders

Once the data model is converted into a star schema, there will be [answer choice] tables.

- 4
- 5
- 6
- 7

Question 72

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage1.

New files are uploaded daily to storage1.

You need to recommend a solution that configures storage1 as a structured streaming source. The solution must meet the following requirements:

- Incrementally process new files as they are uploaded to storage1.
- 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. COPY INTO
- B. Azure Data Factory
- C. Auto Loader
- D. Apache Spark FileStreamSource

Answer: C

  **Nikiboy**  5 months ago

Auto Loader provides a Structured Streaming source called cloudFiles. Plus, it supports schema drift. Hence, Auto Loader is the correct answer.

<https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/>

   upvoted 11 times

  **mr_examers** 3 months, 3 weeks ago

Auto Loader does not support Azure Data Lake Storage Gen2

   upvoted 1 times

Question 73

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible TSV files that do NOT have a header row
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

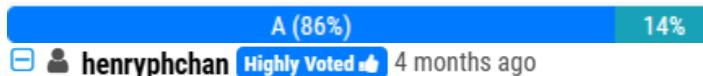
You need to read the TSV files by using ad-hoc queries and the OPENROWSET function.

The solution must assign a name and override the inferred data type of each column.

What should you include in the OPENROWSET function?

- A. the WITH clause
- B. the ROWSET_OPTIONS bulk option
- C. the DATAFILETYPE bulk option
- D. the DATA_SOURCE parameter

ET User:



Selected Answer: A

In the Question "The solution must assign a name and override the inferred data type of each column", so we must need a WITH Clause to define the column names and data types.

thumb up 7 times



I think it's A. WITH CLAUSE

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset>

thumb up 6 times



Agreed - Should be A. "To specify explicit column names and data types, you can override the default column names and inferred data types by providing a schema definition in a WITH clause" (<https://learn.microsoft.com/en-us/training/modules/query-data-lake-using-azure-synapse-serverless-sql-pools/3-query-files>)

thumb up 3 times

✉  **kkk5566** Most Recent 1 week, 2 days ago

A is correct

   upvoted 1 times

✉  **Deeksha1234** 2 weeks, 5 days ago

Selected Answer: A

answer is correct

   upvoted 1 times

✉  **ravigolu** 2 months ago

Selected Answer: A

Ans is A

DATA_SOURCE is used to : Use DATA_SOURCE option and define credential that you want to use to access storage.

The WITH clause allows you to specify columns that you want to read from files.

For CSV data files, to read all the columns, provide column names and their data types.

   upvoted 2 times

✉  **auwia** 2 months, 1 week ago

Selected Answer: A

Yes we need the WITH clause to define the data type for each column.

   upvoted 1 times

✉  **Ankit_Az** 3 months, 1 week ago

Selected Answer: A

A is correct

   upvoted 1 times

✉  **ustefan11** 3 months, 2 weeks ago

Selected Answer: D

The question is 'What should you include in the OPENROWSET function?' with the emphasis on the 'in the function', therefore the correct answer is 'D. the DATA_SOURCE parameter' since it's usually amongst the first arguments when using the OPENROWSET function.

   upvoted 3 times

✉  **peches** 3 months ago

I agree with you, the way the question is written it seems like is asking for a parameter of the OPENROWSET function and that would discard WITH as an answer (even though we need it in the query to accomplish the overriding of data types). ROWSET_OPTIONS is used to avoid query failures due to constantly appended files, which is not the case here. DATAFILETYPE is used to specify the charset (utf-8 or utf-16) which isn't mentioned in the question. That would leave us with option D, but what is strange is that it clearly says that the files are publicly accessible, so we could pass the whole URL using BULK and not use DATA_SOURCE either. Also, we are not using the information regarding the files not having a header row, which should be taken into account by setting the FIRSTROW parameter of the OPENROWSET function.

   upvoted 2 times

✉  **rocky48** 3 months, 2 weeks ago

Selected Answer: A

Option A seems correct answer as With clause helps to overwrite data types and assign names for columns

   upvoted 2 times

✉  **jeroenmouse** 4 months ago

Selected Answer: A

Option A seems correct answer as With clause helps to overwrite data types and assign names for columns

   upvoted 4 times

ET Answer: D

Question 74

You have an Azure Synapse Analytics dedicated SQL pool.

You plan to create a fact table named Table1 that will contain a clustered columnstore index.

You need to optimize data compression and query performance for Table1.

What is the minimum number of rows that Table1 should contain before you create partitions?

- A. 100,000
- B. 600,000
- C. 1 million
- D. 60 million

ET User:



✉️ **Ankit_Az** **Highly Voted** 3 months, 1 week ago

Selected Answer: D

Clustered Column Store will by default have 60 partitions. And to achieve best compression we need at least 1 Million rows per partition, hence Option D 60 Millions (1M per partition)

upvoted 6 times

✉️ **Deeksha1234** **Most Recent** 2 weeks, 5 days ago

should be D

upvoted 1 times

✉️ **akhil5432** 3 weeks, 6 days ago

Selected Answer: C

WHY People mentioned option D..please explain how?

upvoted 1 times

✉️ **akhil5432** 3 weeks, 6 days ago

1 MILLION

upvoted 1 times

✉️  **Lukis92** 1 month, 3 weeks ago

Selected Answer: C

To achieve optimal data compression and query performance with clustered columnstore tables in Azure Synapse Analytics, it is recommended to have a minimum of 1 million rows per distribution and partition.

As Synapse Analytics automatically creates 60 distributions per table, to fulfill the 1 million rows per distribution recommendation, the table should ideally contain 60 million rows if no additional partitions are created.

However, the question is asking about the threshold for creating partitions, not necessarily a table of full 60 million rows. Therefore, you would want to ensure you have at least 1 million rows in each partition to maintain the optimal performance and compression. If the number of rows is less than 1 million, it's better to consider fewer partitions in order to increase the number of rows per partition.

   upvoted 2 times

✉️  **janaki** 3 months, 1 week ago

answer shouls 60 million

The minimum number of rows that Table1 should contain before creating partitions in Azure Synapse Analytics dedicated SQL pool depends on various factors such as data size, query patterns, and performance requirements. However, a commonly recommended threshold is typically around 60 million rows before considering partitioning.

   upvoted 3 times

✉️  **maxstv** 3 months, 2 weeks ago

The minimum number of rows that Table1 should contain before you consider creating partitions in Azure Synapse Analytics dedicated SQL pool depends on multiple factors, such as the size of each row, the expected data growth rate, and the specific requirements of your workload. However, considering the typical guidelines and best practices, a general rule of thumb is to consider creating partitions when the table size reaches around 60 million rows.

Therefore, the minimum number of rows that Table1 should contain before you create partitions is -

D. 60 million.

This is a commonly recommended threshold for optimizing data compression and query performance in large-scale data warehouses. However, it's important to note that this number can vary based on your specific scenario, so it's always advisable to conduct performance testing and consider the characteristics of your data and workload to determine the optimal time for partitioning your table.

   upvoted 1 times

✉️  **RamMovvaa** 3 months, 2 weeks ago

What is the minimum number of rows that Table1 should contain before you create partitions?

Answer : C

   upvoted 1 times

✉️  **ustefan11** 3 months, 2 weeks ago

Selected Answer: D

I've seen in the comments the explanation that this question has something to do with distribution and I don't think this is the case here. It's just that for a partition to have optimal compression, it has to be of at least 1 million rows, and since the idea of having a partition is to divide the data into smaller chunks, you need at least 2 partitions. Therefore, since there's no '2 mil' option, the only option left is '60M'.

   upvoted 1 times

✉️  **rocky48** 3 months, 2 weeks ago

Selected Answer: D

Clustered columnstore has the best compression with 1M rows. So it should be 1M * 60 = 60 million rows

   upvoted 2 times

✉️  **mr_examers** 3 months, 3 weeks ago

Selected Answer: D

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Since Synapse Analytics divides each table into 60 distributions by default, the table should contain at least 1 million rows per distribution or 60 million rows in total before considering partitioning the table.

   upvoted 3 times

✉️  **jeroenmouse** 4 months ago

Selected Answer: D

Hash-distributed tables work well for large fact tables in a star schema and along with that we need to use column store index for better compression and performance. By default it will have 60 distribution before partition and for better performance it is expected to have 1million rows per distribution.

   upvoted 2 times

✉️  **shakes103** 4 months ago

Selected Answer: C

Consider using a clustered columnstore index when:

Each partition has at least a million rows. Columnstore indexes have rowgroups within each partition. If the table is too small to fill a rowgroup within each partition, you won't get the benefits of columnstore compression and query performance.

<https://learn.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-design-guidance?view=sql-server-ver16#use-a-clustered-columnstore-index-for-large-data-warehouse-tables>

   upvoted 2 times

✉️  **peaches** 3 months ago

Exactly, when EACH partition has at least a million rows, hence the whole table should have at least 60 million rows since by default each partition is distributed between 60 nodes ($6.000.000/60 = 1.000.000$).

   upvoted 1 times

✉️  **jlad26** 4 months ago

Selected Answer: D

"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" (<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>)

   upvoted 2 times

✉️  **dksk5** 4 months ago

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.

   upvoted 1 times

 **19c1ee5** 4 months ago

Selected Answer: D

Clustered columnstore has the best compression with 1M rows. So it should be 1M * 60.

   upvoted 2 times

ET Answer: A

Bing: D

The minimum number of rows that Table1 should contain before you create partitions is **60 million**. According to the Azure Synapse Analytics documentation, 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 [1](#). So, the correct answer is **D. 60 million**. Is there anything else you would like to know? 😊

Question 75

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

- RepSourceID
- SalesRepID
- FirstName
- LastName
- StartDate
- EndDate
- Region

You are developing an Azure Synapse Analytics pipeline that includes a mapping data flow named Dataflow1. Dataflow1 will read sales team data from an external source and use a Type 2 slowly changing dimension (SCD) when loading the data into DimSalesPerson.

You need to update the last name of a salesperson in DimSalesPerson.

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

NOTE: Each correct selection is worth one point.

- A. Update three columns of an existing row.
- B. Update two columns of an existing row.
- C. Insert an extra row.
- D. Update one column of an existing row.

ET User:



bakamon Highly Voted 3 months, 1 week ago

Selected Answer: BC

- 1) Insert an extra row with the updated last name and the current date as the StartDate.
- 2) Update two columns of an existing row: set the EndDate of the previous row for that salesperson to the current date and set the current value of the SalesRepID column to inactive.

upvoted 6 times

bakamon 3 months, 1 week ago

This will preserve the history of changes to the salesperson's last name while keeping the most current information in the table

upvoted 1 times

✉  **Ankit_Az**  3 months, 1 week ago

Selected Answer: CD

CD is correct

   upvoted 5 times

✉  **lfss**  2 weeks, 2 days ago

cd is correct

   upvoted 1 times

✉  **Deeksha1234** 2 weeks, 5 days ago

Selected Answer: CD

answer should be CD, since activeRow flag is not present, we need to update only end date.

   upvoted 2 times

✉  **Ram9198** 3 weeks, 3 days ago

Selected Answer: CD

CD is correct

   upvoted 1 times

✉  **Rob77** 3 months, 1 week ago

It's SCD Type 2 - you need to update at least three columns in the original raw:
Surname, StartDate and EndDate. (IsActive if one exists). Then insert new record.
A and C

   upvoted 1 times

✉  **Rob77** 3 months, 1 week ago

* "original row"

   upvoted 1 times

✉  **peches** 3 months ago

but if you update the surname on the original row, don't you lose the previous value?

   upvoted 4 times

✉  **laurasscastro** 3 months, 3 weeks ago

For me this is a little dubious since besides the end date update for the record we could have flg_is_active as well. Making B a possible answer in my opinion

   upvoted 4 times

✉  **ajhak** 3 months, 2 weeks ago

It's saying "update on column of an EXISTING row". AKA you're just changing the
IsCurrent part of the existing row, that's it.

   upvoted 1 times

✉  **henryphchan** 4 months ago

Selected Answer: CD

The answer is correct

   upvoted 3 times

✉  **OfficeSaracus** 4 months ago

Selected Answer: CD

Ans is correct

   upvoted 3 times

✉️  **jeroenmouse** 4 months ago

Selected Answer: CD

SCD Type 2 will have historical changes hence we will have new row and we need to update the existing row's end date. Hence - CD

<https://www.sqlshack.com/implementing-slowly-changing-dimensions-scdfs-in-data-warehouses/>

   upvoted 5 times

✉️  **Yemeral** 4 months ago

Selected Answer: CD

Correct. You need to insert a new row with the updated data and update the EndDate of the old row

   upvoted 4 times

ET Answer: CD

Bing: AC

To update the last name of a salesperson in DimSalesPerson using a Type 2 slowly changing dimension (SCD), you should perform the following two actions: **A. Update three columns of an existing row and C. Insert an extra row.**

In a Type 2 SCD, when a change occurs in the source data, you need to update the EndDate column of the existing row to reflect the date when the change occurred, and insert a new row with the updated information and a new StartDate. In this case, you need to update the EndDate column of the existing row, insert a new row with the updated LastName and a new StartDate, and leave the other columns unchanged. Is there anything else you would like to know? 😊

Question 76

HOTSPOT -

You plan to use an Azure Data Lake Storage Gen2 account to implement a Data Lake development environment that meets the following requirements:

- Read and write access to data must be maintained if an availability zone becomes unavailable.
- Data that was last modified more than two years ago must be deleted automatically.
- Costs must be minimized.

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

NOTE: Each correct selection is worth one point.

Answer Area

For storage redundancy:

- Geo-zone-redundant storage (GZRS)
- Locally-redundant storage (LRS)
- Zone-redundant storage (ZRS)

For data deletion:

- A lifecycle management policy
- Soft delete
- Versioning

Answer:

Answer Area

For storage redundancy:

- Geo-zone-redundant storage (GZRS)
- Locally-redundant storage (LRS)
- Zone-redundant storage (ZRS)

For data deletion:

- A lifecycle management policy
- Soft delete
- Versioning

  **bakamon**  3 months, 1 week ago

Statement 1: For Storage redundancy, you should select ZRS (Zone-redundant storage). This will maintain read and write access to data even if an availability zone becomes unavailable.

Statement 2: For data deletion, you should select A lifecycle management policy. This will allow you to automatically delete data that was last modified more than two years ago

   upvoted 8 times

Question 77

HOTSPOT -

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 and rarely accessed.
- 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.

What should you include in the storage policy for each department? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

HR:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Operations:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Answer:

Answer Area

HR:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Operations:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

✉  **dksks** Highly Voted 4 months ago

correct

   upvoted 10 times

✉  **OfficeSaracus** Highly Voted 4 months ago

The answer for HR depends on the meaning of "rarely" and the duration of "initial processing". If rarely is like once a year and initial processing is complete within 24 h the answer is correct. If rarely is like on a weekly basis, archiv might be the wrong way

   upvoted 9 times

✉  **semauni** 1 month, 1 week ago

I agree, I also felt like I was missing information. In this case however, I'd say go for 'minimizing costs'. So the lowest cost option possible.

   upvoted 1 times

✉  **Ankit_Az** Most Recent 3 months, 1 week ago

Correct

   upvoted 1 times

✉  **Rob77** 3 months, 1 week ago

You can't access data that was archived without rehydration. Rehydration requires either amending blob tier to hot or cold and is likely to incur a fee if stored less than 180 day or copying blob to another location... therefore "rarely" is unlikely a good option...

   upvoted 1 times

Question 78

HOTSPOT -

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 DimCustomer in an Azure Synapse Analytics dedicated SQL pool.

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

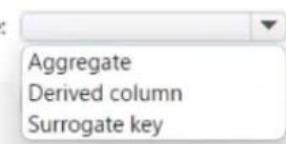
- Detect whether the data of a given customer has changed in the DimCustomer table.
- Perform an upsert to the DimCustomer table.

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

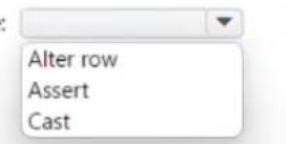
NOTE: Each correct selection is worth one point.

Answer Area

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



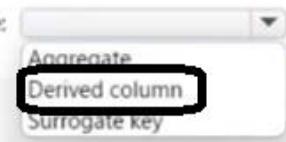
Perform an upsert to the DimCustomer table:



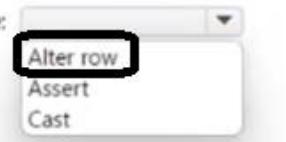
Answer:

Answer Area

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



Perform an upsert to the DimCustomer table:



aemilka Highly Voted 3 months, 3 weeks ago

The answer is correct. Check "Exercise - Design and implement a Type 1 slowly changing dimension with mapping data flows", there is described implementation of the dataflow mentioned in this question.

<https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/4-exercise-design-implement-type-1-dimension>

In the exercise 'Derived column' transformation is used to add InsertedDate and ModifiedDate columns. ModifiedDate column can be used to detect whether the customer data has changed. For Upsert 'Alter row' transformation is used. The answer is definitely correct.

upvoted 13 times

Question 79

DRAG DROP -

You have an Azure Synapse Analytics serverless SQL pool.

You have an Azure Data Lake Storage account named adls1 that contains a public container named container1. The container1 container contains a folder named folder1.

You need to query the top 100 rows of all the CSV files in folder1.

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	SELECT TOP 100 *
DATA_SOURCE	FROM [] (
LOCATION	[] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
OPENROWSET	FORMAT = 'CSV') AS rows

Answer:

Answer Area

```
SELECT TOP 100 *
FROM [OPENROWSET] (
[BULK] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
```

  **rocky48** 3 months, 2 weeks ago

The provided query is correct for Azure Synapse Analytics serverless SQL pool. It selects the top 100 rows from the data in CSV format located at the specified URL: https://adls1.dfs.core.windows.net/container1/folder1/*.csv. The results are returned under the alias rows. Answer is correct.

   upvoted 2 times

Question 80

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

You plan to create a database named DB1 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 pool.

Which format should you use for the tables in DB1?

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

Answer: A

✉  **Deeksha1234** 2 weeks, 5 days ago

Selected Answer: A

parquet. CSV , delta also possible but not an option here.

   upvoted 1 times

✉  **aemilka** 3 months, 3 weeks ago

Selected Answer: A

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>

Supported formats for serverless pool: Delimited/CSV, Parquet, Delta Lake

So Parquet is the correct answer

   upvoted 1 times

✉  **henryphchan** 4 months ago

Selected Answer: A

Parquet is supported by serverless SQL pool

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

   upvoted 2 times

Question 81

You have an Azure Data Lake Storage Gen2 account named storage1.

You plan to implement query acceleration for storage1.

Which two file types support query acceleration? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

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

Answer: AD

  orionduo  2 months, 2 weeks ago

Correct.

Query acceleration supports CSV and JSON formatted data as input to each request.
<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration>

   upvoted 7 times

  vctrhugo  2 months, 3 weeks ago

Selected Answer: AD

Query acceleration supports CSV and JSON formatted data as input to each request.

   upvoted 6 times

  aga444  2 months, 3 weeks ago

Parquet and CSV

   upvoted 1 times

  IanKwok81 2 months, 3 weeks ago

Correct. <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration>

   upvoted 4 times

Question 82

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible JSON files
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

You need to read the files in storage1 by using ad-hoc queries and the OPENROWSET function. The solution must ensure that each rowset contains a single JSON record.

To what should you set the FORMAT option of the OPENROWSET function?

- A. JSON
- B. DELTA
- C. PARQUET
- D. CSV

ET User:

D (94%) 6%
✉ **susbhat** 6 days, 15 hours ago
Selected Answer: D
Ignore my previous comment.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#read-json-files>
👍 ↺ 📈 upvoted 1 times

✉ **susbhat** 6 days, 15 hours ago
Selected Answer: A
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#read-json-files>
👍 ↺ 📈 upvoted 1 times

✉ **kdp203** 1 week, 2 days ago
CSV -> D is the correct answer
👍 ↺ 📈 upvoted 1 times

ET Answer: A

Question 83

HOTSPOT -

You have an Azure subscription that contains the Azure Synapse Analytics workspaces shown in the following table.

Name	Primary storage account
workspace1	datalake1
workspace2	datalake2
workspace3	datalake1

Each workspace must read and write data to datalake1.

Each workspace contains an unused Apache Spark pool.

You plan to configure each Spark pool to share catalog objects that reference datalake1.

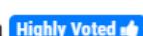
For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	<input type="radio"/>	<input type="radio"/>
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	<input type="radio"/>	<input type="radio"/>
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	<input type="radio"/>	<input type="radio"/>

ET User: YYN/YYY/YNY

  **auwia**  2 months, 1 week ago

Provided answers are correct:

1. Yes:

Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog. When customers want to persist the Hive catalog metadata outside of the workspace, and share catalog objects with other computational engines outside of the workspace, such as HDInsight and Azure Databricks, they can connect to an external Hive Metastore. Only Azure SQL Database and Azure Database for MySQL are supported as an external Hive Metastore.

2. Yes:

And currently we only support User-Password authentication.

3. No:

And currently we only support User-Password authentication. ==> STORAGE BLOB CONTRIBUTOR is an Azure RBAC (Role-Based Access Control) ==> NOT COMPATIBLE (it is supported User-Password authentication ONLY).

ref.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

   upvoted 6 times

- **Deeksha1234** Most Recent 2 weeks, 5 days ago
given answer is correct
<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>
Upvoted 1 times
- **Ram9198** 3 weeks, 3 days ago
Yes , yes , no <https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>
Upvoted 1 times
- **pavankr** 2 months ago
Correct order should be Yes, No, Yes
Upvoted 2 times
- **DataSaM** 2 months ago
<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>
Upvoted 1 times
- **Albeeliu** 2 months ago
What are the correct answers???
Upvoted 1 times
- **Ram9198** 3 weeks, 2 days ago
Yes, Yes , No you can check the document link
Upvoted 1 times
- **Paulkuzzio** 2 months, 1 week ago
Only Azure SQL Database and Azure Database for MySQL are supported as an external Hive Metastore. And currently we only support User-Password authentication.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>
Upvoted 5 times
- **sridat** 2 months, 3 weeks ago
A "Blob Data Contributor" role must be assigned to user in order to access the files in blob storage. So it's a "Yes"
<https://learn.microsoft.com/en-us/azure/storage/blobs/assign-azure-role-data-access?tabs=portal>
Upvoted 6 times

Answer:

Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	<input checked="" type="radio"/>	<input type="radio"/>
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	<input checked="" type="radio"/>	<input type="radio"/>
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	<input type="radio"/>	<input checked="" type="radio"/>

Question 84

DRAG DROP -

You have a data warehouse.

You need to implement a slowly changing dimension (SCD) named Product that will include three columns named ProductName, ProductColor, and ProductSize. The solution must meet the following requirements:

- Prevent changes to the values stored in ProductName.
- Retain only the current and the last values in ProductSize.
- Retain all the current and previous values in ProductColor.

Which type of SCD should you implement for each column? To answer, drag the appropriate types to the correct columns. Each 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.

SCD Type	Answer Area
Type 0	ProductName: <input type="text"/>
Type 1	Color: <input type="text"/>
Type 2	Size: <input type="text"/>
Type 3	

ET User: 023/032

✉ Ram0202 2 months, 3 weeks ago

Product name -type 0

color -type 2

size -type 3

upvoted 27 times

✉ andjurovicela 2 months, 2 weeks ago

ProductName - type 0, as no changes are done. Color - type 3, as with type 3 we have one column for the current value and one for the previous so only these two are preserved. Size - type 2, as it inserts a new row for every change, so we get all historical values.

upvoted 14 times

✉ hiyoww 1 month, 2 weeks ago

Agree. beware that the order of ProductSize, ProductColor in the question, not same as in the graph.

Product name -type 0

color -type 3

size -type 2

upvoted 7 times

✉  **Deeksha1234**  2 weeks, 5 days ago

correct answer is - type 0, type 3,type 2

   upvoted 3 times

✉  **akhil5432** 3 weeks, 6 days ago

type 0

type 3

type 2

   upvoted 2 times

✉  **ravigolu** 2 months ago

Answer is

Product name -type 0

color -type 2

size -type 3

Type 0 – Fixed Dimension

No changes allowed, dimension never changes

A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten.

Type 2 SCD

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.

Type 3 SCD

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.

   upvoted 7 times

✉  **auwia** 2 months, 2 weeks ago

Answer are corrects as reported in the solution: 0, 1, and 2.

   upvoted 2 times

✉  **auwia** 2 months, 2 weeks ago

- Prevent changes to the values stored in ProductName. => TYPE 0
- Retain only the current and the last values in ProductSize. => TYPE 2 (current and last, means all the history)
- Retain all the current and previous values in ProductColor. => TYPE 3 (it includes a column for the previous value)

   upvoted 3 times

✉  **auwia** 2 months, 2 weeks ago

Retain all the current and previous values in ProductColor. => TYPE 2 (because the plural in the requirement: previous values, probably it means all the history). Concluding the answers provided are correct :)

   upvoted 2 times

✉  **auwia** 2 months, 2 weeks ago

• Retain only the current and the last values in ProductSize. => TYPE 1 (current and last, DOESN'T mean all the history, but as written only the last ... meaning the current).

<https://learn.microsoft.com/en-us/training/modules/load-optimize-data-into-relational-data-warehouse/5-load-slowly-changing-dimensions>

   upvoted 2 times

✉  **HimaC5991** 2 months, 2 weeks ago

my answer is 0,2,3

   upvoted 3 times

✉ sridat 2 months, 3 weeks ago

ProductName is Type 2 since it maintains current and last record. Type 1 can only have current value.

Like | Reply | Flag upvoted 2 times

✉ wendy 2 months, 3 weeks ago

Retain ONLY the current and the last values in ProductSize. type2 will include all changes.
type 3 is correct.

Like | Reply | Flag upvoted 4 times

✉ mehroosal 2 months, 3 weeks ago

Correct answer is type 0, type 2, type 3

Like | Reply | Flag upvoted 1 times

✉ abdallaissa 2 months, 3 weeks ago

Correct!

Like | Reply | Flag upvoted 3 times

✉ IanKwok81 2 months, 3 weeks ago

Should be ProductName Type0, ProductColor Type2, ProductSize Type1

Like | Reply | Flag upvoted 1 times

✉ IanKwok81 2 months, 3 weeks ago

ProductSize Type3

Like | Reply | Flag upvoted 1 times

✉ abdallaissa 2 months, 3 weeks ago

Color need all the the values which mean you need all the rows for it, in type 1 any change happened a new row will be added

Like | Reply | Flag upvoted 1 times

✉ RoyP654 2 months, 3 weeks ago

ProductName = Type3? retain only the current and last value on the same row

Like | Reply | Flag upvoted 2 times

ET Answer:

Answer Area

ProductName: Type 0

Color: Type 1

Size: Type 2

Bing: 032, 注意顺序

Based on the requirements, you should implement the following types of Slowly Changing Dimensions (SCD) for each column:

- **ProductName:** This column should be a **Type 0 SCD**. Type 0 SCDs are used when you want to prevent changes to the values stored in a column ¹. In this case, changes to the ProductName are not allowed ¹.
- **ProductSize:** This column should be a **Type 3 SCD**. Type 3 SCDs maintain an old attribute but in the same row by adding a previous value column ¹. In this case, you want to retain only the current and the last values in ProductSize ¹.
- **ProductColor:** This column should be a **Type 2 SCD**. Type 2 SCDs maintain all the old records for the dimension by versioning the row ¹. In this case, you want to retain all the current and previous values in ProductColor ¹.

Question 85

HOTSPOT -

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.

NOTE: Each correct selection is worth one point.

Answer Area

Table distribution:

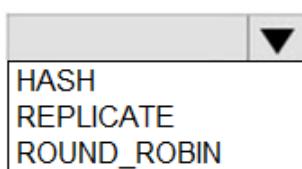


Table structure:



Answer:

Answer Area

Table distribution:

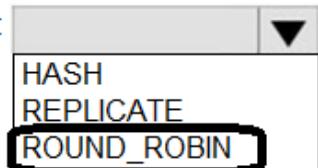


Table structure:



orionduo Highly Voted 2 months, 2 weeks ago

Correct!

The ROUND_ROBIN distribution distributes the data evenly across all distribution nodes in the SQL pool. This distribution type is suitable for loading data quickly into the staging tables because it minimizes the data movement during the loading process.

Use a HEAP table: Instead of creating a clustered index on the staging table, it is recommended to create a HEAP table. A HEAP table does not have a clustered index, which eliminates the need for maintaining the index and improves the data loading performance. It allows for faster insert operations.

upvoted 8 times

Question 86

You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos DB database account named Cosmos1. Cosmos1 contains a container named container1 and ws1 contains a serverless SQL pool.

You need to ensure that you can query the data in container1 by using the serverless 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: ACD

  **gusztimm** 1 week, 4 days ago

Selected Answer: ACD

Correct

   upvoted 1 times

  **Deeksha1234** 2 weeks, 5 days ago

The answer is correct

   upvoted 1 times

  **pramod4lk** 1 month ago

The answer is correct. We need to enable an analytical store in container1.

   upvoted 4 times

Question 87

HOTSPOT -

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
Workspace1	Azure Synapse workspace	Contains the Built-in serverless SQL pool
Pool1	Azure Synapse Analytics dedicated SQL pool	Deployed to Workspace1
storage1	Storage account	Hierarchical namespace enabled

The storage1 account contains a container named container1. The container1 container contains the following files.

```
Webdata <root folder>
    Monthly <folder>
        _monthly.csv
        Monthly.csv
        .testdata.csv
        testdata.csv
```

In Pool1, you run the following script.

```
CREATE EXTERNAL DATA SOURCE Ds1
WITH
    ( LOCATION = 'abfss://container1@storage1.dfs.core.windows.net' ,
      CREDENTIAL = credential1,
      TYPE = HADOOP
    ) ;
```

In the Built-in serverless SQL pool, you run the following script.

```
CREATE EXTERNAL DATA SOURCE Ds2
WITH (
    LOCATION = 'https://storage1.blob.core.windows.net/container1/Webdata/',
    CREDENTIAL = credential2
);
```

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
An external table that uses Ds1 can read the _monthly.csv file.	<input type="radio"/>	<input type="radio"/>
An external table that uses Ds1 can read the Monthly.csv file.	<input type="radio"/>	<input type="radio"/>
An external table that uses Ds2 can read the .testdata.csv file.	<input type="radio"/>	<input type="radio"/>

ET User: NYN

✉  **pramod4lk**  1 month ago

The answer is No, Yes, No

It will ignore "_" and ":"

   upvoted 15 times

✉  **subhraz**  5 days, 7 hours ago

NO

YES

NO

   upvoted 1 times

✉  **pc1337xd** 1 month ago

Both Hadoop(dedicated) and native(serverless) external tables will skip the files with the names that begin with an underline (_) or a period (.).

   upvoted 4 times

✉  **g2000** 1 month ago

The last one is No. File is prefixed with a period and therefore can't be returned.

https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

   upvoted 3 times

ET Answer:

Answer Area

Statements

Yes

No

An external table that uses Ds1 can read the _monthly.csv file.



An external table that uses Ds1 can read the Monthly.csv file.



An external table that uses Ds2 can read the .testdata.csv file.



Question 88

DRAG DROP -

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and a user named User1.

In account1, you create a container named container1. In container1, you create a folder named folder1.

You need to ensure that User1 can list and read all the files in folder1. The solution must use the principle of least privilege.

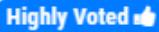
How should you configure the permissions for each folder? To answer, drag the appropriate permissions to the correct folders. Each permission may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Permissions		Answer Area
Execute	None	container1/: <input type="text"/>
Read	Read and Execute	container1/folder1: <input type="text"/>
Read and Write	Write	

Answer Area

container1/: <input type="text"/>
container1/folder1: <input type="text"/>

  g2000  1 month ago

correct!

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control#levels-of-permission>

   upvoted 6 times

Question 89

You have an Azure Data Factory pipeline named pipeline1.

You need to execute pipeline1 at 2 AM every day. The solution must ensure that if the trigger for pipeline1 stops, the next pipeline execution will occur at 2 AM, following a restart of the trigger.

Which type of trigger should you create?

- A. schedule
- B. tumbling
- C. storage event
- D. custom event

ET User: B / A

✉  **yassine70** 1 week, 3 days ago

Answer is Tumbling :

Link : <https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison>

"Retry capability Supported. Failed pipeline runs have a default retry policy of 0, or a policy that's specified by the user in the trigger definition. Automatically retries when the pipeline runs fail due to concurrency/server/throttling limits (that is, status codes 400: User Error, 429: Too many requests, and 500: Internal Server error)"

Retry capability is not supported on Schedule trigger

   upvoted 1 times

✉  **Biswada** 2 weeks, 6 days ago

Tumbling

   upvoted 3 times

✉️  **Matt2000** 3 weeks ago

Sounds like a tumbling trigger with self-dependency: "if the trigger for pipeline1 stops, the next pipeline execution will occur at 2 AM, following a restart of the trigger", implicating that the trigger does not start if the previous run of the trigger is not yet completed.

By using a tumbling trigger with self-dependency, one can let a trigger only start if a previous run of the same trigger has completed. To achieve that maxConcurrency has to be set to '1'.

Ref: <https://learn.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency>

   upvoted 2 times

✉️  **akhil5432** 3 weeks, 6 days ago

Selected Answer: A

Schedule

   upvoted 3 times

ET Answer: A

Bing: A

The correct answer is **A. schedule**. A schedule trigger in Azure Data Factory runs pipelines on a wall-clock schedule. It's ideal for scenarios where you want the pipeline to run at a specific time or specific times of day, such as 2 AM every day  . The schedule trigger ensures that if the trigger stops, the next pipeline execution will occur at the specified time (2 AM in this case), following a restart of the trigger  . Is there anything else you would like to know? 

Question 90

HOTSPOT -

You have an Azure data factory named adf1 that contains a pipeline named ExecProduct. ExecProduct contains a data flow named Product.

The Product data flow contains the following transformations:

1. WeeklyData: A source that points to a CSV file in an Azure Data Lake Storage Gen2 account with 20 columns
2. ProductColumns: A select transformation that selects from WeeklyData six columns named ProductID, ProductDescr, ProductSubCategory, ProductCategory, ProductStatus, and ProductLastUpdated
3. ProductRows: An aggregate transformation
4. ProductList: A sink that outputs data to an Azure Synapse Analytics dedicated SQL pool

The Aggregate settings for ProductRows are configured as shown in the following exhibit.

The screenshot shows the 'Aggregate settings' tab of the ProductRows transformation configuration. It includes fields for 'Output stream name' (ProductRows), 'Incoming stream' (ProductColumns), and a 'Group by' section. The 'Aggregates' tab is selected, showing a table with one row. The row has a 'Column' column containing 'Each column that matches name='ProductID'' and an 'Expression' column containing '\$\$ abc first(\$\$) ANY'. There are also buttons for adding, cloning, deleting, and opening an expression builder.

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
There will be six columns in the output of ProductRows.	<input type="radio"/>	<input type="radio"/>
There will always be one output row for each unique value of ProductDescr.	<input type="radio"/>	<input type="radio"/>
There will always be one output row for each unique value of ProductID.	<input type="radio"/>	<input type="radio"/>

ET User: YNY

Ram9198 Highly Voted 3 weeks, 3 days ago

Yes , no, yes - <https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate>

upvoted 5 times

✉  **KKK5566** **Most Recent** 1 week, 2 days ago

Yes, no, yes

   upvoted 1 times

✉  **mmoayed** 2 weeks, 2 days ago

I have notice that some answers might be wrong. What does this mean? who is confirming the correct answers?

   upvoted 1 times

✉  **mmoayed** 2 weeks, 2 days ago

yes, no, yes

   upvoted 1 times

✉  **MSEExpert** 2 weeks, 5 days ago

Yes No Yes

   upvoted 1 times

✉  **DataEngDP** 2 weeks, 6 days ago

yes, no, yes

   upvoted 2 times

✉  **ClydeZ** 3 weeks, 2 days ago

6 columns from product + the aggregated measurement = 7 columns in total, so answer is NO?

   upvoted 1 times

✉  **Elxaxe** 2 weeks, 4 days ago

The IRS no new aggregated measurement. You're grouping by ProductID, so that makes one column. All the other columns that are not 'ProductID' are grouped by choosing their first row. So you will obtain the same number of columns, like a SELECT DISTINCT.

   upvoted 2 times

✉  **Heringer** 3 weeks, 4 days ago

How can one know that the answer to the third question is either yes or no? In my understanding, you'd have to assume that there is no duplicates in the source table, i.e. there are no rows that share the same values in all columns except for ProductID

   upvoted 2 times

ET Answer:

Answer Area

Statements

There will be six columns in the output of ProductRows.

Yes

No



There will always be one output row for each unique value of ProductDescr.



There will always be one output row for each unique value of ProductID.



Question 91

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. DWU limit
- B. Cache hit percentage
- C. Local tempdb percentage
- D. Data IO percentage

Answer: B

 **MSExpert** 3 weeks, 4 days ago

Correct

   upvoted 4 times

Bing: B

To determine the source of the performance issues in your Azure Synapse Analytics enterprise data warehouse, you should monitor the **Cache hit percentage** metric. This metric measures the percentage of columnstore segments hits in the local SSD cache, where cache hits are 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 ¹. A low Cache hit percentage could indicate that commonly used data is not being cached effectively, leading to slower performance for commonly used queries. So, the correct answer is B. **Cache hit percentage**. Is there anything else you would like to know? 😊

Question 92

HOTSPOT -

You have an Azure Synapse Analytics serverless SQL pool.

You have an Apache Parquet file that contains 10 columns.

You need to query data from the file. The solution must return only two columns.

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,  
N'https://myaccount.dfs.core.windows.net/mycontainer/mysubfolder/data.parquet', FORMAT = 'PARQUET')  
WITH  
(Col1 int, Col2 varchar(20))  
FILEPATH('2')  
PARSER_VERSION = '2.0'  
SINGLE_BLOB
```

Answer:

Answer Area

```
SELECT * FROM  
OPENROWSET(BULK,  
N'https://myaccount.dfs.core.windows.net/mycontainer/mysubfolder/data.parquet', FORMAT = 'PARQUET')  
WITH  
(Col1 int, Col2 varchar(20))  
FILEPATH('2')  
PARSER_VERSION = '2.0'  
SINGLE_BLOB
```

✉  **pramod4lk** 1 month ago

Correct, Serverless SQL pool uses BULK.

   upvoted 2 times

✉  **g2000** 1 month ago

correct!

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset#data-source>

   upvoted 4 times

Common Question 2

Question 1

HOTSPOT -

You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area

Answer Area

Input type:

Stream
Reference

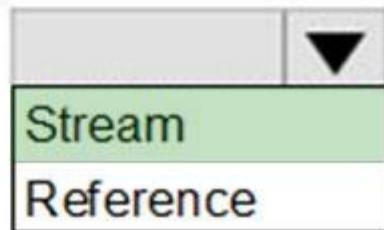
Function:

Aggregate
Geospatial
Windowing

Answer:

Answer Area

Input type:



Function:



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.

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

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

Question 2

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.

ET User: 推荐 CF



✉ **manquak** 2 years ago

Partition input and output.

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

upvoted 56 times

✉ **kolakone** 1 year, 11 months ago

Agree. And partitioning Input and output with same number of partitions gives the best performance optimization..

upvoted 12 times

✉ **Lio95** 1 year, 11 months ago

No event consumer was mentioned. Therefore, partitioning output is not relevant. Answer is correct

upvoted 14 times

✉️  **nicolas1999** 1 year, 9 months ago

Stream analytics ALWAYS has at least one output. There is no need to mention that. So correct answer is input and output

   upvoted 3 times

✉️  **Boompiee** 1 year, 3 months ago

The stream analytics job is the consumer.

   upvoted 1 times

✉️  **dp_learner** **Most Recent** 3 months, 1 week ago

"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. This parallelism has the following requirements:

...

4. The number of input partitions must equal the number of output partitions."

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

   upvoted 3 times

✉️  **bakamon** 3 months, 1 week ago

Selected Answer: CF

input and output

   upvoted 3 times

✉️  **rocky48** 3 months, 2 weeks ago

Selected Answer: CF

C. Implement query parallelization by partitioning the data output.

F. Implement query parallelization by partitioning the data input.

   upvoted 3 times

✉️  **dksks** 4 months ago

Selected Answer: CD

C. Implement query parallelization by partitioning the data output.

D. Scale the SU count for the job up.

Explanation:

A higher SU count provides more processing power and can improve the performance of the Azure Stream Analytics job. Scaling up the job by increasing the SU count can reduce query latency and improve throughput.

Partitioning the data output allows for query parallelization, which can improve the performance of the job. By dividing the output into partitions, the job can process data simultaneously, reducing the time required to complete the job.

   upvoted 2 times

✉️  **esaade** 6 months ago

To optimize the performance of the Azure Stream Analytics job, you should perform the following two actions:

C. Implement query parallelization by partitioning the data output. Partitioning the data output helps to distribute query processing across multiple partitions, which can improve performance for queries that require a large amount of processing power.

D. Scale the SU count for the job up. Scaling up the number of Streaming Units (SU) will provide more processing power for the job, which can improve performance.

Therefore, the correct answers are C and D. Implement query parallelization by partitioning the data output, and scale the SU count for the job up.

   upvoted 3 times

□  **akk_1289** 7 months ago

- C. Implement query parallelization by partitioning the data output.
- D. Scale the SU count for the job up.

By partitioning the data output, the query processing can be split into smaller, parallel tasks which can lead to better performance. Scaling up the SU count for the job increases the processing power available for the job, which can also lead to improved performance.

Note: The specific optimizations required may vary based on the specific requirements and nature of the data analysis solution.

   upvoted 1 times

□  **VivekMadas** 8 months, 2 weeks ago

Already 120 SU used (6 per node = 20 nodes) - Adding extra wont be any use.

Answer would be Partitioning Input & Output

   upvoted 4 times

□  **vigilante89** 8 months, 2 weeks ago

Selected Answer: CD

D - Scale the SU count for the job up

Because the assigned streaming units might be low as compared to the streaming data analytics requirements.

C - Implement parallelization by partitioning the output data

Because any data to be analyzed should be partitioned for query optimization.

   upvoted 1 times

□  **OldSchool** 9 months, 2 weeks ago

Selected Answer: CF

Because there is no mention of any bottleneck timestamping of stream data my answer is CF

   upvoted 1 times

□  **cosarac** 9 months, 2 weeks ago

My answer is CF.

As per the shared document we can't utilize parallelism with just partitioning the input when we don't partition the output. If you can't utilize parallelism what is the point of partitioning the input

"Most of the outputs supported in Stream Analytics can take advantage of partitioning. If you use an output type that doesn't support partitioning your job won't be embarrassingly parallel. For Event Hubs output, ensure Partition key column is set to the same partition key used in the query. Refer to the output section for more details."

   upvoted 2 times

□  **dmitriypo** 10 months, 1 week ago

The answer is CF.

Read section "Embarrassingly parallel jobs" of the article:

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

   upvoted 3 times

□  **Deeksha1234** 1 year, 1 month ago

Selected Answer: DF

D and F seems to be correct based on the reference article

   upvoted 3 times

✉️  **HebaN** 1 year, 3 months ago

Selected Answer: DF

As the event consumer is not mentioned, partitioning output is not relevant. For instance, Power BI output doesn't currently support partitioning

   upvoted 4 times

✉️  **Aditya0891** 1 year, 2 months ago

A stream analytics job must include at least 1 input, 1 query and 1 output. Check this it's mentioned in the first part itself "<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>". So by default we need to take the output as well in consideration or else the whole streaming analytics won't work

   upvoted 1 times

✉️  **Aditya0891** 1 year, 2 months ago

sorry my bad you are right, all the output don't have partition so we can't decide the output on our own. The answer should be DF

   upvoted 1 times

✉️  **Andushi** 1 year, 4 months ago

Selected Answer: CF

I agree with @manquak.

   upvoted 1 times

✉️  **DingDongSingSong** 1 year, 5 months ago

I think the answer is correct. The two things you do is: 1. Scale up SU and 2. partition input. If this doesn't work, THEN you could partition output as well.

   upvoted 4 times

ET Answer: DF

Bing: CF

To optimize performance for the Azure Stream Analytics job, you should perform the following two actions: C.

Implement query parallelization by partitioning the data output and **F. Implement query parallelization by partitioning the data input**.

Query parallelization can be achieved by partitioning the data input and output **1**. Partitioning the data input allows you to divide the workload of queries into multiple streaming nodes, which can reduce the overall execution time **2**. Partitioning the data output allows each shard to be processed independently, which enables you to linearly scale out your streaming pipeline **3**.

Implementing event ordering (A) or Azure Stream Analytics user-defined functions (UDF) (B) may improve the accuracy of your results, but they are not directly related to performance optimization. Scaling the SU count for the job up (D) or down (E) may affect performance, but it is not a direct way to optimize performance. Is there anything else you would like to know? 😊

Question 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

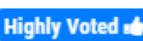
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion 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>

  **jv2120**  1 year, 9 months ago

Correct. C

Azure Event Grids – Event-driven publish-subscribe model (think reactive programming)

Azure Event Hubs – Multiple source big data streaming pipeline (think telemetry data)

In this case its more suitable vs Event Hubs.

   upvoted 26 times

Question 4

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: B

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

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

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

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

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

Reference:

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

<https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs>

Question 5

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
WITH LastInWindow AS
(
    SELECT
        

|        |                  |
|--------|------------------|
| (Time) | AS LastEventTime |
| COUNT  |                  |
| MAX    |                  |
| MIN    |                  |
| TOPONE |                  |


    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        

|                |
|----------------|
| (minute, 10)   |
| HoppingWindow  |
| SessionWindow  |
| SlidingWindow  |
| TumblingWindow |


)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON 

|                                                |
|------------------------------------------------|
| (minute, Input, LastInWindow) BETWEEN 0 AND 10 |
| DATEADD                                        |
| DATEDIFF                                       |
| DATENAME                                       |
| DATEPART                                       |


    AND Input.Time = LastInWindow.LastEventTime
```

Answer:

Answer Area

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

Box 1: MAX -

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

Query:

```
WITH LastInWindow AS -
```

```
(
```

```
SELECT -
```

```
MAX(Time) AS LastEventTime -
```

```
FROM -
```

```
Input TIMESTAMP BY Time -
```

```
GROUP BY -  
TumblingWindow(minute, 10)  
)
```

```
SELECT -  
Input.License_plate,  
Input.Make,  
  
Input.Time -  
  
FROM -  
  
Input TIMESTAMP BY Time -
```

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

Box 2: TumblingWindow -

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

Box 3: DATEDIFF -

DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.

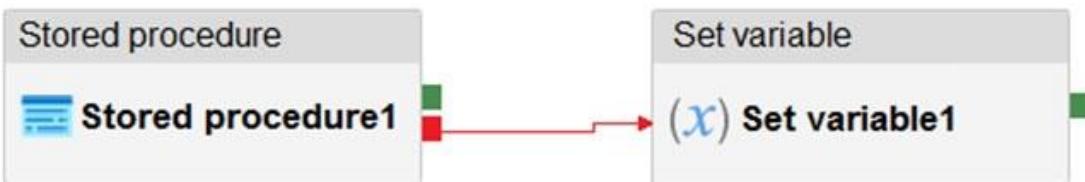
Reference:

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

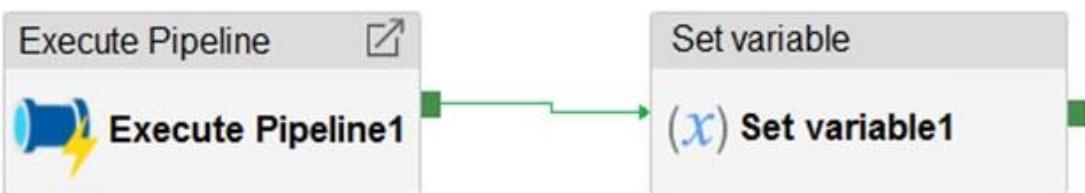
Question 6

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?

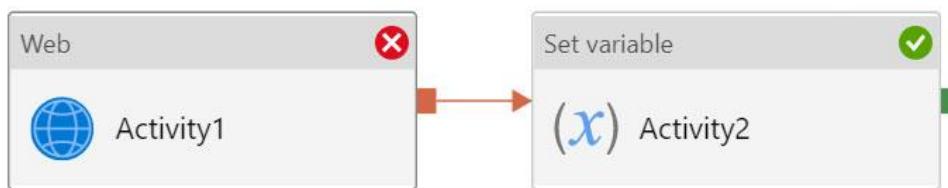
- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

Answer: A

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

Consider Pipeline1:

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



The failure dependency means this pipeline reports success.

Note:

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

<https://datasawy.me/category/azure-data-factory/>

SaferSephys Highly Voted 1 year, 12 months ago

Correct answer is A. The trick is the fact that pipeline 1 only has a Failure dependency between de activity's. In this situation this results in a Succeeded pipeline if the Stored procedure failed.

If also the success connection was linked to a follow up activity, and the SP would fail, the pipeline would be indeed marked as failed.

So A.

upvoted 42 times

Question 7

HOTSPOT -

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

Ingest:

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

Store:

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

Prepare and Train:

- ☞ Provide a fully-managed and interactive workspace for exploration and visualization.
- ☞ Provide the ability to program in R, SQL, Python, Scala, and Java.
- ☞ Provide seamless user authentication with Azure Active Directory.

Model & Serve:

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

You need to build the data integration pipeline.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Answer:

Answer Area

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

Ingest: Azure Data Factory -

Azure Data Factory pipelines can execute SSIS packages.

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

Store: Data Lake Storage -

Data Lake Storage Gen1 provides unlimited storage.

Note: Data at rest includes information that resides in persistent storage on physical media, in any digital format. Microsoft Azure offers a variety of data storage solutions to meet different needs, including file, disk, blob, and table storage. Microsoft also provides encryption to protect Azure SQL Database, Azure Cosmos DB, and Azure Data Lake.

Prepare and Train: Azure Databricks

Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.

With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace.

Azure Databricks supports Python, Scala, R, Java and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch and scikit-learn.

Model and Serve: Azure Synapse Analytics

Azure Synapse Analytics/ SQL Data Warehouse stores data into relational tables with columnar storage.

Azure SQL Data Warehouse connector now offers efficient and scalable structured streaming write support for SQL Data Warehouse. Access SQL Data Warehouse from Azure Databricks using the SQL Data Warehouse connector.

Note: As of November 2019, Azure SQL Data Warehouse is now Azure Synapse Analytics.

<https://docs.microsoft.com/bs-latn-ba/azure/architecture/data-guide/technology-choices/pipeline-orchestration-data-movement>

<https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks>

Question 8

DRAG DROP -

You have the following table named Employees.

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

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
CASE	*
ELSE	WHEN hire_date >= '2019-01-01' THEN 'New'
OVER	ELSE 'standard'
PARTITION BY	END AS employee_type
ROW_NUMBER	FROM
	employees

Answer:

Values	Answer Area
CASE	*
ELSE	CASE
OVER	WHEN hire_date >= '2019-01-01' THEN 'New'
PARTITION BY	ELSE 'standard'
ROW_NUMBER	END AS employee_type
	FROM
	employees

Box 1: CASE -

CASE evaluates a list of conditions and returns one of multiple possible result expressions. CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.

Syntax: Simple CASE expression:

CASE input_expression -

WHEN when_expression THEN result_expression [...n]

[ELSE else_result_expression]

END -

Box 2: ELSE -

Reference:

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

Question 9

DRAG DROP -

You have an Azure Synapse Analytics workspace named WS1.

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

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

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

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.

You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
	select*
	FROM
opendatasource	openrowset (
openjson	BULK 'https://contoso.blob.core.windows.net/contosodw', FORMAT= 'CSV', fieldterminator = '0x0b', fieldquote = '0x0b', rowterminator = '0x0b'
openquery) with (id varchar(50), contextdateventTime 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'
openrowset) 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')

Answer:

Values	Answer Area
	select*
	FROM
opendatasource	openrowset (
	BULK 'https://contoso.blob.core.windows.net/contosodw', FORMAT= 'CSV', fieldterminator = '0x0b', fieldquote = '0x0b', rowterminator = '0x0b'
openquery) with (id varchar(50), contextdateventTime 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')

Box 1: openrowset -

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

Example:

```
SELECT *
FROM OPENROWSET(
BULK 'csv/population/population.csv',
DATA_SOURCE = 'SqlOnDemandDemo',
FORMAT = 'CSV', PARSER_VERSION = '2.0',
```

```
FIELDTERMINATOR = ',',
ROWTERMINATOR = '\n'
```

Box 2: openjson -

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

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

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file>

<https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server>

Question 10

DRAG DROP -

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

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

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

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

How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.

You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values 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
```

Values:

- CAST
- COLLATE
- CONVERT
- FLATTEN
- PIVOT
- UNPIVOT

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 ( CAST (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
```

Box 1: PIVOT -

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

Incorrect Answers:

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

Box 2: CAST -

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

Example:

SELECT -

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

Here is the result:

decimal_value

12.00

Reference:

<https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/>

<https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot>

Question 11

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

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.

<https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/>

Question 12

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	<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:

Answer Area

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

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.

<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>

✉ GameLift Highly Voted 1 year, 10 months ago

FROM DP-201, thanks to rmk4ever ::

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.

ref: <https://adatis.co.uk/databricks-cluster-sizing/>

2. NO

recommended: New Job Cluster.

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.

ref: <https://docs.microsoft.com/en-us/azure/databricks/jobs>

Scheduled batch workload- Launch new cluster via job

ref: <https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html#plan-capacity-and-control-cost>

3.YES

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

ref: <https://docs.databricks.com/delta/index.html>

   upvoted 47 times

Question 13

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs.

You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.

What should you recommend?

- A. Azure Synapse Analytics
- B. Azure Databricks
- C. Azure Stream Analytics
- D. Azure SQL Database

ET User:

B (75%) **C (25%)**

kolakone 1 year, 11 months ago

My answer will be B

Stream Analytics supports "extending SQL language with JavaScript and C# user-defined functions (UDFs)". There is no mention of Python support; hence Stream Analytics is not correct.

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

Azure Databricks supports near real-time data from Azure Event Hubs. And includes support for R, SQL, Python, Scala, and Java. So I will go for option B.

upvoted 81 times

anto69 1 year, 6 months ago

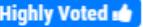
But Python runs on Event Hubs why the other service does should support Python too?

upvoted 2 times

Aditya0891 1 year, 2 months ago

It's mentioned that "python runs on real time data from event hubs not on event hubs". Also event hub is to gather that data and after that it is analyzed by either databricks stream analytics. And since stream analytics doesn't support python so the answer is databricks

upvoted 2 times

  **anto69**  1 year, 6 months ago

I'm sure it's Stream Analytics cause Event Hubs already supports Python (<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-python-get-started-send>). We don't need the other service to support it. We just need to lower costs. Hence ASA is the correct solution

   upvoted 13 times

  **RoyP654** 2 months, 3 weeks ago

the question does not ask which service can run Python, it's asking where to send the data for analytics since Python can run with Event Hubs libraries

   upvoted 1 times

ET Answer: C

<https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics>

Bing: B

For a statistical analysis solution that uses custom proprietary Python functions on near real-time data from Azure Event Hubs, I would recommend using **Azure Databricks**. Azure Databricks is a fully managed, fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure ¹. It provides a collaborative environment where data scientists, data engineers, and business analysts can work together to build and productionize machine learning models using custom Python code. Azure Databricks also integrates with Azure Event Hubs to enable real-time data ingestion and processing, which can help minimize latency in your solution. So, the correct answer is **B**.
Azure Databricks. Is there anything else you would like to know? 😊

Question 14

HOTSPOT -

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

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

- ☞ Create four partitions based on the order date.
 - ☞ Ensure that each partition contains all the orders placed during a given calendar year.
- How should you complete the T-SQL command? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Answer:

Answer Area

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



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

✉ **Canary_2021** 1 year, 8 months ago

Answer is correct.

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

upvoted 22 times

✉ **victor90** 1 year, 8 months ago

I think the box 2 should be 20090101,2010101,20110101,20120101 since the question asked about 4 partitions.

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15#c-creating-a-range-right-partition-function-on-a-datetime-column>

upvoted 14 times

✉ **TestMitch** 1 year, 8 months ago

No! That's wrong! Number of partitions created = Number of partition boundaries specified + 1.

upvoted 23 times

Question 15

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

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

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

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

NOTE: Each correct selection is worth one point.

- A. [EffectiveStartDate] [datetime] NOT NULL,
- B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,
- C. [EffectiveEndDate] [datetime] NULL,
- D. [ProductCategory] [nvarchar] (100) NOT NULL,
- E. [OriginalProductCategory] [nvarchar] (100) NOT NULL,

Answer: BE, 重复题

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.

Question 16

HOTSPOT -

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Answer:

Answer Area

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

Box 1: DATEDIFF -

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

Syntax: DATEDIFF (datepart , startdate, enddate)

Box 2: LAST -

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

Example:

```
SELECT -  
[user],  
feature,  
DATEDIFF(  
second,  
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,  
1) WHEN Event = 'start'),
```

Time) as duration -

FROM input TIMESTAMP BY Time -

WHERE -

Event = 'end'

Reference:

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

 **Miris**  2 years, 2 months ago

correct

   upvoted 42 times

 **ohana**  1 year, 10 months ago

Took the exam today. This question came out.

Ans: DateDiff, Last

   upvoted 23 times

Question 17

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.

The data flow already contains the following:

- ☞ A source transformation.
- ☞ A Derived Column transformation to set the appropriate types of data.
- ☞ A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements:

- ☞ All valid rows must be written to the destination table.
- ☞ Truncation errors in the comment column must be avoided proactively.
- ☞ Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

NOTE: Each correct selection is worth one point.

- A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
- B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
- C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
- D. Add a select transformation to select only the rows that will cause truncation errors.

Answer: AB

B: Example:

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

Any row that is larger than five will go into the BadRows stream.

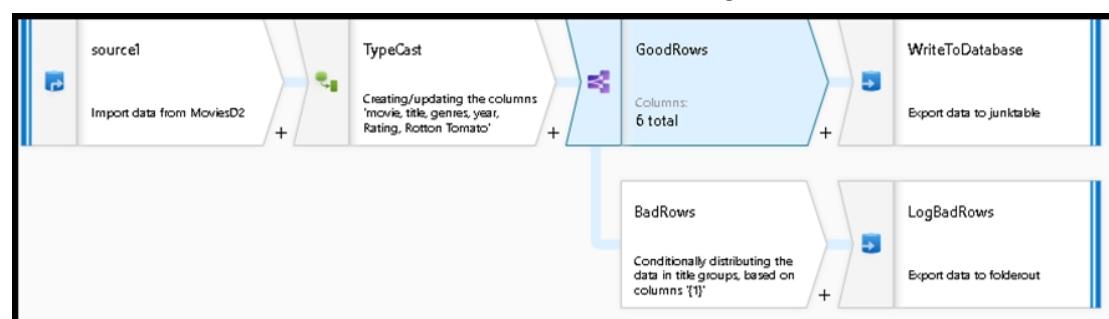
2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

A:

3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".

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

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



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

Question 18

DRAG DROP -

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
all, ecommerce, retail, wholesale	CleanData split(
dept=='ecommerce', dept=='retail', dept=='wholesale'	[]
dept=='ecommerce', dept== 'wholesale', dept=='retail'	[]
disjoint: false) ~> SplitByDept@([])
disjoint: true	
ecommerce, retail, wholesale, all	

ET User:

Box 2: Disjoint=False

Alekx42 Highly Voted 2 years, 3 months ago

I think "disjoint" should be True, so that data can be sent to all matching conditions. In this way the "all" output can get the data from every department, which ensures that "data can also be processed by the entire company".

upvoted 73 times

Stevyke 2 years, 2 months ago

I concur with @Alekx42 thought. Since we want to process for each dept (3 streams), then we must ensure we can still process for ALL depts at the same time (4th or default stream), hence DISJOINT:TRUE. Else, DISJOINT:FALSE.

upvoted 9 times

ET Answer:

Values	Answer Area
all, ecommerce, retail, wholesale	CleanData split(
dept=='ecommerce', dept=='retail', dept=='wholesale'	dept=='ecommerce', dept=='retail', dept=='wholesale'
dept=='ecommerce', dept== 'wholesale', dept=='retail'	disjoint: false
disjoint: false) ~> SplitByDept@([ecommerce, retail, wholesale, all])
disjoint: true	
ecommerce, retail, wholesale, all	

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>

Question 19

DRAG DROP -

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

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

You create the following components:

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

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

Select and Place:

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.	Read the file into a data frame.
Perform transformations on the file.	Perform transformations on the data frame.
Specify a temporary folder to stage the data.	Specify a temporary folder to stage the data.
Write the results to Data Lake Storage.	Write the results to a table in Azure Synapse.
Read the file into a data frame.	
Drop the data frame.	
Perform transformations on the data frame.	

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>

✉️  **sagga**  2 years, 3 months ago

I think the correct order is:

- 1) mount onto DBFS
- 2) read into data frame
- 3) transform data frame
- 4) specify temporary folder
- 5) write to table in SQL data warehouse

About temporary folder, there is a note explain this:

<https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse#load-data-into-azure-synapse>

Discussions about this question:

<https://www.examtopics.com/discussions/microsoft/view/11653-exam-dp-200-topic-2-question-30-discussion/>

   upvoted 163 times

✉️  **andylop04** 2 years, 2 months ago

Today I received this question in my exam. Only appeared the 5 options of this response.
I only had to order, not choice. This solutions is the correct. Thanks sagga.

   upvoted 38 times

Question 20

HOTSPOT -

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

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

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Type:	<input type="button" value="▼"/>				
<table border="1"><tr><td>Event</td></tr><tr><td>On-demand</td></tr><tr><td>Schedule</td></tr><tr><td>Tumbling window</td></tr></table>		Event	On-demand	Schedule	Tumbling window
Event					
On-demand					
Schedule					
Tumbling window					
Additional properties:	<input type="button" value="▼"/>				
<table border="1"><tr><td>Prefix: /in/, Event: Blob created</td></tr><tr><td>Recurrence: 30 minutes, Start time: 2021-01-01T00:00</td></tr><tr><td>Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes</td></tr><tr><td>Recurrence: 32 minutes, Start time: 2021-01-15T01:45</td></tr></table>		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
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:

Answer Area

Type:	<input type="button" value="▼"/>				
<table border="1"><tr><td>Event</td></tr><tr><td>On-demand</td></tr><tr><td>Schedule</td></tr><tr><td>Tumbling window</td></tr></table>		Event	On-demand	Schedule	Tumbling window
Event					
On-demand					
Schedule					
Tumbling window					
Additional properties:	<input type="button" value="▼"/>				
<table border="1"><tr><td>Prefix: /in/, Event: Blob created</td></tr><tr><td>Recurrence: 30 minutes, Start time: 2021-01-01T00:00</td></tr><tr><td>Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes</td></tr><tr><td>Recurrence: 32 minutes, Start time: 2021-01-15T01:45</td></tr></table>		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
Prefix: /in/, Event: Blob created					
Recurrence: 30 minutes, Start time: 2021-01-01T00:00					
Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes					
Recurrence: 32 minutes, Start time: 2021-01-15T01:45					

Box 1: Tumbling window -

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

Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

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

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

Reference:

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

Question 21

HOTSPOT -

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

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

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

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Azure Stream Analytics input type:

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

Azure Stream Analytics output type:

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

Aggregation query location:

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

Answer:

Answer Area

Azure Stream Analytics input type:

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

Azure Stream Analytics output type:

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

Aggregation query location:

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

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

Question 22

DRAG DROP -

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

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

Select and Place:

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

ET User:

- 1 Add an Azure Stream Analytics Customer Deserializer Project(.net) project to the Solution
- 2 Add .net deseriliaizer Code to ProtoBuf to customer deserializer project
3. Change the event Serialization format to protobuf in the input.json File of the job and reference the DLL.

 **zarga** Highly Voted 2 years, 1 month ago

The third one is wrong because the stream analytics application already exist in the project. The goal is to modify the current stream analytics application in order to read protobuf data.

I think the right answer is the first one in the list (update input.json file and reference dll)

   upvoted 60 times

ET Answer:

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

Question 23

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

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

Incorrect Answers:

C: Self-hosted integration runtime is to be used On-premises.

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

Question 24

HOTSPOT -

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

HubA:

Stream
Reference

HubB:

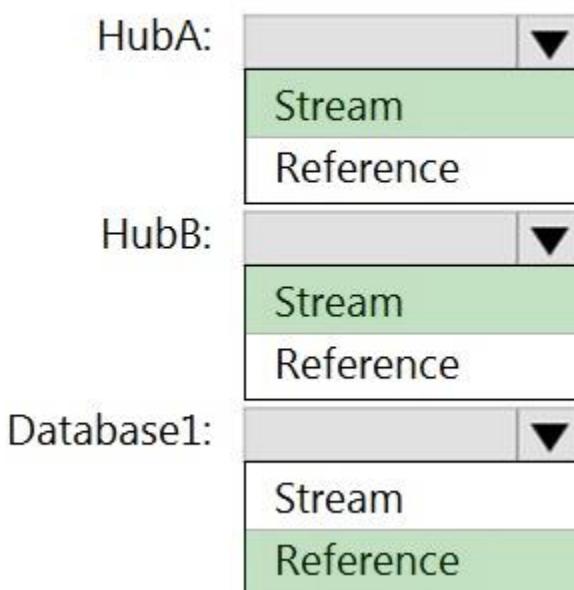
Stream
Reference

Database1:

Stream
Reference

Answer:

Answer Area



HubA: Stream -

HubB: Stream -

Database1: Reference -

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

Reference:

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

Question 25

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

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.

Incorrect Answers:

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

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

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

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

Question 26

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT sensorId,
       growth = reading -
               ▼ (reading) OVER (PARTITION BY sensorId ▼ (hour,1))
               □ LAG
               □ LAST
               □ LEAD
               □ LIMIT DURATION
               □ OFFSET
               □ WHEN
FROM input
```

Answer:

Answer Area

```
SELECT sensorId,
       growth = reading -
               ▼ (reading) OVER (PARTITION BY sensorId ▼ (hour,1))
               □ LAG
               □ LAST
               □ LEAD
               □ LIMIT DURATION
               □ OFFSET
               □ WHEN
FROM input
```

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>

Question 27

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. event

Answer: D

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account.

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger>

Question 28

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

In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.

Note: The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.

1. In Azure DevOps, open the project that's configured with your data factory.
2. On the left side of the page, select Pipelines, and then select Releases.
3. Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
4. In the Stage name box, enter the name of your environment.
5. Select Add artifact, and then select the git repository configured with your development data factory. Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.
6. Select the Empty job template.

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

Question 29

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

Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.

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

 **trungngonptit**  2 years, 2 months ago

correct, blob storage or azure sql database

   upvoted 30 times

Question 30

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: C

Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consist of three parameters: the timeunit, the windowsize (how long each window lasts) and the hopsize (by how much each window moves forward relative to the previous one).

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

Question 31

HOTSPOT -

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Service:	<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

ET User:

1. Azure Stream Analytics
2. No Window
3. Point within Polygon

 **Alekx42** Highly Voted 2 years, 3 months ago

You do not need a Window function. You just process the data and perform the geospatial check as it arrives. See the same example here:

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

   upvoted 77 times

✉  **JackArmitage**  2 years, 2 months ago

1. Azure Stream Analytics

2. No Window

3. Point within Polygon

   upvoted 69 times

ET Answer:

Answer Area

Service:

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

Window:

- Hopping**
- No window
- Session
- Tumbling

Analysis type:

- Event pattern matching
- Lagged record comparison
- Point within polygon**
- Polygon overlap

Question 32

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.

ET User: ABC 都有人选，还有人认为应该是不存在的选项 Auto Loader，可能 A

B (84%)

A (16%)

✉  **bc5468521**  2 years, 3 months ago

The ABS-AQS source is deprecated. For new streams, we recommend using Auto Loader instead.

   upvoted 22 times

✉  **manquak**  2 years ago

Why not partition by date? What does the auto loader have to do with streaming jobs?

   upvoted 14 times

✉  **akhil5432**  3 weeks, 6 days ago

Selected Answer: B

option B

   upvoted 1 times

✉  **vctrhugo** 2 months, 1 week ago

Sinking to Azure Queue storage is not necessary for persisting the events in the Azure Databricks table. Azure Queue storage is typically used for decoupling and asynchronous messaging scenarios and may not directly contribute to minimizing storage costs or incremental load times for the Databricks table.

   upvoted 1 times

✉  **auwia** 2 months, 2 weeks ago

Selected Answer: B

Probably it is B:

Partition by date&time is not the best, imagine events with each single partition because of (day, hour, minute, second) => the requirement is clear, minimize the space, etc..

You use Watermark when you need to reduce the amount of state data to improve latency during a long-running streaming operation.

JSON I would exclude because how it is formulated.

My answer is B, even if it's deprecated, it's clear that this question is an old one, but looking at the comments, we can still get in the exam.

   upvoted 3 times

✉  **dksks** 4 months ago

Selected Answer: A

A. Partition by DateTime fields: Partitioning the table on frequently used columns such as DateTime fields can improve query performance and reduce incremental load times.

Partitioning by DateTime can help to reduce the amount of data scanned during query execution and facilitate incremental loading.

   upvoted 2 times

✉  **hiyoww** 5 months, 1 week ago

is the question outdated?

   upvoted 2 times

✉  **haidebelognime** 6 months, 3 weeks ago

Selected Answer: A

I'm sure it is A. Partition by DateTime!!

   upvoted 1 times

✉  **kckalahasthi** 9 months ago

<https://learn.microsoft.com/en-us/azure/databricks/structured-streaming/aqs>

   upvoted 2 times

✉  **Igor85** 9 months, 2 weeks ago

question is deprecated, AutoLoader is the way to do the incremental loads

   upvoted 3 times

✉  **RajashekharC** 1 year ago

As per requirement: "You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks."

What I understood from this is, dataset which will be stored would be used by Databricks and load type is incremental. Considering this, I see "watermark column" makes more sense.

   upvoted 3 times

✉  **Deeksha1234** 1 year, 1 month ago

B is correct

   upvoted 1 times

✉  **Aurelkb** 1 year, 3 months ago

Selected Answer: B

Correct

   upvoted 2 times

✉  **kamil_k** 1 year, 5 months ago

Do we want to persist this data in a table or in a message queue? From what the question asks it has to be a table. Why would we use queue storage for this task?

   upvoted 3 times

✉️  **Canary_2021** 1 year, 8 months ago

Selected Answer: B

A. Partition by DateTime field

Each partition will generate a file. Loading latency may reduce, but feel storage cost will increase because generate more folders and files for different partition. Is it right???

B. Sink to Azure Queue Storage.

Read this document. Spark table files are stored in DBFS. Mount Azure Blob storage containers to Databricks File System (DBFS). If The Databricks ABS-AQS provides these two benefits, sounds like it is a correct answer.

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

C. Include a watermark column: For sure it is not correct

Watermarks define how long your aggregate should wait around for data delay.

D. User a Json format for physical data storage. - ???

Don't find any documents to compare physical data storage of JSON, CSV, and Parquet.

   upvoted 10 times

✉️  **berserksap** 1 year, 10 months ago

I think ABS-AQS would be right answer though deprecated. If not it would be water mark or auto loader. As per my understanding Water mark columns use more space and hit the API accordingly many times when compared to ABS-AQS and is costly.

   upvoted 3 times

✉️  **kilowd** 1 year, 10 months ago

Why not C ??

   upvoted 1 times

✉️  **Amalbenrebai** 1 year, 12 months ago

I hesitate between A OR C

   upvoted 4 times

✉️  **belha** 2 years, 2 months ago

TRUE ???

   upvoted 2 times

ET Answer: B

The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files. This provides two major advantages:

⇒ Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive.

⇒ Lower costs: no more costly LIST API requests made to ABS.

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

To minimize storage costs and incremental load times for an Azure Databricks table that ingests an average of 20 million streaming events per day, you should **partition by DateTime fields**. Partitioning the data by DateTime fields can help reduce the amount of data that needs to be read during incremental loads, as only the relevant partitions need to be scanned. This can help minimize both storage costs and incremental load times ¹. So, the correct answer is

A. Partition by DateTime fields. Is there anything else you would like to know? 😊

Question 33

HOTSPOT -

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

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

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

The integration runtime has the following node details:

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

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

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

▼
raised
lowered
left as is

Answer: 有争议, 但没看懂

Answer Area

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

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

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

raised
lowered
left as is

Box 1: fail until the node comes back online

We see: High Availability Enabled: False

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

Box 2: lowered -

We see:

Concurrent Jobs (Running/Limit): 2/14

CPU Utilization: 6%

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

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

[Exam DP-203 topic 2 question 35 discussion - ExamTopics](#)

✉ **Sunnyb** Highly Voted 2 years, 3 months ago

$1/14 = 0.07$

$6\% = 0.06$

should be lowered.

👍 ↪ ⚡ upvoted 29 times

✉ **romanzdk** 1 year, 7 months ago

$0.06/2 = 0.03$

$0.03 * 14 = 0.42$ = maximally 42% of cpu for all jobs
isn't this better?

👍 ↪ ⚡ upvoted 9 times

✉ **semauni** 1 month ago

Why is this the calculation you make? I see 6% utilization, so 94% to go, so the amount can be raised.

👍 ↪ ⚡ upvoted 2 times

✉ **shachar_ash** 1 year, 3 months ago

The question mentions 2/14 which is 0.14, therefore it can be increased.

👍 ↪ ⚡ upvoted 4 times

✉️  **Mirandal**  2 years, 2 months ago

"We recommend that you increase the concurrent jobs limit only when you see low resource usage with the default values on each node."
<https://docs.microsoft.com/en-us/azure/data-factory/monitor-integration-runtime>

   upvoted 20 times

✉️  **Omkarrokee**  2 months, 1 week ago

Based on the information provided, the CPU Utilization is 6% and the Concurrent Jobs (Running/Limit) is 2/14. This indicates that the integration runtime is utilizing only 6% of the available CPU capacity and currently running 2 out of a maximum limit of 14 concurrent jobs.

Given this information, the appropriate answer choice for the completion statement would be:

Concurrent Job should be scaled up

Since the current CPU utilization is relatively low at 6% and there is still capacity available for running additional jobs, scaling up the concurrent job limit would allow for more jobs to run simultaneously and make better use of the available resources.

   upvoted 4 times

✉️  **auwia** 2 months, 2 weeks ago

We are talking about max number of job running in parallel!
If you have available resource of course it is recommended to raise up the current limit to afford future load.

Also Microsoft recommended that:

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

We recommend that you increase the concurrent jobs limit only when you see low resource usage with the default values on each node. I think this is the case, also the question doesn't tell you it's mandatory, what should! So I think we should follow recommendation and raise up the limit.

   upvoted 2 times

✉️  **pavankr** 3 months, 2 weeks ago

when the explanation is "scale up by increasing the number" then why the answer is "Lowered"???

   upvoted 2 times

✉️  **norbitek** 7 months, 3 weeks ago

I would leave it as it is.

See:

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

"The default value of the concurrent jobs limit is set based on the machine size. The factors used to calculate this value depend on the amount of RAM and the number of CPU cores of the machine. So the more cores and the more memory, the higher the default limit of concurrent jobs.

You scale out by increasing the number of nodes. When you increase the number of nodes, the concurrent jobs limit is the sum of the concurrent job limit values of all the available nodes. For example, if one node lets you run a maximum of twelve concurrent jobs, then adding three more similar nodes lets you run a maximum of 48 concurrent jobs (that is, 4 x 12). We recommend that you increase the concurrent jobs limit only when you see low resource usage with the default values on each node."

   upvoted 3 times

✉️  **martcery** 8 months, 2 weeks ago

the concurrent jobs limit is the sum of the concurrent job limit values of all the available nodes

   upvoted 1 times

✉️  **OldSchool** 9 months, 3 weeks ago

Here is my thinking of this. High availability is False, so no scaling. The 2nd Q is what should be done with the number of concurrent jobs, not scaling up CPU. Since there are only 2 running jobs of possible 14 and CPU utilization is only 6% the number of concurrent jobs should be increased. If left as is we are overspending, if decreased we are still overspending even more since CPU utilization will be lowered too.

   upvoted 3 times

✉️  **OldSchool** 9 months, 1 week ago

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.

<https://learn.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime?tabs=data-factory#scale-up>

   upvoted 1 times

✉️  **kamil_k** 1 year, 5 months ago

according to this article <https://docs.microsoft.com/en-us/azure/data-factory/monitor-integration-runtime#:~:text=When%20you%20increase%20the%20number,is%2C%204%20x%2012>.

it is not advisable to touch the default calculated limits unless we encounter issues..

   upvoted 1 times

✉️  **VyshakhUnnikrishnan** 1 year, 6 months ago

The CPU is only used 6% with 2 parallel jobs running. This gives the opportunity for the cluster to scale up the number of concurrent jobs. The number of parallel/concurrent jobs should hence be increased

   upvoted 4 times

✉️  **kamil_k** 1 year, 5 months ago

This question is poorly written, it doesn't paint the whole picture. You would need to monitor resource utilisation over a prolonged period of time e.g. 24 hours to see what happens at peak times when you can have all 14 spots taken. Each job can take different amount of compute power. For instance you may find that at some times one job consumes 50% CPU.

   upvoted 1 times

✉️  **edba** 1 year, 8 months ago

It sounds many people were confused regarding 2nd question. after check the link below, I think it means to lower or raise limit concurrent jobs. Apparently 14 is too high as usage is 2 only, so it should be "lowered". ref:<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime?tabs=data-factory#scale-considerations>

   upvoted 16 times

✉️  **Billybob0604** 9 months ago

Yes, this is the right explanation. the question is about lowering the limit not the running jobs.

   upvoted 4 times

✉️  **Dusica** 9 months, 2 weeks ago

agree with this; practically paying more capacity then they need

   upvoted 1 times

- sdokmak 1 year, 3 months ago
maybe they worded the question poorly deliberately to test you on what you can scale up or down, which would be the limit not the number of jobs.
 upvoted 3 times
- onyerleft 1 year, 8 months ago
1) Fail until the node comes back online
2) left as is
- For all those who are extrapolating CPU based on the 2 jobs that are running - you have no idea what the other 12 concurrent jobs could look like. You could have one additional job that maxes the CPU. You could have 12 easy jobs that bring it up to 10% utilization. Since we don't know, leave things as they are until one of the values becomes a bottleneck.
 upvoted 6 times
- Davico93 1 year, 2 months ago
You're right, a professional would say that, but I think Ms is making it easier. I would say also left as is, but that they have to monitor
- clement_ 1 year, 9 months ago
The question is poorly written. The value is expressed as a ratio, so lowering the value consists in increasing the denominator which is the Limit. Which is what I assume is done in practice when your 14/14 jobs would only consume 42% of your CPU.
 upvoted 2 times
- Marcus1612 1 year, 11 months ago
There is no reason to scale down the concurrency limit, because there is no cost impact. So, from an administration perspective, there is no benefit. On the other hand, if the resources were topped, you should scale up the node, but it is not the case in this scenario. The answer is keep this configuration as is.
 upvoted 4 times
- A1000 2 years ago
left as-is
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.
so there is no need to scale it up as we did not reach the limit and there is no benefit from scale it down.
 upvoted 5 times
- Jacob_Wang 2 years, 2 months ago
It might be the ratio. For instance, 2/14 might should be lowered to 2/20.
 upvoted 2 times
- saty_nl 2 years, 2 months ago
Concurrent jobs limit must be raised, as we are under-utilizing the provisioned capacity.
 upvoted 3 times
- damaldon 2 years, 2 months ago
A) is correct because of HA is set to FALSE
<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime#high-availability-and-scalability>
 upvoted 1 times
- terajuana 2 years, 2 months ago
the limit should be left as is to allow capacity for more jobs - a single job could use 20% CPU if it is running intensive work. The pricing model isn't by concurrency so there is no budget rationale to lower it.
 upvoted 1 times
- bc5468521 2 years, 3 months ago
2 jobs/node, but the CPU is not fully utilized; based on the workload, don't need too many concurrent jobs, so lower to 1 job/node
 upvoted 1 times
- dfdsfdsfsd 2 years, 3 months ago
I might be misunderstanding this but the way I look at it is that if 2 concurrent jobs use 6% of the CPU, then 1 job requires 3% CPU and you could have approximately 100/3=33 concurrent jobs. So you can raise the limit. What makes me insecure is that I imagine not every job would be equal in CPU-load.
 upvoted 8 times
- Alekx42 2 years, 3 months ago
I agree with your explanation.
I think lowering the limit makes no sense: the system is underloaded, why should you limit the parallelism that you could have when many jobs eventually get executed at the same time?
Maintaining the current value could be an option: there are no issues with the current configuration with respect to the maximum concurrent job value.
Increasing the value is good if we take as true your hypothesis that every job requires the same CPU %.
 upvoted 5 times
- AssilAbdulrahim 2 years, 3 months ago
CPU Utilization: 6%
Concurrent Jobs (Running/Limit): 2/14
I am also confused but I tend to adjust the explanation because the system still has very low utilization 6% and only 2 out of 14 concurrent jobs are there... Hence I might think it should be lowered...
Can you please explain why both of you think it should be raised?
 upvoted 1 times
- AssilAbdulrahim 2 years, 3 months ago
I meant the scalability of nodes should be lowered...
 upvoted 1 times
- tanza 2 years, 3 months ago
Concurrent jobs limit should be raised , no?
 upvoted 10 times

- ✉️  **MacronfromFrance** 2 years, 3 months ago
for me, it should be raised. I don't find explanation in the given link... :(
   upvoted 3 times
- ✉️  **alain2** 2 years, 3 months ago
IMO, it should be lowered because:
. Concurrent Jobs (Running/Limit): 2/14
. CPU Utilization: 6%
   upvoted 1 times
- ✉️  **Preben** 2 years, 2 months ago
If you eat 1 ice cream a day, but you buy 5 new ones every day -- should you increase the amount of ice cream you buy, or lower it? This is the same. You are paying for 14 concurrent jobs, but you are only using 2. You are only using 6 % of the CPU you have purchased, so you are paying for 94 % that you do not use.
   upvoted 14 times
- ✉️  **bsa_2021** 2 years, 2 months ago
The question is about the action w.r.t. concurrent jobs value. Concurrent jobs should be raised to make full use of resources. Also, (if possible) the resources should be lowered so that it is not wasted. I think the choice of answer raised/lowered should be based on the context and the context here is about the concurrent jobs, not resources. Hence, I think raised would be correct.
   upvoted 6 times
- ✉️  **Banach** 2 years, 1 month ago
I understand your point of view, and I understood the question in the same way you did at first. But after reading carefully the sentence it asks (as you said) about the limit value (or the settings) of concurrent jobs, knowing that you only use 6% of your CPU with only 2 concurrent jobs. Therefore, considering the waste of resources, "lowered" is, imo, the correct answer here (although the formulation of the question is a bit confusing, I admit).
   upvoted 5 times
- ✉️  **terajuana** 2 years, 2 months ago
data factory pricing is based on activity runs and not concurrency
   upvoted 2 times

Bing:

If the X-M node becomes unavailable, the executed pipelines will **fail until the node comes back online**. Since the integration runtime has only one node and high availability is not enabled, there are no other nodes or integration runtimes to switch to in case of a failure. So, the correct answer is **A. fail until the node comes back online**. Is there anything else you would like to know? 😊

Based on the information you provided, the current CPU utilization of the X-M node is **6%** and the number of concurrent jobs running is **2 out of a limit of 14**. This suggests that the node is not being fully utilized and there is room for increasing the number of concurrent jobs to improve performance. However, it's important to monitor the CPU utilization, memory usage, and network usage to ensure that they do not become bottlenecks as the number of concurrent jobs is increased. So, the correct answer is **A. raised**. Is there anything else you would like to know? 😊

Question 34

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

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.

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

Question 35

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account.

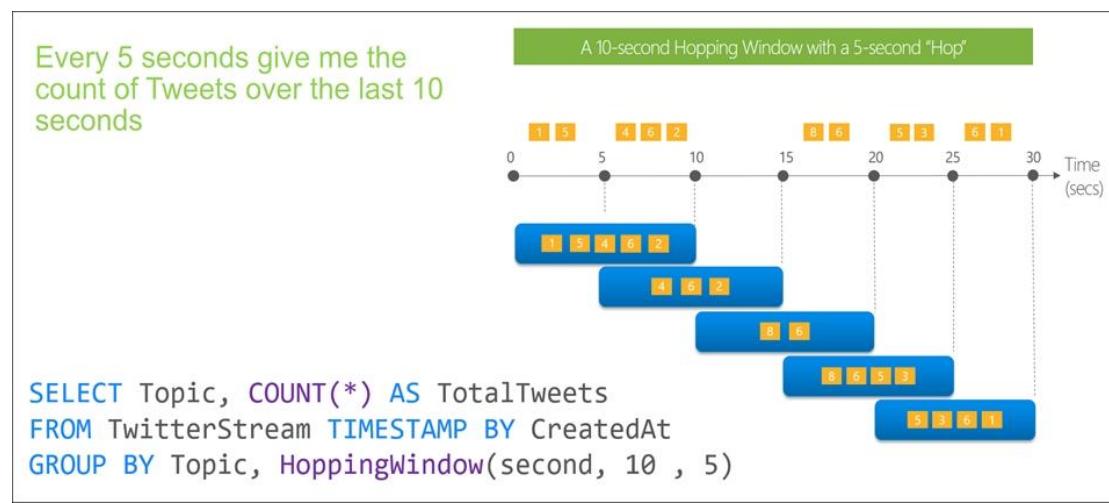
You need to output the count of records received from the last five minutes every minute.

Which windowing function should you use?

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

Answer: D

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



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

Question 36

HOTSPOT -

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

Git repository
Git 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.

Hot Area:

Answer Area

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

/	▼
adf_publish	
main	
Parameterization template	

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

/	▼
contososales	
/dwh_batchetl/adf_publish/contososales	
/main	

Answer:

Answer Area

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

/	▼
adf_publish	
main	
Parameterization template	

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

/	▼
contososales	
/dwh_batchetl/adf_publish/contososales	
/main	

Box 1: adf_publish -

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows.

You can create a new repository or use an existing repository that's already in your project.

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

Question 37

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

▼
LAST
OVER
SYSTEM.TIMESTAMP()
TIMESTAMP BY

CreatedAt

GROUP BY TimeZone,

▼
HOPPINGWINDOW
SESSIONWINDOW
SLIDINGWINDOW
TUMBLINGWINDOW

(second,15)

Answer:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

▼
LAST
OVER
SYSTEM.TIMESTAMP()
TIMESTAMP BY

CreatedAt

GROUP BY TimeZone,

▼
HOPPINGWINDOW
SESSIONWINDOW
SLIDINGWINDOW
TUMBLINGWINDOW

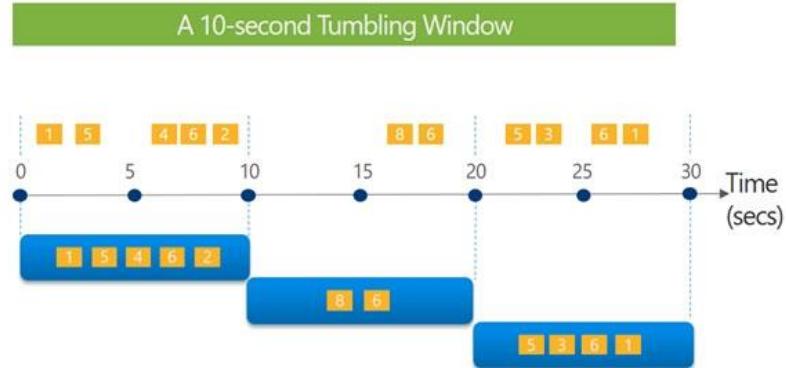
(second,15)

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.

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)
```

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

Question 38

HOTSPOT -

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

ADF1 contains the following pipelines:

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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 | ▼ |

ET User:

Set the Partition option to Dynamic range

Set the Copy method to PolyBase

✉️  **marcin1212**  1 year, 8 months ago

how to use PolyBase when copy data from Synapse to file ? I don't have idea.
Moreover PolyBase option is available only when the target is Synapse

it should be

P1: Set the partition option to "Dynamic range "

P2: PolyBase

regarding to P1

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#parallel-copy-from-synapse-analytics>

Scenario: "Full load from large table, without physical partitions.."->

Suggested settings: Partition options: Dynamic range partition.

   upvoted 80 times

✉️  **Canary_2021**  1 year, 8 months ago

P1: Copy data from SQL to Data Lake.

- Bulk insert and PolyBase are not a choice in Sink tab if target is Data Lake. So they are not correct.

- Isolation level can be setup if SQL database is the source. Repeatable Read means that locks are placed on all data that is used in a query. Don't think it maximizes parallelism and performance.

- Set the Partition option to Dynamic range

Can be setup if source is SQL in copy activity. And it maximizes parallelism and performance. So I select this option.

P2: Copy data from Data Lake to SQL. It is for sure to select PolyBase.

   upvoted 41 times

ET Answer:

Answer Area

P1:

- | | |
|---|---|
| <p>Set the Copy method to Bulk insert</p> <p>Set the Copy method to PolyBase</p> <p>Set the Isolation level to Repeatable read</p> <p>Set the Partition option to Dynamic range</p> |  |
|---|---|

P2:

- | | |
|---|---|
| <p>Set the Copy method to Bulk insert</p> <p>Set the Copy method to PolyBase</p> <p>Set the Isolation level to Repeatable read</p> <p>Set the Partition option to Dynamic range</p> |  |
|---|---|

Question 39

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Load methodology:

Full Load	▼
Incremental Load	
Load individual files as they arrive	

Trigger:

Fixed schedule	▼
New file	
Tumbling window	

Answer:

Answer Area

Load methodology:

Full Load	▼
Incremental Load	
Load individual files as they arrive	

Trigger:

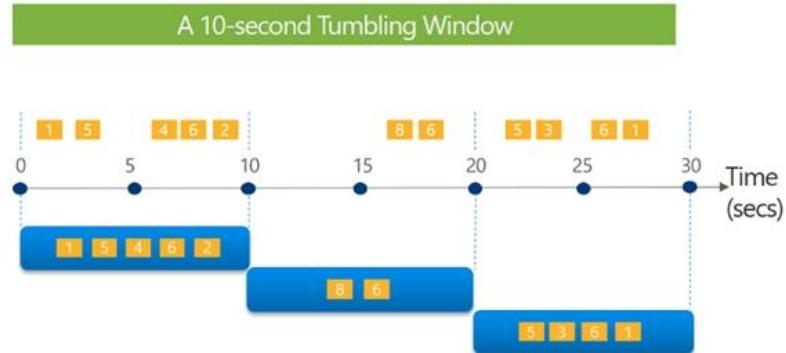
Fixed schedule	▼
New file	
Tumbling window	

Box 1: Incremental load -

Box 2: Tumbling window -

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

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



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

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

Question 40

You have the following Azure Data Factory pipelines:

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

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System2. Populate Facts must execute after Populate Dimensions pipeline. All the pipelines must execute every eight hours.

What should you do to schedule the pipelines for execution?

- A. Add an event trigger to all four pipelines.
- B. Add a schedule trigger to all four pipelines.
- C. Create a parent pipeline that contains the four pipelines and use a schedule trigger.
- D. Create a parent pipeline that contains the four pipelines and use an event trigger.

Answer: C

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

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

  **onyerleft**  1 year, 8 months ago

Selected Answer: C

C is correct, but with poor wording. Should be 'parent pipeline' with a schedule trigger.

The parent pipeline has 4 execute pipeline activities. Ingest 1 and Ingest 2 have no dependencies. Dimension pipeline has two dependencies from 'on completion' outputs of both Ingest 1 and Ingest 2 pipelines. Fact pipeline has one 'on completion' dependency on the Dimension pipeline. Absolutely nothing to do with a tumbling window trigger

   upvoted 58 times

Question 41

DRAG DROP -

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

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

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

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

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

Select and Place:

Components	Answer Area
a database scoped credential	
an asymmetric key	
an external data source	
a database encryption key	
an external file format	

ET User: 有 3 个投票比较多

alexleonvalencia 1 year, 8 months ago

- 1.- A database scoped credential
- 2.- an External data sorce
- 3.- a external file format

upvoted 166 times

engrbrain 1 year, 8 months ago

According to the documentation, the first thing you are to create is
CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'S0me!nfo';
I don't think this is means an asymmetric key. It is simply a database encryption key. So I think the answer is

- 1- Create a Database Encryption Key
- 2 - Create a Database Scoped Credential
- 3 - Create an External Data Source

upvoted 52 times

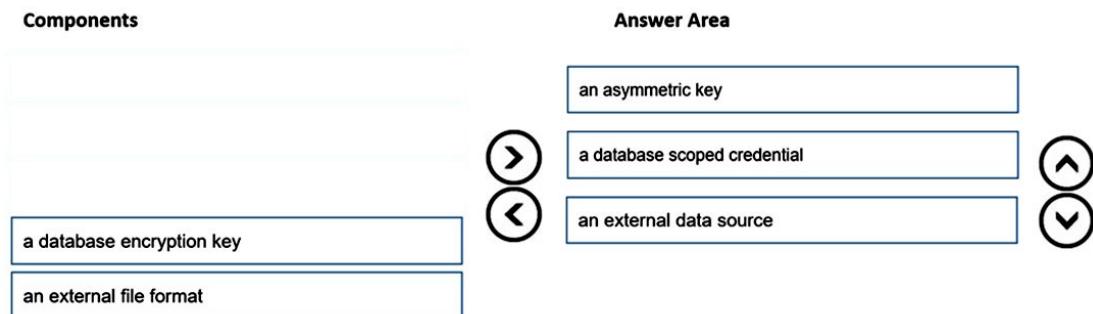
 **Franz58** 1 year, 1 month ago

you need to connect to the DW, not to a specific file. Therefore :

- 1- Create a Database Encryption Key
- 2 - Create a Database Scoped Credential
- 3 - Create an External Data Source

   upvoted 18 times

ET Answer:



Question 42

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

Watermark Delay indicates the delay of the streaming data processing job.

There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

1. Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
2. Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
3. Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.

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

Question 43

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

SELECT

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))

ET User:

TopOne()

Tumbling

  alexleonvalencia Highly Voted 1 year, 8 months ago

TopOne() / Tumbling

   upvoted 114 times

  gf2tw Highly Voted 1 year, 8 months ago

Syntax for Hopping window requires 3 arguments, seems this should be Tumbling Window which fulfills the exact same requirements.

   upvoted 38 times

ET Answer:

Answer Area

```
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))
```

Bing:

To ensure that an Azure Stream Analytics job returns the highest scoring record for each five-minute time interval of each game, you can use a **Tumbling Window** function in your Stream Analytics query. A tumbling window is a type of windowing function that returns distinct, non-overlapping windows of time. You can specify the size of the tumbling window as five minutes to divide the data into five-minute time intervals. Then, you can use an aggregate function such as `MAX` to return the highest score within each window. Here's an example of how you can complete the Stream Analytics query:

SQL

⌚ 此代码由 AI 生成。仔细查看和使用。有关详细信息，请[访问我们的常见问题解答](#)。

```
SELECT  
    System.Timestamp() AS WindowEnd,  
    MAX(score) AS MaxScore  
FROM input TIMESTAMP BY timestamp  
GROUP BY TUMBLINGWINDOW(minute, 5)
```

In this example, `input` is the name of the input stream and `timestamp` is the name of the timestamp column in the input data. The query uses a tumbling window of size five minutes to divide the data into non-overlapping windows and returns the maximum score within each window. Is there anything else you would like to know? 😊

Question 44

You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:

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

Which cluster type should you recommend?

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

ET User:



✉ **Canary_2021** 1 year, 8 months ago

Selected Answer: B

B is correct answer.

High concurrency cluster cannot terminate automatically by default, so C is wrong.

Standard cluster cannot be shared by multiple tasks, so A and D are wrong.

upvoted 14 times

✉ **HaBroNounen** 1 year, 8 months ago

"High Concurrency clusters do not terminate automatically by default."

but you can change that default so your argument about C is incorrect..

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

upvoted 14 times

✉ **_lene_** 4 months, 4 weeks ago

Selected Answer: C

The cluster does auto-scaling by default. Auto-termination should be set up manually

upvoted 5 times

ET Answer: B

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

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

Incorrect Answers:

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

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

High Concurrency clusters do not terminate automatically by default.

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

Bing:

To design an Azure Databricks cluster that runs user-defined local processes and meets the requirements of minimizing query latency, maximizing the number of users that can run queries on the cluster at the same time, and reducing overall costs without compromising other requirements, you should recommend a **High Concurrency with Autoscaling** cluster type. A High Concurrency cluster is optimized for running multiple concurrent queries and is designed to provide fine-grained resource sharing for maximum resource utilization ¹. Autoscaling allows the cluster to dynamically scale up or down based on the workload, which can help reduce costs while still providing good performance ¹. So, the correct answer is **B. High Concurrency with Autoscaling**. Is there anything else you would like to know? 😊

Question 45

HOTSPOT -

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Parameter:

<input type="checkbox"/>	▼

Naming pattern:

<input type="checkbox"/>	▼

Copy behavior:

<input type="checkbox"/>	▼

ET User:

- 1) `@trigger().outputs.windowStartTime`
- 2) `/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json`
- 3) Merge files

✉️  **ItHYMeRish**  1 year, 8 months ago

The correct copy behavior is merge - not flatten hierarchy.

The question starts with a folder structure as the following:

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

It indicates there are multiple device ID JSON files per deviceType. Those need to be merged to get the target naming pattern - "one file per device type per hour."

The target naming pattern is the following:

`/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}_{{deviceType}}.json`

The correct copy behavior is "Merge" because there are multiple files in the source folder that are merged into a single folder per device type per hour.

   upvoted 85 times

✉️  **onyerleft**  1 year, 8 months ago

- 1) `@trigger().outputs.windowStartTime` - this output is from a tumbling window trigger, and is required to identify the correct directory at the `/{HH}/` level. Using `windowStartTime` will give the hour with complete data. The `@trigger().startTime` is for a schedule trigger, which corresponds to the hour for which data has not arrived yet.
- 2) `/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}_{{deviceType}}.json` is the naming pattern to achieve an hourly dataset for each device type.
- 3) Multiple files for each device type will exist on the source side, since the naming pattern starts with `{deviceID}...` so the files must be merged in the sink to create a single file per device type.

   upvoted 78 times

ET Answer:

Answer Area

Parameter:

<code>@pipeline(),TriggerTime</code>
<code>@pipeline(),TriggerType</code>
<code>@trigger().outputs.windowStartTime</code>
<code>@trigger().startTime</code>

Naming pattern:

<code>/{{deviceID}}/out/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}.json</code>
<code>/{{YYYY}}/{{MM}}/{{DD}}/{{deviceType}}.json</code>
<code>/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}.json</code>
<code>/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}_{{deviceType}}.json</code>

Copy behavior:

<code>Add dynamic content</code>
<code>Flatten hierarchy</code>
<code>Merge files</code>

Question 46

DRAG DROP -

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

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

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

How should you recommend completing the structure? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.

You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
{deviceID}	/ <input type="text"/> Value / <input type="text"/> Value / <input type="text"/> Value .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceID}	
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	
raw/{regionID}	

Answer:

Values	Answer Area
{deviceID}	/ <input type="text"/> raw/{regionID} / <input type="text"/> {YYYY}/{MM}/{DD}/{HH}/{mm} / <input type="text"/> {deviceID} .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceID}	
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	
raw/{regionID}	

 **ItHYMeRish**  1 year, 8 months ago

The correct answer is
{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json

{raw/regionID} is the first level because raw is the container name for the raw data. RegionID follows it for ease of managing security.

{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json instead of
{deviceID}/{YYYY}/{MM}/{DD}/{HH}/{mm}.json. The primary reason is that you want your namespace structure to have as few folders as high up and narrow those down as you get deeper into your structure.

For example, if you have 1 year worth of data and 25 million devices, using
{YYYY}/{MM}/{DD}/{HH}/{mm} results in 2.1 million folders (1 year * 12 months * 30 days [estimate] * 24 hours * 60 minutes). If you start your folder structure with {deviceID}, you end up with 25 million folders - one for each device - before you even get to including the date in the hierarchy.

   upvoted 182 times

Question 47

OTSPOT -

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

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

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

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

You need to calculate the uptime between the faults.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT  
  
DeviceID,  
  
MIN(EventTime) as StartTime,  
  
MAX(EventTime) as EndTime,  
  
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds  
  
FROM input TIMESTAMP BY EventTime
```

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

GROUP BY

DeviceID

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

ET User:

1. Where EventType = 'HeartBeat'
2. SessionWindow

✉ Fer079 **Highly Voted** 1 year, 8 months ago

I think the right answers should be WHERE EventType='HeartBeat' and Session window. If we want to calculate the uptime between the faults, we must use session window for each device, we know that will be receiving events for each 5 seconds if there is no error, so when an error occurs (or if we reach the maximum size of the window) then a new event will not be received within the next 5 seconds and the window will close, calculating the uptime. However if We use Tumbling window, it's not possible to calculate the uptime beyond 5 seconds

thumb up, thumb down, flag upvoted 74 times

✉ Canary_2021 **Highly Voted** 1 year, 8 months ago

My answer is:

Question 1: B. Use LAG function as a filter to only filter out the events that switch from 'HeartBeat' to fault or switch from fault to 'HeartBeat'.

Question 2: C. No matter if there is a fault, device always sends message every 5min. Calculate the uptime between the faults don't need any window here. Any duration > 5s should between fault line and heartbeat line should be part of items that need to count into to calculate duration.

thumb up, thumb down, flag upvoted 21 times

✉ Fer079 1 year, 7 months ago

You cannot use the LAG function here because the "partition by" by deviceld is not included here, so the change between the status could be between different devices. This LAG function is evaluated before the "group by" clause of the query.

If you see the Microsoft documentation:

<https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics>
It says clearly that "LAG isn't affected by predicates in the WHERE clause, join conditions in the JOIN clause, or grouping expressions in the GROUP BY clause of the current query because it's evaluated before those clauses."

thumb up, thumb down, flag upvoted 8 times

ET Answer:

Answer Area

```
SELECT  
  
DeviceID,  
  
MIN(EventTime) as StartTime,  
  
MAX(EventTime) as EndTime,  
  
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds  
  
FROM input TIMESTAMP BY EventTime
```

WHERE EventType='HeartBeat'
WHERE LAG(EventType, 1) OVER (PARTITION BY DeviceID ORDER BY EventTime) <> 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

Question 48

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: A

To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.

%python //or r, scala, sql

<https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure-databricks>

Question 49

You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account.

Data to be loaded is identified by a column named LastUpdatedDate in the source table.

You plan to execute the pipeline every four hours.

You need to ensure that the pipeline execution meets the following requirements:

- ☞ Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits.
- ☞ Supports backfilling existing data in the table.

Which type of trigger should you use?

- A. event
- B. on-demand
- C. schedule
- D. tumbling window

Answer: D

In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.

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

👤 Canary_2021 Highly Voted 1 year, 8 months ago

Selected Answer: D

D is correct answer.

<https://www.sqlshack.com/how-to-schedule-azure-data-factory-pipeline-executions-using-triggers/>

Azure Data Factory pipeline executions using Triggers:

- Schedule Trigger: The schedule trigger is used to execute the Azure Data Factory pipelines on a wall-clock schedule.
- Tumbling Window Trigger: Can be used to process history data. Also can define Delay, Max concurrency, retry policy etc.
- Event-Based Triggers : The event-based trigger executes the pipelines in response to a blob-related event, such as creating or deleting a blob file, in an Azure Blob Storage

   upvoted 23 times

Question 50

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.

The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.

You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.

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

NOTE: Each correct selection is worth one point

- A. Specify a file naming pattern for the destination.
- B. Delete the files in the destination before loading the data.
- C. Filter by the last modified date of the source files.
- D. Delete the source files after they are copied.

ET User:

AC (73%) **AD (17%)** **10%**

✉ **Philipp** Highly Voted 1 year, 7 months ago

Selected Answer: AC

AC is correct, there is no point about deletion in source and might be the case that the data should stay in source too.

1 like, 1 comment, 12 upvotes

✉ **necktru** Highly Voted 1 year, 4 months ago

Selected Answer: AC

I think the C option has impact in data transfer, B are incorrect, D is irrelevant for the question, and A is a complement of the task

1 like, 1 comment, 7 upvotes

✉ **Spinazabubble** Most Recent 4 months ago

A. Specify a file naming pattern for the destination:

By specifying a file naming pattern for the destination files in the Azure Data Lake Storage Gen2 account, you can ensure that the files are organized and stored in a structured manner. This can help with data management and subsequent processing.

C. Filter by the last modified date of the source files:

By filtering the source files based on the last modified date, you can select only the files that have been modified on the current day. This reduces the amount of data transferred and improves the efficiency of the data load process.

1 like, 1 comment, 4 upvotes

ET Answer: AC

Copy only the daily files by using filtering.

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

Bing: AC

To design a daily Azure Data Factory data load that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account and minimize the data transfer between the two accounts, you should include the following two configurations in the design: **A. Specify a file naming pattern for the destination** and **C. Filter by the last modified date of the source files**.

Specifying a file naming pattern for the destination can help ensure that the data is organized in a way that makes it easy to find and access, and can also help minimize data transfer by avoiding unnecessary overwrites of existing data **1**. Filtering by the last modified date of the source files can help ensure that only new or updated files are copied, which can help minimize data transfer by avoiding unnecessary transfers of unchanged data **2**.

So, the correct answers are **A. Specify a file naming pattern for the destination** and **C. Filter by the last modified date of the source files**. Is there anything else you would like to know? 😊

Question 51

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

Which output mode should you use?

- A. update
- B. complete
- C. append

Answer: C

Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.

Incorrect Answers:

B: Complete Mode: The entire updated result table is written to external storage. It is up to the storage connector to decide how to handle the writing of the entire table.

A: Update Mode: Only the rows that were updated in the result table since the last trigger are written to external storage. This is different from Complete Mode in that Update Mode outputs only the rows that have changed since the last trigger. If the query doesn't contain aggregations, it is equivalent to Append mode.

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

Question 52

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

Answer: D

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.

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

Question 53

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.

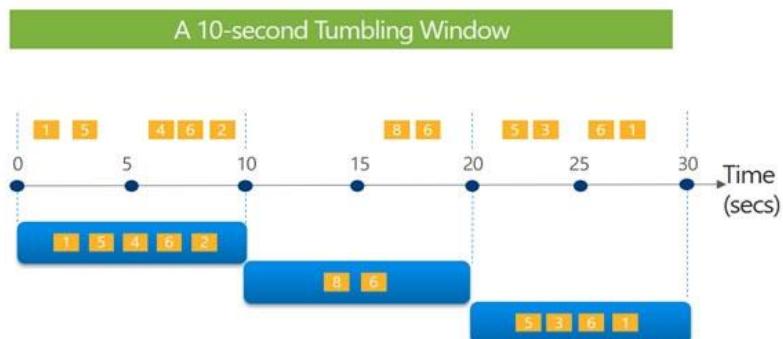
Which windowing function should you use?

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

Answer: D

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.

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)
```

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

Question 54

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. Update
- B. Complete
- C. Append

ET User: 有可能是 C

C (60%) A (40%)

[Exam DP-203 topic 2 question 68 discussion - ExamTopics](#)

👤 **necktru** Highly Voted 1 year, 4 months ago

Selected Answer: A

I think Update is correct, because " new rows will be added to adjust a sale" , that means that in the course of a day you must update de daily import with the new sales, the group by process generates new amounts, keep in mind that when it say "sales transactions will never be updated" its about the online store, not the aggregated rows.

👍👎FLAG upvoted 15 times

👤 **vctrhugo** 2 months, 1 week ago

Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

👍👎FLAG upvoted 4 times

👤 **[Removed]** Highly Voted 7 months, 2 weeks ago

Selected Answer: C

Using chatgpt : Append

👍👎FLAG upvoted 9 times

👤 **Tightbot** Most Recent 2 weeks, 2 days ago

Selected Answer: C

Append is the right choice . Update is for modifications and append is to add new rows

👍👎FLAG upvoted 1 times

👤 **Ram9198** 3 weeks, 3 days ago

Selected Answer: A

When say adjusted, it means update because we need to reaggregate to get the latest total amount after adjustment, also there is a hint to minimise duplicates.

Append will ignore the updated state records only emit new records

👍👎FLAG upvoted 1 times

👤 **pavankr** 2 months ago

The requirement says "not to update"; your answer says "update"??????

👍👎FLAG upvoted 1 times

✉ vctrhugo 2 months, 3 weeks ago

Selected Answer: C

C. Append

For the given scenario, where sales transactions are never updated but new rows are added to adjust a sale, the recommended output mode for the dataset processed by using Structured Streaming in Azure Databricks is "Append".

The "Append" output mode ensures that only new rows are added to the output data as they arrive in the streaming data source. It appends the new rows to the existing result without modifying or updating previously processed data. This mode is suitable when you want to continuously append new records to the output data without duplicating or modifying existing data.

In this case, as new rows are added to adjust a sale, the "Append" mode will capture these new rows and include them in the output data, allowing you to aggregate the line total sales amount and line total tax amount in Databricks while minimizing duplicate data.

Like 1 Upvote 3 times

✉ Ankit_Az 3 months ago

I feel Append is correct here

Like 1 Upvote 3 times

✉ janaki 3 months, 1 week ago

It's Append as the 3rd instruction says

Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.
So it's not UPDATE but an APPEND

Like 1 Upvote 3 times

✉ esaade 6 months ago

Selected Answer: C

I would recommend using the "Append" output mode for the dataset processed by using Structured Streaming in this scenario.

The "Append" output mode is appropriate when the output dataset is a set of new records and does not include any updates or deletions. It will only append new rows to the output dataset, which means there will be no duplicate data created as a result of the streaming data solution. Since the solution will never update existing rows, but rather add new rows, the "Append" mode is the best choice to meet the requirements.

Like 1 Upvote 6 times

✉ Rakrah 6 months, 4 weeks ago

Very Correct Answer is "APPEND" MODE - Because Sales transaction never be updated using Update Mode, would not provide any benefits, rather "Append" mode will be add new row to the output dataset and correctly aggregate the line total sales amount and line total tax amount without any duplicates. So Append mode 200% meet the requirement.

Like 1 Upvote 5 times

✉ Okea 7 months, 1 week ago

Update mode is the answer:

It involves writing the data records that are either new or for which the old value is updated. So this mode can be used when it is required to have the "upsert" mode of operation doing some aggregation. If no aggregation is applied, the update mode works the same as the append mode.

<https://medium.com/analytics-vidhya/spark-streaming-output-modes-600c689b6bf9>

Like 1 Upvote 1 times

✉ kkk5566 1 week ago

"Line total sales amount and line total tax amount will be aggregated in Databricks."

Like 1 Upvote 1 times

✉ kkk5566 1 week ago

Append is correct.

Like 1 Upvote 1 times

✉ agold96 7 months, 3 weeks ago

Selected Answer: C

New rows will be added suggest "Append", correct is C for me

Like 1 Upvote 3 times

✉ hanzocuk 8 months ago

Correct is A

Focus on the task-> "Sales >> transactions will never be updated<<. Instead, >>new rows will be added<< to adjust a sale" (Yes, very poorly formulated as usual, who is responsible for this adjustment??)

From spark docs -> "append: Only the new rows in the streaming DataFrame/Dataset will be written to the sink"

update would be similar to append if no aggregations were involved, but in our case we have aggregations.

<https://spark.apache.org/docs/latest/api/python/reference/pyspark.ss/api/pyspark.sql.streaming.DataStreamWriter.outputMode.html#pyspark.sql.streaming.DataStreamWriter.outputMode>

Like 1 Upvote 1 times

✉ hanzocuk 8 months ago

Sorry made a typo, C: Append....

Like 1 Upvote 1 times

✉ hanzocuk 8 months ago

Moderator please dont include any of above, I feel it could mislead people as I am not even sure myself...

- 1) transactions never updated -> suggests Append
- 2) new rows added -> suggests Append.... to adjust a sale -> suggests Update
- 3) rows minimized - suggests Update

This is altogether poorly formulated... I think as a whole A: Update is a better choice

Like 1 Upvote 2 times

👤 **vikaspt07** 8 months ago

As per below article answer should be complete mode
<https://medium.com/analytics-vidhya/spark-streaming-output-modes-600c689b6bf9>

👍 ↵ 💬 upvoted 3 times

👤 **vikaspt07** 8 months, 1 week ago

update in outputmode in streaming does not work without watermark on timestamp ,append is the answer
👍 ↵ 💬 upvoted 1 times

👤 **Igor85** 9 months, 2 weeks ago

the fact that aggregations are mentioned here is clearly pointing to 'update', without them it would be obviously 'append' mode
👍 ↵ 💬 upvoted 2 times

👤 **Dusica** 9 months, 2 weeks ago

APPEND
👍 ↵ 💬 upvoted 1 times

👤 **bokLuci** 10 months, 1 week ago

Selected Answer: C
Certainly 'C'. Line total will be aggregated at reporting time and those aggs will be run on the cumulative delta of sales transaction amounts. It's normal design of transactional delta for end of period reporting.
You don't have anything to Update, you are only appending the delta from the previous transaction.

👍 ↵ 💬 upvoted 4 times

👤 **rzeng** 10 months, 1 week ago

A - update, this reduce duplicate data
👍 ↵ 💬 upvoted 2 times

👤 **ads5891** 1 year ago

Selected Answer: A
I think this should be "Update" mode because the key is "minimize duplicates". Please check <https://sparkbyexamples.com/spark/spark-streaming-outputmode/>
👍 ↵ 💬 upvoted 2 times

👤 **Fidel_104** 1 year ago

Selected Answer: C
I'm still not convinced about 'update'. The DB docs doesn't even mention append for writeStream outputmodes, just complete and append:
<https://docs.databricks.com/delta/delta-streaming.html>
Lmk if I'm missing something!
👍 ↵ 💬 upvoted 1 times

👤 **Genere** 1 year ago

Mode

Example

Notes

Complete

.outputMode("complete")

The entire updated Result Table is written to the sink. The individual sink implementation decides how to handle writing the entire table.

Append

.outputMode("append")

Only the new rows appended to the Result Table since the last trigger are written to the sink.

Update

.outputMode("update")

Only the rows in the Result Table that were updated since the last trigger will be outputted to the sink. Since Spark 2.1.1

👍 ↵ 💬 upvoted 1 times

👤 **wendyy** 1 year ago

Update is Correct.
"The output data will contain items purchased, quantity, line total sales amount, and line total tax amount" means output data has been aggregated, they are not Sales transactions(input). You don't want the aggregated data has duplicates, so update is correct.

👍 ↵ 💬 upvoted 1 times

👤 **Deeksha1234** 1 year ago

Selected Answer: C
sorry, but when I check all the documentation available , I find only append and complete modes for sink, append will add only new rows so the answer should be append.
👍 ↵ 💬 upvoted 4 times

👤 **Deeksha1234** 1 year ago

Update - to minimize duplicates and Output can be updated (given that transactions can't be updated but we need to find a solution for output)
👍 ↵ 💬 upvoted 1 times

👤 **dsp17** 1 year, 1 month ago

Selected Answer: A
Update is Correct - "The solution must minimize duplicate data"
👍 ↵ 💬 upvoted 3 times

- ✉️ **Saim8711** 1 year, 2 months ago
Selected Answer: C
Append looks good to me
👍 ↗️ 📈 upvoted 1 times
- ✉️ **Remedios79** 1 year, 2 months ago
For me is C, append, as it would be like INSERT in SQL statement
👍 ↗️ 📈 upvoted 2 times
- ✉️ **upliftinghut** 1 year, 3 months ago
Selected Answer: C
Correct answer should be "Append". Detail here: <https://docs.microsoft.com/en-us/azure/stream-analytics/sql-database-upsert>
👍 ↗️ 📈 upvoted 1 times
- ✉️ **Aditya0891** 1 year, 2 months ago
check the question first. It talks about Azure Databricks and not Stream Analytics
👍 ↗️ 📈 upvoted 1 times
- ✉️ **Rickmundo** 1 year, 3 months ago
For me it's not clear why it is update instead of append. Does have anyone a explanation?
👍 ↗️ 📈 upvoted 1 times
- ✉️ **Davico93** 1 year, 2 months ago
The key word is "OUTPUT" and this is updated, transactions are "INPUT" and this won't be updated
👍 ↗️ 📈 upvoted 3 times
- ✉️ **oakmm** 1 year, 4 months ago
should it be "complete" output mode?
👍 ↗️ 📈 upvoted 2 times
- ✉️ **chuckas** 1 year, 4 months ago
Initially I thought append as well but the line that says "minimise duplicates" indicates to me that you don't want to continually add records if you can update existing ones
👍 ↗️ 📈 upvoted 2 times
- ✉️ **jackttt** 1 year, 4 months ago
Why not append?
👍 ↗️ 📈 upvoted 3 times

ET Answer:

Answer: A

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

Incorrect Answers:

B: Complete mode: replace the entire table with every batch.

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

Bing: C

In this scenario, you should recommend using the **Append** output mode for the dataset that will be processed by using Structured Streaming in Azure Databricks. The Append output mode is used when you want to write only the new rows appended to the Result Table since the last trigger ¹. Since sales transactions will never be updated and new rows will be added to adjust a sale, the Append output mode is appropriate for minimizing duplicate data. So, the correct answer is **C. Append**. Is there anything else you would like to know? 😊

Question 55

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.

ET User:

D (79%)	A (21%)
<p>✉ Saransundar <small>Highly Voted</small> 1 year, 2 months ago</p> <p>The question asks for transaction log size on each distribution. The correct answer is D: Link below: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor</p> <p>-- Transaction log size</p> <pre>SELECT instance_name as distribution_db, cntr_value*1.0/1048576 as log_file_size_used_Gb, pdw_node_id FROM sys.dm_pdw_nodes_os_performance_counters WHERE instance_name like 'Distribution_%' AND counter_name = 'Log File(s) Used Size (KB)'</pre> <p>1 upvoted 16 times</p>	
<p>✉ Davico93 1 year, 2 months ago</p> <p>but you don't need it from master, just DW1</p> <p>1 upvoted 4 times</p>	
<p>✉ Saim8711 <small>Highly Voted</small> 1 year, 2 months ago</p> <p><small>Selected Answer: D</small></p> <p>D is totally correct. Link has this very clearly mentioned</p> <p>https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor</p> <p>1 upvoted 8 times</p>	

ET Answer: A

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.

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

Question 56

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

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

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

Question 57

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: C

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.

Incorrect Answers:

A: Deserialization issues are caused when the input stream of your Stream Analytics job contains malformed messages.

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

Question 58

HOTSPOT -

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

You plan to have a fact table for website visits. The table will be approximately 5 GB.

You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Distribution:

Hash
Round robin
Replicated

Index:

Clustered columnstore
Clustered
Nonclustered

Answer:

Answer Area

Distribution:

Hash
Round robin
Replicated

Index:

Clustered columnstore
Clustered
Nonclustered

Box 1: Hash -

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

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

Box 2: Clustered columnstore -

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

Reference:

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

Question 59

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: AB

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.

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

Question 60

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

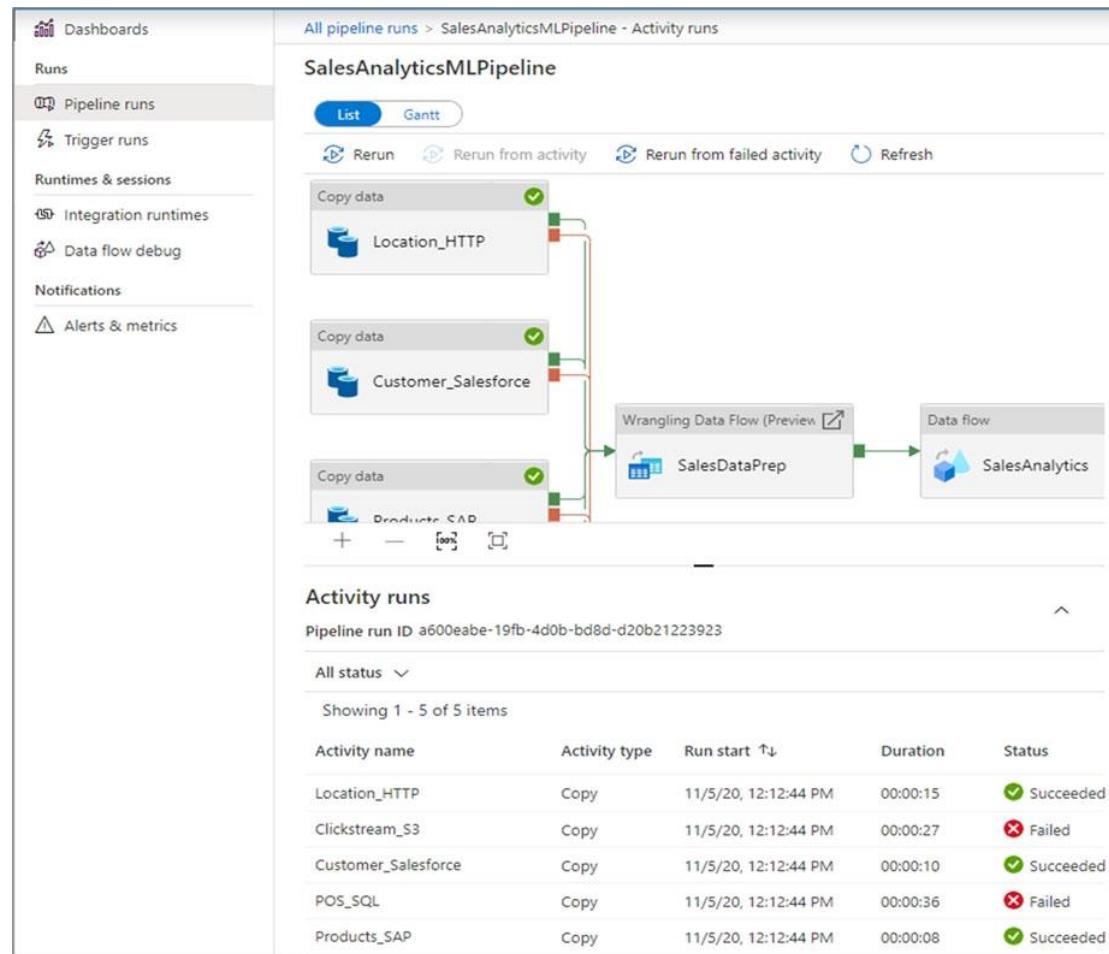
Answer: A

Monitor activity runs. To get a detailed view of the individual activity runs of a specific pipeline run, click on the pipeline name.

Example:

Pipeline runs			
Triggered		Debug	
		Rerun	Cancel
<input type="text"/> Search by run ID or name		Pacific Time (US & C... : Last 7 days	<input type="button"/>
Showing 1 - 21 items			
<input type="checkbox"/> Pipeline name		Run start ↑	Run end
<input type="checkbox"/> S3ToDataLakeCopy		11/5/20, 6:00:18 AM	11/5/20, 6:03:18 AM
<input type="checkbox"/> DatabricksJarPipeline		11/4/20, 6:04:11 PM	11/4/20, 6:10:11 PM
<input type="checkbox"/> S3ToDataLakeCopy		11/4/20, 6:00:18 PM	11/4/20, 6:03:18 PM
<input type="checkbox"/> S3ToDataLakeCopy		11/4/20, 6:00:19 AM	11/4/20, 6:04:19 AM

The list view shows activity runs that correspond to each pipeline run. Hover over the specific activity run to get run-specific information such as the JSON input, JSON output, and detailed activity-specific monitoring experiences.



You can check the Duration.

Incorrect Answers:

C: sys.dm_pdw_wait_stats holds information related to the SQL Server OS state related to instances running on the different nodes.

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

Question 61

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.

ET User:

B (94%) 6%

KashRaynardMorse Highly Voted 1 year, 3 months ago
Selected Answer: B

The error message says a missing file, which matches with answer B: missing data from 06:00. The process had re-tried three times, 15 mins apart, which explains that the error was generated 07:45.

ET Answer: A

A file is missing.

Bing:

A possible cause of the error is that from 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON. The error message indicates that the specified path does not exist, which suggests that the data for the specified time period is missing or not available. So, the correct answer is **B. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON**. Is there anything else you would like to know? 😊

Question 62

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 Synapse Studio, select the workspace. From Monitor, select SQL requests.
- B. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
- C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.
- D. From Azure Monitor, run a Kusto query against the SparkLoggingEvent_CL table.

Answer: C

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.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications>

Question 63

DRAG DROP -

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

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

You create the following components:

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

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.

Select and Place:

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.
	Read the file into a data frame.
	Perform transformations on the data frame.
	Specify a temporary folder to stage the data.
	Write the results to a table in Azure Synapse.

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.

Question 64

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. From the UX Authoring canvas, select Set up code repository.
- B. Create a Git repository.
- C. Create a GitHub action.
- D. Create an Azure Data Factory trigger.
- E. From the UX Authoring canvas, select Publish.
- F. From the UX Authoring canvas, run Publish All.

ET User:



Selected Answer: AB

They are asking to "implement version control".

- B Create Git repo
- A From the UX Set up code repository

👍 ↺ ⚡ upvoted 16 times

ET Answer: BF

Bing:

To implement version control for the changes made to pipeline artifacts in an Azure data factory named ADF1, you should perform the following two actions: **A. From the UX Authoring canvas, select Set up code repository** and **B. Create a Git repository**. Setting up a code repository from the UX Authoring canvas allows you to configure version control for your data factory resources. You can use a Git repository to store and manage your code, and track changes to your pipeline artifacts. So, the correct answers are **A. From the UX Authoring canvas, select Set up code repository** and **B. Create a Git repository**. Is there anything else you would like to know? 😊

Question 65

DRAG DROP -

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.

Select and Place:

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.
	Associate the schedule trigger with pipeline1.
	> Merge the changes from branch1 into main.
Switch to Synapse live mode.	< Publish the contents of main.

  **MJSnail**  7 months ago

If it's hard to remember, memorize it as CAMP.

   upvoted 27 times

Question 66

HOTSPOT -

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.

Hot Area:

Answer Area

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

The image shows two dropdown menus. The first dropdown, under 'Location1', has the following options: abfs, abfss, wasb, and wasbs. The second dropdown, under 'TYPE', has the following options: BLOB_STORAGE, HADOOP, RDBMS, and SHARP MAP MANAGER.

ET User:

abfss and Hadoop

✉️ **Hema_V** Highly Voted 1 year ago

Answer: abfss and Hadoop
Hint: Storage1 requires secure transfers --> The default option is to use enable secure SSL connections when provisioning Azure Data Lake Storage Gen2. When this is enabled, you must use abfss when a secure TLS/SSL connection is selected.

Reference: <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated>

1 upvoted 26 times

✉️ **vigilante89** Highly Voted 8 months, 3 weeks ago

abfss and Hadoop

1 upvoted 6 times

ET Answer:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore  
WITH  
( Location1 ` ://data@newyorktaxidataset.dfs.core.windows.net` ,  
    abfs  
    abfss  
    wasb  
    wasbs  
credential = ADLS_credential ,  
TYPE -  
);  
    BLOB_STORAGE  
    HADOOP  
    RDBMS  
    SHARP MAP MANAGER
```

Question 67

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.

ET User: C

C (85%)

B (15%)

👤 **yogiazaad** Highly Voted 7 months, 1 week ago

Selected Answer: C

You wont be able to create restore point when the SQL pool is paused. The the correct answer is Result SQL Pool. See below from Microsoft documentation.

User-defined restore points can also be created through Azure portal.

Sign in to your Azure portal account.

Navigate to the dedicated SQL pool (formerly SQL DW) that you want to create a restore point for.

Select Overview from the left pane, select + New Restore Point. If the New Restore Point button isn't enabled, make sure that the dedicated SQL pool (formerly SQL DW) isn't paused.

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

12 upvoted

ET Answer: B

Bing: C

To restore the current state of an Azure Synapse Analytics dedicated SQL pool named SQLPool1 to a new SQL pool, you should first **resume SQLPool1**. You cannot restore a paused SQL pool, so you need to resume it before you can create a user-defined restore point or restore its state to a new SQL pool. So, the correct answer is **C. Resume SQLPool1**. Is there anything else you would like to know? 😊

Question 68

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 correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. an X.509 certificate
- B. an RSA key
- C. an Azure virtual network that has a network security group (NSG)
- D. an Azure Policy initiative
- E. an Azure key vault that has purge protection enabled

Answer: BE

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.

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

👤 **allagowf** Highly Voted 10 months, 1 week ago

Selected Answer: BE

Answer is correct : BE

1 like, 1 reply, 1 flag, upvoted 5 times

👤 **Pradeep2675** Most Recent 1 month, 1 week ago

Selected Answer: BE

Answer: BE

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>

1 like, 1 reply, 1 flag, upvoted 1 times

👤 **vctrhugo** 2 months, 1 week ago

Selected Answer: AE

A. Including an X.509 certificate in the solution can be used to provide encryption at rest for the data in Azure Synapse Analytics. X.509 certificates are widely used for securing data and communications.

E. An Azure Key Vault with purge protection enabled can be utilized to securely store and manage encryption keys. By storing the encryption keys in Azure Key Vault, you can ensure that the keys are well protected and access to them is tightly controlled.

1 like, 1 reply, 1 flag, upvoted 1 times

Bing: AE

To provide double encryption of all the data at rest in an Azure Synapse Analytics workspace, you should include an

Azure key vault that has purge protection enabled and an **X.509 certificate** in your recommendation. Azure

Synapse Analytics supports double encryption of data at rest, which means that data is encrypted using two separate keys: one key is managed by Microsoft and the other key is managed by the customer. You can use an Azure key vault to store and manage your customer-managed key, and enable purge protection to prevent accidental deletion of the key. An X.509 certificate can be used to encrypt the data using the customer-managed key. So, the correct answers are

A. an X.509 certificate and **E. an Azure key vault that has purge protection enabled**. Is there anything else you would like to know? 😊

Question 69

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowBlobPublicAccess 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 storage from Pool1.

What should you create first?

- A. an external resource pool
- B. an external library
- C. database scoped credentials
- D. a remote service binding

Answer: C

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.

<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>

Question 70

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

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.

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

Question 71

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

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",
        "start": "2023-01-01T00:00:00Z",
        "end": "2023-01-01T01:00:00Z",
        "size": "01:00:00"
    }
}
```

```
"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>

Question 72

HOTSPOT -

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.

Hot Area:

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:

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

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>

Question 73

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 pools
- ⇒ Supports fast data retrieval for data from the current month
- ⇒ Simplifies data security management by department

Which folder structure should you recommend?

- A. \Department\DataSource\YYYY\MM\DataFile_YYYYMMDD.parquet
- B. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet
- C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
- D. \YYYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet

Answer: A

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.

Bing: A

The recommended folder structure for the files in an Azure Data Lake Storage Gen2 account that supports partition elimination for queries by Azure Synapse Analytics serverless SQL pools, supports fast data retrieval for data from the current month, and simplifies data security management by department is: **A.**

\Department\DataSource\YYYY\MM\DataFile_YYYYMMDD.parquet.

This folder structure organizes the data by department first, then by data source, and then by year and month. This allows for partition elimination when querying by department or data source, and also allows for fast data retrieval of data from the current month. Additionally, organizing the data by department first simplifies data security management by allowing you to set permissions at the department level. Is there anything else you would like to know? 😊

Question 74

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.

ET User: DE?

BD (70%)

DE (26%)

4%

[Exam DP-203 topic 2 question 90 discussion - ExamTopics](#)

✉️ **esaade** Highly Voted 5 months, 4 weeks ago

Selected Answer: BD

- B. Convert the avg_c column into a calculated column.
- D. Enable result set caching.

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

👉💡 upvoted 6 times

✉️ **kkk5566** Most Recent 6 days, 4 hours ago

Selected Answer: DE

correct
👉💡 upvoted 1 times

✉️ **kkk5566** 6 days, 4 hours ago

Selected Answer: BD

correct
👉💡 upvoted 1 times

✉️ **kkk5566** 6 days, 4 hours ago

D & E should be correct

👉💡 upvoted 1 times

✉️ **Matt2000** 3 weeks, 5 days ago

Calculated columns exist in Power BI, not dedicated SQL pools. Computed columns are not supported in dedicated SQL pools.
Ref: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview>

👉💡 upvoted 2 times

✉️ **dumbled** 4 months, 1 week ago

Selected Answer: BD

correct
👉💡 upvoted 2 times

✉️ **Lestrang** 7 months, 2 weeks ago

Selected Answer: AB

With that point by erhard being made (caching does work with queries using UDF), the most commonly voted D is wrong, so B and what now?
Replicated cannot be right because it received date everyday and has aggregations so not a dim table and we have no clue about its size,
by elimination that leaves us A and C

Indexing is less useful with no joins but it does improve some performance being on where clause target. so I'd go with A and B.
👉💡 upvoted 1 times

✉️ **Lestrang** 7 months, 2 weeks ago

Creating an index on the avg_f column will improve the performance of the query, as it will allow the query to find the relevant data more quickly. Converting the avg_c column into a calculated column will allow the query to return the temperature in Celsius without the need to perform the calculation at runtime, which will also improve the performance of the query.

👉💡 upvoted 1 times

✉️ **Lestrang** 7 months, 1 week ago

After re-considering, I am unsure whether the indexing would help. That would only leave Replication as the viable option even though it is not viable design
but the request is to minimize query time and that is what it will do, so I guess final answer is
BE

👉💡 upvoted 1 times

✉️ **Karforcerts** 9 months, 2 weeks ago

Selected Answer: BD

need to first change UDF to a calculated column and then enable result set caching. agreed with the answer

👉💡 upvoted 3 times

✉️ **erhard** 9 months, 3 weeks ago

Queries using user defined functions are not cached.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching>

👉💡 upvoted 3 times

✉️ **kl8585** 9 months, 3 weeks ago

Selected Answer: DE

A,C not right since index don't help if join are not involved.
D for sure help query performance.

I don't get why B:

"A computed column is a virtual column whose value is calculated from other values in the table. By default, the expression's outputted value is not physically stored. Instead, SQL Server runs the expression when the column is queried and returns the value as part of the result set ... In many cases, non-persistent computed columns put too much burden on the processor, resulting in SLOWER QUERIES and unresponsive applications"

Since the only requirements is faster execution times for queries, i don't think calculated columns will improve performance.

Si second option for me would be D (replicate). Although it will cause more effort writing, because updates should be written to every partition, optimized writes aren't a requirement in the question.

👉💡 upvoted 3 times

- ✉ **rzeng** 10 months, 1 week ago
pool ingest data once per 24 hrs, while query happens every 15mins, caching result can definitely avoid the some duplicate calculation, I'll go with BD.
 upvoted 1 times
- ✉ **Xinyuehong** 10 months, 3 weeks ago
Selected Answer: DE
I think should be DE.
since "the query is executed once every 15 minutes and the @parameter value is set to the current date", and the it receives new data once every 24 hours, it means the query result isn't change in one day even you run it every 15 mins. The data is static within a day. Replication could help the performance.
 upvoted 2 times
- ✉ **anks84** 12 months ago
Selected Answer: BD
Answer is Correct !
 upvoted 4 times

ET Answer: BD

D: 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.

Incorrect:

Not A, not C: No joins so index not helpful.

Not E: What is a replicated table?

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. If the data is static and does not change, you can replicate larger tables.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching>

Bing: DE

To minimize the time it takes for the query to return results, you should perform the following two actions: **D. Enable result set caching** and **E. Change the table distribution to replicate**. Enabling result set caching can help improve query performance by caching the results of intermediate query steps, which can help reduce the amount of data that needs to be processed when the query is executed again. Changing the table distribution to replicate can help improve query performance by replicating the data across all compute nodes, which can help reduce data movement and improve query performance for certain types of queries. So, the correct answers are **D. Enable result set caching** and **E. Change the table distribution to replicate**. Is there anything else you would like to know?

Question 75

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

ET User: A?

B (61%)	A (39%)
---------	---------

✉  esaade  5 months, 3 weeks ago

Selected Answer: B

B. Structured Streaming in Azure Databricks is the best option for this scenario as it allows for processing of streaming data and outputting it to Azure Data Lake Storage, while also providing the ability for analysts to interactively query the data using Databricks notebooks.

Azure Stream Analytics and Azure Synapse notebooks (option A) can also process streaming data and output to Data Lake Storage, but they may not provide the same level of interactivity for analysts.

Event triggers in Azure Data Factory (option C) can help automate data movement between Event Hubs and Data Lake Storage, but they do not provide the necessary functionality for processing and querying streaming data.

Azure Queue Storage and read-access geo-redundant storage (RA-GRS) (option D) are not relevant for this scenario as they do not provide capabilities for processing and querying streaming data.

Like    upvoted 8 times

✉  kkk5566  6 days, 3 hours ago

Selected Answer: A

should be correct

Like    upvoted 1 times

✉  auwia 2 months, 2 weeks ago

Selected Answer: A

What streaming sources and sinks does Azure Databricks support?
Databricks recommends using Auto Loader to ingest supported file types from cloud object storage into Delta Lake. For ETL pipelines, Databricks recommends using Delta Live Tables (which uses Delta tables and Structured Streaming). You can also configure incremental ETL workloads by streaming to and from Delta Lake tables.

In addition to Delta Lake and Auto Loader, Structured Streaming can connect to messaging services such as Apache Kafka.
<https://learn.microsoft.com/en-us/azure/databricks/structured-streaming/>
I don't see data lake in the list, so probably the answer is A.

Like    upvoted 1 times

✉  vadiminski_a 5 months ago

Selected Answer: B

I am in favour of B because of this piece of information I have encountered:
<https://www.databricks.com/spark/getting-started-with-apache-spark/streaming>

Like    upvoted 3 times

✉  vadiminski_a 5 months ago

On the other hand, there is this: <https://learn.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics>
So I believe both to be valid, Azure Stream Analytics seems to be more straightforward

Like    upvoted 1 times

✉️  **Kate0204** 6 months ago

Selected Answer: A

An Azure Stream Analytics job consists of an input, query, and an output.

   upvoted 1 times

✉️  **Karl_Cen** 7 months, 1 week ago

"The solution must ensure that analysts can interactively query the streaming data"

Streaming analysis can't query streaming data interactively

   upvoted 2 times

✉️  **Lestrang** 7 months, 2 weeks ago

Selected Answer: A

B. Structured Streaming in Azure Databricks is incorrect because while it allows you to process streaming data using Spark's structured streaming API, it is not designed to directly output the data to Azure Data Lake Storage. Instead, it typically outputs the data to storage systems like HDFS, S3, or Cosmos DB. Additionally, Databricks is a separate service that does not integrate with Azure Synapse for interactive querying. While it's possible to use Databricks to read the data from Data Lake Storage and use Spark to process the data and then write it back to Data Lake Storage, it will not be as efficient as using Azure Stream Analytics for this use case as it is specifically designed for streaming data processing and also has built-in connectors to various data storage and analytics services like Data Lake Storage

   upvoted 2 times

✉️  **Lestrang** 7 months, 1 week ago

Although this might be true, after some pondering, the given solution A. Azure Stream Analytics and Azure Synapse notebooks requires a Synapse workspace which is not implied.

So I guess it would be databricks.

   upvoted 1 times

✉️  **Mal2002** 2 months, 3 weeks ago

It's implied. Solutions said Azure Stream Analytics and Azure Synapse Notebook, Azure Synapse notebook cannot be created without Azure Synapse Workspace.

   upvoted 1 times

✉️  **alexnicolita** 7 months, 3 weeks ago

Selected Answer: A

Why not Azure Stream Analytics and Azure Synapse Analytics?

   upvoted 2 times

ET Answer: B

Bing: A

To process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage while ensuring that analysts can interactively query the streaming data, you can use **Azure Stream Analytics and Azure Synapse notebooks**. Azure Stream Analytics is a real-time event processing engine that can read data from an Azure Event Hub and write the processed data to various outputs, including Azure Data Lake Storage. Azure Synapse notebooks provide a collaborative environment for data exploration, data transformation, and machine learning development, allowing analysts to interactively query the streaming data stored in Azure Data Lake Storage. So, the correct answer is **A. Azure Stream Analytics and Azure Synapse notebooks**. Is there anything else you would like to know? 😊

Question 76

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

✉️  **Rajcse03** Highly Voted  7 months, 3 weeks ago

Selected Answer: A

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

   upvoted 6 times

✉️  **kkk5566** Most Recent  6 days, 4 hours ago

Selected Answer: A

A is correct

   upvoted 1 times

✉️  **[Removed]** 7 months, 3 weeks ago

Selected Answer: A

correct

   upvoted 3 times

Question 77

DRAG DROP -

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.

NOTE: Each correct selection is worth one point.

Values	Answer Area
agg(col('SalesPerson'))	df_sales.filter(col('Region')=='HQ').
filter(col('SalesPerson'))	.agg(sum('Amount').alias('TotalAmount')).
groupBy(col('SalesPerson'))	limit(3)
groupBy(col('TotalAmount'))	
orderBy(col('TotalAmount'))	
orderBy(desc('TotalAmount'))	

Answer:

Answer Area

```
df_sales.filter(col('Region')=='HQ').groupBy(col('SalesPerson'))
.agg(sum('Amount').alias('TotalAmount')).orderBy(desc('TotalAmount')) limit(3)
```

✉  **esaade**  5 months, 4 weeks ago

```
df_sales.filter(col("Region") == "HQ")
.groupBy(col("SalesPerson"))
.agg(sum('Amount').alias('TotalAmount'))
.orderBy(desc('TotalAmount'))
.limit(3)
```

   upvoted 10 times

✉  **kkk5566**  6 days, 3 hours ago

```
.groupBy(col('SalesPerson')) and .orderBy(desc('TotalAmount'))
.groupBy(col('SalesPerson')) and .orderBy(desc('TotalAmount'))
```

   upvoted 1 times

✉  **aurorafang** 7 months, 1 week ago

for the sequence, group by usually put before the order by operations

   upvoted 3 times

✉  **[Removed]** 7 months, 3 weeks ago

correct

   upvoted 2 times

Question 78

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

Bing: D

To schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container, you should use a **storage event trigger**. A storage event trigger runs a pipeline in response to an event in Azure Blob storage or Azure Data Lake Storage Gen2. You can configure the trigger to fire when a new file is created in the specified container, which will cause the pipeline to execute automatically when a new file arrives. So, the correct answer is **D. storage event**. Is there anything else you would like to know? 😊

Question 79

DRAG DROP -

You have a project in Azure DevOps that contains a repository named Repo1. Repo1 contains a branch named main.

You create a new Azure Synapse workspace named Workspace1.

You need to create data processing pipelines in Workspace1. The solution must meet the following requirements:

- Pipeline artifacts must be stored in Repo1
- Source control must be provided for pipeline artifacts.
- All development must be performed in a feature branch.

Which four actions should you perform in sequence in Synapse Studio? 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
Create pipeline artifacts and save them in the main branch.	>
Set the main branch as the collaboration branch.	^
Create a pull request to merge the contents of the main branch into the new branch.	<
Create pipeline artifacts and save them in the new branch.	▼
Create a new branch.	
Configure a code repository and select Repo1.	

ET User:

Configure a code repo and select Repo1

Set the main branch as the collaboration branch

Create a new branch

Create pipeline artifacts and save them in the new branch

✉ SinSS Highly Voted 3 months, 2 weeks ago

Configure a code repo and select Repo1
Set the main branch as the collaboration branch
Create a new brach
Create pipeline artifacts and save them in the new branch

1 upvoted 13 times

✉ mhi Highly Voted 3 months, 3 weeks ago

Shouldn't you merge the new branch into the main branch?
1 upvoted 8 times

✉ peches 2 months, 3 weeks ago

Agree, you create a feature branch from the collaboration branch, work on it, and after you finished you merge back to the collaboration branch (by default is main). Source: <https://learn.microsoft.com/en-us/azure/synapse-analytics/cicd/source-control#version-control>

1 upvoted 3 times

✉ kkk5566 Most Recent 6 days, 3 hours ago

correct
1 upvoted 1 times

✉ kkk5566 6 days, 3 hours ago

Configure a code repo and select Repo1
Set the main branch as the collaboration branch
Create a new brach
and PR

1 upvoted 1 times

✉ AlviraTony 1 week, 1 day ago

Given solution is correct
1 upvoted 1 times

✉ Matt2000 3 weeks, 5 days ago

"Configure a code repository and select Repo2" is not required as you already have a repo Repo1 with main as branch.
1 upvoted 1 times

✉ FRANCIS_A_M 5 months ago

Correct
1 upvoted 6 times

ET Answer:

Answer Area

- Configure a code repository and select **Repo1**.
- Create a new branch.
- Create pipeline artifacts and save them in the new branch.
- Create a pull request to merge the contents of the main branch into the new branch.

Bing:

To create data processing pipelines in an Azure Synapse workspace named Workspace1, store pipeline artifacts in a repository named Repo1 in an Azure DevOps project, provide source control for pipeline artifacts, and perform all development in a feature branch, you should perform the following four actions in sequence in Synapse Studio:

1. **Configure a code repository and select Repo1.** This will allow you to store your pipeline artifacts in Repo1 and provide source control for them.
2. **Set the main branch as the collaboration branch.** This will allow you to collaborate with others on the development of your pipelines.
3. **Create a new branch.** This will allow you to perform all development in a feature branch, separate from the main branch.
4. **Create pipeline artifacts and save them in the new branch.** This will allow you to develop your pipelines in the feature branch and save your changes to Repo1.

After you have completed these steps, you can create a pull request to merge the contents of the new branch into the main branch when you are ready to publish your changes. So, the correct sequence of actions is: Configure a code repository and select Repo1, Set the main branch as the collaboration branch, Create a new branch, Create pipeline artifacts and save them in the new branch. Is there anything else you would like to know? 😊

Question 80

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

ET User: BC

FRANCIS_A_M Highly Voted 5 months ago

Selected Answer: BC

It's BC. Use the LookUp Activity to read the .txt file. ForEach to Loop though making sure Sequential is off (which off by default) for parallelization

upvoted 6 times

aemilka Highly Voted 4 months, 3 weeks ago

Selected Answer: BC

Lookup activity reads and returns the content of a configuration file or table. It also returns the result of executing a query or stored procedure. The output can be a singleton value or an array of attributes, which can be consumed in a subsequent copy, transformation, or control flow activities like ForEach activity.

<https://learn.microsoft.com/en-us/azure/data-factory/control-flow-lookup-activity>

upvoted 5 times

ET Answer: AC

Bing: BC

To use an Azure Synapse pipeline to copy data from the selected tables in an Azure SQL database named DB1 to the files in a storage account named storage1, based on the data in a file named File1.txt, you should include the following two pipeline activities in the pipeline: **B. Lookup** and **C. ForEach**.

The **Lookup** activity can be used to read the data from File1.txt and pass it as an input to the next activity in the pipeline. The **ForEach** activity can be used to iterate over the data from File1.txt and execute a Copy activity for each item, using the data to parameterize the source and destination of the copy. This will allow you to copy data from the selected tables in DB1 to the files in storage1, based on the data in File1.txt. The ForEach activity can also be configured to execute its inner activities in parallel, which can help improve performance by allowing multiple copy activities to occur in parallel.

So, the correct answers are **B. Lookup** and **C. ForEach**. Is there anything else you would like to know? 😊

Question 81

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 should you do first?

- A. Disconnect the Microsoft Purview account from the data factory.
- B. Execute the pipeline.
- C. Execute an Azure DevOps build pipeline.
- D. Locate the related asset in the Microsoft Purview portal.

Answer: B

 **Sibaprasad** Highly Voted 5 months ago

B. Execute the Pipeline is correct answer.

Refer link : <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview> and <https://learn.microsoft.com/en-us/azure/data-factory/connect-data-factory-to-azure-purview>

   upvoted 10 times

 **aemilka** 3 months, 3 weeks ago

Correct.

"The lineage data will automatically be captured during the activities execution."

   upvoted 2 times

 **kkk5566** Most Recent 6 days, 3 hours ago

Selected Answer: B

executing the pipeline

   upvoted 1 times

 **vctrhugo** 2 months, 2 weeks ago

Selected Answer: B

By executing the pipeline, the Data Factory will generate the lineage information and propagate it to the connected Microsoft Purview account. This will update the lineage in Purview and reflect any changes made in the pipeline.

   upvoted 2 times

 **Gopinath123** 4 months, 1 week ago

Selected Answer: B

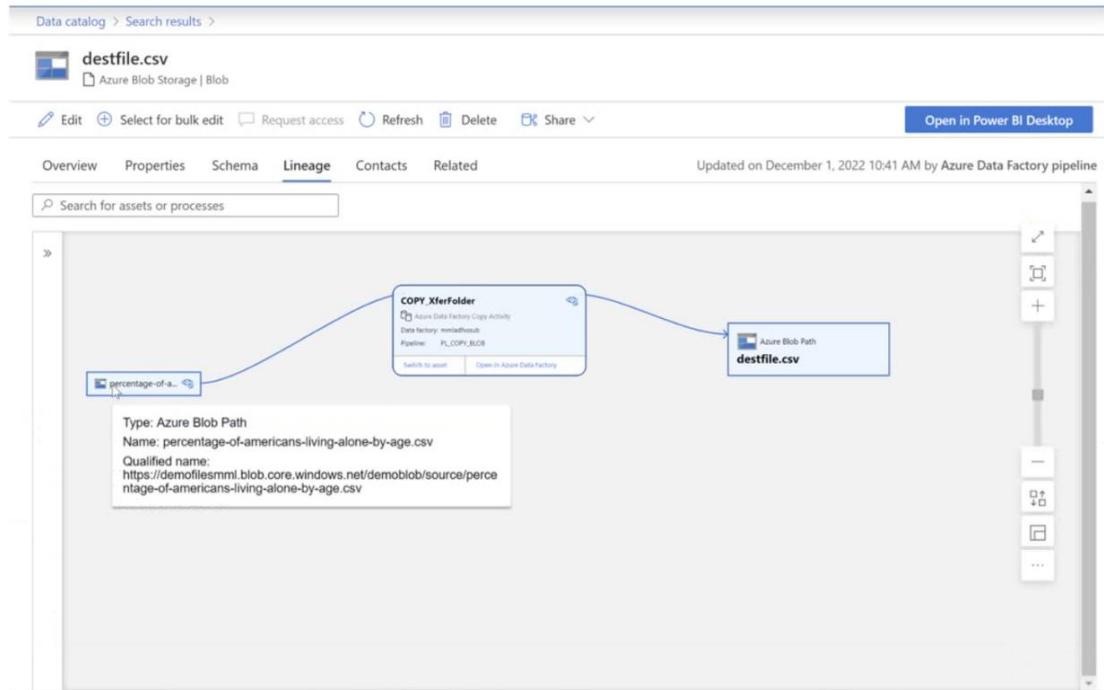
<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview>

   upvoted 2 times

Question 82

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: C

shakes103 Highly Voted 4 months, 2 weeks ago

Selected Answer: C

Answer is C

Find reason here: <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview#run-pipeline-and-push-lineage-data-to-microsoft-purview>

upvoted 5 times

shakes103 4 months, 2 weeks ago

The answer is also displayed on the top right corner of the image displayed.

upvoted 11 times

kkk5566 Most Recent 6 days, 3 hours ago

Selected Answer: C

C is correct

upvoted 1 times

Question 83

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage1. MP1 is configured to scan storage1. DF1 is connected to MP1 and contains a dataset named DS1. DS1 references a file in storage1.

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 search bar in the Microsoft Purview governance portal
- B. the Storage browser of storage1 in the Azure portal
- C. the search bar in the Azure portal
- D. the search bar in Azure Data Factory Studio

ET User: AD

✉️👤 **Sibaprasad** Highly Voted 5 months ago

From ChatGPT :

- A. the search bar in the Microsoft Purview governance portal
- D. the search bar in Azure Data Factory Studio

To review the schema and lineage information in MP1 for the data referenced by DS1, you can use the following two features:

The search bar in the Microsoft Purview governance portal: You can search for the file in storage1 that is referenced by DS1 in the search bar of the Purview governance portal. Once you locate the file, you can view the schema and lineage information for it.

The search bar in Azure Data Factory Studio: You can search for the dataset DS1 in the Azure Data Factory Studio search bar. Once you locate the dataset, you can view the schema and lineage information for the data it references in storage1, which can also be viewed in Purview.

👉👉👉 upvoted 11 times

✉️👤 **chryckie** Highly Voted 4 months, 2 weeks ago

Selected Answer: AD

You need lineage info. Lineage is in Purview. Also, the lineage is all based off what the Data Factory pipeline is doing. I'd say A and D.

<https://learn.microsoft.com/en-us/azure/purview/how-to-search-catalog#searching-microsoft-purview-in-connected-services>

👉👉👉 upvoted 5 times

✉️👤 **kkk5566** Most Recent 6 days, 3 hours ago

Selected Answer: AD

is correct

👉👉👉 upvoted 1 times

✉️👤 **peches** 3 months ago

Selected Answer: AD

If the Data Factory resource is connected to a Purview account there will be a column in the monitoring view of the Pipeline with the lineage status.

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview#step-3-monitor-lineage-reporting-status>

👉👉👉 upvoted 3 times

ET Answer: AB

Bing: AD

To review the schema and lineage information in a Microsoft Purview account named MP1 for the data referenced by a dataset named DS1 in an Azure data factory named DF1, you can use the **search bar in the Microsoft Purview governance portal** or the **search bar in Azure Data Factory Studio**.

In the Microsoft Purview governance portal, you can use the search bar to search for assets and view their schema and lineage information. In Azure Data Factory Studio, you can use the search bar to search for assets in your Purview account and view their schema and lineage information directly within Data Factory Studio.

So, the correct answers are **A. the search bar in the Microsoft Purview governance portal** and **D. the search bar in Azure Data Factory Studio**. Is there anything else you would like to know? 😊

Question 84

HOTSPOT -

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 in which format? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Storage:

Multiple blob storage accounts	▼
Multiple containers in the blob storage account	
Timeslice partitioning in the folders	

Format:

Apache Parquet	▼
CSV	
JSON	

ET User:

Timeslice partitioning in the folders

Apache Parquet

 ababatunde_hs  5 months ago

Time partitioning is correct as the fastest way to load only new files, but requires that the timeslice information be part of the file or folder name (<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-overview>)

However, Parquet is the correct file format since it's a columnar format
   upvoted 33 times

 kkk5566  6 days, 3 hours ago

Time partitioning and parquet
   upvoted 1 times

 vctrhugo 2 months, 2 weeks ago

You need to minimize how long it takes to perform the incremental loads. With Parquet, which is a columnar format, it is way faster to select a few columns than csv.
   upvoted 2 times

 vegeta379 3 months, 1 week ago

we can do incremental load just with deltatable for a parquet file which supported by datarbricks or synapse spark and here he didn't give details so I think it will be CSV
   upvoted 1 times

 pavankr 3 months, 1 week ago

I think the requirement is to select specific columns, hence CSV?

   upvoted 1 times

 verisdev 3 months, 3 weeks ago

it supposed to be Parquet instead of CSV

   upvoted 3 times

ET Answer:

Answer Area

Storage:

Multiple blob storage accounts	▼
Multiple containers in the blob storage account	
Timeslice partitioning in the folders	

Format:

Apache Parquet	▼
CSV	
JSON	

Question 85

DRAG DROP -

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

ET User:

BEGIN DISTRIBUTED TRANSACTION / BEGIN TRAN
ROLLBACK TRAN

OfficeSaracus Highly Voted 4 months ago

Given answer is wrong. It should be BEGIN TRAN as SQL pool in Azure Synapse Analytics does not support distributed transaction.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-develop-transactions>

"Limitations

SQL pool does have a few other restrictions that relate to transactions.

They are as follows:

No distributed transactions
No nested transactions permitted
No save points allowed
No named transactions
No marked transactions
No support for DDL such as CREATE TABLE inside a user-defined transaction
"

Distributed Transactions are only allowed in SQL Server and Azure SQL Managed Instance:

<https://learn.microsoft.com/de-de/sql/t-sql/language-elements/begin-distributed-transaction-transact-sql?view=sql-server-ver16>

upvoted 16 times

janaki Highly Voted 3 months ago

Its BEGIN TRAN
then ROLLBACK TRAN

upvoted 8 times

kkk5566 Most Recent 6 days, 3 hours ago

BEGIN TRAN
ROLLBACK TRAN

upvoted 1 times

ET Answer:

Answer Area

```
BEGIN DISTRIBUTED TRANSACTION

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

```

Question 86

HOTSPOT -

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table1. 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.

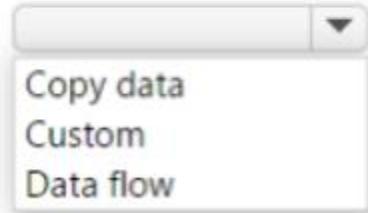
NOTE: Each correct answer is worth one point.

Answer Area

To retrieve the watermark value, use:



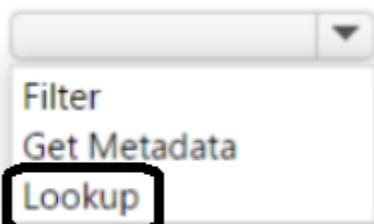
To perform the upload, use:



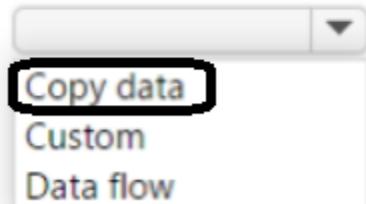
Answer:

Answer Area

To retrieve the watermark value, use:



To perform the upload, use:



OfficeSaracus [Highly Voted] 4 months ago

Seems correct to me

1 like 1 comment 1 upvoted 11 times

DarKru [Highly Voted] 3 months ago

Correct. The example is here

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-portal>

1 like 1 comment 1 upvoted 10 times

kkk5566 [Most Recent] 6 days, 3 hours ago

lookup & copy activity

1 like 1 comment 1 upvoted 1 times

haythemsi 3 months, 4 weeks ago

Filter not lookup, because we have to "Minimize the effort to author the pipeline" and we have only the LastModifiedOn column as information, we are not sure for lookup.

1 like 1 comment 1 upvoted 3 times

auwia 2 months, 1 week ago

The Filter activity in Azure Data Factory is used to filter an array of objects from a previous activity's output (typically from a Lookup activity). It cannot directly query a database or compare a value from a database (watermark in this case) against data in another database.

1 like 1 comment 1 upvoted 4 times

Question 87

HOTSPOT -

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 =  
        'CSV'  
        'DELTA'  
        'JSON' ✓  
        'PARQUET'  
    FIELDTERMINATOR = '0x0b',  
    FIELDQUOTE =  
        '0x09'  
        '0x0a'  
        '0x0b' ✓  
        '0x0c'  
    ROWTERMINATOR = '0x0b'  
)  
WITH (jsondoc nvarchar(max) AS JsonDocuments
```

Answer:

Answer Area

```
SELECT *
FROM OPENROWSET
(
    BULK
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',
    FORMAT = 'CSV'
    FIELDTERMINATOR = '0x0b',
    FIELDQUOTE = '0x09'
    ROWTERMINATOR = '0x0d'
)
WITH (jsondoc nvarchar(max) AS JsonDocuments
```

✉  **Yemeral**  4 months ago

Correct. It's weird but best way to open a json is as a csv and with 0x0b for fieldterminator and fieldquote.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files>

   upvoted 22 times

✉  **kkk5566**  6 days, 3 hours ago

Correct

   upvoted 1 times

Question 88

You use Azure Data Factory to create data pipelines.

You are evaluating whether to integrate Data Factory and GitHub for source and version control.

What are two advantages of the integration? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. additional triggers
- B. lower pipeline execution times
- C. the ability to save without publishing
- D. the ability to save pipelines that have validation issues

Answer: CD

✉️ akk_1289 **Highly Voted** 4 months ago

- C. the ability to save without publishing
- D. the ability to save pipelines that have validation issues

When you integrate Data Factory and GitHub, you can save your pipelines to a GitHub repository without publishing them to Azure. This allows you to work on your pipelines in a development environment and then publish them to Azure when you are ready.

You can also save pipelines that have validation issues. This is because GitHub does not validate your pipelines when you save them. This allows you to work on your pipelines and fix the validation issues before you publish them to Azure.

👍 ↗️ 🚩 upvoted 15 times

Question 89

DRAG DROP -

You have an Azure Synapse Analytics workspace named Workspace1.

You perform the following changes:

- Implement source control for Workspace1.
- Create a branch named Feature based on the collaboration branch.
- Switch to the Feature branch.
- Modify Workspace1.

You need to publish the changes to Azure Synapse.

From which branch should you perform each change? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Branches	Answer Area
Collaboration	Create a pull request: <input type="text"/>
Publish	Publish the changes: <input type="text"/>
Feature	

Answer:

Answer Area

Create a pull request: Feature

Publish the changes: Collaboration

Question 90

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. Run the Copy Data tool and select Metadata-driven copy task.
- B. Create a pipeline that contains a Data Flow activity.
- C. Create a pipeline that contains a flowlet.
- D. Run the Copy Data tool and select Built-in copy task.

ET User: D

✉️  Sabbath  2 months, 3 weeks ago

Selected Answer: D

Just use Built-in copy task, according to: <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-lastmodified-copy-data-tool>

   upvoted 10 times

✉️  kkk5566  6 days, 3 hours ago

Selected Answer: D

D is correct

   upvoted 1 times

✉️  auwia 2 months, 1 week ago

Selected Answer: D

Create a data factory.

Use the Built-in Copy Data tool to create a pipeline.

Monitor the pipeline and activity runs.

   upvoted 2 times

✉️  vctrhugo 2 months, 2 weeks ago

Selected Answer: D

"[...] use the Copy Data tool to create a pipeline that incrementally copies new and changed files only, from Azure Blob storage to Azure Blob storage. It uses LastModifiedDate to determine which files to copy."

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-lastmodified-copy-data-tool>

   upvoted 3 times

ET Answer: A

Bing: A

To create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from one Azure Blob Storage account to another, and ensure that the pipeline only copies blobs that were created or modified since the most recent replication event while minimizing the effort to create the pipeline, you should run the

Copy Data tool and select Metadata-driven copy task. The Metadata-driven copy task allows you to configure the pipeline to copy only new or changed files based on their last modified time, which can help ensure that only the relevant blobs are copied. The Copy Data tool provides a user-friendly interface for creating and configuring the pipeline, which can help minimize the effort required to create the pipeline. So, the correct answer is **A. Run the Copy Data tool and select Metadata-driven copy task.** Is there anything else you would like to know? 😊

Question 91

You have an Azure Data Factory pipeline named pipeline1 that contains a data flow activity named activity1.

You need to run pipeline1.

Which runtime will be used to run activity1?

- A. Azure Integration runtime
- B. Self-hosted integration runtime
- C. SSIS integration runtime

Answer: A

✉️  **kkk5566** 6 days, 3 hours ago

Selected Answer: A

correct

   upvoted 1 times

✉️  **auwia** 2 months, 1 week ago

Selected Answer: A

Probably the correct answer.

   upvoted 2 times

Bing: A

The runtime that will be used to run a data flow activity named activity1 in an Azure Data Factory pipeline named pipeline1 is the **Azure Integration runtime**. The Azure Integration runtime is the compute infrastructure used by Azure Data Factory to execute data flow activities. It provides a fully managed, serverless environment for running data flows at scale. So, the correct answer is **A. Azure Integration runtime**. Is there anything else you would like to know? 😊

Question 92

HOTSPOT -

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQLPool1 and an Apache Spark pool named sparkpool1. Sparkpool1 contains a DataFrame named pyspark_df.

You need to write the contents of pyspark_df to a table in SQLPool1 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

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")  
%%local  
%%spark  
%%sql  
  
val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")  
scala_df.write.  
    jdbc ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)  
    saveAsTable  
    synapsesql
```

Answer:

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")  
%%local  
%%spark  
%%SQL  
  
val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")  
scala_df.write.  
    jdbc ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)  
    saveAsTable  
    synapsesql
```

  **Azure_2023**  2 months, 3 weeks ago

Correct

<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/synapse-spark-sql-pool-import-export?tabs=scala%2Cscala1%2Cscala2%2Cscala3%2Cscala4%2Cscala5>

   upvoted 5 times

  **kkk5566**  6 days, 3 hours ago

%%spark

&&

df.write.synapsesql

   upvoted 1 times

Question 93

You have an Azure data factory named ADF1 and an Azure Synapse Analytics workspace that contains a pipeline named SynPipeline1. SynPipeline1 includes a Notebook activity. You create a pipeline in ADF1 named ADPPipeline1.

You need to invoke SynPipeline1 from ADPPipeline1.

Which type of activity should you use?

- A. Web
- B. Spark
- C. Custom
- D. Notebook

Answer: A

✉  **ludaka** Highly Voted  2 months, 2 weeks ago

Selected Answer: A

Web Activity

<https://learn.microsoft.com/en-us/azure/data-factory/solution-template-synapse-notebook>

   upvoted 5 times

✉  **kkk5566** Most Recent  6 days, 2 hours ago

Selected Answer: A

is correct

   upvoted 1 times

✉  **Mani_V** 2 months, 1 week ago

its a notebook activity

<https://learn.microsoft.com/en-us/azure/data-factory/solution-template-synapse-notebook>

   upvoted 1 times

✉  **vctrhugo** 2 months, 1 week ago

Selected Answer: A

Web calls a Synapse pipeline with a notebook activity.

   upvoted 2 times

Question 94

HOTSPOT -

You have an Azure data factory that contains the linked service shown in the following exhibit.

Edit linked service

 Azure SQL Database [Learn more](#)

 To avoid publishing immediately to Data Factory, please use Azure Key Vault to retrieve secrets securely. Learn more [here](#)

Name *

AzureSqlDatabase1

Description

Connect via integration runtime * 

AutoResolveIntegrationRuntime

[Connection string](#)

[Azure Key Vault](#)

Account selection method 

From Azure subscription Enter manually

Fully qualified domain name *

ssio2022.database.windows.net

Database name *

Contoso

Authentication type *

SQL authentication

User name *

SQLAdmin

[Password](#)

[Azure Key Vault](#)

Password *

.....

Always encrypted 

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 answer is worth one point.

Answer Area

When working in a feature branch, changes to the linked service will be published to the live service

upon publishing the changes
upon saving the changes
when the changes are merged into the collaboration branch

A Copy activity that uses the linked service as the source will perform the Copy activity

in the region of the data factory
in the region of the selected external compute
in the region of the source database

ET User:

upon publishing to the change / upon saving the changes

In the region of the data factory

peches Highly Voted 2 months, 3 weeks ago

According to Microsoft, AutoResolveIntegrationRuntime will attempt to use the sink location to get an IR in the same region (or the closest available) to execute the Copy activity, not the source location. I would go with the region of data factory, since that is the default option when the sink's location is not detectable. Source: <https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#azure-ir-location>

upvoted 11 times

kkk5566 Most Recent 6 days, 2 hours ago

upon publishing changes to the service

upvoted 1 times

Galvanir 1 month, 2 weeks ago

the first one should be "upon publishing changes to the service". See <https://learn.microsoft.com/en-us/azure/data-factory/source-control>

upvoted 4 times

abarakadabra200 1 month, 4 weeks ago

Shouldn't we choose the **Upon saving the changes** option in the first dropdown?
Link: <https://learn.microsoft.com/en-us/azure/data-factory/source-control#stale-publish-branch>

upvoted 2 times

andjurovicela 2 months, 2 weeks ago

I did not manage to find a clear answer to this one, but based on cross-reading a few articles, I think "in the region of data factory" should be the correct answer, and this article explains it a bit better than the others I found: <https://asankap.wordpress.com/2021/10/26/why-you-shouldnt-use-auto-resolve-integration-runtime-in-azure-data-factory-or-synapse/>

upvoted 2 times

JG1984 2 months, 2 weeks ago

When using the AutoResolveIntegrationRuntime with a Copy activity in Azure Data Factory that uses a linked service as the source, the copy operation will be performed in the region of the source data store.

The AutoResolveIntegrationRuntime is a system-assigned integration runtime that automatically routes data movement and activity dispatch to the optimal region based on the location of the source and sink data stores. When using a linked service as the source, the service will attempt to detect the location of the source data store and use an Integration Runtime in the same region to perform the copy operation.

upvoted 1 times

vctrhugo 2 months, 2 weeks ago

For copy activity, a best effort is made to automatically detect your sink data store's location, then use the IR in either the same region, if available, or the closest one in the same geography, otherwise; if the sink data store's region is not detectable, the IR in the instance's region is used instead.

upvoted 1 times

mehroosali 2 months, 3 weeks ago

correct

upvoted 1 times

ET Answer:

Answer Area

When working in a feature branch, changes to the linked service will be published to the live service

upon publishing the changes
upon saving the changes
when the changes are merged into the collaboration branch

A Copy activity that uses the linked service as the source will perform the Copy activity

in the region of the data factory
in the region of the selected external compute
in the region of the source database

Question 95

HOTSPOT -

In Azure Data Factory, you have a schedule trigger that is scheduled in Pacific Time.

Pacific Time observes daylight saving time.

The trigger has the following JSON file.

```
{  
    "name": "Trigger 1",  
    "properties": {  
        "annotations": [],  
        "runtimeState": "Started",  
        "pipelines": [],  
        "type": "ScheduleTrigger",  
        "typeProperties": {  
            "recurrence": {  
                "frequency": "Week",  
                "interval": 1,  
                "startTime": "2022-08-05T04:00:00",  
                "timeZone": "Pacific Standard Time",  
                "schedule": {  
                    "minutes": [  
                        0  
                    ],  
                    "hours": [  
                        3,  
                        21  
                    ],  
                    "weekDays": [  
                        "Sunday",  
                        "Saturday"  
                    ]  
                }  
            }  
        }  
    }  
}
```

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.

Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

one time
two times
zero times

The trigger [answer choice] daylight saving time.

is unaffected by
will automatically adjust for
will require an adjustment for

ET User: 有争议

1. two times
2. will automatically adjust for / will require an adjustment for

✉ **ludaka** Highly Voted 2 months, 2 weeks ago

1. two times
2. will automatically adjust

"For time zones that observe daylight saving, trigger time will auto-adjust for the twice a year change, if the recurrence is set to Days or above. To opt out of the daylight saving change, please select a time zone that does not observe daylight saving, for instance UTC."

<https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger?tabs=data-factory#azure-data-factory-and-synapse-portal-experience>

1 upvoted 8 times

✉ **JezWalters** Most Recent 1 month ago

There's a catch here, as daylight savings actually starts on the SECOND Sunday of March, and March 3 2024 is before this date.

1 upvoted 1 times

✉ **vctrhugo** 2 months, 2 weeks ago

"[...] we are also adding support for Daylight Saving auto-adjustment: for time zones that observe Daylight Saving, auto change schedule trigger time twice a year (e.g. 8AM daily trigger will fire at 8AM, whether it's PST or PDT)"

<https://techcommunity.microsoft.com/t5/azure-data-factory-blog/time-zone-and-daylight-saving-support-for-schedule-trigger/ba-p/1840199>

1 upvoted 2 times

✉ **iVath** 2 months, 2 weeks ago

2nd answer should be : will require an adjustment for

ref to : <https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger?tabs=data-factory>

The timeZone element specifies the time zone that the trigger is created in. This setting affects both startTime and endTime.

1 upvoted 2 times

✉ **peches** 2 months, 3 weeks ago

Agree, as of 2020 ADF supports auto-adjustment for Daylight Saving in Schedule Triggers for time zones that aren't UTC. Since here we are using Pacific time, answer seems correct. Source: <https://techcommunity.microsoft.com/t5/azure-data-factory-blog/time-zone-and-daylight-saving-support-for-schedule-trigger/ba-p/1840199>

1 upvoted 4 times

✉ **wendy** 2 months, 3 weeks ago

Azure Data Factory only supports time zones UTC. I think should require the adjustment.

1 upvoted 1 times

✉ **vctrhugo** 2 months, 2 weeks ago

Incorrect. As of 2020 you can create schedule triggers in your local time zone, without the need to convert timestamps to Coordinated Universal Time (UTC) first.

1 upvoted 2 times

ET Answer:

Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

one time
two times
zero times

The trigger [answer choice] daylight saving time.

is unaffected by
will automatically adjust for
will require an adjustment for

Question 96

You have an Azure Synapse Analytics dedicated SQL pool.

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 for a downstream activity. The solution must minimize development effort.

Which type of activity should you use in the pipeline?

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

ET User:

C (68%)

B (32%)

✉ **ludaka** Highly Voted 2 months, 2 weeks ago

Selected Answer: C

For me the correct answer is C.

The store procedure activity doesn't return any data.

In the description of the script activity is written that it can be used for : "Run stored procedures. If the SQL statement invokes a stored procedure that returns results from a temporary table, use the WITH RESULT SETS option to define metadata for the result set."

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

👉 ↵ 🗞 upvoted 7 times

✉ **andjurovicela** 2 months, 2 weeks ago

I also think this one is correct. One of the things script activity can do is "...Save the rowset returned from a query as activity output for downstream consumption." which is pretty much what is needed here. This is not viable with 'execute SP' activity as it doesn't return any data.

👉 ↵ 🗞 upvoted 2 times

✉ **auwia** Highly Voted 2 months, 1 week ago

Selected Answer: C

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

The script may contain either a single SQL statement or multiple SQL statements that run sequentially. You can use the Script task for the following purposes:

Truncate a table in preparation for inserting data.

Create, alter, and drop database objects such as tables and views.

Re-create fact and dimension tables before loading data into them.

Run stored procedures. If the SQL statement invokes a stored procedure that returns results from a temporary table, use the WITH RESULT SETS option to define

metadata for the result set.

Save the rowset returned from a query as activity output for downstream consumption.

👉 ↵ 🗞 upvoted 5 times

✉ **KKK5566** Most Recent 6 days, 2 hours ago

Selected Answer: C

C is correct

👉 ↵ 🗞 upvoted 1 times

✉ **CoinUmbrella** 2 months ago

Selected Answer: B

B. Chat GPT says the given answer is correct. Stored Procedure is specifically designed to execute stored procedures within Azure Synapse Analytics and is the most suitable option for the scenario, minimizing development effort.

👉 ↵ 🗞 upvoted 1 times

✉️  **auwia** 2 months, 1 week ago

Selected Answer: B

The "Script" activity in Azure Data Factory is primarily used to run HDInsight scripts such as Hive, Pig, MapReduce, and Spark. These are typically used for big data processing tasks.

   upvoted 1 times

✉️  **auwia** 2 months, 1 week ago

False, finally I've found the link, it's C:
<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>

   upvoted 3 times

✉️  **vctrhugo** 2 months, 2 weeks ago

Selected Answer: B

"In Azure Synapse Analytics, you can use the SQL pool Stored Procedure Activity to invoke a stored procedure in a dedicated SQL pool."

<https://learn.microsoft.com/en-us/azure/synapse-analytics/data-integration/sql-pool-stored-procedure-activity>

   upvoted 2 times

✉️  **sandpl203** 2 months, 3 weeks ago

Selected Answer: B

<https://learn.microsoft.com/en-us/azure/synapse-analytics/data-integration/sql-pool-stored-procedure-activity>

   upvoted 2 times

ET Answer: B

Bing: B

To create a pipeline that will execute a stored procedure in an Azure Synapse Analytics dedicated SQL pool and use the returned result set as the input for a downstream activity while minimizing development effort, you should use a

Stored Procedure activity in the pipeline. The Stored Procedure activity allows you to execute a stored procedure in a dedicated SQL pool and use the returned result set as the input for a downstream activity. This can help minimize development effort by allowing you to reuse existing stored procedures and easily integrate them into your pipeline.

So, the correct answer is **B. Stored Procedure**. Is there anything else you would like to know? 😊

Question 97

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline1.

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 pipeline1 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? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

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

ET User: 好像是 CD 的最多，这里没统计

AC (57%)

AD (43%)

Mal2002 1 day, 20 hours ago

Here is an example of how we can use the Script activity to execute SP1:

```
Script activity (name: "ExecuteSP1Script")
{
    ScriptSource = "<![CDATA[
var results = SqlCommand('EXEC SP1', connection);
var myVar = results[0];
]]>"}
In this example, the ScriptSource property specifies the script that is used to execute SP1. The script first executes the SQL statement EXEC SP1. The script then stores the results of SP1 in the variable myVar.
```

Correct Answers are: A & D

upvoted 1 times

kkk5566 6 days, 2 hours ago

Selected Answer: AC

is correct

upvoted 1 times

andie123 1 month, 1 week ago

why not CD?

A Lookup activity can be used to execute a query or stored procedure against a data source and retrieve a single row of data. The returned values can then be stored as pipeline variables and used in subsequent activities.

A Stored Procedure activity can be used to directly execute a stored procedure against a data source. The returned values can be captured as output parameters and stored as pipeline variables for use in subsequent activities.

upvoted 4 times

✉️ **Ram9198** 2 months ago
Selected Answer: AC
<https://learn.microsoft.com/en-us/answers/questions/925742/how-to-process-output-from-stored-procedure-in-azu>
SP Activity does not capture result.. use lookup instead of script
👍 ↗️ 🎉 upvoted 3 times

✉️ **Ram9198** 2 months ago
sorry, Answer - AC
👍 ↗️ 🎉 upvoted 1 times

✉️ **Ram9198** 2 months ago
Answer is wrong - <https://learn.microsoft.com/en-us/answers/questions/925742/how-to-process-output-from-stored-procedure-in-azu>
Answer is CD. SP act cannot emit result..
👍 ↗️ 🎉 upvoted 2 times

✉️ **Mani_V** 2 months, 1 week ago
CD is the rite answer
👍 ↗️ 🎉 upvoted 1 times

✉️ **vctrhugo** 2 months, 2 weeks ago
Selected Answer: AD
<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure>
<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>
👍 ↗️ 🎉 upvoted 3 times

✉️ **WayOps** 2 months, 3 weeks ago
C. Lookup
D. Stored Procedure
Explanation:
C. Lookup: The Lookup activity is used to retrieve a dataset from a data source and the output can be used in subsequent activities. It is often used to fetch a small amount of data to be used as parameters in other activities. In this case, it can be used to execute the stored procedure and capture the result into pipeline variables.
D. Stored Procedure: The Stored Procedure activity is used specifically to execute stored procedures. You can capture the output of the stored procedure and assign it to pipeline variables. This activity is designed specifically for executing stored procedures, making it a direct option for this requirement.
A. Script: There is no "Script" activity in Azure Data Factory.
B. Copy: The Copy activity is primarily used for copying data from a source to a destination and is not suitable for executing a stored procedure and capturing its output into pipeline variables.
👍 ↗️ 🎉 upvoted 3 times

✉️ **vctrhugo** 2 months, 2 weeks ago
There is Script activity in ADF.
"The script may contain either a single SQL statement or multiple SQL statements that run sequentially. You can use the Script task for the following purposes:
[...] Run stored procedures. [...]"
<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>
👍 ↗️ 🎉 upvoted 3 times

✉️ **mehroosali** 2 months, 3 weeks ago
I think the correct answer is C and D.
👍 ↗️ 🎉 upvoted 1 times

✉️ **vctrhugo** 2 months, 2 weeks ago
You use lookup to consume, not to get.
👍 ↗️ 🎉 upvoted 1 times

ET Answer: AD

Bing: CD

To execute a stored procedure named SP1 in an Azure SQL database named DB1 from an Azure Data Factory data pipeline named pipeline1 and store the values in the columns as pipeline variables, you can use a **Lookup** activity or a **Stored Procedure** activity.

The **Lookup** activity can be used to execute a stored procedure and retrieve a dataset that can be used as input for other activities in the pipeline.

The **Stored Procedure** activity allows you to call a stored procedure in Azure SQL Database. You can use an output parameter to return the result set from the stored procedure and store it as a pipeline variable.

So, the correct answers are **C. Lookup** and **D. Stored Procedure**. Is there anything else you would like to know? 😊

Question 98

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 Azure Data Factory Studio 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. From the Azure Data Factory Studio, run Publish All.
- B. Create an Azure Data Factory trigger.
- C. Create a Git repository.
- D. Create a GitHub action.
- E. From the Azure Data Factory Studio, select Set up code repository.
- F. From the Azure Data Factory Studio, select Publish.

Answer: CE

  **Paulkuzzio** 1 week, 4 days ago

Answer is highly correct. I know this one for sure.

   upvoted 2 times

To implement version control for the changes made to pipeline artifacts in an Azure data factory named ADF1, you should **create a Git repository** and **set up a code repository from the Azure Data Factory Studio**.

Creating a Git repository provides a platform where multiple users can collaborate on the same project at the same time. It allows you to track changes, create branches, and more.

Setting up a code repository from the Azure Data Factory Studio allows you to connect your data factory with the Git repository. This way, all the changes made to pipeline artifacts can be saved and version controlled in the Git repository.

So, the correct answers are **C. Create a Git repository** and **E. From the Azure Data Factory Studio, select Set up code repository**. Is there anything else you would like to know? 😊

Question 99

You have an Azure data factory named ADF1 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: D

✉  **kkk5566** 6 days, 2 hours ago

Selected Answer: D

seem to correct

   upvoted 1 times

✉  **Paulkuzzio** 1 week, 4 days ago

It seems correct but the 15mins offset is throwing me off. Somebody please explain. Thanks

   upvoted 2 times

Question 100

HOTSPOT -

You have a Microsoft 365 E5 subscription that contains the users shown in the following table.

Name	Member of
User1	Group1, Group2
User2	Group2
User3	Group3

In Microsoft Endpoint Manager, you have the Policies for Office apps configurations shown in the following table.

Name	Priority	Assigned to	Microsoft Word default file format
Policy1	0	Group1	OpenDocument Text (*.odt)
Policy2	1	Group2	Word XML Document (*.xml)

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

Answer Area

Statements	Yes	No
User1 saves new Word documents by default in the ODT file format.	<input type="radio"/>	<input type="radio"/>
User2 saves new Word documents by default in the XML file format.	<input type="radio"/>	<input type="radio"/>
User3 saves new Word documents by default in the DOCX file format.	<input type="radio"/>	<input type="radio"/>

Answer:

Answer Area

Statements	Yes	No
User1 saves new Word documents by default in the ODT file format.	<input checked="" type="checkbox"/>	<input type="radio"/>
User2 saves new Word documents by default in the XML file format.	<input checked="" type="checkbox"/>	<input type="radio"/>
User3 saves new Word documents by default in the DOCX file format.	<input checked="" type="checkbox"/>	<input type="radio"/>

  shaden2000  7 months, 3 weeks ago

YYY

Policy 1 applies correctly on user1 (odt)

Policy 2 applies correctly on user2 (XML)

No policy for user 3 so the default of Docx will be sustained.

   upvoted 8 times

Common Question 3

Question 1

DRAG DROP -

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.

Select and Place:

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 user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.
Create a database role named Role1 and grant Role1 SELECT permissions to dw1.	Create a database role named Role1 and grant Role1 SELECT permissions to schema1.
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 Role1 to the Group1 database user.	

  Rob77  2 years, 3 months ago

1. create user from external provider for Group1
2. create Role1 with select on schema1
3. add user to the Role1

   upvoted 89 times

Question 2

HOTSPOT -

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.

Hot Area:

Answer Area

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:

Answer Area

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.

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>

  **Francesco1985**  2 years, 2 months ago

Guys the answers are correct: <https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview>

   upvoted 55 times

  **Siena** 1 year, 11 months ago

Agreed. "Link each server with two key vaults that reside in different regions and hold the same key material, to ensure high availability of encrypted databases. Mark only the key from the key vault in the same region as a TDE protector. System will automatically switch to the key vault in the remote region if there is an outage affecting the key vault in the same region."

<https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview>

   upvoted 5 times

Question 3

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

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

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

NOTE: Each correct selection is worth one point.

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

Answer: AC

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

The screenshot shows the Azure portal interface for a database named 'MySampleDatabase2'. On the left, there's a sidebar with various management options like Power Platform, Settings, Integrations, Security, and Intelligent Performance. The main area is titled 'Classification' under the 'Data Discovery & Classification' section. It displays a list of 15 columns from the 'SalesLT' schema in the 'Customer' table, each with a recommended classification. The columns listed are FirstName, LastName, EmailAddress, Phone, PasswordHash, PasswordSalt, UserName, AddressLine1, AddressLine2, City, PostalCode, AddressType, AccountNumber, CreditCardApprovalCode, and 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
	7.125	Microsoft SQL Server Management Studio - Query	1	847	847	C244A066-2271-...	Confidential - GDPR
	7.125	Microsoft SQL Server Management Studio - Query	2	32	32	C244A066-2271-...	Confidential
	7.125	Microsoft SQL Server Management Studio - Query	41	32	32	A7088FD4-759E-...	Confidential, Confidential - GDPR

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

Question 4

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, DP-300 好像也遇到这道题了，也是争议很大

Column-level security simplifies the design and coding of security in your application, allowing you to restrict column access to protect sensitive data.

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

<https://www.examtopics.com/discussions/microsoft/view/18788-exam-dp-201-topic-3-question-12-discussion/>

<https://www.examtopics.com/discussions/microsoft/view/52626-exam-dp-203-topic-3-question-4-discussion/>

- ✉ **Aleksx42** 2 years, 2 months ago
C is the right answer. Check the discussion here:
<https://www.examtopics.com/discussions/microsoft/view/18788-exam-dp-201-topic-3-question-12-discussion/>
 upvoted 32 times
- ✉ **ant069** 1 year, 7 months ago
yeah, from ms docs: "ensuring that specific users can access only certain columns of a table pertinent to their department"
 upvoted 2 times
- ✉ **mikerss** 2 years, 2 months ago
the key word is 'infer'. as listed in the below documentation, data masking is not used to protect against malicious intent to infer the underlying data. I would therefore choose C
 upvoted 10 times
- ✉ **Marcus1612** 1 year, 11 months ago
I agree with mikerss, the key word is 'infer'. Data masking is a kind of column-level security but it is only partial. A malicious person could infer the credit card number. The good answer is C
 upvoted 3 times
- ✉ **Deeksha1234** 1 year ago
I agree with the logic provided
 upvoted 2 times
- ✉ **Tracy_Anderson** 2 years, 1 month ago
The link below show how you can infer a column that is data masked. It is also referenced in the 201 topic, <https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver15>
 upvoted 2 times
- ✉ **FredNo** 1 year, 9 months ago
Selected Answer: C
Data masking does not protect against inferring with the data
 upvoted 10 times
- ✉ **tyler4kn** 2 years, 10 months ago
I just took the exam with 100% on the "Design for data security and compliance (25-30%)" category, from which this question is clearly a part of.
The question asks for viewing all entries, which means All Rows, not all columns from a row. Also, because the requirement is that the salespeople can't view or infer data, the answer CANNOT be Data Masking.
The correct answer is "Column-level security".
 upvoted 37 times
- ✉ **Neendum** 2 years, 6 months ago
I seem to agree with you that credit Card data masking function may not satisfy the requirements of this question as the credit card can be inferred from the last four digits. However, I can still use other masking functions like default or custom to ensure that the credit card details cannot be inferred thereby achieving the same objectives
 upvoted 2 times
- ✉ **hokigir** 3 years, 4 months ago
No I disagree it says "You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers.". Key word here is => all entries. Data masking is correct.
 upvoted 25 times
- ✉ **Leonido** 3 years, 4 months ago
The question is "The solution must prevent all the salespeople from viewing or inferring the credit card information." I'm not sure what they meant by "inferring", but masking doesn't prevent you from manipulating data of the credit card field, so you can do some assumptions. So I would go with CLS.
 upvoted 4 times
- ✉ **Kashan_Ali** 2 years, 11 months ago
It means they can view the Credit card information but don't get the actual information as it is encrypted. So the correct answer is Data masking. This way Sales people can view entire table but credit card information is masked.
 upvoted 1 times

Bing: A

The best solution to provide salespeople with the ability to view all the entries in Customers while preventing them from viewing or inferring the credit card information would be to use **data masking**. Data masking is a way to create a fake, but realistic version of your organizational data. The goal is to protect sensitive data, while providing a functional alternative when real data is not needed—for example, in user training, sales demos, or software testing. Data masking processes change the values of the data while using the same format ¹. This would allow salespeople to view all entries in the Customers table without being able to view or infer the credit card information. So, the correct answer is **A. data masking**.

Question 5

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

Answer: ACE

AC: Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 accounts.

E: Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system

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

✉ vctrhugo 2 months, 2 weeks ago

Selected Answer: ACE

Create security groups in Azure Active Directory (Azure AD) and add project members: Start by creating the necessary security groups in Azure AD and adding the project members to these groups. This step allows you to organize users and manage their permissions collectively.

Configure access control lists (ACL) for the Azure Data Lake Storage account: Next, configure the access control lists (ACL) for the Azure Data Lake Storage account. ACLs provide granular control over permissions at the individual file or folder level within the storage. By setting up ACLs, you can define specific access rights for different data assets.

Assign Azure AD security groups to Azure Data Lake Storage: Once the security groups and ACLs are set up, assign the Azure AD security groups to the Azure Data Lake Storage account. This step associates the security groups with the storage resources and enables you to grant permissions based on group membership rather than individually managing permissions for each user.

Like 14 Upvote 4 times

Question 6

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 vault1.
- B. Enable Azure role-based access control on vault1.
- C. Remove the linked service from Df1.
- D. Create a self-hosted integration runtime.

Answer: C

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

Incorrect Answers:

D: A self-hosted integration runtime copies data between an on-premises store and cloud storage.

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>

  **gnulf69**  1 year, 11 months ago

I believe this is correct, based on the question: What should you do FIRST?
A DF needs to be empty to be encrypted: <https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key#post-factory-creation-in-data-factory-ui>
So FIRST we need to empty the DF - then we can move on.

   upvoted 38 times

Question 7

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. column-level security
- B. dynamic data masking
- C. row-level security (RLS)
- D. sensitivity classifications

Answer: D

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.

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

Question 8

HOTSPOT -

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.

Hot Area:

Answer Area

Use	▼	to authenticate by using	▼
<input type="checkbox"/>		Azure Active Directory (Azure AD)	<input type="checkbox"/>
<input type="checkbox"/>		a shared access signature (SAS)	<input type="checkbox"/>
<input type="checkbox"/>		a shared key	<input type="checkbox"/>
			<input type="checkbox"/>
			<input type="checkbox"/>
			<input type="checkbox"/>

Answer:

Answer Area

Use	▼	to authenticate by using	▼
<input checked="" type="checkbox"/>		Azure Active Directory (Azure AD)	<input checked="" type="checkbox"/>
<input type="checkbox"/>		a shared access signature (SAS)	<input type="checkbox"/>
<input type="checkbox"/>		a shared key	<input type="checkbox"/>
			<input checked="" type="checkbox"/>
			<input checked="" type="checkbox"/>
			<input checked="" type="checkbox"/>

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>

✉️  **Podavenna**  1 year, 11 months ago

Correct Answer!

   upvoted 29 times

✉️  **kkk5566**  6 days, 13 hours ago

correct

   upvoted 1 times

✉️  **Deeksha1234** 1 year ago

correct

   upvoted 3 times

✉️  **juanlu46** 1 year, 4 months ago

It's make sense. You only authorised Data Factory instance by Azure Active Directory, you don't need to share keys that can be retriever by users. And this option meet all of the requirements.

   upvoted 2 times

Question 9

HOTSPOT -

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.

Hot Area:

Answer Area

Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	<input type="radio"/>	<input type="radio"/>
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	<input type="radio"/>	<input type="radio"/>
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	<input type="radio"/>	<input type="radio"/>

ET User: NNN/YNY

- ✉ **steeee** 2 years ago
The Answer should be No, No, No. Analysts have access to in-region sensitive data, so the first one should be No. Engineers have access to all numeric sensitive data, Height is patient's height in CM, so the second and third one should also No.
 upvoted 119 times
- ✉ **auwia** 2 months, 1 week ago
The first 2 are false because the Patient table is not present in Region A. NO Medical for region A. Probably the third one is Yes because of comma between table name and height, so probably it means give access to other columns too.
 upvoted 1 times
- ✉ **Amalbenrebai** 2 years ago
I agree: NO NO NO
 upvoted 13 times
- ✉ **Seansmyrke** 1 year, 6 months ago
Its no,yes,yes
Engineers only have access to numeric data. the contact email is considered sensitive in the regions and is not numeric
 upvoted 10 times
- ✉ **HaBroNounen** 1 year, 11 months ago
the solution is correct: Yes, no, yes. Just because somebody has access, doesn't mean that they don't need any dynamic masking. It just means that they have access and a policy is required. If they had no access, then obviously no data masking is required.
Statement 1: Analysts in Region A have access to (all) the following sensitive data in region A: CardOnFile, Height and ContactEmail. Since financial (CardOnFile) and PII (ContactEmail) are considered sensitive data you need dynamic data masking: so Yes.
Statement 2 & 3: Engineers have access to all numeric sensitive data (which means in every region). So they have access to height. Height is medical and therefore only sensitive in Region B according to the second table, but not in Region A. So Statement 2 is "No" and Statement 3 is "Yes"
 upvoted 74 times
- ✉ **Julius7000** 1 year, 11 months ago
I think You are correct
 upvoted 6 times
- ✉ **noranathalie** 1 year, 10 months ago
I would go for this answer as well.. otherwise the double question 2 and 3 would be useless..
 upvoted 2 times
- ✉ **YLiu** 1 year, 10 months ago
But for statement 1, [height] is not considered sensitive data for Region A, so it should not require data mask on [height]. -> A is NO
Also I am confused about whether we should apply the policy of sensitive data based on the region of data or the region of the requester (eg engineer from region C requesting data of region A)?
 upvoted 2 times

ET Answer:

Answer Area

Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	<input checked="" type="radio"/>	<input type="radio"/>
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	<input type="radio"/>	<input checked="" type="radio"/>
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	<input checked="" type="radio"/>	<input type="radio"/>

Question 10

DRAG DROP -

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.

Select and Place:

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

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 -

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

✉ Sudheer_K Highly Voted 1 year, 11 months ago

Answer is right!

Like Share Flag upvoted 16 times

✉ Liz42 Highly Voted 1 year, 10 months ago

Shouldn't the last two be switched? Enable TDE then configure the key?

Like Share Flag upvoted 7 times

✉ anto69 1 year, 7 months ago

I also think so, but not sure

Like Share Flag upvoted 1 times

✉ noranathalie 1 year, 10 months ago

I think the correct answer is the one provided.

Please see the link below:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-configure?tabs=azure-powershell>

Like Share Flag upvoted 9 times

✉ shoottheduck 6 months, 1 week ago

Checked this link and it supports the answer given

Like Share Flag upvoted 2 times

✉ kkk5566 Most Recent 6 days, 13 hours ago

correct

Like Share Flag upvoted 1 times

✉ hanzocuk 8 months ago

1. Get a KV

2. Add key to KV

3. Assign MI to server

4. Enable TDE

5. Config TDE

Like Share Flag upvoted 3 times

✉ Okea 7 months, 1 week ago

4 and 5 should be swapped

Like Share Flag upvoted 2 times

Bing:

To implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1, you should perform the following actions in sequence:

1. Create an Azure key vault and grant the managed identity permissions to the key vault [1](#).
2. Add key1 to the Azure key vault [2](#).
3. Assign a managed identity to Server1 [3](#).
4. Configure key1 as the TDE protector for Server1 [4](#).
5. Enable TDE on Pool1 [5](#).

Question 11

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

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.

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

Question 12

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: A

You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

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

  Canary, 2021  1 year, 8 months ago

Selected Answer: A

A is the correct Answer.

You can specify the number of partitions at the time of creating an event hub. In some scenarios, you may need to add partitions after the event hub has been created. This article describes how to dynamically add partitions to an existing event hub.

Dynamic additions of partitions is available only in premium and dedicated tiers of Event Hubs.

<https://docs.microsoft.com/en-us/azure/event-hubs/dynamically-add-partitions>

   upvoted 12 times

Question 13

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 during queries?

- A. HASH
- B. REPLICATE
- C. ROUND_ROBIN

Answer: B

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.

Incorrect Answers:

A: A hash distributed table is designed to achieve high performance for queries on large tables.

C: A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods.

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

Question 14

HOTSPOT -

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.

Hot Area:

Answer Area

```
df.write  
    .bucketBy  
    .format  
    .partitionBy  
    .sortBy  
    .mode ("append")
```

(“*”)
("GeographyRegionID")
("GeographyRegionID", "Year", "Month", "Day")
("Year", "Month", "Day", "GeographyRegionID")

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

ET User:

1. Partition by
2. GeographyRegionID, Year, Month, Day as the pipelines are per region this seems right choice
3. Parquet

✉️ PallaviPatel **Highly Voted** 1 year, 8 months ago

1. Partition by
2. GeographyRegionID, Year, Month, Day as the pipelines are per region this seems right choice
3. Parquet

👍👎FLAG upvoted 81 times

✉️ uzairahm 1 year, 2 months ago

regarding point 2 Solution needs to support daily incremental load so having Year, Month, Day first would be more useful
👍👎FLAG upvoted 3 times

✉️ petulda **Highly Voted** 2 years ago

I suggest storing the data in parquet
👍👎FLAG upvoted 50 times

ET Answer:

Answer Area

```
df.write  
.bucketBy  
.format  
.partitionBy  
.sortBy  
.mode("append")
```

```
(“*”)  
("GeographyRegionID")  
("GeographyRegionID", "Year", "Month", "Day")  
("Year", "Month", "Day", "GeographyRegionID")
```

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

Question 15

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 is worth one point.

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

ET User:

AC (48%)	AB (48%)	3%
----------	----------	----

 **alexleonvalencia**  1 year, 8 months ago
Selected Answer: AC
Respuesta A/C
   upvoted 15 times

 **VJPR** 1 year, 8 months ago
why not RBAC?
   upvoted 6 times

 **sensaint** 8 months ago
Assuming RBAC is already in place, predicate function for row-level security would be next step. However, it's not clearly stated in question which makes it confusing.
   upvoted 1 times

 **zizonesol** 5 months, 3 weeks ago
That's why I went with AB instead because it wasn't mentioned. Therefore, we should assume that the system does not already have the RBAC already in place.
   upvoted 4 times

 **lukeonline**  1 year, 8 months ago
Selected Answer: AB
A and B
   upvoted 14 times

 **kkk5566**  6 days, 13 hours ago
Selected Answer: AC
correct
   upvoted 1 times

✉ **kkk5566** 6 days, 13 hours ago

Selected Answer: AB

a and B
Like upvoted 1 times

✉ **kkk5566** 6 days, 13 hours ago
sorry a and C
Like upvoted 1 times

✉ **pavankr** 2 months ago

RBAC is for to use "internal" company. So 100% wrong.
Like upvoted 2 times

✉ **auwia** 2 months, 1 week ago

Selected Answer: AC

<https://learn.microsoft.com/en-gb/training/modules/implement-compliance-controls-sensitive-data/5-implement-row-level-security>

Create the users or groups you want to isolate access.
Create the inline table-valued function that will filter the results based on the predicate defined.
Create a security policy for the table, assigning the function created above.

Like upvoted 1 times

✉ **klaytech** 2 months, 3 weeks ago

Option C, a predicate function, is not wrong. It can be a helpful tool for fine-grained control over data access. However, it is not strictly necessary for the solution described in the question. A security policy and a custom RBAC role can be used to achieve the desired outcome without a predicate function.

Here is an example of how you could use a security policy and a custom RBAC role to control access to data in a Synapse Analytics dedicated SQL pool:

Create a security policy that defines who can access data, what data they can access, and how they can access it.
Create a custom RBAC role that gives users specific permissions to data in the pool.
Assign the custom RBAC role to users based on their company name.
With this configuration in place, users will only be able to access data that they are authorized to access.

The use of a predicate function can be helpful in some scenarios, but it is not always necessary. In the case of the question, a security policy and a custom RBAC role are sufficient to achieve the desired outcome.

Like upvoted 3 times

✉ **mamahani** 4 months, 1 week ago

I think A/C as per examples in docs:
<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#CodeExamples>
i dont think its RBAC; according to documentation Synapse RBAC is used to manage who can:

Publish code artifacts and list or access published code artifacts,
Execute code on Apaches Spark pools and Integration runtimes,
Access linked (data) services protected by credentials
Monitor or cancel job execution, review job output, and execution logs."
I do not see the direct link with limiting retrieved data here;
<https://learn.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-synapse-rbac>
Like upvoted 2 times

✉ **esaade** 5 months, 3 weeks ago

Selected Answer: BC

Answer: B and C

To ensure that users from each company can view only the data of their respective company in an Azure Synapse Analytics dedicated SQL pool, you can use custom role-based access control (RBAC) roles to define specific permissions for each company, and use predicate functions to apply row-level security (RLS) to restrict access based on company membership. By doing this, you can limit the scope of access to the appropriate company data.

A security policy is a mechanism for implementing automatic security controls to enforce compliance requirements, which may not be directly related to company-specific data access.

A column encryption key is used for encrypting sensitive data, but it does not necessarily restrict access based on company membership.

Asymmetric keys are used for secure communication and authentication, but they do not directly relate to company-specific data access control.

Like upvoted 2 times

✉ **jz10** 5 months, 1 week ago
ChatGPT isn't always reliable
Like upvoted 4 times

✉ **janaki** 3 months, 1 week ago
@jz10 you're correct. After ChatGPT answers any of your certification exam questions, you then type -- sure? ChatGPT will change its answer...so 'Yes' ChatGPT is not reliable.
Like upvoted 2 times

✉ **AHUI** 6 months ago

A, C
<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#CodeExamples>
Like upvoted 2 times

✉ **haidebelogime** 6 months, 1 week ago

Selected Answer: B

The answer is B

Like upvoted 1 times

- ✉  **yogiazaad** 7 months, 3 weeks ago
Given answer is correct.
Below from Microsoft documentation:
"A multi-tenant application can create a policy to enforce a logical separation of each tenant's data rows from every other tenant's rows. Efficiencies are achieved by the storage of data for many tenants in a single table. Each tenant can see only its data rows."
<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16>
👉 ↵ 🗞 upvoted 1 times
- ✉  **Taou** 8 months, 1 week ago
Selected Answer: AC
A and C must go together, so I think the right answer is AC
👉 ↵ 🗞 upvoted 1 times
- ✉  **juamrd** 9 months ago
According to Microsoft documentation:
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
ON Sales.Orders
WITH (STATE = ON);
GO
So the answer are A and C
👉 ↵ 🗞 upvoted 6 times
- ✉  **AzureJobsTillRetire** 9 months ago
Selected Answer: AB
Given answer is correct.
D & E are obviously wrong.
C (in Row Level Security) is not necessary and may not be the right solution either. The best way to secure data is not to allow users to access the data at all. For example, we can store data in different databases or schemas and use RBAC to control user access. Row level security first gives users access to the data (in the table that contains all the data for all users) and then restrict data access to a particular part of the table. This is always less secure than not giving user access to the tables that do not contain any data the user should not have access to. Furthermore, Row Level Security may be breached by guessing work queries. I have done that before and I'm quite confident that I can breach any Row Level Security in SQL database but do not want to elaborate here.
👉 ↵ 🗞 upvoted 4 times
- ✉  **AzureJobsTillRetire** 8 months, 2 weeks ago
Please disregard my previous comments. For purpose of the exam, the answer to the question is AC.
👉 ↵ 🗞 upvoted 1 times
- ✉  **dmitriypo** 10 months ago
Selected Answer: AC
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
ON Sales.Orders
WITH (STATE = ON);
GO
<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=azure-sqldw-latest>
👉 ↵ 🗞 upvoted 1 times
- ✉  **allagowf** 10 months, 4 weeks ago
Selected Answer: AB
for those who select C : this function is to apply RLS on inserting and updating, but not selecting so other user can read the data.
[https://azure.microsoft.com/en-gb/blog/sql-database-row-level-security-block-predicates/#:~:text=Block%20predicates%20address%20a%20common,SQL%20Database%20\(V12\)%20server.](https://azure.microsoft.com/en-gb/blog/sql-database-row-level-security-block-predicates/#:~:text=Block%20predicates%20address%20a%20common,SQL%20Database%20(V12)%20server.)
👉 ↵ 🗞 upvoted 5 times
- ✉  **k18585** 9 months, 3 weeks ago
the link you provided talks about BLOCK predicates, the option in this question says "filter predicates". And in the article you linked it states "Whereas filter predicates apply to read operations, block predicates apply to write operations". Please don't give wrong answer/reference
👉 ↵ 🗞 upvoted 2 times

ET Answer: AB

A: Row-Level Security (RLS) enables you to use group membership or execution context to control access to rows in a database table. Implement RLS by using the CREATE SECURITY POLICYTransact-SQL statement. B: Azure Synapse provides a comprehensive and fine-grained access control system, that integrates:

Azure roles for resource management and access to data in storage,
 ☈ Synapse roles for managing live access to code and execution,
 ☈ SQL roles for data plane access to data in SQL pools.

Bing: AC

To ensure that users from each company can view only the data of their respective company in an Azure Synapse Analytics dedicated SQL pool, you should include **a security policy** and **a predicate function** in the solution. A security policy is a database object that groups together one or more security predicates. A predicate function is a user-defined function that returns a Boolean value. You can use a security policy to filter rows based on a predicate function, which determines whether a row should be visible to the user or not. So, the correct answers are **A. a security policy** and **C. a predicate function**.

Question 16

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.

ET User:

A (68%)	B (32%)
---------	---------

edba Highly Voted 1 year, 8 months ago
I think it's a terrible question, both A(using T-SQL) and B (via GUI) can do the job.
 upvoted 18 times

rzeng Highly Voted 10 months, 1 week ago
Selected Answer: A
Go with A, reason for not B, if email column is string type ,default masking will make it as xxxxxxxx, so here I go with email mask on email column.
<https://learn.microsoft.com/en-us/azure/sql-database/dynamic-data-masking-overview?view=azuresql>
 upvoted 10 times

andjurovicela 1 month ago
but you don't have to opt for the default and the add masking rule step from the link shows the exact same format as in the task. Therefore I would go with B to avoid overthinking :D
 upvoted 1 times

kkk5566 Most Recent 6 days, 13 hours ago
Selected Answer: B
A or B
 upvoted 1 times

Ram9198 3 weeks, 1 day ago
Selected Answer: A
default masking will make it as xxxxxxxx,
 upvoted 1 times

Ram9198 2 months ago
Selected Answer: A
B says just mask and not email mask
 upvoted 2 times

- ✉ **auwia** 2 months, 1 week ago
Selected Answer: B
The link provided in the solution is correctly pointing to the solution: Dynamic Data Masking, that is done from the Azure Portal, so the correct answer is B! :)
 upvoted 3 times
- ✉ **Shanmahi** 9 months ago
Selected Answer: B
email masking option via ssms
 upvoted 1 times
- ✉ **OldSchool** 9 months, 3 weeks ago
Selected Answer: B
Vote for B because of "You set up a dynamic data masking policy in the Azure portal by selecting the Dynamic Data Masking blade under Security in your SQL Database configuration pane."
Source: <https://learn.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview?view=azuresql#:~:text>You%20set%20up%20a%20dynamic%20data%20masking%20policy%20in%20the%20Azure%20portal%20by%20selecting%20the%20Dynamic%20Data%20Masking%20blade%20under%20Security%20in%20your%20SQL%20Database%20configuration%20pane>.
 upvoted 1 times
- ✉ **amitshinde14** 11 months, 2 weeks ago
B correct
 upvoted 1 times
- ✉ **Deeksha1234** 1 year ago
both A and B are correct
 upvoted 1 times
- ✉ **ROLLINGROCKS** 1 year, 1 month ago
Selected Answer: A
Occams razor with this one
 upvoted 2 times
- ✉ **Glen711** 1 year, 1 month ago
Selected Answer: A
There are lots of comments here saying that the question does not ask for the default masking format. I'd be interested in hearing from people who saw this question on the exam. Because the way I read this question - it IS asking for the default format. There's just a line break in the question. The text says "in a format of a XXX@XXXX.com" it's just that someone with less command of English put a space between the "a" and the "XXX" so the space got turned into a line break.
So I think that if the question is actually the default format, then "A".
 upvoted 3 times
- ✉ **StudentFromAus** 1 year, 2 months ago
The answer should be B as it's not the default email mask format.
 upvoted 1 times
- ✉ **Navthing** 1 year, 2 months ago
Selected Answer: A
Both A & B are correct But I will prefer A.
 upvoted 2 times
- ✉ **NamitSehgal** 1 year, 2 months ago
Sorry A is my proffered way, I can not edit my earlier comment.
 upvoted 2 times
- ✉ **NamitSehgal** 1 year, 2 months ago
Both A and B are correct
For SQLMI, it can not be done for portal from for Azure SQL and Azure Synapse, we just need to click Add Mask for one or more columns and then Save to apply a mask for these fields.
Yes, using SSMS is my preferred way as we have handoff policy from portal, all should be automated. So I go for B.
 upvoted 1 times

ET Answer: A

The Email masking method, which exposes the first letter and replaces the domain with XXX.com using a constant string prefix in the form of an email address. aXX@XXXX.com

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

Bing: B

To prevent nonadministrative users from seeing the full email addresses in the Email column of the dbo.Customers table in a SQL pool in Azure Synapse, you should set a mask on the Email column from the Azure portal. This will allow you to define a custom mask format for the Email column, such as `aXXX@XXXX. com`, which will replace the original email addresses with masked values when nonadministrative users query the table. So, the correct answer is B. From the Azure portal, set a mask on the Email column.

Question 17

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. an Azure Active Directory (Azure AD) user
- B. a shared key
- C. a shared access signature (SAS)
- D. a managed identity

Answer: D

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

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

👤 ploer **Highly Voted** 1 year, 7 months ago

Selected Answer: D

D is the way we do it in our company. So it works at least.
Upvoted 9 times

👤 PallaviPatel **Highly Voted** 1 year, 8 months ago

the answer and explanation given is correct.
Upvoted 7 times

👤 kkk5566 **Most Recent** 6 days, 13 hours ago

Selected Answer: D

Vnet = managed identity
Upvoted 1 times

👤 vctrhugo 2 months, 2 weeks ago

Selected Answer: D

VNet = managed identity
Upvoted 3 times

👤 janaki 3 months, 1 week ago

Vnet = managed identity
Upvoted 2 times

👤 yogiazaad 7 months, 3 weeks ago

Answer is correct.

The blow link has more details.

"Analytic capabilities such as Dedicated SQL pool and Serverless SQL pool use multi-tenant infrastructure that is not deployed into the managed virtual network. In order for traffic from these capabilities to access the secured storage account, you must configure access to your storage account based on the workspace's system-assigned managed identity by following the steps below."
<https://learn.microsoft.com/en-us/azure/synapse-analytics/security/connect-to-a-secure-storage-account#grant-your-azure-synapse-workspace-access-to-your-secure-storage-account-as-a-trusted-azure-service>

Upvoted 2 times

Question 18

HOTSPOT -

You have an Azure Synapse Analytics SQL pool named Pool1. In Azure Active Directory (Azure AD), you have a security group named Group1.

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.

NOTE: Each correct selection is worth one point.

Hot Area:

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:

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	

Box 1: GRANT -

You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.

Box 2: CREATE SECURITY POLICY -

Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.

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

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

✉ **RajBathani** Highly Voted 1 year, 8 months ago

Correct Answer

1 upvoted 19 times

✉ **HaBroNouen** Highly Voted 1 year, 8 months ago

Answer is correct.

for Row Level Security: <https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver15>

For Column Level Security: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security>

1 upvoted 10 times

✉ **janaki** 3 months, 1 week ago

You are correct! :-)

1 upvoted 2 times

✉ **kkk5566** Most Recent 6 days, 13 hours ago

Row = CREATE SECURITY POLICY

Column = GRANT

1 upvoted 1 times

✉ **vctrhugo** 2 months, 2 weeks ago

Row = CREATE SECURITY POLICY

Column = GRANT

1 upvoted 2 times

✉ **vrodriguesp** 6 months, 3 weeks ago

correct, as documentation claims:

to control access to the columns-->Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.

to control access to the rows) -->You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.

1 upvoted 3 times

✉ **vctrhugo** 2 months, 1 week ago

It should be swaped.

Row = CREATE SECURITY POLICY

Column = GRANT

1 upvoted 2 times

✉ **Deeksha1234** 1 year ago

correct

1 upvoted 1 times

✉ **Remedios79** 1 year, 2 months ago

Correct!

1 upvoted 1 times

✉ **juanlu46** 1 year, 4 months ago

Totally correct!

1 upvoted 1 times

Question 19

HOTSPOT -

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.

Hot Area:

Answer Area

Tier:	<input type="checkbox"/>	▼
	<input checked="" type="checkbox"/>	Premium
	<input checked="" type="checkbox"/>	Standard

Advanced option to enable:	<input type="checkbox"/>	▼
	<input checked="" type="checkbox"/>	Azure Data Lake Storage Credential Passthrough
	<input checked="" type="checkbox"/>	Table Access Control

Answer:

Answer Area

Tier:	<input type="checkbox"/>	▼
	<input checked="" type="checkbox"/>	Premium
	<input checked="" type="checkbox"/>	Standard

Advanced option to enable:	<input type="checkbox"/>	▼
	<input checked="" type="checkbox"/>	Azure Data Lake Storage Credential Passthrough
	<input checked="" type="checkbox"/>	Table Access Control

Box 1: Premium -

Credential passthrough requires an Azure Databricks Premium Plan

Box 2: Azure Data Lake Storage credential passthrough

You can access Azure Data Lake Storage using Azure Active Directory credential passthrough.

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.

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

Question 20

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

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

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

✉️  **PallaviPatel**  1 year, 8 months ago

correct

   upvoted 11 times

✉️  **Jerrie86**  7 months, 2 weeks ago

Whenever you see Vnet , answer is usually managed Identity

   upvoted 5 times

✉️  **kkk5566**  5 days, 21 hours ago

Selected Answer: A

correct

   upvoted 1 times

Question 21

You are developing an application that uses Azure Data Lake Storage Gen2.

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

Answer: B

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.

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

  **bad_atitude**  1 year, 8 months ago

Agree with the answer => B

   upvoted 18 times

  **kkk5566**  5 days, 21 hours ago

Selected Answer: B

correct

   upvoted 1 times

Question 22

HOTSPOT -

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.

Hot Area:

Answer Area

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

ET User:

Azure Active Directory credential Passthrough

SAS

 **ItHYMeRish** Highly Voted  1 year, 8 months ago

Accessing the ADLS via Databricks should be using Azure Active Directory with Passthrough. Accessing the files in ADLS should be SAS, based on the options provided.

The explanation provided for this question is incorrect.

   upvoted 50 times

 **Billybob0604** 8 months, 3 weeks ago

This is it. Correct

   upvoted 2 times

 **edba** 1 year, 7 months ago

To be more clear, for box it shall be user delegation SAS which is secured with ADD credentials.

   upvoted 2 times

 **vivekazure** Highly Voted  1 year, 7 months ago

1. Accessing the Databricks should be using Personal Tokens
2. Accessing the ADLS should be using Shared Access Signatures. (Because of controlled access to project folders they work).

   upvoted 13 times

ET Answer:

Answer Area

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	

Question 23

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

Answer: E

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.

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

Question 24

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

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.

<https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine>

 juanlu46  1 year, 4 months ago

Selected Answer: C

By discard, is C, you can create a symmetric key to encrypt a data, for example one column, and then use this data as feature of the model
<https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/encrypt-a-column-of-data?view=sql-server-ver15>

The other options that not meet the requirements:

- TDE encrypt data, but decrypt when you query <https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-tde-overview?tabs=azure-portal>
- RLS is for row restriction, not meet the requirement
- Azure AD pass-through is for authentication

   upvoted 14 times

Question 25

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 contains a folder named Folder1 that has a URI of <https://adls1.dfs.core.windows.net/container1>.

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

ET User:



✉ ki8585 **Highly Voted** 9 months, 3 weeks ago

Selected Answer: CD

Phrased different, the question for me says: if you create "Folder3" inside Folder2, you should be able to read files created in Folder3.

This means that you for sure need Executive and Read permissions to Folder2 (Executive to traverse child folder, read to read the files).

Now, starting from the least privilege, suppose you give "Access" permission both for read and execute. In this case, you can't read files created in Folder3. This is a requirement ("child items that are created in Folder2"), so you need Default Read access.

You don't need Default Execute, otherwise you would have access to a Folder created in Folder3 (say Folder 4) and this is not required so for the least privilege you must give Access Execute and not Default Execute.

Like Share Flag upvoted 14 times

✉ yogiazaad 7 months, 3 weeks ago

Requirement 1 says Traverse child items that are created in Folder2. Means that you need to be able to traverse the subFolders under Folder2. So Default:Execute is a required permission.

Like Share Flag upvoted 2 times

✉ bokLuci **Highly Voted** 10 months, 1 week ago

Selected Answer: CD

C - You need to traverse the Folder2 only and no potential children folders - Principals of least privilege.

D- You need to pass on the READ access to the files in Folder2. Default ACLs are not passed to files but we are not setting the permission on a file level, we are setting it on Folder2.

Like Share Flag upvoted 8 times

✉ kkk5566 **Most Recent** 5 days, 21 hours ago

"Default - Read" and "Default - Execute"

Like Share Flag upvoted 1 times

✉ Ram9198 3 weeks, 1 day ago

Selected Answer: CD

Traverse require access execute, file reads need default read

Like Share Flag upvoted 1 times

✉ Ram9198 2 months ago

Selected Answer: DF

Default Execute is mandatory to traverse child items through cascade.. Default Read by process of elimination

Like Share Flag upvoted 2 times

✉ auwia 2 months, 1 week ago

Selected Answer: AF

⇒ Traverse child items that are created in Folder2. => DEFAULT EXECUTE

⇒ Read files that are created in Folder2. => ACCESS READ (that was already given).

Like Share Flag upvoted 2 times

✉ esaade 5 months, 3 weeks ago

Selected Answer: DF

Based on the permissions table provided, the ServicePrincipal1 has "Access - Execute" permission on container1, "Access - Execute" permission on Folder1, and "Access - Read" permission on Folder2. To allow ServicePrincipal1 to traverse child items that are created in Folder2 and read files created in Folder2, you should grant the "Default - Read" and "Default - Execute" permissions on Folder2. The "Default - Read" permission allows ServicePrincipal1 to read files created in Folder2, and the "Default - Execute" permission allows ServicePrincipal1 to traverse child items that are created in Folder2.

Therefore, the correct answer is:

D. Default - Read

F. Default - Execute

Like Share Flag upvoted 5 times

✉ yogiazaad 7 months, 3 weeks ago

Traverse child items that are created in Folder2.

This needs Default:Execute Because user needs to traverse any child items(Sub Folders) created under under Folder2.

Read files that are created in Folder2.

Since the The Access:read ACL is already set on Folder2.Any files that are created under Folder2 can be access by User. But to see (or list) the items/files under Folder2 we need Access:Execute .

SO the answer is Access: Execute and Default: Execute

Like Share Flag upvoted 3 times

✉ AzureJobsTillRetire 9 months ago

Selected Answer: DF

Default Read and Execute are required. The reason is as below.

In the POSIX-style model that's used by Data Lake Storage Gen2, permissions for an item are stored on the item itself. In other words, permissions for an item cannot be inherited from the parent items if the permissions are set after the child item has already been created. Permissions are only inherited if default permissions have been set on the parent items before the child items have been created.

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

Like Share Flag upvoted 5 times

✉ Deeksha1234 1 year ago

Selected Answer: DF

so the answer is correct

Like Share Flag upvoted 3 times

✉ Deeksha1234 1 year ago

I think the given answer is correct. Since we should be able to traverse and read the child items from the folder 2 .

From one of the DP 203 Microsoft lab exercise -

Access ACLs control access to an object. Files and directories both have access ACLs.

Default ACLs are templates of ACLs associated with a directory that determine the access ACLs for any child items that are created under that directory. Files do not have default ACLs.

Like Share Flag upvoted 3 times

✉️  **Davico93** 1 year, 2 months ago

Selected Answer: AF

Default is not related to files so, if we want to read files, we need ACCESS - READ

   upvoted 3 times

✉️  **Aditya0891** 1 year, 2 months ago

Please make a note how the sentence is framed "Traverse child items that are created in Folder2". Access ACL doesn't propagate the permissions to child items but default ACL does. So it is obvious that new files or folders can be created in Folder2 and that requires default ACL. So according to me default execute and default read on folder2 should be the correct answer

   upvoted 1 times

✉️  **Aditya0891** 1 year, 2 months ago

Please ignore this. It's not correct. Examtopics should provide a delete option here.

   upvoted 4 times

✉️  **sdokmak** 1 year, 3 months ago

Following principal of least privilege, isn't Access Execute and Default Read enough? You only need to traverse the files in Folder2, not the folders within Folder2 (even though there aren't any)

   upvoted 4 times

✉️  **virendrapsingh** 1 year, 3 months ago

Agreed with your comment on least privilege as it is mentioned specifically in the question.

Choices A & F should be the answer.

   upvoted 5 times

✉️  **Aditya0891** 1 year, 2 months ago

sdokmak not sure but it's not mentioned that there are only files inside folder2 and in the next line it specifically mentioned that to read files inside folder 2. I think the answers are correct as per requirement. Please correct me if I'm wrong

   upvoted 1 times

✉️  **MadEgg** 1 year, 3 months ago

Selected Answer: DF

Correct

   upvoted 1 times

✉️  **juanlu46** 1 year, 4 months ago

Selected Answer: DF

Is correct!

   upvoted 2 times

ET Answer: DF

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.

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

Question 26

HOTSPOT -

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.

Hot Area:

Answer Area

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:

Answer Area

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

Box 1: Azure AD authentication -

Azure AD authentication has the option to include MFA.

Box 2: Contained database users -

Azure AD authentication uses contained database users to authenticate identities at the database level.

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

- Skeinofi** 1 year, 8 months ago
Correct
 upvoted 20 times
- Amsterliese** 1 year, 4 months ago
"SQL Database and Azure Synapse Analytics support Azure Active Directory identities as contained database users"
<https://docs.microsoft.com/en-us/sql/relational-databases/security/contained-database-users-making-your-database-portable?view=sql-server-ver15#contained-database-user-model>
 upvoted 6 times
- kkk5566** 5 days, 21 hours ago
Correct
 upvoted 1 times
- JG1984** 2 months, 2 weeks ago
Azure Synapse Analytics supports two types of database-level authentication:
Azure Active Directory (Azure AD) authentication: This uses your Azure AD identity to authenticate to Synapse SQL. This is the recommended authentication method, as it provides a single sign-on experience and allows you to manage permissions using Azure AD groups.
SQL Server authentication: This uses a traditional SQL Server username and password to authenticate to Synapse SQL. This authentication method is less secure than Azure AD authentication, but it may be necessary if you are using legacy applications that do not support Azure AD.
 upvoted 1 times
- Deeksha1234** 1 year ago
answer is correct
 upvoted 3 times
- dev2dev** 1 year, 7 months ago
B is wrong. Contained users not supported by synapse analytics. D is correct ('MS SQL Server logins')
 upvoted 4 times
- PallaviPatel** 1 year, 7 months ago
<https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-aad-overview> this document says contained users are supported by synapse analytics, so this is correct answer.
 upvoted 14 times
- dev2dev** 1 year, 7 months ago
correct
 upvoted 1 times

Question 27

DRAG DROP -

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.

Select and Place:

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.	

ET User:

评论多的一：

Step 1: Create a Log Analytics workspace that has Data Retention set to 120 days.

Step 2: From Azure Portal, add a diagnostic setting.

Step 3: Select the PipelineRuns Category

Step 4: Send the data to a Log Analytics workspace.

评论多的二：

step 1. From Azure Portal, add a diagnostic setting.

step 2. Send data to a Log analytics workspace.

step 3. Create a Log Analytics workspace that has Data Retention set to 120 days.

step 4. Select the PipelineRuns Category.

✉️👤 **Sunnyb** **Highly Voted** 2 years, 3 months ago

Step 1: Create a Log Analytics workspace that has Data Retention set to 120 days.
Step 2: From Azure Portal, add a diagnostic setting.
Step 3: Select the PipelineRuns Category
Step 4: Send the data to a Log Analytics workspace.

👍👎FLAG upvoted 165 times

✉️👤 **hercilian_effort** **Highly Voted** 2 years, 1 month ago

step 1. From Azure Portal, add a diagnostic setting.
step 2. Send data to a Log analytics workspace.
step 3. Create a Log Analytics workspace that has Data Retention set to 120 days.
step 4. Select the PipelineRuns Category.

The video in the below link walks you through the process step by step, start watching at 2min 30sec mark

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor#keeping-azure-data-factory-metrics-and-pipeline-run-data>

👍👎FLAG upvoted 39 times

Answer:

Actions

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 Area

Create an Azure Storage account that has a lifecycle policy.

Create a Log Analytics workspace that has Data Retention set to 120 days.

From the Azure portal, add a diagnostic setting.

Send the data to a Log Analytics workspace.

Question 28

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

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.

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

✉️  **damaldon**  2 years, 2 months ago

Correct!

   upvoted 39 times

✉️  **kkk5566**  5 days, 21 hours ago

 **Selected Answer: B**

correct

   upvoted 1 times

✉️  **vctrhugo** 2 months, 1 week ago

 **Selected Answer: B**

Transparent Data Encryption (TDE) is a feature provided by Azure SQL Database and Azure Synapse Analytics that encrypts the database files at rest. It performs real-time I/O encryption and decryption of the database files, ensuring that the data is encrypted on disk. TDE operates transparently and does not require any changes to the application code or queries.

By enabling TDE for the dedicated SQL pool in Azure Synapse Analytics, you can achieve encryption at rest for the data stored in the pool without impacting the applications that access the data.

   upvoted 1 times

Question 29

DRAG DROP -

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.

Select and Place:

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"/>
Write	directory1: <input type="text"/>
Execute	file1: <input type="text"/>

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.

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

✉️ **ank84** Highly Voted 12 months ago

-Execute
-Execute
-Write
 upvoted 9 times

✉️ **Matt2000** 3 weeks, 5 days ago

Supported by the following two references:

without additional permissions: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>
with additional permissions such as storage blob data reader: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model#permissions-table-combining-azure-rbac-abac-and-acls>

upvoted 1 times

✉️ **AlviraTony** 1 week ago

In the above link, the use case is given for appending to Data.txt file, then the answers would be
-Execute
-Execute
-Read and Write
 upvoted 1 times

✉️ **dom271219** Highly Voted 11 months, 4 weeks ago

Correct : Execute to traverse the folders and Write to append the file
 upvoted 5 times

✉️ **kkk5566** Most Recent 5 days, 21 hours ago

correct
 upvoted 1 times

✉️ **bakamon** 3 months, 1 week ago

container1 : Read access [by default because User1 that is assigned the Storage Blob Data Reader role for storage1]

directory1: Execute [since requirement is only to append file1 so traverse (execute) permission will be enough for it]

file1 : Write [because execute cannot append the file in Azure Data Lake Storage Gen2]
only write permission can append a file.

upvoted 1 times

✉️ **OldSchool** 9 months ago

Can't remember if the wording on actual exam was the same or very similar but instead of Append was Delete and the Q was like this:
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 delete file1. The solution must use the principle of least privilege.
Permission:

--WX
---X

Answer Area and my answers:
container1 ---X
directory1 ---X
file1 --WX
 upvoted 4 times

✉️ **mamahani** 4 months, 1 week ago

i dont think you gave correct answers;
see this doc: <https://learn.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control#common-scenarios-related-to-permissions>
to delete a file you dont need any permissions on the file itself; only on the folder where it resides (read + execute)

upvoted 1 times

✉️ **Matt2000** 3 weeks, 5 days ago

mamahani is correct. See the following references:
<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>
<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model#permissions-table-combining-azure-rbac-abac-and-acls>
 upvoted 1 times

✉️ **renukahouse** 1 month, 2 weeks ago

if you give write access to entire folder , the user can delete/modify other folders , whihc is not correct
 upvoted 1 times

✉️ **vctrhugo** 2 months, 1 week ago

The solution must use the principle of least privilege. You shouldn't do -WX on folder, only on file.
 upvoted 2 times

Question 30

HOTSPOT -

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.

Hot Area:

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

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.

1. Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service.
2. Set up Hive Metastore linked service
3. Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.
4. Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.
5. 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.

6. Provide User name and Password to set up the connection.

7. Test connection to verify the username and password.

8. Click Create to create the linked service.

Box 2: A Hive Metastore -

<https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

✉ **federc** Highly Voted 12 months ago

I would say:

1. sql - this is correct
2. managed hive metastore

👉 ↵ 🗞 upvoted 7 times

✉ **federc** 12 months ago

scrath that, given anwers are correct. sql + hive metastore

👉 ↵ 🗞 upvoted 14 times

✉ **kkk5566** Most Recent 5 days, 20 hours ago

1. sql db
2. hive metastore

👉 ↵ 🗞 upvoted 1 times

✉ **andjurovicela** 1 month, 3 weeks ago

1 - definitely correct per documentation TestingCRM provided.
2 - I think the devil's in the detail here :/ documentation says "Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog".
The word managed may sway you towards the answer managed hive metasotre SERVICE but the docs don't mention "service" at all, which is why I would go with Hive metastore

👉 ↵ 🗞 upvoted 1 times

✉ **TestingCRM** 3 months ago

1. sql - this is correct
2. managed hive metastore

See <https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

👉 ↵ 🗞 upvoted 1 times

Question 31

HOTSPOT -

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.
- * Administrative effort 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.

Hot Area:

To minimize costs:

Locally-redundant storage (LRS)

The Archive access tier

The Cool access tier

Zone-redundant storage (ZRS)

To delete blobs:

Azure Automation runbooks

Azure Storage lifecycle management

Soft delete

ET User:

LRS and Lifecycle storage mgt

✉ goxxx **Highly Voted** 11 months, 2 weeks ago

If u choose premium storage account, there is no possibility to choose tiers (hot, cool, archive), its always hot, so LRS and lifecycle storage mngt

1 like 1 reply 1 upvoted 31 times

✉ mambahani 4 months, 1 week ago

its not the same as hot; see this microsoft article: <https://azure.microsoft.com/nl-nl/blog/azure-premium-block-blob-storage-is-now-generally-available/>
"Premium Blob Storage is a new performance tier in Azure Blob Storage for block blobs and append blobs, complimenting the existing Hot, Cool, and Archive access tiers."

1 like 1 reply 1 upvoted 1 times

✉ allagowf 10 months, 1 week ago

Agree no mention for tiering in the question so LRS is the best option to minimize the cost

1 like 1 reply 1 upvoted 4 times

✉ dom271219 **Highly Voted** 11 months, 4 weeks ago

The statement doesn't mention requirement for a tiering storage archive nor cool nor hot before deletion.

Then I think it is LRS and lifecycle storage mngt

1 like 1 reply 1 upvoted 14 times

✉ kkk5566 **Most Recent** 5 days, 20 hours ago

LRS and LCM

1 like 1 reply 1 upvoted 1 times

✉ Ram9198 3 weeks, 1 day ago

As per the response from the Microsoft <https://github.com/MicrosoftDocs/azure-docs/issues/100695> tiering is not supported for premium but delete through LCM is supported.. but still not clearly mentioned in this document <https://learn.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

Answer LRS and LCM

1 like 1 reply 1 upvoted 1 times

ET Answer:

To minimize costs:

Locally-redundant storage (LRS)
The Archive access tier
The Cool access tier
Zone-redundant storage (ZRS)

To delete blobs:

Azure Automation runbooks
Azure Storage lifecycle management
Soft delete

Question 32

HOTSPOT -

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.

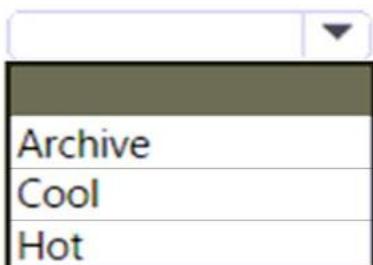
You need to recommend a data retention solution. The solution must minimize costs.

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.

Hot Area:

First 30 days:



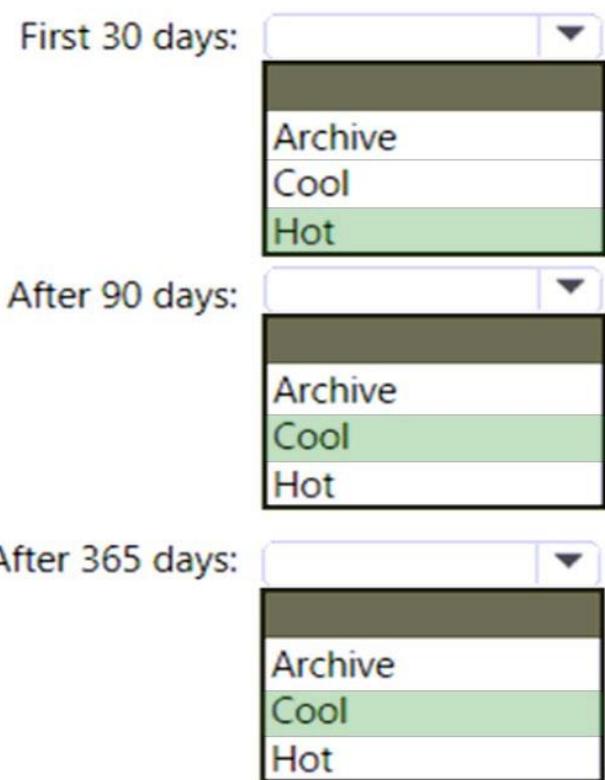
After 90 days:



After 365 days:



Answer:



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.

Incorrect:

Not Archive:

While a blob is in the Archive access tier, it's considered to be offline and can't be read or modified. In order to read or modify data in an archived blob, you must first rehydrate the blob to an online tier, either the Hot or Cool tier.

Rehydration priority -

When you rehydrate a blob, you can set the priority for the rehydration operation via the optional `x-ms-rehydrate-priority` header on a Set Blob Tier or Copy Blob operation.

Rehydration priority options include:

Standard priority: The rehydration request will be processed in the order it was received and may take up to 15 hours.

High priority: The rehydration request will be prioritized over standard priority requests and may complete in less than one hour for objects under 10 GB in size.

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

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

✉️  **OdogwuSaina**  7 months, 2 weeks ago

Hot, Cool, Cool is correct.
Ref: <https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

   upvoted 8 times

✉️  **kkk5566**  5 days, 20 hours ago

Hot, Cool, Cool is correct.
   upvoted 1 times

✉️  **Sima.al** 7 months, 3 weeks ago

1. Hot - because of the 99.9% availability.
2. Hot - because Cool tier needs several minutes to give back an answer (but 30 sec. is asked for).
3. Cool - because the answer is needed within 5 minutes. Thats what cool tier does.

   upvoted 1 times

✉️  **kkk5566** 5 days, 20 hours ago

<https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview#summary-of-access-tier-options>
   upvoted 1 times

✉️  **shoottheduck** 6 months, 1 week ago

Cool has a response time of Milliseconds. So Hot, Cool, Cool
   upvoted 6 times

✉️  **gabrys1997** 11 months, 2 weeks ago

I think that 'cool' tier is just enough, it provides availability on 99.9%
   upvoted 2 times

✉️  **hanzocuk** 8 months ago

Keep this in mind --> "The data will be accessed several times a day during the first 30 days". Cool tier is more expensive to read from.
hot, cool, cool looks correct.

   upvoted 3 times

✉️  **Marcohem** 10 months, 4 weeks ago

Cool Tier provides 99.9% availability only on RA-GRS. For ZRS, it should be 99%.
<https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview#summary-of-access-tier-options>
   upvoted 4 times

✉️  **Strix** 1 year ago

Correct!

   upvoted 2 times

Question 33

DRAG DROP -

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	Answer Area
Access control lists (ACLs)	To grant permissions at the storage account level: <input type="text"/>
Role-based access control (RBAC) roles	To grant permissions at the object level: <input type="text"/>
Shared access signatures (SAS)	
Shared account keys	

Answer:

Answer Area
To grant permissions at the storage account level: <input type="text"/> Role-based access control (RBAC) roles
To grant permissions at the object level: <input type="text"/> Access control lists (ACLs)

SannPro 7 months, 1 week ago

Correct

upvoted 7 times

aemilka 4 months, 2 weeks ago

Correct.

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).

Azure RBAC scope are storage accounts and containers.

ACL scope are directories and files.

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

upvoted 5 times

kkk5566 5 days, 20 hours ago

correct

upvoted 1 times

Venub28 7 months, 3 weeks ago

Given answer is correct

upvoted 3 times

Question 34

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_securitypredicate(@SalesRep AS sysname)
    RETURNS TABLE
    WITH SCHEMABINDING
    AS
        RETURN SELECT 1 AS fn_securitypredicate_result
    WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';
```

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: D

✉️ **mamahani** Highly Voted 4 months, 1 week ago

Selected Answer: D

here is the same example directly from microsoft docs:
<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#Typical>
its definitely D

upvoted 5 times

✉️ **kkk5566** Most Recent 5 days, 19 hours ago

Selected Answer: D

D is correct , see link
<https://learn.microsoft.com/en-us/training/modules/secure-data-warehouse-azure-synapse-analytics/6-exercise-manage-authorization-through-column-row-level-security>

upvoted 1 times

✉️ **kkk5566** 5 days, 19 hours ago

<https://learn.microsoft.com/en-us/training/modules/secure-data-warehouse-azure-synapse-analytics/6-exercise-manage-authorization-through-column-row-level-security>

upvoted 1 times

✉️ **AHUI** 5 months ago
Ans is C.
The function returns 1 when a row in the SalesRep column is the same as the user executing the query (@SalesRep = USER_NAME()) or if the user executing the query is the Manager user (USER_NAME() = 'Manager').
ref: <https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16>
 upvoted 2 times

✉️ **zekescokies** 4 months, 3 weeks ago
It's D. It clearly states that the user querying the table is SalesUser1. I feel they should have mentioned it being a manager if it's C.
 upvoted 7 times

✉️ **shakes103** 4 months, 3 weeks ago
I have looked it up too. Answer is C
 upvoted 2 times

✉️ **Mal2002** 4 weeks, 1 day ago
If you really looked up then what did you understand from this?

```
EXECUTE AS USER = 'SalesRep1';
SELECT * FROM Sales.Orders;
REVERT;
```

```
EXECUTE AS USER = 'SalesRep2';
SELECT * FROM Sales.Orders;
REVERT;
```

```
EXECUTE AS USER = 'Manager';
SELECT * FROM Sales.Orders;
REVERT;
```

The manager should see all six rows. The Sales1 and Sales2 users should only see their own sales.

<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#Typical>

It's clearly D
 upvoted 1 times

✉️ **aemilka** 4 months, 2 weeks ago
In the "Scenario for users who authenticate to the database" there is the same code snippet and it's clearly stated that after applying security policy adding the function as a filter predicate "the manager should see all rows. The Sales1 and Sales2 users should only see their own sales."

So the answer is D.
 upvoted 6 times

Bing: D

When SalesUser1, who is assigned the db_datareader role for Pool1, queries the Sales table in an Azure Synapse Analytics dedicated SQL pool named Pool1 that has row-level security (RLS) applied, only the rows for which the value in the SalesRep column is SalesUser1 will be returned. This is because the RLS predicate filter defined in the question specifies that only rows where the value of the SalesRep column is equal to the current user's name (`@SalesRep = USER_NAME()`) or where the current user's name is 'Manager' (`USER_NAME() = 'Manager'`) should be returned. Since SalesUser1 is not a Manager, only rows where the value of the SalesRep column is equal to 'SalesUser1' will be returned. So, the correct answer is **D. only the rows for which the value in the SalesRep column is SalesUser1**.

Question 35

HOTSPOT -

You have an Azure Data Lake Storage Gen2 account named account1 that contains the resources shown in the following table.

Name	Type	Description
container1	Container	A container
Directory1	Directory	A directory in container1
File1	File	A file in Directory1

You need to configure access control lists (ACLs) to allow a user named User1 to delete File1. User1 is NOT assigned any role-based access control (RBAC) roles for account1. The solution must use the principle of least privilege.

Which type of ACL should you configure for each resource? To answer select the appropriate options in the answer area.

Answer Area

container1:

	▼
--- permissions -WX permissions --X permissions	

Directory1:

	▼
--- permissions -WX permissions --X permissions	

File1:

	▼
--- permissions -WX permissions --X permissions	

ET User:

--x/ -wx/ ---

- ✉  **BPW**  4 months, 2 weeks ago
Answer is
--x/ -wx/ ---
<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>
   upvoted 25 times
- ✉  **Ahmad_Abukhater**  4 months, 4 weeks ago
last box file1 should be --- (First option)
<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>
   upvoted 7 times
- ✉  **vctrhugo** 2 months, 2 weeks ago
"So long as the previous two conditions are true."
   upvoted 1 times
- ✉  **DataEngineer7331** 4 months, 2 weeks ago
According to this your Link, the Directory should have "-WX" and the File "---"
   upvoted 1 times
- ✉  **kkk5566**  5 days, 19 hours ago
--x/ --x/ ---
   upvoted 1 times
- ✉  **Ram9198** 2 months ago
--x/ -wx/ ---
   upvoted 1 times
- ✉  **vctrhugo** 2 months, 2 weeks ago
The solution must use the principle of least privilege!!!
You shouldn't grant -WX to the entire Directory1. Instead, do -x / --w
   upvoted 2 times
- ✉  **mamahani** 4 months, 1 week ago
you do not need any permissions on a file itself to delete it; you only need permissions on the folder where the file resides;
<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control#common-scenarios-related-to-acl-permissions>
so answer -x / -wx / ---
   upvoted 4 times
- ✉  **andjurovicela** 1 month, 3 weeks ago
I agree with everything except the write&execute permission for directory. According to the "famous" link on ACLs the directory permissions should be only execute for deleting actions.
   upvoted 1 times
- ✉  **DataEngDP** 1 day, 17 hours ago
You need write and execute in order to create child items in a directory. And for deleting you dont need permissions so ---.
   upvoted 1 times

Question 36

You have an Azure subscription that is linked to a tenant in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra. The tenant that contains a security group named Group1. The subscription contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and container2. 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 Table Data Reader for myaccount1
- B. Storage Blob Data Reader for container1
- C. Storage Blob Data Reader for myaccount1
- D. Storage Table Data Reader for container1

Answer: B

  **iamthealpha** Highly Voted  4 months, 1 week ago

The appropriate role to assign to Group1 to grant read access to container1 with the principle of least privilege is option B, Storage Blob Data Reader for container1.

Option A, Storage Table Data Reader for myaccount1, is incorrect because it grants read access to all tables in the storage account, not just container1.

Option C, Storage Blob Data Reader for myaccount1, is incorrect because it grants read access to all containers in the storage account, not just container1.

Option D, Storage Table Data Reader for container1, is incorrect because it grants read access to tables in the specified container only, not blobs in container1.

Therefore, option B, Storage Blob Data Reader for container1, is the most appropriate role to assign Group1 to grant read access to container1 with the principle of least privilege.

   upvoted 8 times

  **xymtyk** Most Recent  4 months, 1 week ago

Selected Answer: B

Correct.

   upvoted 3 times

Question 37

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

Answer: A, 还是老堵点

✉️ lamthealpha **Highly Voted** 4 months, 1 week ago

The appropriate feature to use to prevent a group of users from reading user email addresses from dbo.Users in an Azure Synapse Analytics dedicated SQL pool is option A, column-level security.

Option B, row-level security (RLS), is used to filter rows in a table based on the user executing a query, but it cannot prevent certain columns from being read by a group of users.

Option C, Transparent Data Encryption (TDE), encrypts data at rest and does not prevent a group of users from reading specific columns in a table.

Option D, dynamic data masking, is used to mask sensitive data in query results, but it does not prevent a group of users from reading the actual values in a column.

Therefore, option A, column-level security, is the most appropriate feature to use to prevent a group of users from reading user email addresses from dbo.Users in an Azure Synapse Analytics dedicated SQL pool. Column-level security can be used to deny read access to specific columns in a table based on a user or group's permissions.

👍👎🚩 upvoted 11 times

✉️ shakes103 **Highly Voted** 4 months, 3 weeks ago

Selected Answer: A

A is correct. Column-level security simplifies the design and coding of security in your application, allowing you to restrict column access to protect sensitive data. For example, ensuring that specific users can access only certain columns of a table pertinent to their department.

👍👎🚩 upvoted 5 times

✉️ kkk5566 **Most Recent** 5 days, 19 hours ago

Selected Answer: A

correct

👍👎🚩 upvoted 1 times

✉️ RoyP654 3 months, 2 weeks ago

I guess i missed reading about it, but how do you implement column-level security? If via view, folks still have access to the underlying table. Let me know.

👍👎🚩 upvoted 1 times

✉️ RoyP654 2 months, 3 weeks ago

sorry, pls ignore my comment here

👍👎🚩 upvoted 1 times

✉️ halamgir15 4 months, 3 weeks ago

I think it should be D:

dynamic data masking

👍👎🚩 upvoted 4 times

Bing: D

To prevent a group of users from reading user email addresses from the dbo.Users table in an Azure Synapse Analytics dedicated SQL pool, you should use **dynamic data masking**. Dynamic data masking is a feature that automatically hides sensitive data in the result set of a query, based on a set of masking rules that you define. This allows you to limit exposure of sensitive data to non-privileged users, while still allowing them to perform their tasks. So, the correct answer is **D. dynamic data masking**.

Question 38

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool that hosts a database named DB1.

You need to ensure that DB1 meets the following security requirements:

- When credit card numbers show in applications, only the last four digits must be visible.
- Tax numbers must be visible only to specific users.

What 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

Credit card numbers:

- Column-level security
- Dynamic Data Masking
- Row-level security (RLS)

Tax numbers:

- Column-level security
- Row-level security (RLS)
- Transparent Database Encryption (TDE)

Answer:

Answer Area

Credit card numbers:

- Column-level security
- Dynamic Data Masking
- Row-level security (RLS)

Tax numbers:

- Column-level security
- Row-level security (RLS)
- Transparent Database Encryption (TDE)

- ✉  **haythemsi**  3 months, 4 weeks ago
Correct
   upvoted 5 times
- ✉  **kim32**  3 months, 3 weeks ago
It should be Row Level security not column since limited for some users
   upvoted 5 times
- ✉  **Matt2000** 3 weeks, 5 days ago
yes. that is what row-level security is designed for.
   upvoted 1 times
- ✉  **francocalvo** 3 months, 3 weeks ago
I think the answer is correct. Imagine a team where all have access to the table, but just one person needs access to the tax numbers, you can use column-level to disable access for all the other people except the one that needs it
   upvoted 9 times
- ✉  **kkk5566**  5 days, 19 hours ago
masking ,row Level
   upvoted 1 times
- ✉  **kkk5566** 5 days, 19 hours ago
should be Column-level security
   upvoted 1 times
- ✉  **eternalamit5** 2 weeks, 4 days ago
It should be Column-level security as it ensures those specific users can access only certain columns of a table.
Whereas, RLS can help you to create a group membership or execution context in order to control not just columns in a database table, but actually, the rows.
   upvoted 1 times

Question 39

You have an Azure subscription that contains a storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool. The storage1 account contains a CSV file that requires an account key for access.

You plan to read the contents of the CSV file by using an external table.

You need to create an external data source for the external table.

What should you create first?

- A. a database role
- B. a database scoped credential
- C. a database view
- D. an external file format

Answer: B

✉  **cloud_lady**  4 months ago

Given answer is correct.

Refer this link - <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

   upvoted 10 times

✉  **kkk5566**  5 days, 19 hours ago

Selected Answer: B

correct

   upvoted 1 times

✉  **MSExpert** 3 weeks, 1 day ago

Correct

   upvoted 1 times

Question 40

HOTSPOT -

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 container. 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.

Answer Area

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

Answer:

Answer Area

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

 **kkk5566** 5 days, 19 hours ago

correcr

   upvoted 1 times

 **MSExpert** 3 weeks, 1 day ago

Correct

   upvoted 1 times

Common Question 4

Question 1

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub 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 Data Factory instance using Azure Portal
- B. Azure Data Factory instance using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal
- D. Azure Data Factory instance using Microsoft Visual Studio

Answer: C

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. Your Stream Analytics job would ingest these events from Event Hubs and run real-time analytics queries against the streams.

Create a Stream Analytics job:

In the Azure portal, select + Create a resource from the left navigation menu. Then, select Stream Analytics job from Analytics.

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

Question 2

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

ET User:

B (64%) **C (18%)** Other

<https://www.examtopics.com/discussions/microsoft/view/54909-exam-dp-203-topic-4-question-2-discussion/>

✉ **damaldon** Highly Voted 2 years, 2 months ago

correct!

upvoted 29 times

✉ **Miris** Highly Voted 2 years, 2 months ago

correct

upvoted 14 times

✉ **kkk5566** Most Recent 5 days, 19 hours ago

Selected Answer: D

f your table size is less than the recommended 60 million rows for clustered columnstore indexing, consider using heap or clustered index tables.

upvoted 1 times

✉ **kkk5566** 5 days, 19 hours ago

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

upvoted 1 times

✉ **Ram9198** 3 weeks ago

Selected Answer: C

It asks which index to add answer c

upvoted 1 times

✉ Ram9198 3 weeks ago

Selected Answer: A

A heap is a table without a clustered index. One or more nonclustered indexes can be created on tables stored as a heap. Question says already it's a heap table and asks what to add? So answer is A

👍 ↵ 🎉 upvoted 2 times

✉ matianda1 1 month, 2 weeks ago

why not a NCCI - why not A?

Nonclustered columnstore index on a disk-based heap or B-tree index Use for:

1) An OLTP workload that has some analytics queries. You can drop B-tree indexes created for analytics and replace them with one nonclustered columnstore index.

2) Many traditional OLTP workloads that perform Extract Transform and Load (ETL) operations to move data to a separate data warehouse. You can eliminate ETL and a separate data warehouse by creating a nonclustered columnstore index on some of the OLTP tables. NCCI is an additional index that requires 10% more storage on average.

R: <https://learn.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-design-guidance?view=sql-server-ver16#choose-the-best-columnstore-index-for-your-needs>

Enjoy!

👍 ↵ 🎉 upvoted 1 times

✉ Matt2000 3 weeks, 5 days ago

it is currently a heap, thus clustered columnstore makes most sense.

👍 ↵ 🎉 upvoted 1 times

✉ Ram9198 2 months ago

Selected Answer: B

Only 2 columns returned

👍 ↵ 🎉 upvoted 1 times

✉ auwia 2 months, 1 week ago

Selected Answer: B

B of course, there are a few scenarios where clustered columnstore may not be a good option:

Columnstore tables do not support varchar(max), nvarchar(max), and varbinary(max). Consider heap or clustered index instead.

Columnstore tables may be less efficient for transient data. Consider heap and perhaps even temporary tables.

Small tables with less than 60 million rows. Consider heap tables.

👍 ↵ 🎉 upvoted 1 times

✉ vctrhugo 2 months, 2 weeks ago

Selected Answer: B

B. clustered columnstore index.

Given the large fact table with 50 columns and 5 billion rows, and the fact that most queries aggregate values from approximately 100 million rows and return only two columns, a clustered columnstore index would be the most suitable choice. Clustered columnstore indexes are designed for large-scale data warehousing scenarios and provide excellent compression and query performance for analytical workloads.

A clustered columnstore index stores the data in columnar format, enabling efficient data compression and batch-based query execution. It allows for significant query performance improvements, especially for aggregations and large-scale data retrieval.

👍 ↵ 🎉 upvoted 2 times

✉ mamahani 4 months, 1 week ago

im really baffled by all the answers here; noone is even considering clustered index, which is what microsoft is recommending for this particular user case scenario; <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet#index-your-table> for a table up to 100 mln records and using heavily 1-2 columns and performing queries with lots of joins and aggregations (group by clause) microsoft recommends clustered index; why is this recommendation not applicable here? could someone explain?

👍 ↵ 🎉 upvoted 2 times

✉ mamahani 4 months ago

ignore pls; instead of reading watch out if...i read just if, must have been tired; so clustered index is NOT good when group by operations; its good if you need to retrieve 1 single row or few rows (but aggregate is not just few rows -> its many many rows aggregating to 1 row, which is not the same); by this i believe its indeed clustered columnstore index so the given answer is correct

👍 ↵ 🎉 upvoted 2 times

✉ AHUI 5 months ago

Selected Answer: B

correct

👍 ↵ 🎉 upvoted 3 times

✉ Rakrah 7 months ago

Answer is correct (B) clustered columnstore - This index reordered the physical table data with columnar format which is stored with index and compressed. All the query will fetch from index columnstored data and it is designed specially Data warehouse complex query and aggregated data too.

👍 ↵ 🎉 upvoted 3 times

✉ OldSchool 9 months, 1 week ago

Selected Answer: B

It's B

"Do not use a heap when ranges of data are frequently queried from the table. A clustered index on the range column will avoid sorting the entire heap."
<https://learn.microsoft.com/en-us/sql/relational-databases/indexes/heaps-tables-without-clustered-indexes?toc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Ftoc.json&bc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Fbreadcrumb%2Ftoc.json&view=sql-server-ver15&preserve-view=true#when-not-to-use-a-heap>

👍 ↵ 🎉 upvoted 1 times

✉ stunner85_ 11 months, 2 weeks ago

Selected Answer: C

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

👍 ↵ 🎉 upvoted 2 times

dom271219 11 months, 4 weeks ago

Selected Answer: B

"return only two columns" => don't be confused. It's 2 col and not 2 rows => then Clustered columnstore

upvoted 6 times

Ast999 1 year ago

Selected Answer: C

<https://docs.microsoft.com/en-us/sql/relational-databases/indexes/heaps-tables-without-clustered-indexes?view=sql-server-ver16>

upvoted 1 times

proserv 1 year ago

It's Option C , Non Clustered .

Reason : Heap table is a table without clustered columns store index and we can create non cluster index on heap table. Furthermore , there is not non cluster column store index exist

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

Thanks

upvoted 3 times

ET Answer: B

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.

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

Bing: B

To provide the fastest query times for a large fact table in an Azure Synapse Analytics dedicated SQL pool, you should add a **clustered columnstore index**. A clustered columnstore index is a type of columnstore index that physically reorganizes the data in the table and stores it in a columnar format. This allows for high levels of data compression and improved query performance, especially for large-scale data warehousing and analytics workloads that involve aggregating large amounts of data. So, the correct answer is **B. clustered columnstore**.

Question 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

ET User:



✉️  **Dizzystar**  1 year, 10 months ago

I should say Cluster Event logs:
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>

   upvoted 30 times

✉️  **dragos_dragos62000**  2 years, 2 months ago

Correct
   upvoted 11 times

✉️  **kkk5566**  5 days, 18 hours ago

Legacy global init scripts and cluster-named init scripts are deprecated and cannot be used in new workspaces starting February 21, 2023. On September 1st, 2023, Azure Databricks will disable legacy global init scripts for all workspaces.
   upvoted 1 times

✉️  **kkk5566** 5 days, 18 hours ago

Selected Answer: C
should be C
   upvoted 1 times

✉️  **Ram9198** 2 months ago

Selected Answer: B
Cluster event logs
   upvoted 2 times

✉ vctrhugo 2 months, 1 week ago

Selected Answer: B

Cluster event logs in Azure Databricks provide detailed information about the cluster's lifecycle events, including the installation and initialization of libraries. By reviewing the cluster event logs, you can examine the events related to library installation and determine if any errors or issues occurred during the process.

↑ ↵ ← upvoted 2 times

✉ auwia 2 months, 1 week ago

Selected Answer: C

Cluster event logs do not log init script events for each cluster node; only one node is selected to represent them all.

<https://learn.microsoft.com/en-us/azure/databricks/clusters/init-scripts>

↑ ↵ ← upvoted 1 times

✉ vctrhugo 2 months, 1 week ago

Installation and initialization of libraries is not part of init scripts.

↑ ↵ ← upvoted 2 times

✉ bch9994 3 weeks, 6 days ago

That's incorrect. It is a part of init scripts.

Some examples of tasks performed by init scripts include:

Set system properties and environment variables used by the JVM.

Modify Spark configuration parameters.

Modify the JVM system classpath in special cases.

Install packages and libraries not included in Databricks Runtime. To install Python packages, use the Azure Databricks pip binary located at /databricks/python/bin/pip to ensure that Python packages install into the Azure Databricks Python virtual environment rather than the system Python environment. For example, /databricks/python/bin/pip install <package-name>.

[https://learn.microsoft.com/en-us/azure/databricks/init-scripts/](https://learn.microsoft.com/en-us/azure/databricks/init-scripts)

↑ ↵ ← upvoted 1 times

✉ aemilka 4 months, 2 weeks ago

Selected Answer: C

Additional libraries are installed in global init scripts, so correct answer is C.

Some examples of tasks performed by init scripts include:

- Install packages and libraries not included in Databricks Runtime. To install Python packages, use the Azure Databricks pip binary located at /databricks/python/bin/pip to ensure that Python packages install into the Azure Databricks Python virtual environment rather than the system Python environment. For example, /databricks/python/bin/pip install <package-name>.
- Modify the JVM system classpath in special cases.
- Set system properties and environment variables used by the JVM.
- Modify Spark configuration parameters.

ref: <https://learn.microsoft.com/en-us/azure/databricks/clusters/init-scripts>

↑ ↵ ← upvoted 2 times

✉ vctrhugo 2 months, 1 week ago

There are two primary ways to install a library on a cluster:

- Install a workspace library that has been already been uploaded to the workspace.
- Install a library for use with a specific cluster only.

↑ ↵ ← upvoted 1 times

✉ kornat 5 months ago

Selected Answer: C

correct

↑ ↵ ← upvoted 2 times

✉ esaade 5 months, 3 weeks ago

Selected Answer: B

the best option in this scenario would be to review the cluster event logs to identify the cause of the issue where an additional library is not found in the Azure Databricks cluster.

↑ ↵ ← upvoted 3 times

✉ lafita 6 months, 4 weeks ago

Answer C.

A global init script runs on every cluster created in your workspace. Global init scripts are useful when you want to enforce organization-wide library configurations or security screens. Only admins can create global init scripts. You can create them using either the UI or REST API.

↑ ↵ ← upvoted 2 times

✉ youngbug 7 months, 1 week ago

Selected Answer: C

cluster event logs only record start and finish event, so C is right, init script logs record the details of running.

↑ ↵ ← upvoted 2 times

✉ gerrie1979 9 months, 3 weeks ago

Selected Answer: B

<https://learn.microsoft.com/en-us/azure/databricks/clusters/init-scripts>:

Init script start and finish events are captured in cluster event logs. Details are captured in cluster logs. Global init script create, edit, and delete events are also captured in account-level diagnostic logs.

Cluster event logs capture two init script events: INIT_SCRIPTS_STARTED and INIT_SCRIPTS_FINISHED, indicating which scripts are scheduled for execution and which have completed successfully. INIT_SCRIPTS_FINISHED also captures execution duration.

Global init scripts are indicated in the log event details by the key "global" and cluster-scoped init scripts are indicated by the key "cluster".

↑ ↵ ← upvoted 2 times

- ✉ **dmitriypo** 10 months ago
Selected Answer: C
Agree with the given answer - C
Database customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.
Reference:
<https://www.databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>
 upvoted 2 times
- ✉ **Raghul08** 1 year, 7 months ago
My Answer is B
 upvoted 1 times
- ✉ **edba** 1 year, 8 months ago
I think answer is B - Cluster Event logs. Because there are 3 ways to install a new library (<https://docs.microsoft.com/en-us/azure/databricks/libraries/#--install-a-library-on-a-cluster>), using init script is just one of them.
 upvoted 6 times
- ✉ **Canary_2021** 1 year, 8 months ago
B 'Cluster event logs' is the correct answer.
 upvoted 5 times
- ✉ **rashjan** 1 year, 9 months ago
Selected Answer: B
I would go with Cluster Event Logs.
 upvoted 6 times
- ✉ **Sudheer_K** 1 year, 11 months ago
Shouldn't it be cluster-scoped Init scripts rather than global init scripts.
 upvoted 3 times
- ✉ **auwia** 2 months, 1 week ago
Cluster-scoped: run on every cluster configured with the script. This is the recommended way to run an init script.
Global: run on every cluster in the workspace. They can help you to enforce consistent cluster configurations across your workspace. Use them carefully because they can cause unanticipated impacts, like library conflicts.
 upvoted 1 times
- ✉ **gf2tw** 1 year, 9 months ago
Yep, the answer even specifies cluster-scoped init scripts so it seems somehow the question doesn't match up.
 upvoted 1 times

ET Answer: C

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.

<https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>

Bing: B

To identify the cause of the issue where an additional library specified during the creation of an Azure Databricks cluster is not found when attempting to load it to a notebook, you should review the **cluster event logs**. Cluster event logs provide information about events that occur on a cluster, including information about library installation. By reviewing the cluster event logs, you can determine whether there were any issues or errors during the installation of the library, and take appropriate action to resolve them. So, the correct answer is **B. cluster event logs**.

Question 4

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, 看评论像 B

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

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

✉ erssiws **Highly Voted** 2 years, 2 months ago

Activity logs show only activities, e.g., trigger the pipeline, stop the pipeline, ...

Resource health check shows only the healthiness of the resource.

The monitor app indeed contains the pipeline run failure information. But it keep the data only for 45 days.

👍 ↗️ 📈 upvoted 29 times

✉ snna4 1 year, 8 months ago

"Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time."

👍 ↗️ 📈 upvoted 7 times

✉ damaldon **Highly Voted** 2 years, 2 months ago

Correct!

👍 ↗️ 📈 upvoted 7 times

Bing: B

To examine pipeline failures from the last 60 days in an Azure data factory, you should use the **Monitor & Manage app in Data Factory**. The Monitor & Manage app provides a centralized view of all your data factory activity, including pipeline runs, activity runs, and trigger runs. You can use the app to view the status of your pipelines, including any failures that have occurred, and troubleshoot issues by viewing detailed error messages and logs. So, the correct answer is **B. the Monitor & Manage app in Data Factory**.

Question 5

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

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.

<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>

✉️ MinionVII **Highly Voted** 2 years, 2 months ago

Correct.

"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."

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

👍👎FLAG upvoted 18 times

Question 6

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. Pin the cluster.
- B. Create an Azure runbook that starts the cluster every 90 days.
- C. Terminate the cluster manually when processing completes.
- D. Clone the cluster after it is terminated.

Answer: A

Azure Databricks retains cluster configuration information for up to 70 all-purpose clusters terminated in the last 30 days and up to 30 job clusters recently terminated by the job scheduler. To keep an all-purpose cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

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

👤 markpumc 5 months, 4 weeks ago

To ensure that the cluster configuration is retained indefinitely after the cluster is terminated while minimizing costs, you should pin the cluster.

Pinning a cluster in Azure Databricks prevents it from being terminated by the auto-termination feature. This means that the cluster configuration and installed libraries will be retained even if the cluster is not being used. This is the most efficient and cost-effective way to ensure that the cluster configuration is retained indefinitely after the cluster is terminated.

Creating an Azure runbook to start the cluster every 90 days would require additional resources and would not be a cost-effective solution. Terminating the cluster manually when processing completes would not retain the cluster configuration. Cloning the cluster after it is terminated would create a new cluster with the same configuration, but this would also result in additional costs. Should be A

👍👎💡 upvoted 5 times

Question 7

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

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.

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

Question 8

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_status.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Answer: D

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

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

✉ **wuespe** Highly Voted 1 year, 11 months ago

The right answer is D, I tested it in Synapse and it's the only one that actually runs without an error

👍 ↗️ 🚩 upvoted 28 times

✉ **wijaz789** Highly Voted 1 year, 12 months ago

-- Find data skew for a distributed table
DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales');

👍 ↗️ 🚩 upvoted 17 times

✉ **IHYMeRsh** 1 year, 8 months ago

This will only work if you connect to the dedicated pool. The answer you've chosen says you are connecting to the built-in (serverless) pool.
👍 ↗️ 🚩 upvoted 10 times

Bing: D

To identify the extent of data skew in Table1, a fact table in a database named DB1 in an Azure Synapse Analytics dedicated SQL pool named Pool1, you should connect to Pool1 and query the

`sys.dm_pdw_nodes_db_partition_stats` dynamic management view in Synapse Studio. This view returns information about the distribution of data across the compute nodes for each table and index in a database. By querying this view, you can determine the distribution of rows for Table1 across the compute nodes and identify any data skew that may be present. So, the correct answer is D. **Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.**

Question 9

HOTSPOT -

You need to collect application metrics, streaming query events, and application log messages for an Azure Databricks 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.

Hot Area:

Answer Area

Library:

Azure Databricks Monitoring Library
Microsoft Azure Management Monitoring Library
PyTorch
TensorFlow

Workspace:

Azure Databricks
Azure Log Analytics
Azure Machine Learning

Answer:

Answer Area

Library:

Azure Databricks Monitoring Library
Microsoft Azure Management Monitoring Library
PyTorch
TensorFlow

Workspace:

Azure Databricks
Azure Log Analytics
Azure Machine Learning

Question 10

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_request_steps
- B. sys.dm_pdw_nodes_tran_database_transactions
- C. sys.dm_pdw_waits
- D. sys.dm_pdw_exec_sessions

Answer: B

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]
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback
```

Question 11

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

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.

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

Question 12

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

Answer: D

D (77%)

C (23%)

✉ Lio95 Highly Voted 1 year, 11 months ago

It is recommended to have at least 1 million rows per partition and distribution. Since there are 60 distributions, the number of rows for each partition must exceed 60 millions. Answer is correct

1 like 1 reply 1 upvoted 23 times

✉ yassine70 1 year, 11 months ago

I fully Agree! Answer is correct
Link below :<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

"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."

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."

1 like 1 reply 1 upvoted 7 times

✉ LiamRT 1 year, 9 months ago

Partitioning by EventDate does not mean a partition for each day. Partitioning by quarter years would be effective.

- ✉ **Canary_2021** 1 year, 8 months ago
Selected Answer: D
D is the correct answer.
Analysts will most commonly analyze transactions for a warehouse. This means that warehouseID is always in the WHERE clause. Partition field should be in WHERE clause to improve query performance.
Upvoted 15 times
- ✉ **Canary_2021** 1 year, 8 months ago
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>
Upvoted 1 times
- ✉ **Matt2000** 3 weeks, 1 day ago
However the cheat sheet says: "In 99 percent of cases, the partition key should be based on date"
Upvoted 1 times
- ✉ **kkk5566** 5 days, 21 hours ago
Selected Answer: D
is correct
Upvoted 1 times
- ✉ **Karl_Cen** 7 months, 1 week ago
Selected Answer: C
I don't think the answer is right, the answer should be C, EventDate .
The total row number in this inventory updates table is determined before it's created. And here the question is asking us to choose the partition column, not distribution column.
Upvoted 2 times
- ✉ **dmitriypo** 10 months ago
Selected Answer: C
I would go for a DATE column since positions are most often created for a DATE column
Upvoted 2 times
- ✉ **dmitriypo** 10 months ago
Forget it. I agree with the provided answer D
Upvoted 1 times
- ✉ **dom271219** 1 year ago
Selected Answer: D
Tables ? These are the columns, aren't they ?
Upvoted 1 times
- ✉ **Deeksha1234** 1 year ago
D is right
Upvoted 1 times
- ✉ **nefarious_smalls** 1 year, 3 months ago
Selected Answer: C
I will go C. We are querying about warehouses. Therefore I think the distribution column would have to be warehouse. If not then we would most likely have to do a shuffle to aggregate all the transactions for the same warehouse which would be spread out amongst the 60 distributions.
Upvoted 1 times
- ✉ **Aditya0891** 1 year, 2 months ago
It's about partition not distribution. Read the question carefully first
Upvoted 2 times
- ✉ **Dizzystar** 1 year, 10 months ago
I agree on date column. "In most cases, table partitions are created on a date column." <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>
Upvoted 1 times
- ✉ **Dizzystar** 1 year, 10 months ago
I agree on date column. "In most cases, table partitions are created on a date column." <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>
Upvoted 1 times
- ✉ **ploer** 1 year, 7 months ago
But only in most cases. In most cases old data is not needed so date column often shows up in the WHERE clause. This is why partitioning often makes sense on date columns. In this case the "Analysts will most commonly analyze transactions for a warehouse", so WarehouseID will be in the WHERE clause and therefore we should partition on this column.
Upvoted 2 times
- ✉ **sreejani** 1 year, 10 months ago
Aren't partition supposed to be done on columns of group by? So here it's product type on which analysts summarise so partition should be on product type
Upvoted 2 times
- ✉ **Samanda** 1 year, 10 months ago
are you thinking of hash distributions instead of partitions?
Upvoted 5 times
- ✉ **rikku33** 1 year, 11 months ago
For effective partitions it's good to have one million rows per partitions for an ideal optimized scenario. This is also mentioned in the Microsoft documentation. C
Upvoted 2 times
- ✉ **Samanda** 1 year, 10 months ago
You don't have to put each warehouse into its own partition though so the sizing argument doesn't make sense....Answer is D as you will benefit from partition elimination when you use the warehouseID in the WHERE clause
Upvoted 2 times

- ✉  **sachabess79** 1 year, 11 months ago
WHERE is applied on the WarehouseID, so D
   upvoted 6 times
- ✉  **YipingRuan** 1 year, 10 months ago
Nope, don't use WHERE
   upvoted 2 times
- ✉  **mbl** 1 year, 10 months ago
it does : "Analysts will most commonly analyze transactions for a warehouse"
   upvoted 3 times
- ✉  **AppleVan** 1 year, 11 months ago
I think it faster to go by date (C)....Otherwise, the query time will be extremely long since it has wrangled here and there...
   upvoted 2 times
- ✉  **Amalbenrebai** 1 year, 11 months ago
can someone confirm this ?
   upvoted 1 times
- ✉  **Samanda** 1 year, 10 months ago
It's 100% D
   upvoted 3 times
- ✉  **rav009** 1 year, 11 months ago
I will go C
   upvoted 3 times

Question 13

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 YYYYMMDD.
- D. In the fact table, use integer columns for the date fields.
- E. Use DateTime columns for the date fields.

ET User:

CD (90%)	5%
<p>✉ echerish <small>Highly Voted</small> 2 years ago</p> <p>Should be C and D</p> <p>1 upvoted 60 times</p>	
<p>✉ anto69 1 year, 7 months ago</p> <p>Yup, that makes sense</p> <p>1 upvoted 4 times</p>	
<p>✉ GervasioMontaNelas <small>Highly Voted</small> 2 years ago</p> <p>100% CD</p> <p>1 upvoted 14 times</p>	

ET Answer: BD

Bing: CD

To ensure that the design of a star schema for a dataset that contains records of online orders provides the fastest query times when querying for arbitrary date ranges and aggregating by fiscal calendar attributes, you should perform the following two actions:

1. Create a date dimension table that has an integer key in the format of YYYYMMDD. This will allow you to efficiently join the fact table with the date dimension table and filter records based on arbitrary date ranges.
2. In the fact table, use integer columns for the date fields. This will allow you to store the dates as integers in the fact table, which can improve query performance when joining with the date dimension table.

So, the correct answers are **C. Create a date dimension table that has an integer key in the format of YYYYMMDD** and **D. In the fact table, use integer columns for the date fields**.

Question 14

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub 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 Analysis Services using Azure Portal
- B. Azure Analysis Services using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal
- D. Azure Data Factory instance using Microsoft Visual Studio

Answer: C

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. Your Stream Analytics job would ingest these events from Event Hubs and run real-time analytics queries against the streams.

Create a Stream Analytics job:

In the Azure portal, select + Create a resource from the left navigation menu. Then, select Stream Analytics job from Analytics.

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

Question 15

You have a SQL pool in Azure Synapse.

A user reports that queries against the pool take longer than expected to complete. You determine that the issue relates to queried columnstore segments.

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. Snapshot Storage Size
- B. Cache used percentage
- C. DWU Limit
- D. Cache hit percentage

Answer: BD

D: Cache hit percentage: $(\text{cache hits} / \text{cache miss}) * 100$ where 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

B: $(\text{cache used} / \text{cache capacity}) * 100$ where 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

Incorrect Answers:

C: DWU limit: Service level objective of the data warehouse.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

Question 16

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. DWU percentage
- B. Cache hit percentage
- C. DWU limit
- D. Data IO percentage

Answer: B

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

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

Question 17

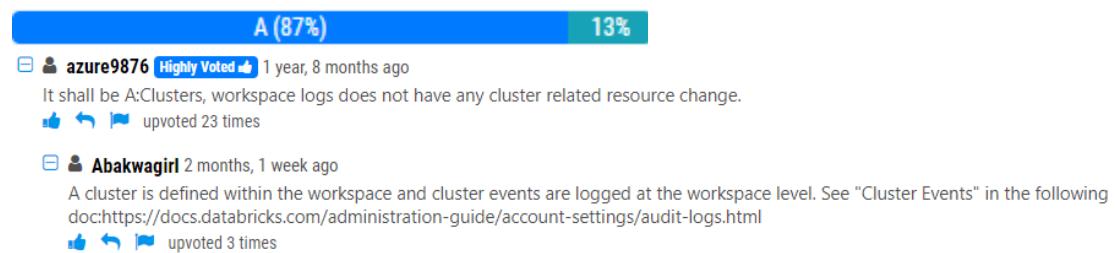
You have an Azure Databricks resource.

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

ET User:



ET Answer: B

Bing: A

To log actions that relate to changes in compute for an Azure Databricks resource, you should log the **clusters** service. The clusters service in Azure Databricks is responsible for managing the creation, configuration, and termination of clusters. By logging the clusters service, you can track changes to the compute resources used by your Databricks resource, including the creation and deletion of clusters, changes to cluster configurations, and scaling events. So, the correct answer is **A. clusters**.

Question 18

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. 5xx: Server Error errors
- B. Average Success E2E Latency
- C. availability
- D. Last Sync Time

Answer: D

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.

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get>

Question 19

You configure monitoring for an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Storage Gen2 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 file.'
- B. 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.
- C. 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 file.'
- D. 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.'

Answer: B

Error message: Cannot execute the query "Remote Query"

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.
<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-errors-and-possible-solutions>

Question 20

You have an Azure Synapse Analytics dedicated SQL pool.

You run PDW_SHOWSPACEUSED('dbo.FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_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	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	40	1992	1	9
3008	3016	960	32	2024	1	10
...
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1437	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
384	2632	560	32	2040	1	59
225	2768	544	40	2184	1	60

Which statement accurately describes the dbo.FactInternetSales table?

- A. All distributions contain data.
- B. The table contains less than 10,000 rows.
- C. The table uses round-robin distribution.
- D. The table is skewed.

Answer: D

Data skew means the data is not distributed evenly across the distributions.

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

Question 21

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

Hash-distribution improves query performance on large fact tables.

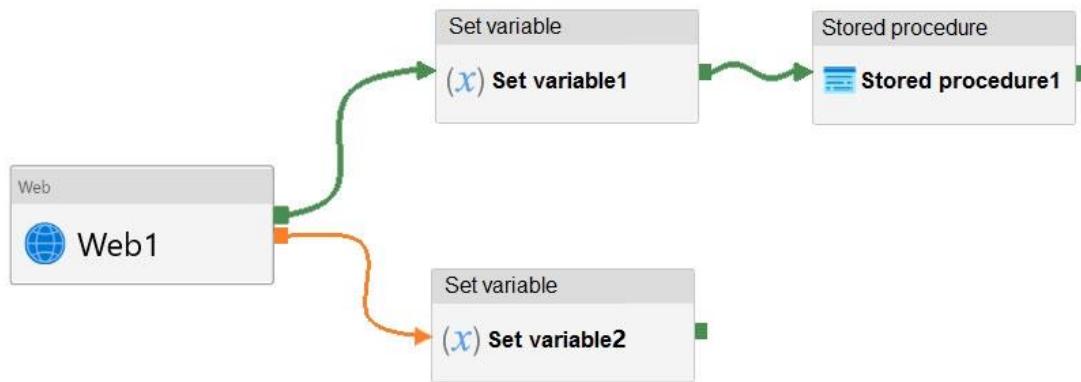
Incorrect Answers:

- A: Do not use a date column for hash distribution. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

Question 22

HOTSPOT -

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.

Hot Area:

Answer Area

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: Box 2 大争议

Answer Area

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

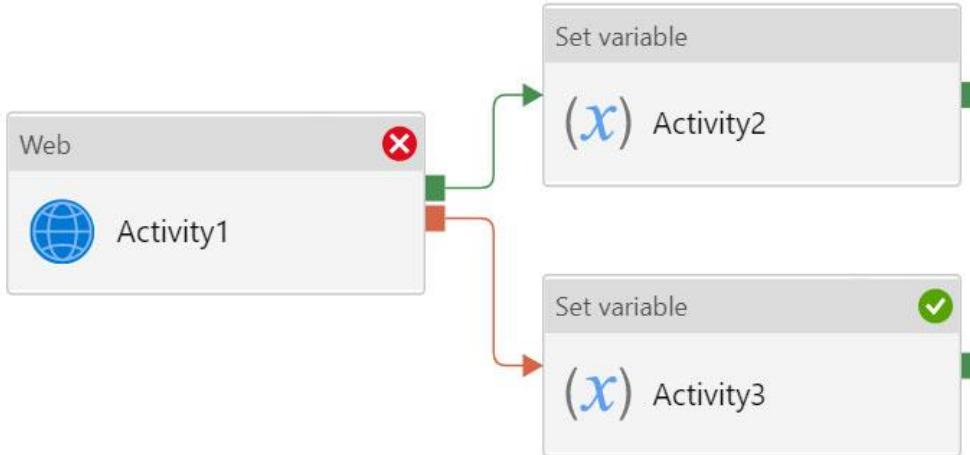
Box 1: succeed -

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.

<https://datasawy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/>

□ **IHYMeRish** Highly Voted 1 year, 8 months ago

The answers are correct.

The second question is "failed" because web1 has both a success and failed path. web1 would have to have only a failed path for the second question to be considered successful.

1 upvoted 38 times

□ **XiltroX** 9 months ago

The second answer should be "Succeeded". You are providing false information to other members. The reason why it is a success is because Set Variable 2 happened because of the failure of Web 1. Therefore, this red pipeline is deemed a success.

1 upvoted 13 times

□ **Avi_Bdj** 1 year, 5 months ago

Second should also be succeeded.
1 upvoted 7 times

□ **a03** 1 year, 8 months ago

Agree. Second is "Fail" because Success connector presented.
1 upvoted 4 times

□ **HaBroNounen** 1 year, 8 months ago

I just tested it myself. Provided answers are correct.
1 upvoted 8 times

□ **RajBathani** Highly Voted 1 year, 8 months ago

The second answer should be Succeeded as 'Set Variable 2' has failed dependency on Web1.
1 upvoted 35 times

□ **kkk5566** Most Recent 5 days, 21 hours ago

1. Success and 2. Failed
1 upvoted 1 times

□ **chryckie** 4 months, 1 week ago

The answer is correct! It's actually pretty neat how ADF determines that.

If an activity fails but there was a subsequent OnSuccess activity that never runs, it's a fail. To handle that, you also need an OnSkipped activity to follow the OnSuccess activity in case it never ran!

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#do-if-else-block>

1 upvoted 5 times

- ✉ AHUI 5 months ago
second box should be succeeded
<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#do-if-skip-else-block>
upvoted 4 times
- ✉ vrodriguesp 7 months, 3 weeks ago
Using this microsot doc: <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#try-catch-block>
that claims:
""We determine pipeline success and failures as follows:
-)Evaluate outcome for all leaves activities. If a leaf activity was skipped, we evaluate its parent activity instead
-)Pipeline result is success if and only if all nodes evaluated succeed""
I used this logic
When web1 activity fails: node setVariable2 succeeds and setVariable1 is skipped and its parent node web1 failed; overall pipeline fails
upvoted 6 times
- ✉ csd 1 year ago
In any scenario pipeline will show success status, cause we are catching the failure
upvoted 2 times
- ✉ StudentFromAus 1 year, 2 months ago
The answers are correct.
upvoted 2 times
- ✉ datnguye 1 year, 8 months ago
It should be Succeeded in both.
The reference article says: The failure dependency means this pipeline reports success.
upvoted 14 times
- ✉ datnguye 1 year, 8 months ago
Updated: Correct ans as 1. Success and 2. Failed
The failure dependency means this pipeline reports success.
But, the presence of the success path alongside the failure path changes the outcome reported by the pipeline: Web-1 fails, Set-var-1 is skipped, and Set-var-2 succeeds --> The pipeline reports failure.
upvoted 13 times
- ✉ Remedios79 1 year, 2 months ago
I agree with you too
upvoted 1 times
- ✉ ladywhiteadder 1 year, 5 months ago
See <https://docs.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#do-if-else-block>
upvoted 6 times
- ✉ ROLLINGROCKS 1 year, 1 month ago
This is all you need for the right answer. Its well explained in the link.
upvoted 1 times
- ✉ Yohannesmulu 1 year, 5 months ago
Agreed!
upvoted 1 times

Bing for Box 2:

If the Web1 activity fails and the Set variable2 activity succeeds, the pipeline will be in a **Failed** status. This is because the failure of any activity within a pipeline will cause the entire pipeline to fail, even if other activities within the pipeline have succeeded. So, the correct answer is **Failed**.

Question 23

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

ET User: DE

✉️ ⚒ KrishIC **Highly Voted** 1 year, 8 months ago

Selected Answer: DE

Notebook- azure databricks, managing activities in pipeline-datafactory
👍 ↗️ 🗃️ upvoted 35 times

✉️ ⚒ ElHomo2222 **Highly Voted** 1 year, 7 months ago

Selected Answer: DE

D & E; Databricks for Wrangling and Notebooks; ADF for Copy and Jar
👍 ↗️ 🗃️ upvoted 15 times

✉️ ⚒ kilowd 1 year, 6 months ago

Wrangling and Copy = ADF
Jar and Notebooks = Databricks
👍 ↗️ 🗃️ upvoted 8 times

ET Answer: AC

Bing: DE

To debug the activities in several Azure Data Factory pipelines that contain a mix of Wrangling data flow, Notebook, Copy, and Jar activities, you should use **Azure Data Factory** and **Azure Databricks**. Azure Data Factory provides a rich set of tools for monitoring and debugging pipeline activities, including the ability to view activity run details, input and output data, and error messages. Azure Databricks is an Apache Spark-based analytics platform that can be used to execute Notebook and Jar activities. By using the Databricks workspace, you can interactively debug these activities by running code snippets, viewing results, and inspecting logs. So, the correct answers are **D. Azure Data Factory** and **E. Azure Databricks**.

Question 24

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 sys.dm_pdw_nodes_db_partition_stats.
- 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.

ET User:



Lotusss 1 year, 4 months ago

Correct. See Question 12 topic 4

upvoted 10 times

kkk5566 5 days, 21 hours ago

Question 8 topic 4

upvoted 1 times

ET Answer: D

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

<https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-partition-stats-transact-sql>

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

Bing: D

To identify the extent of data skew in Table1, a fact table in a database named DB1 in an Azure Synapse Analytics dedicated SQL pool named Pool1, you should connect to Pool1 and query the

`sys.dm_pdw_nodes_db_partition_stats` dynamic management view in Synapse Studio. This view returns information about the distribution of data across the compute nodes for each table and index in a database. By querying this view, you can determine the distribution of rows for Table1 across the compute nodes and identify any data skew that may be present. So, the correct answer is D. Connect to Pool1 and query `sys.dm_pdw_nodes_db_partition_stats`.

Question 25

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. Local tempdb percentage
- B. Cache used percentage
- C. Data IO percentage
- D. CPU percentage

Answer: B

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

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

Question 26

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. Pipeline runs in the Azure Data Factory user experience
- C. the Resource health blade for the Data Factory resource
- D. Azure Data Factory activity runs in Azure Monitor

Answer: D

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

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

Question 27

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub 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 Analysis Services using Azure PowerShell
- B. Azure Stream Analytics Edge application using Microsoft Visual Studio
- C. Azure Analysis Services using Microsoft Visual Studio
- D. Azure Data Factory instance using Azure Portal

Answer: B

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.

You can use Stream Analytics tools for Visual Studio to author, debug, and create your Stream Analytics Edge jobs. After you create and test the job, you can go to the Azure portal to deploy it to your devices.

Incorrect:

Not A, not C: Azure Analysis Services is a fully managed platform as a service (PaaS) that provides enterprise-grade data models in the cloud. Use advanced mashup and modeling features to combine data from multiple data sources, define metrics, and secure your data in a single, trusted tabular semantic data model.

<https://docs.microsoft.com/en-us/azure/iot-hub/monitor-iot-hub>

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-tools-for-visual-studio-edge-jobs>

Question 28

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

Answer: C/D 争议

Use sys.pdw_nodes_column_store_row_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt.

Note: sys.pdw_nodes_column_store_row_groups provides clustered columnstore index information on a per-segment basis to help the administrator make system management decisions in Azure Synapse Analytics. sys.pdw_nodes_column_store_row_groups has a column for the total number of rows physically stored

(including those marked as deleted) and a column for the number of rows marked as deleted.

Incorrect:

Not A: You can join sys.pdw_nodes_column_store_segments with other system tables to determine the number of columnstore segments per logical table.

Not B: Use sys.dm_db_column_store_row_group_operational_stats to track the length of time a user query must wait to read or write to a compressed rowgroup or partition of a columnstore index, and identify rowgroups that are encountering significant I/O activity or hot spots.

C (86%)

14%

✉️ **kkk5566** 5 days, 21 hours ago

Selected Answer: D

D is correct

upvoted 1 times

✉️ **kkk5566** 5 days, 21 hours ago

change to C

<https://learn.microsoft.com/en-us/sql/relational-databases/system-catalog-views/sys-pdw-nodes-column-store-row-groups-transact-sql?view=aps-pdw-2016-au7>

upvoted 1 times

- ✉ andie123 1 month, 1 week ago
Selected Answer: D
D is correct answer
Upvoted 1 times
- ✉ andie123 1 month, 1 week ago
The sys.dm_db_column_store_row_group_physical_stats dynamic management view provides information about the physical characteristics of row groups in columnstore indexes, including the number of deleted rows in each row group. You can use this view to identify tables that have a high percentage of deleted rows by calculating the ratio of deleted rows to total rows for each table. -> D is the answer
Upvoted 1 times
- ✉ vctrhugo 2 months, 2 weeks ago
Selected Answer: C
Use sys.pdw_nodes_column_store_row_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt.
<https://learn.microsoft.com/en-us/sql/relational-databases/system-catalog-views/sys-pdw-nodes-column-store-row-groups-transact-sql?view=aps-pdw-2016-au7>
Upvoted 2 times
- ✉ dimbrici 9 months, 3 weeks ago
Selected Answer: C
C is the correct Answer!
Upvoted 3 times
- ✉ greenlever 10 months, 3 weeks ago
Selected Answer: C
has a column for the total number of rows physically stored (including those marked as deleted) and a column for the number of rows marked as deleted. Use sys.pdw_nodes_column_store_row_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt
Upvoted 4 times
- ✉ anks84 12 months ago
Selected Answer: C
C is the correct Answer!
Upvoted 3 times

Bing: D

To identify tables that have a high percentage of deleted rows in an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1, you should run the `sys.dm_db_column_store_row_group_physical_stats` dynamic management view. This view returns information about the physical characteristics of row groups in columnstore indexes, including the number of deleted rows. By querying this view, you can determine the percentage of deleted rows for each table and identify tables that have a high percentage of deleted rows. So, the correct answer is D.

sys.dm_db_column_store_row_group_physical_stats.

Question 29

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. DWU used
- B. CPU percentage
- C. DWU percentage
- D. Data IO percentage

ET User:



✉️👤 chryckie Highly Voted 4 months, 1 week ago

Selected Answer: C

It must be DWU percentage. e.g. 95% is bad and 99% is very bad, and you don't need to look at anything else.

If you looked at DWU used, what can you infer without also knowing the DWU limit (or DWU percentage)?

👍👎🚩 upvoted 8 times

✉️👤 markpumc Highly Voted 5 months, 3 weeks ago

C. DWU percentage is the best metric to monitor to identify whether you must scale up to a higher service level to accommodate the current workloads in Azure Synapse Analytics. DWU percentage measures the percentage of Data Warehouse Units (DWUs) in use, which indicates how much processing power is being used. If the DWU percentage consistently exceeds a certain threshold, it may be necessary to scale up to a higher service level to accommodate the workload. DWU used, CPU percentage, and Data IO percentage are also important metrics to monitor, but they do not directly reflect the overall processing power available in the data warehouse.

👍👎🚩 upvoted 6 times

ET Answer: A

Bing: C

To monitor an enterprise data warehouse in Azure Synapse Analytics and identify whether you must scale up to a higher service level to accommodate the current workloads, the best metric to monitor is **DWU percentage**. DWU (Data Warehouse Unit) percentage is a measure of the resources being used by the data warehouse, expressed as a percentage of the maximum resources available at the current service level. Monitoring this metric can help you determine whether your data warehouse is approaching its resource limits and whether you need to scale up to a higher service level to accommodate the current workloads. So, the correct answer is **C. DWU percentage**.

Question 30

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub 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 Analysis Services using Azure PowerShell
- B. Azure Data Factory instance using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal
- D. Azure Data Factory instance using Microsoft Visual Studio

Answer: C

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. Your Stream Analytics job would ingest these events from Event Hubs and run real-time analytics queries against the streams.

Create a Stream Analytics job:

In the Azure portal, select + Create a resource from the left navigation menu. Then, select Stream Analytics job from Analytics.

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

Question 31

HOTSPOT -

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.

Hot Area:

Answer Area

Number of partitions:

1
8
16
32

Partition key:

Fraud indicator
Fraud score
Individual line items
Payment details
Transaction ID

Answer:

Answer Area

Number of partitions:

1
8
16
32

Partition key:

Fraud indicator
Fraud score
Individual line items
Payment details
Transaction ID

Box 1: 16 -

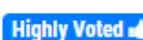
For Event Hubs you need to set the partition key explicitly.

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>

 **Preben**  2 years, 2 months ago

Correct.

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

Embarrassingly parallel jobs

Step 3 and 4.

   upvoted 41 times

Question 32

HOTSPOT -

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

Hot Area:

Answer Area

Table	Distribution type	Distribution column
Sales:	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> Hash-distributed <input type="checkbox"/> Round-robin</div>	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> DateKey <input type="checkbox"/> ProductKey <input type="checkbox"/> RegionKey</div>
Invoices:	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> Hash-distributed <input type="checkbox"/> Round-robin</div>	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><input type="checkbox"/> DateKey <input type="checkbox"/> ProductKey <input type="checkbox"/> RegionKey</div>

Answer:

Answer Area

Table	Distribution type	Distribution column
Sales:	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><div style="background-color: #e0f2e0; border-bottom: 1px solid black; padding: 2px;">Hash-distributed</div><div style="border-bottom: 1px solid black; padding: 2px;">Round-robin</div></div>	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><div style="background-color: #e0f2e0; border-bottom: 1px solid black; padding: 2px;">DateKey</div><div style="background-color: #e0f2e0; border-bottom: 1px solid black; padding: 2px;">ProductKey</div><div style="padding: 2px;">RegionKey</div></div>
Invoices:	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><div style="background-color: #e0f2e0; border-bottom: 1px solid black; padding: 2px;">Hash-distributed</div><div style="border-bottom: 1px solid black; padding: 2px;">Round-robin</div></div>	<div style="border: 1px solid black; padding: 5px; display: inline-block;"><div style="background-color: #e0f2e0; border-bottom: 1px solid black; padding: 2px;">DateKey</div><div style="background-color: #e0f2e0; border-bottom: 1px solid black; padding: 2px;">ProductKey</div><div style="background-color: #e0f2e0; padding: 2px;">RegionKey</div></div>

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: Hash-distributed -

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.

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

 **lara_mia1**  2 years, 3 months ago

1. Hash Distributed, ProductKey because >2GB and ProductKey is extensively used in joins
2. Hash Distributed, RegionKey because "The table size on disk is more than 2 GB." and you have to chose a distribution column which: "Is not used in WHERE clauses. This could narrow the query to not run on all the distributions."

source: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#choosing-a-distribution-column>
   upvoted 98 times

Question 33

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

- elimey **Highly Voted** 2 years, 1 month ago
correct
upvoted 8 times
- SG1705 **Highly Voted** 2 years, 2 months ago
Why ??
upvoted 6 times
- okechi 2 years, 2 months ago
Why ?? Because When you add the "WHERE" clause to your T-SQL query it allows the query optimizer accesses only the relevant partitions to satisfy the filter criteria of the query - which is what partition elimination is all about.
upvoted 40 times
- noranathalie 1 year, 10 months ago
In question 2, we just mentionned to not use the where condition columns to create partitions.. so the logic is unclear for me..
upvoted 2 times
- noranathalie 1 year, 10 months ago
please disregard my comment above. Partitioning is different from hash-column, so the criterias are different
upvoted 4 times

Bing: B

To maximize the benefits of partition elimination when querying a partitioned table in an Azure Synapse Analytics dedicated SQL pool, you should include a **WHERE** clause in your Transact-SQL queries. The WHERE clause allows you to specify filter conditions that can be used to eliminate partitions that do not contain data relevant to the query. By carefully designing your filter conditions to align with the partitioning scheme of the table, you can significantly reduce the amount of data that needs to be scanned by the query, improving its performance. So, the correct answer is B.

WHERE.

Question 34

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

Answer: BD

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

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

👤 allagowf Highly Voted 10 months, 1 week ago

Selected Answer: BD

key word: contains 10,000 distinct values for a column named clusterID --> PARTITION.
reduce the latency --> Increase SU + it refer to PARTITION too.

👍 ↗️ 🎉 upvoted 6 times

👤 vctrhugo Highly Voted 2 months, 2 weeks ago

Selected Answer: BD

To reduce latency in an Azure Stream Analytics job with a query returning a result set containing 10,000 distinct values for a column named clusterID, you should perform the following actions:

B. Increase the number of streaming units:

Increasing the number of streaming units allocates more resources to your Stream Analytics job, allowing it to handle higher data volumes and processing loads. By increasing the streaming units, you can improve the job's throughput and reduce latency.

D. Scale out the query by using PARTITION BY:

Using the PARTITION BY clause in your query allows you to distribute the workload across multiple partitions or parallel processes. By partitioning the data based on relevant criteria, such as clusterID in this case, you can distribute the processing load and reduce latency by enabling parallel processing.

👍 ↗️ 🎉 upvoted 5 times

👤 rzeng Most Recent 10 months, 1 week ago

correct

👍 ↗️ 🎉 upvoted 3 times

Question 35

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 sys.dmv_nodes_db_partition_stats.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dmv_node_status.
- D. Connect to Pool1 and query sys.dmv_nodes_db_partition_stats.

Answer: D. 重复题

Question 36

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 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 Pool1 and DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dm_pdw_sys_info.

ET User: A

👤 Leyya1111 Highly Voted 12 months ago

Selected Answer: A

<https://github.com/rgl/azure-content/blob/master/articles/sql-data-warehouse/sql-data-warehouse-manage-distributed-data-skew.md>

1 upvoted 8 times

👤 anks84 12 months ago

Correct, answer is A !

1 upvoted 4 times

👤 kkk5566 Most Recent 5 days, 2 hours ago

Selected Answer: A

concept repeated , A is correct

1 upvoted 1 times

👤 vctrhugo 2 months, 2 weeks ago

Selected Answer: A

sys.dm_pdw_sys_info actually provides a set of appliance-level counters that reflect overall activity on the appliance. DBCC PDW_SHOWSPACEUSED should be used instead since it displays the number of rows, disk space reserved, and disk space used for a specific table, or for all tables in a Azure Synapse Analytics or Analytics Platform System (PDW) database.

1 upvoted 1 times

ET Answer: D

Bing: A

To identify the extent of data skew in Table1, a fact table in an Azure Synapse Analytics dedicated SQL pool named Pool1, you should connect to Pool1 and run `DBCC PDW_SHOWSPACEUSED`. This command returns information about space usage for each table in the database, including information about the distribution of data across the compute nodes. By running this command and examining the results, you can determine the distribution of rows for Table1 across the compute nodes and identify any data skew that may be present. So, the correct answer is A. **Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.**

Question 37

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 Azure Storage resource provider.
- B. Create a storage policy that is scoped to a container.
- C. Reregister the Microsoft Data Lake Store resource provider.
- D. Create a storage policy that is scoped to a container prefix filter.
- E. Register the query acceleration feature.

ET User:



Exam DP-203 topic 4 question 37 discussion - ExamTopics

✉ **Sima_af** Highly Voted 8 months ago

- E. Register the query acceleration feature.
- D. Create a storage policy that is scoped to a container prefix filter.

To filter data at the time it is read from disk, you need to use the query acceleration feature of Azure Data Lake Storage Gen2. To enable this feature, you need to register the query acceleration feature in your Azure subscription.

In addition, you can use storage policies scoped to a container prefix filter to specify which files and directories in a container should be eligible for query acceleration. This can be used to optimize the performance of the queries by only considering a subset of the data in the container.

◀ ▶ ⌂ upvoted 17 times

✉ **esaade** Highly Voted 5 months, 3 weeks ago

Selected Answer: BE

Option A, reregistering the Azure Storage resource provider, and Option C, reregistering the Microsoft Data Lake Store resource provider, are not necessary to enable filter predicates and column projections in Azure Data Lake Storage Gen2.

Option D, creating a storage policy that is scoped to a container prefix filter, is not a valid option as Azure Data Lake Storage Gen2 does not support storage policies scoped to container prefix filters.

◀ ▶ ⌂ upvoted 5 times

✉ **Nidie** 1 month ago

It has, I think

◀ ▶ ⌂ upvoted 1 times

✉ **kkk5566** Most Recent 5 days, 2 hours ago

Selected Answer: DE

should be correct

◀ ▶ ⌂ upvoted 1 times

✉ **Ast999** 6 months ago

Selected Answer: DE

D + E = correct

◀ ▶ ⌂ upvoted 3 times

✉ **nicky87654** 7 months, 3 weeks ago

Selected Answer: DE

E. Register the query acceleration feature.

D. Create a storage policy that is scoped to a container prefix filter.

◀ ▶ ⌂ upvoted 3 times

ET Answer: AE

Prerequisites -

To access Azure Storage, you'll need an Azure subscription. If you don't already have a subscription, create a free account before you begin.

A general-purpose v2 storage account.

Query acceleration accepts filtering predicates and column projections which enable applications to filter rows and columns at the time that data is read from disk.

Only the data that meets the conditions of a predicate are transferred over the network to the application. This reduces network latency and compute cost.

Note: Query acceleration enables applications and analytics frameworks to dramatically optimize data processing by retrieving only the data that they require to perform a given operation. This reduces the time and processing power that is required to gain critical insights into stored data.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration-how-to>

Bing: BDE, 这个 NB

To ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk when using Azure Data Lake Storage Gen2, you should perform the following two actions:

1. Register the query acceleration feature. This feature enables the use of filter predicates and column projections to push down filtering to the storage layer, reducing the amount of data that needs to be read from disk.
2. Create a storage policy that is scoped to a container or a container prefix filter. This policy defines the conditions under which data will be filtered at the storage layer, based on the filter predicates and column projections specified in the query.

So, the correct answers are **E. Register the query acceleration feature** and either **B. Create a storage policy that is scoped to a container** or **D. Create a storage policy that is scoped to a container prefix filter**.

Question 38

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 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 Pool1 and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to Pool1 and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dm_pdw_sys_info.

ET User: A

Jerrie86 Highly Voted 7 months, 2 weeks ago

This is repeated way too many times.

upvoted 6 times

kkk5566 Most Recent 5 days, 2 hours ago

Selected Answer: A

A is correct

upvoted 1 times

ET Answer: B

Bing: A

To identify the extent of data skew in Table1, a fact table in an Azure Synapse Analytics dedicated SQL pool named Pool1, you should connect to Pool1 and run `DBCC PDW_SHOWSPACEUSED` in Synapse Studio. This command returns information about space usage for each table in the database, including information about the distribution of data across the compute nodes. By running this command and examining the results, you can determine the distribution of rows for Table1 across the compute nodes and identify any data skew that may be present. So, the correct answer is A.
Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.

Question 39

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.

You receive the following error.

Operation on target Copy_sks failed: Failure happened on 'Sink' side.

ErrorCode=DelimitedTextMoreColumnsThanDefined,

'Type=Microsoft.DataTransfer.Common.Snared.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,'

What should you do to resolve the error?

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

ET User:



✉️ **Yemeral** 4 months ago

Selected Answer: A

Correct answer is A. We are just copying files between folders. Selecting binary copy, ADF will not check schema. With D we would discard data. With C we would change file contents

upvoted 9 times

✉️ **chryckie** 4 months, 1 week ago

Selected Answer: A

It's tricky.

Not D, because you don't just throw away data.

Likely not C, because it doesn't solve for future schema variability. (Avro formats are usually chosen in situations where the schema may evolve over time, because they store both the data and schema in the file itself.)

A makes most sense, since you're just trying to move files over. Binary preserves everything as-is, and you can read/interpret them as ASCII/UTF-8/whatever later.

upvoted 7 times

✉️ **chryckie** 4 months, 1 week ago

Oh! Also, the message says it's trying to process the Avro file as a Csv/Tsv Format Text. That's likely the issue.

upvoted 1 times

✉️ **kkk5566** 5 days, 2 hours ago

Selected Answer: A

A is correct

upvoted 1 times

✉️ **Tightbot** 2 weeks, 4 days ago

Selected Answer: C

I would go with Option C- Add an explicit mapping.
Laying out possible derivations from the question
1. the actual error says - column mismatch .
2. Even though the filename is "filename.avro" , it could just be a filename, the source file type is CSV/TSV.
Possible answers
1. Add an explicit mapping
2. Enabling Fault tolerance to skip incompatible rows
I think both would be a possible solution, but to me, skipping incompatible rows is more of a temporary solution and explicit mapping would be more permanent for this error. I'm also excluding future schema issues that arise after this as there is no information about it.

upvoted 1 times

✉️ **Ram9198** 3 weeks ago

Selected Answer: A

It says CSV/tsv source but file is avro so A is the answer

upvoted 1 times

✉️ **andjurovicela** 2 months, 1 week ago

Selected Answer: D

I was pondering a bit about this one, and decided to go with D. Reasoning behind this is because the question was "how to resolve this error?" and 100% preservation of source data hasn't been a condition, hence D is the most straightforward.

upvoted 1 times

✉️ **vctrhugo** 2 months, 2 weeks ago

Selected Answer: C

Binary Copy is a setting that can be used in Azure Data Factory to improve performance when copying binary data, such as Avro or Parquet files. It optimizes the data transfer by copying the data as-is without parsing or transforming it. However, in this case, the error is related to the mismatch in the column structure, which cannot be resolved by changing the copy setting to Binary Copy.

upvoted 1 times

✉️ **azure_user11** 3 months, 2 weeks ago

Selected Answer: A

I think the purpose here is to just copy files as-is from one folder to another. <https://learn.microsoft.com/en-us/azure/data-factory/format-binary>

upvoted 1 times

✉️ **levto** 3 months, 3 weeks ago

Selected Answer: A

agree with Yemeral

upvoted 1 times

✉️ **sk20** 4 months, 2 weeks ago

Correct Answer D . It makes sense to use Fault Tolerance . Refer link below.
<https://learn.microsoft.com/en-us/answers/questions/1178682/found-more-columns-than-expected-column-count-35>

upvoted 2 times

✉️ **shakes103** 4 months, 3 weeks ago

Selected Answer: C

Correct answer is C

upvoted 2 times

✉️ **AscentAcademy** 6 months ago

It appears we're trying to copy an avro file. This should be done as a binary copy, so we should select A. In fact, you I found someone who had this exact issue here: <https://sqlwithmancj.com/2020/07/29/azure-data-factory-adf-pipeline-failure-found-more-columns-than-expected-column-count-delimitedtextmorecolumnsdefined/>

upvoted 5 times

✉️ **shoottheduck** 6 months, 1 week ago

Selected Answer: D

I have checked this in ADF. Also see doc:

<https://learn.microsoft.com/nl-nl/azure/data-factory/copy-activity-fault-tolerance#copying-tabular-data>

upvoted 4 times

✉️ **raydoneean** 6 months, 3 weeks ago

C is correct

upvoted 1 times

✉️ **vrodriguesp** 6 months, 3 weeks ago

Selected Answer: D

mapping is correct because error is only on one row (row number 53) so the only acceptable should be D

upvoted 4 times

✉️ **Jerrie86** 7 months, 2 weeks ago

Selected Answer: D

The answer should be D. The error 'there are more columns in the source file '0_2020_11_09_11_43_32.avro' than expected could be because of one extra column delimiter. And that leads to the error.

Extra column error would have occurred at row 1 if there was actually an extra column.

Answer should be D to skip that row because the data coming from the source is not correct.

upvoted 4 times

✉️ **Lestrang** 7 months, 1 week ago

Who said it is not correct? It just has extra columns for this particular record. Why discard potentially valuable data when you can keep it by defining an explicit mapping?

Sure this seems like a 1 row only but you have no guarantee that this won't happen again.

upvoted 4 times

✉ agold96 7 months, 2 weeks ago

Selected Answer: D

As the error happens on only one row, I guess the mapping is done correctly, there is just a mistake on the row 53. Then, the answer should be D.

Like Reply Flag upvoted 4 times

✉ Lestrang 7 months, 2 weeks ago

Selected Answer: C

C. Add an explicit mapping.

The error message indicates that there are more columns in the source file '0_2020_11_09_11_43_32.avro' than expected. One way to resolve this issue is to add an explicit mapping in the Copy activity settings, which specifies the columns in the source file and their corresponding columns in the destination. This ensures that the correct columns are being copied and can help prevent issues with incompatible column counts.

Like Reply Flag upvoted 4 times

✉ youngbug 7 months, 2 weeks ago

Why not D?

Like Reply Flag upvoted 2 times

✉ vrodriguesp 7 months, 2 weeks ago

Selected Answer: C

I Agree

Like Reply Flag upvoted 2 times

✉ Stefan94 7 months, 3 weeks ago

Correct

Like Reply Flag upvoted 3 times

ET Answer: C

Bing: D

To resolve the error you received while using Azure Data Factory to copy multiple files from Folder1 to Folder2 in an Azure Data Lake Storage Gen2 account, you should **enable fault tolerance to skip incompatible rows**. This error message indicates that there is an issue with the data in one of the files being copied, where a row was found to have more columns than expected. By enabling fault tolerance, you can configure the Copy activity to skip rows that are incompatible with the schema, allowing the copy operation to continue without failing. So, the correct answer is D.

Enable fault tolerance to skip incompatible rows.

Question 40

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

Answer: D

✉  **Mouli10**  7 months, 3 weeks ago

Selected Answer: D

Azure databricks

   upvoted 6 times

✉  **kkk5566**  5 days, 2 hours ago

Selected Answer: D

correct

   upvoted 1 times

✉  **vctrhugo** 2 months, 2 weeks ago

Selected Answer: D

By leveraging Azure Databricks, you can easily perform advanced analytics on the intrusion detection data using Spark's powerful distributed processing capabilities. Databricks provides an interactive and collaborative environment where you can write Spark code, explore and visualize data, and build machine learning models. It also integrates with popular data sources, including Azure Data Lake Storage, for efficient data ingestion and processing.

   upvoted 1 times

✉  **Stefan94** 7 months, 3 weeks ago

Correct

   upvoted 3 times

Question 41

HOTSPOT -

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

ET User:

1. sys.dm_pdw_exec_requests
2. sys.dm_pdw_waits

  **auwia** Highly Voted  2 months, 1 week ago

The sys.dm_pdw_lock_waits view is specific to SQL Server and is used to monitor lock waits and lock resources in regular SQL Server environments, not in Azure Synapse Analytics dedicated SQL pools.

My answers are:

1. sys.dm_pdw_exec_requests
2. sys.dm_pdw_waits

There is a similar question in the microsoft official practice assessment and the explaination is the following:

The sys.dm_pdw_waits view holds information about all wait stats encountered during the execution of a request or query, including locks and waits on a transmission queue

   upvoted 7 times

  **bp_a_user** Highly Voted  4 months, 1 week ago

Its dm_pdw_waits:

Queries in the Suspended state can be queued due to a large number of active running queries. These queries also appear in the sys.dm_pdw_waits waits query with a type of UserConcurrencyResourceType from the official learning path: <https://learn.microsoft.com/en-us/training/modules/manage-monitor-data-warehouse-activities-azure-synapse-analytics/6-use-dynamic-management-views-to-identify-troubleshoot-query-performance>

   upvoted 7 times

ET 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

Question 42

You have an Azure Data Factory pipeline named pipeline1 that includes a Copy activity named Copy1. Copy1 has the following configurations:

- The source of Copy1 is a table in an on-premises Microsoft SQL Server instance that is accessed by using a linked service connected via a self-hosted integration runtime.
- The sink of Copy1 uses a table in an Azure SQL database that is accessed by using a linked service connected via an Azure integration runtime.

You need to maximize the amount of compute resources available to Copy1. The solution must minimize administrative effort.

What should you do?

- A. Scale out the self-hosted integration runtime.
- B. Scale up the data flow runtime of the Azure integration runtime and scale out the self-hosted integration runtime.
- C. Scale up the data flow runtime of the Azure integration runtime.

ET User:



✉️ **azure_user11** Highly Voted 3 months, 2 weeks ago
Why not B?

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

Azure integration runtime provides the native compute to move data between cloud data stores in a secure, reliable, and high-performance manner. You can set how many data integration units to use on the copy activity, and the compute size of the Azure IR is elastically scaled up accordingly without requiring you to explicitly adjust the size of the Azure Integration Runtime.

For high availability and scalability, you can scale out the self-hosted IR by associating the logical instance with multiple on-premises machines in active-active mode.
🕒 9 upvoted 9 times

✉️ **BillMyI** Highly Voted 3 months, 1 week ago
I would answer A.

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

Copying between a cloud data source and a data source in a private network: if either the source or sink linked service points to a self-hosted IR, the copy activity is executed on the self-hosted IR.
🕒 8 upvoted 8 times

- ✉ **kkk5566** Most Recent 5 days, 2 hours ago
Selected Answer: A
<https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#self-hosted-ir-compute-resource-and-scaling>
A should be correct.
👍 ↗️ 🎉 upvoted 1 times
- ✉ **KK5566** 5 days, 1 hour ago
correct it , C, Option A, "Scale out the self-hosted integration runtime," is not the best solution to maximize the amount of compute resources available to Copy1 because it would not minimize administrative effort. Scaling out the self-hosted integration runtime would involve adding more nodes to the runtime pool, which would require allocating new virtual machines and registering new nodes on the integration runtime. This process can be time-consuming and would require additional administrative effort1.
👍 ↗️ 🎉 upvoted 1 times
- ✉ **Ram9198** 3 weeks ago
Selected Answer: A
if either the source or sink linked service points to a self-hosted IR, the copy activity is executed on the self-hosted IR.
👍 ↗️ 🎉 upvoted 1 times
- ✉ **alegiordx** 1 month, 4 weeks ago
Selected Answer: A
My answer is A due to the precedence criteria among Integration runtimes selection when source and sink linked services are linked to different IRs, as described here
<https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#determining-which-ir-to-use>
👍 ↗️ 🎉 upvoted 2 times
- ✉ **tsmk** 2 months ago
Selected Answer: C
The important point is - "The solution must minimize administrative effort."
Azure integration runtime - Cloud-based service - Not as scalable as a SHIR
Self-hosted integration runtime - More scalable - Requires more administrative effort
Scaling up the Azure IR will give us more compute resources without increasing the administrative effort.
👍 ↗️ 🎉 upvoted 1 times
- ✉ **andjurovicela** 2 months ago
Selected Answer: A
According to the MS document, A seems to be the correct answer.
👍 ↗️ 🎉 upvoted 2 times
- ✉ **Azure_2023** 3 months ago
Selected Answer: B
I would answer B
👍 ↗️ 🎉 upvoted 1 times
- ✉ **JG1984** 2 months, 2 weeks ago
scaling out the self hosted will not increase the amount of compute resources becoz its already running on the physical machine (on-premises) . However, the Azure integration runtime is a managed service, so scaling up its data flow runtime will increase the amount of compute resources available to Copy1.
👍 ↗️ 🎉 upvoted 1 times

ET Answer: C

Bing: A

To maximize the amount of compute resources available to Copy1, a Copy activity in an Azure Data Factory pipeline named pipeline1, and minimize administrative effort, you should **scale out the self-hosted integration runtime**. The self-hosted integration runtime is responsible for moving data between the on-premises Microsoft SQL Server instance and the Azure Data Factory service. By scaling out the self-hosted integration runtime, you can increase the number of nodes that are used to execute data movement activities, allowing more data to be moved in parallel and improving the performance of the Copy activity. So, the correct answer is A. **Scale out the self-hosted integration runtime.**

Question 43

You are designing a 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.

NOTE: Each correct selection is worth one point.

- A. the clone command
- B. Z-Ordering
- C. Apache Spark caching
- D. dynamic file pruning (DFP)

Answer: BD

OfficeSaracus **Highly Voted** 3 months, 4 weeks ago

Selected Answer: BD

Seems correct:

<https://learn.microsoft.com/en-us/azure/databricks/optimizations/dynamic-file-pruning>

<https://learn.microsoft.com/en-us/azure/databricks/delta/data-skipping>

upvoted 7 times

kkk5566 **Most Recent** 5 days, 2 hours ago

Selected Answer: BD

correct

upvoted 1 times

vctrhugo 2 months, 1 week ago

Selected Answer: BD

Dynamic file pruning, can significantly improve the performance of many queries on Delta Lake tables. Dynamic file pruning is especially efficient for non-partitioned tables, or for joins on non-partitioned columns. The performance impact of dynamic file pruning is often correlated to the clustering of data so consider using Z-Ordering to maximize the benefit.

upvoted 3 times

Question 44

You have an Azure Data Lake Storage Gen2 account named account1 that contains a container named container1.

You plan to create lifecycle management policy rules for container1.

You need to ensure that you can create rules that will move blobs between access tiers based on when each blob was accessed last.

What should you do first?

- A. Configure object replication
- B. Create an Azure application
- C. Enable access time tracking
- D. Enable the hierarchical namespace

Answer: C

cloud_lady Highly Voted 3 months, 4 weeks ago

Selected Answer: C

Answer is correct.

Customers stores huge amount of data in Azure blob storage. Sometimes this data is accessed frequently and other times infrequently. Last access time tracking integrates with the lifecycle of Azure blob storage to allow automatic tiering and deletion of data based on when individual blobs are accessed last.

upvoted 8 times

kkk5566 Most Recent 5 days, 1 hour ago

Selected Answer: C

correct

upvoted 1 times

Question 45

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. DWU limit
- B. Data IO percentage
- C. Cache hit percentage
- D. CPU percentage

Answer: C

✉  **kkk5566** 5 days, 1 hour ago

Selected Answer: C

Repeated

   upvoted 1 times

✉  **darshilparmar** 2 months, 4 weeks ago

Repeat Questions

   upvoted 4 times

✉  **henryphchan** 3 months, 3 weeks ago

Selected Answer: C

Answer is C, and it's a repeated question

   upvoted 4 times

Question 46

HOTSPOT -

You have an Azure data factory named DF1 that contains 10 pipelines. The pipelines are executed hourly by using a schedule trigger. All activities are executed on an Azure integration runtime.

You need to ensure that you can identify trends in queue times across the pipeline executions and activities. The solution must minimize administrative effort.

How should you configure the Diagnostic settings for DF1? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Collect:

Pipeline activity runs log
Pipeline runs log
Trigger runs log

Send to:

Event hub
Log Analytics workspace
Storage account

Answer:

Answer Area

Collect:

Pipeline activity runs log
Pipeline runs log
Trigger runs log

Send to:

Event hub
Log Analytics workspace
Storage account

✉ **kkk5566** 5 days, 1 hour ago

correct

upvoted 1 times

✉ **vk8880** 3 months, 3 weeks ago

1. To identify trends in queue times, you should focus on the Pipeline activity run logs rather than the Pipeline run logs. Pipeline activity run logs allow you to track the queue times for individual activities within the pipeline. While Pipeline run logs may provide some information about queue times, they do not provide granular details for each activity within the pipeline.

upvoted 4 times

✉ **henryphchan** 3 months, 3 weeks ago

so the provided answer is correct.

upvoted 5 times

Question 47

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. DWU percentage
- B. Cache hit percentage
- C. Data Warehouse Units (DWU) used
- D. Data IO percentage

Answer: B. 重复题

Question 48

HOTSPOT -

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 run Spark notebooks in ws1. The solution must ensure that you can retrieve 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 ws1.
Create a linked service to kv1.

In Synapse Studio:

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.

ET User:

Box1. Add a role-based access control (RBAC) role to kv1

Box2. Create a linked service to kv1

kkk5566 5 days, 1 hour ago

Box1. Add a role-based access control (RBAC) role to kv1
Box2. Create a linked service to kv1

upvoted 1 times

Lukas 1 month ago

In the Azure portal:
Add a role-based access control (RBAC) role to kv1 - You need to assign the 'Key Vault Secrets User' role to UAMI1 on kv1. This will grant the managed identity the necessary permissions to retrieve secrets from Key Vault.

In Synapse Studio:

Create a linked service to kv1 - You need to create a linked service in Azure Synapse Studio to connect to kv1. The linked service will use the User-Assigned Managed Identity (UAMI1) to authenticate to the Azure Key Vault.

upvoted 4 times

ET 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 ws1.
Create a linked service to kv1.

In Synapse Studio:

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.

Question 49

HOTSPOT -

You have an Azure Data Factory pipeline shown in 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 checked="" type="radio"/>

Answer: 就一个评论，不知道对不对

Answer Area

Statements

The `Retry` property of the `Web_GetIP` activity is set to 1.

The `waitForCompletion` property of the `Exec_COPY_BLOB` activity is set to `true`.

The `Exec_COPY_BLOB` activity was skipped during the second run due to pipeline dependencies.

 **Matt2000** 3 weeks, 1 day ago

No, No, No

The `Retry` Property is not set to one for `Web_GetIP`: Otherwise, we would see a retry of that activity in the first run.

`waitForCompletion` property is not set to `true`: In the second run, `Exec_COPY_BLOB` takes as long as in the first one, despite being skipped. So, it could not have been waiting for the pipeline that it had triggered to complete.

`Exec_COPY_BLOB` cannot be skipped due to a pipeline dependency since it is the first activity in the pipeline. Most likely, its activity state was manually set to `'skipped'`.

   upvoted 1 times

Question 50

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 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 sys.dm_pdw_nodes_db_partition_stats.
- B. Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to Pool1 and query sys.dm_pdw_node_status.
- D. Connect to the built-in pool and query sys.dm_pdw_sys_info.

Answer: B, 重复题

Question 51

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

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

Answer: BE

  **Sdevi49** 1 month ago

The Copy activity is native to Azure Data Factory, so you would use Azure Data Factory to debug it.

The Notebook and Jar activities are related to Databricks jobs, so you would use Azure Databricks to debug them.

Power Query is more associated with data wrangling and transformation, and while it can be used in various services, in the context of Azure Data Factory, you'd likely be debugging within Azure Data Factory or Azure Synapse Analytics. However, given the other activities listed, Azure Data Factory is the more probable choice for this scenario.

Therefore, the correct answers are:

- B. Azure Data Factory
- E. Azure Databricks

   upvoted 1 times

Question 52

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub 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 Analysis Services using Microsoft Visual Studio
- B. Azure Data Factory instance using Azure PowerShell
- C. Azure Analysis Services using Azure PowerShell
- D. Azure Stream Analytics cloud job using Azure Portal

Answer: D

 **kkk5566** 5 days, 1 hour ago

Selected Answer: D

repeated

   upvoted 1 times

Question 53

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
49448	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

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.

<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>