Question 1: Skipped

In Azure Synapse Studio, where would you view the contents of the primary data lake store?

- ○ None of the listed options.
- ○ In the workspace tab of the Integrate hub.
- ○ In the Integration section of the Monitor hub.
- ○ In the workspace tab of the Data hub.
- ○ In the linked tab of the Data tab. **(Correct)**

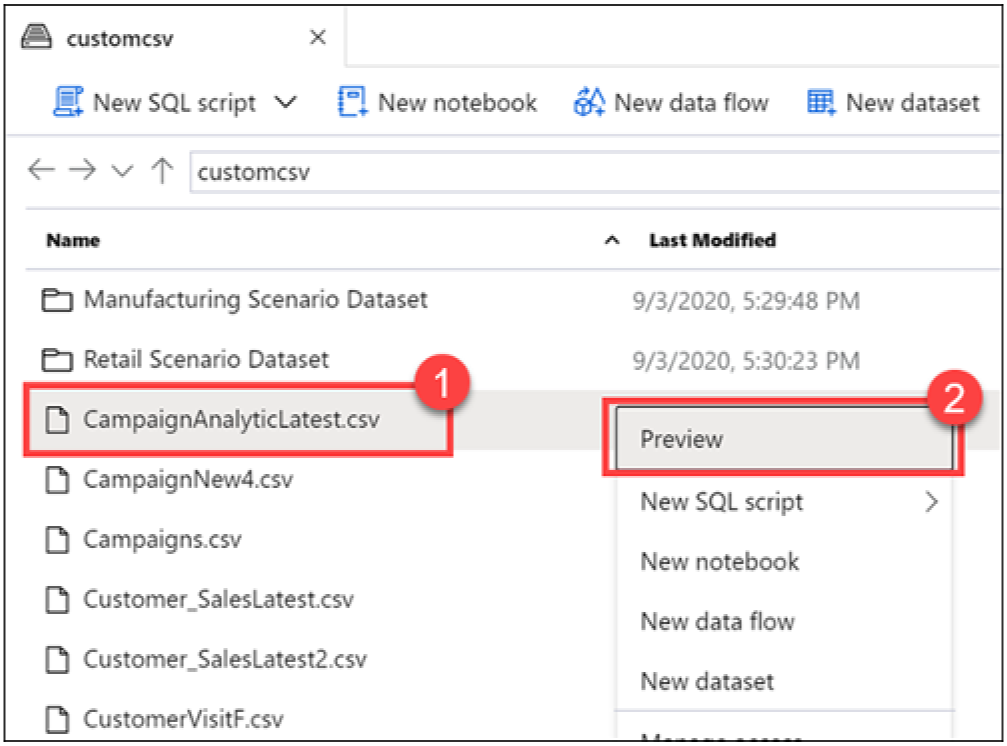
Explanation

The linked tab of the data hub is where you can view the contents of the primary data lake store.

In Azure Synapse Studio, the Data hub is where you access your provisioned SQL pool databases and SQL serverless databases in your workspace, as well as external data sources, such as storage accounts and other linked services.

Every Synapse workspace has a primary ADLS Gen2 account associated with it. This serves as the data lake, which is a great place to store flat files, such as files copied over from on-premises data stores, exported data or data copied directly from external services and applications, telemetry data, etc. Everything is in one place.

The file explorer capabilities allow you to quickly find files and perform actions on them, like preview file contents, generate new SQL scripts or notebooks to access the file, create a new data flow or dataset, and manage the file.



Name	Last Modified
Manufacturing Scenario Dataset	9/3/2020, 5:29:48 PM
Retail Scenario Dataset	9/3/2020, 5:30:23 PM
CampaignAnalyticLatest.csv	
CampaignNew4.csv	
Campaigns.csv	
Customer_SalesLatest.csv	
Customer_SalesLatest2.csv	
CustomerVisitF.csv	

https://azure.microsoft.com/en-us/blog/quickly-get-started-with-samples-in-azure-synapse-analytics/

Question 2: Skipped

Within creating a notebook, you need to specify the pool that needs to be attached to the notebook that is, a SQL or Spark pool. When it comes to the languages, a notebook has to be set with a primary language.

Which of the following are primary languages available within the notebook environment? (Select four)

- □ .NET Spark (C#) **(Correct)**
- □ JSspark (JavaScript)
- □ Spark (Scala) **(Correct)**
- □ Spark SQL

(Correct)

- □ PySpark (Python)
  (Correct)
- □ JVspark (Java)

Explanation
Within creating a notebook, you need to specify the pool that needs to be attached to the notebook that is, a SQL or Spark pool. When it comes to the languages, a notebook has to be set with a primary language.

The primary languages available within the notebook environment are:

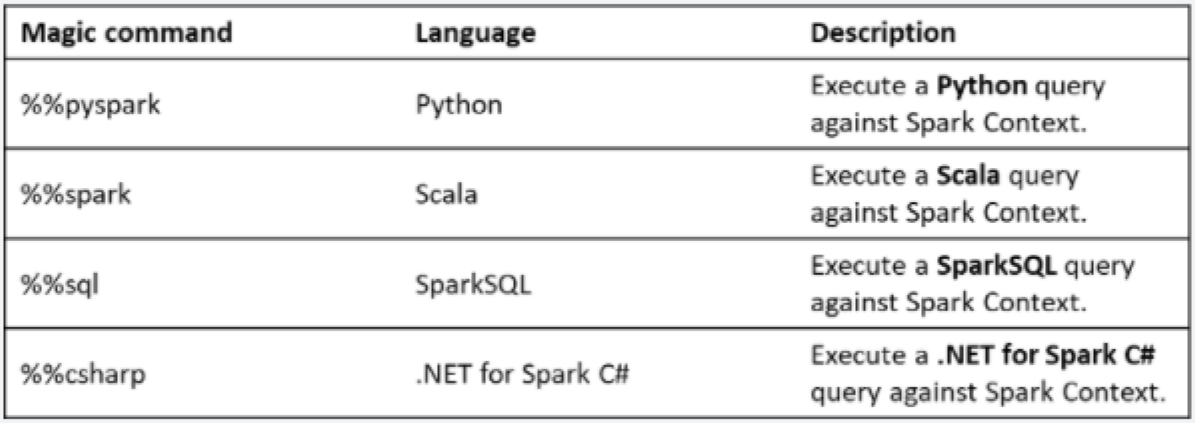
• PySpark (Python)

• Spark (Scala)

• .NET Spark (C#)

• Spark SQL

However, it is possible to use multiple languages in one notebook by specifying the language using a magic command at the beginning of a cell. The following table lists the magic commands to switch cell languages:



Magic command	Language	Description
%%pyspark	Python	Execute a **Python** query against Spark Context.
%%spark	Scala	Execute a **Scala** query against Spark Context.
%%sql	SparkSQL	Execute a **SparkSQL** query against Spark Context.
%%csharp	.NET for Spark C#	Execute a **.NET for Spark C#** query against Spark Context.

It is not possible to reference data or variables directly across different languages in a Synapse Studio notebook. In Spark, it is possible to reference a temporary table across languages.

https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-development-using-notebooks?tabs=classical

Question 3: Skipped

Scenario: You are working as a consultant at **Advanced Idea Mechanics** (**A.I.M.**) who is a privately funded think tank organized of a group of brilliant scientists whose sole dedication is to acquire and develop power through technological means. Their goal is to use this power to overthrow the governments of the world. They supply arms and technology to radicals and subversive organizations in order to foster a violent technological revolution of society while making a profit.

The company has 10,000 employees. Most employees are located in Europe. The company supports teams worldwide.

AIM has two main locations: a main office in London, England, and a manufacturing plant in Berlin, Germany.

At the moment, you are leading a Workgroup meeting with the IT Team where the topic of discussion is the implementation of a process which copies data from an instance on the company's on-prem MS SQL Server to Azure Blob storage.

Required:

• The process must orchestrate and manage the data lifecycle.

• Configuration of Azure Data Factory to connect to the SQL Server instance.

Several ideas have been tabled as action items, which are listed below:

a. Configure a linked service to connect to the SQL Server instance.

b. From the on-prem network, install and configure a self-hosted runtime.

c. From the SQL Server, backup the database and then copy the database to Azure Blob storage.

d. Deploy and Azure Data Factory.

e. From the SQL Server, create a database master key.

The IT Team looks to you as for direction as the Azure SME and you need to advise them on which of the ideas tabled, need to be executed and in which order.

Which of the following calls for the correct action items in the correct order?

- ○ d → e → b → c
- ○ a → c → b → e → d
- ○ e → b → a
- ○ b → c → d → a
- ○ d → b → a (Correct)

Explanation
Step 1: From the on-premises network, install and configure a self-hosted runtime. To use copy data from a SQL Server database that isn't publicly accessible, you need to set up a self-hosted integration runtime.

Step 2: Configure a linked service to connect to the SQL Server instance.

Step 3: Deploy an Azure Data Factory. You need to create a data factory and start the Data Factory UI to create a pipeline in the data factory. With out source and sink we cannot create a Pipeline in Data factory.

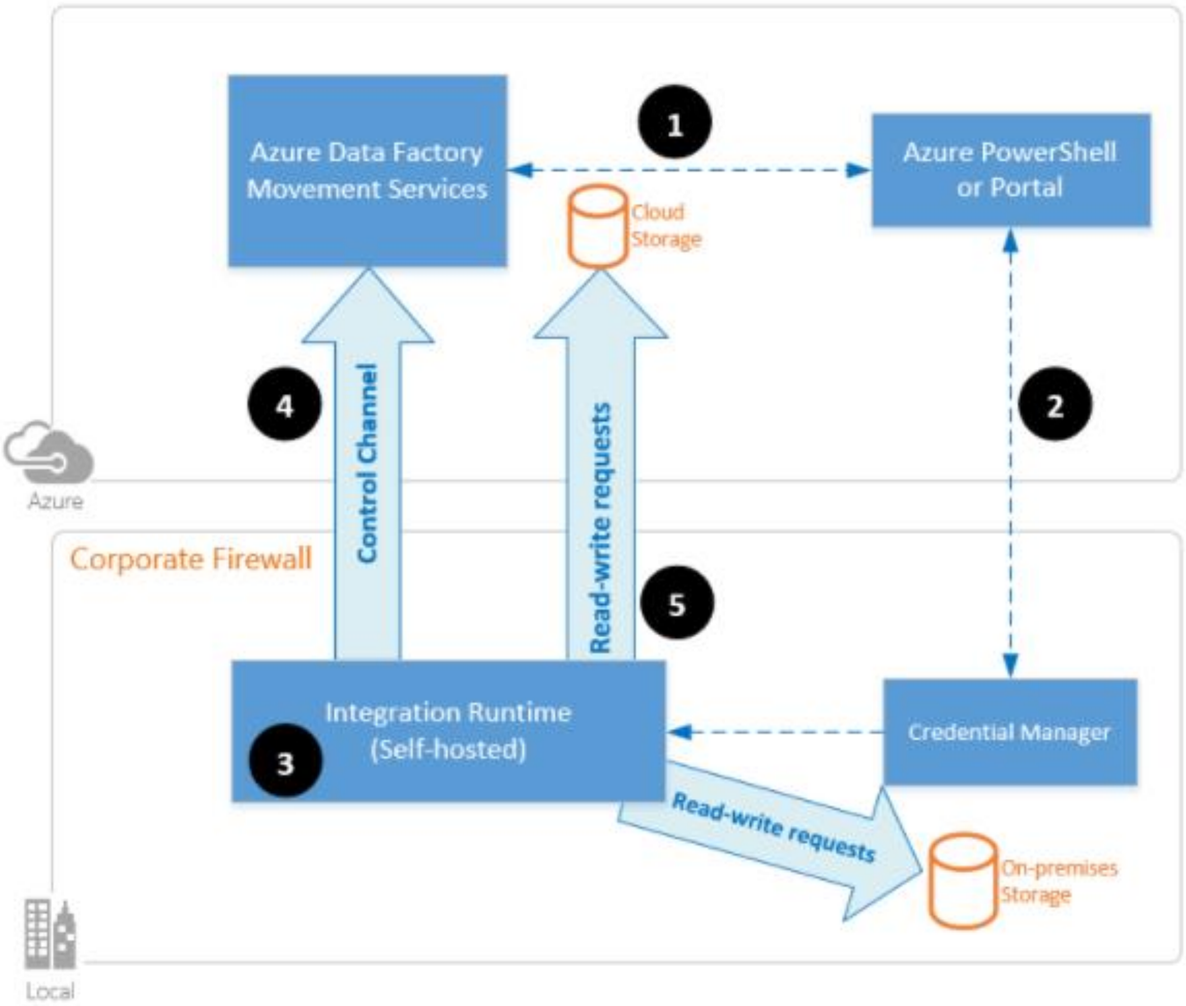
Create and configure a self-hosted integration runtime

The integration runtime (IR) is the compute infrastructure that Azure Data Factory uses to provide data-integration capabilities across different network environments. For details about IR, see Integration runtime overview.

A self-hosted integration runtime can run copy activities between a cloud data store and a data store in a private network. It also can dispatch transform activities against compute resources in an on-premises network or an Azure virtual network. The installation of a self-hosted integration runtime needs an on-premises machine or a virtual machine inside a private network.

When you move data between on-premises and the cloud, the activity uses a self-hosted integration runtime to transfer the data between an on-premises data source and the cloud.

Here is a high-level summary of the data-flow steps for copying with a self-hosted IR:



1. A data developer creates a self-hosted integration runtime within an Azure data factory by using the Azure portal or the PowerShell cmdlet.

2. The data developer creates a linked service for an on-premises data store. The developer does so by specifying the self-hosted integration runtime instance that the service should use to connect to data stores.

3. The self-hosted integration runtime node encrypts the credentials by using Windows Data Protection Application Programming Interface (DPAPI) and saves the credentials

locally. If multiple nodes are set for high availability, the credentials are further synchronized across other nodes. Each node encrypts the credentials by using DPAPI and stores them locally. Credential synchronization is transparent to the data developer and is handled by the self-hosted IR.

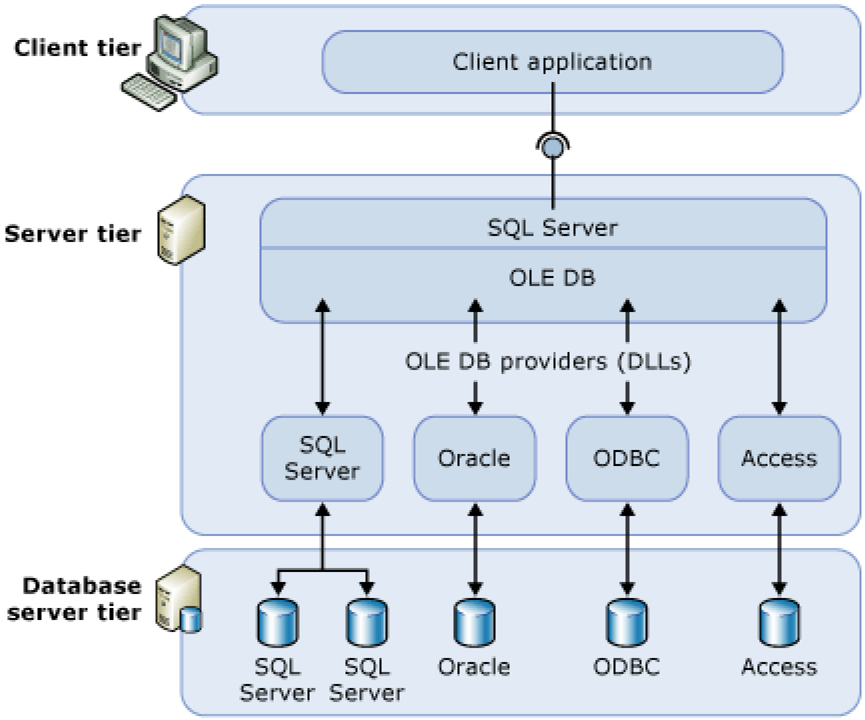
4. Azure Data Factory communicates with the self-hosted integration runtime to schedule and manage jobs. Communication is via a control channel that uses a shared Azure Relay connection. When an activity job needs to be run, Data Factory queues the request along with any credential information. It does so in case credentials aren't already stored on the self-hosted integration runtime. The self-hosted integration runtime starts the job after it polls the queue.

5. The self-hosted integration runtime copies data between an on-premises store and cloud storage. The direction of the copy depends on how the copy activity is configured in the data pipeline. For this step, the self-hosted integration runtime directly communicates with cloud-based storage services like Azure Blob storage over a secure HTTPS channel.

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

Creating SQL Server Linked Servers with Azure

Configure a linked server to enable the SQL Server Database Engine to execute commands against data sources outside of the local instance of SQL Server.



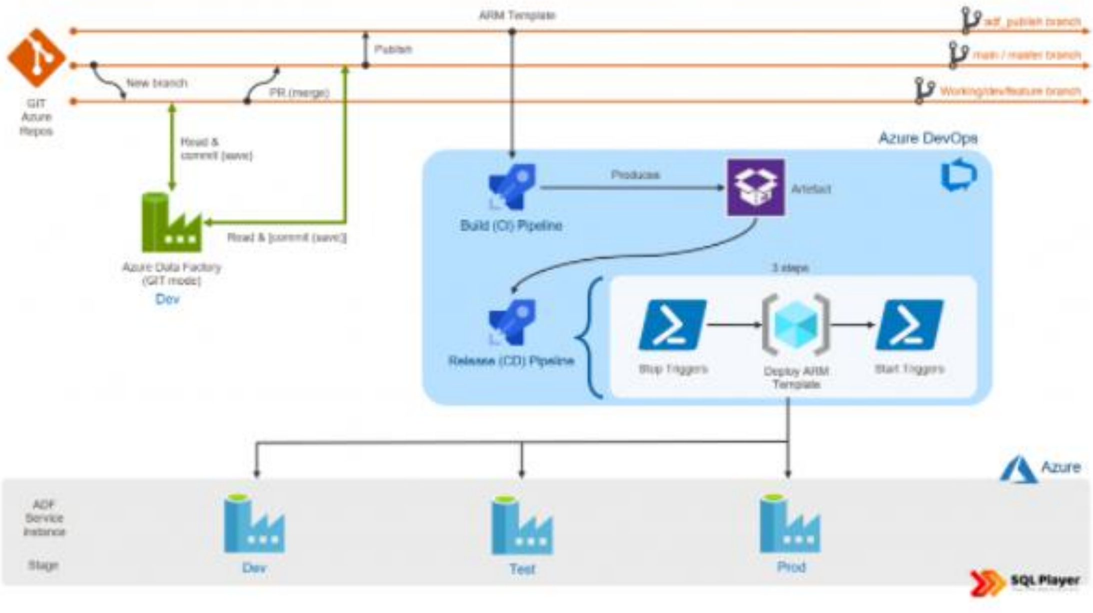
https://www.mssqltips.com/sqlservertip/3630/creating-sql-server-linked-servers-with-azure/

Deploy an Azure Data Factory: Microsoft approach (ARM template)

Set up GIT integration to assign your ADF service to the selected repository. If you are not sure how to achieve that – it is described here: Setting up Code Repository for Azure Data Factory v2. As a developer, you can work with your own branch and can switch ADF between multiple branches (including master/main). How this is possible? It's because having one ADF instance you can switch between two modes: GIT integrated (for development purposes) and real instance.
However, if you want to publish the changes (or new version) to another environment (or instance) – you must **Publish** the changes first. This performs to actions:

• Publishes the code from a developer version of code to real ADF instance. This can be done only from one branch: "collaboration" branch ("master" by default)

• Creates or updates ARM Template files into "adf_publish" branch. This branch will be used as a source for deployment.



Then you can build your own CI/CD process for deployment of ADF, using Azure DevOps, for instance. I don't want to dig deeper about how to deploy ADF with this approach as I already described it in the post: Deployment of Azure Data Factory with Azure DevOps.
Why many people do not like this approach?

• Semi-manual process, as at some point someone has to hit "Publish" button

• Full ADF (all artefacts) can be deployed only (no selective deployment)

• Limitation to one publish branch only (thankfully, you can name it now)

• Parametrize elements exposed within the ARM Template Parameter

• Restriction of 256 parameters maximum

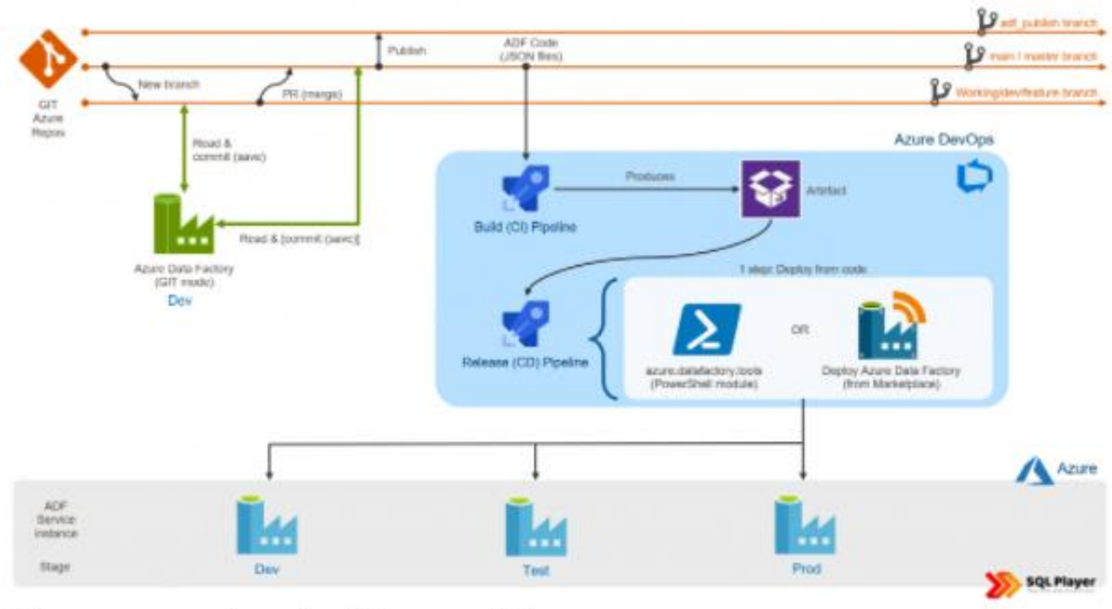
• Building a release pipeline is not an easy thing

• Will not delete any existing ADF objects in the target instance, when the object has been deleted from the source ADF

• Must use a few tasks in Release Pipeline (Azure DevOps) to deploy ADF (including PowerShell script)

Deploy an Azure Data Factory: Custom approach (JSON files, via REST API)

There is another approach in opposite to ARM templates located in ‘ADF_Publish’ branch. Many companies leverage that workaround and it works great. In this scenario, you don’t have to Publish the changes to update ARM Template. With this approach, we can fully automate CI/CD process as collaboration branch will be our source for deployment. This is the reason why the approach is also known as (direct) deployment from code (JSON files). In all branches, ADF is stored as multiple JSON files (one file per object), whereas in ADF_Publish branch – ADF is kept as ARM (Azure Resource Manager) Template file(s).



Why some people prefer this approach?

It’s much more natural and similar to managing the code of other applications

Eliminates enforcement of using only one (`adf_publish`) branch (helpful if the company’s branches policy is much complex)

You can parameterize any single property and artefact of the Data Factory

Selectively deploy a subset of artefacts is possible

Only one task in Release pipeline (Azure DevOps) covers all the needs of deploying ADF from code (more details below)

What both have in common?

In both cases, you must manage ADF triggers properly. Before deployment of any (active) trigger onto target ADF, it must be stopped, then deploy everything and start triggers again. This requires additional steps in a Release pipeline in order to do so. Microsoft offers PowerShell script to start/stop triggers as pre/post-deployment activity.

https://sqlplayer.net/2021/01/two-methods-of-deployment-azure-data-factory/

Question 4: Skipped
What sort of pipeline is required in Azure DevOps for creating artifacts used in releases?

- ○ An Artifact pipeline
- ○ YAML pipelines
- ○ A Build pipeline (Correct)
- ○ A Release pipeline

Explanation
The output of a Build pipeline is one or more artifacts that can be used within release pipelines for automated deployments in Azure DevOps.

In Azure DevOps, before there was the multi stage yaml pipelines (now known as "Pipelines", you usually used the Build Pipeline to build / create your software binaries (e. g. dotnet publish or ng build --prod) and stored these artifacts in the Azure DevOps drop location.

Then you normally had a Releasee Pipeline that gets triggered with these build artifacts (software binaries) and deploys them to one or many stages.

The reason to separate these two pipelines (build and release) is that you want to build a specific version of your software only once and then use the same binaries in each of your target environment (e. g. dev / test / production).

With the new pipeline, you usually use the first Stage to build your artifacts, and the next Stages to deploy it - similar as before but in one module.

If you have previously used the build & release pipeline, you will see the old build definition inside the new Pipeline module, and the old release definition in the old release module. However, they never brought YAML to the Release Pipelines because they know that they will replace them with the multi stage pipelines anyway.

Conclusion: If you use the new multi-stage "Pipeline" module, you shouldn't use the classic Release Pipelines anymore.

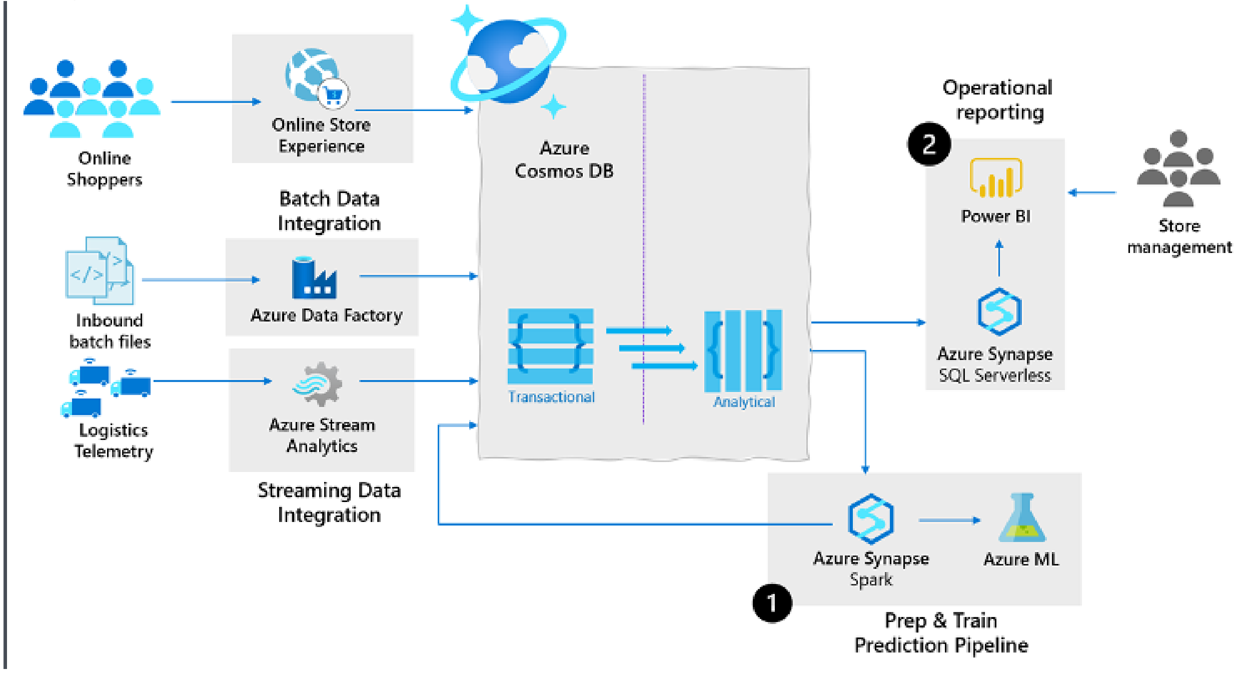
https://stackoverflow.com/questions/58813608/whats-the-difference-between-a-build-pipeline-and-a-release-pipeline-in-azure-de

Question 5: Skipped

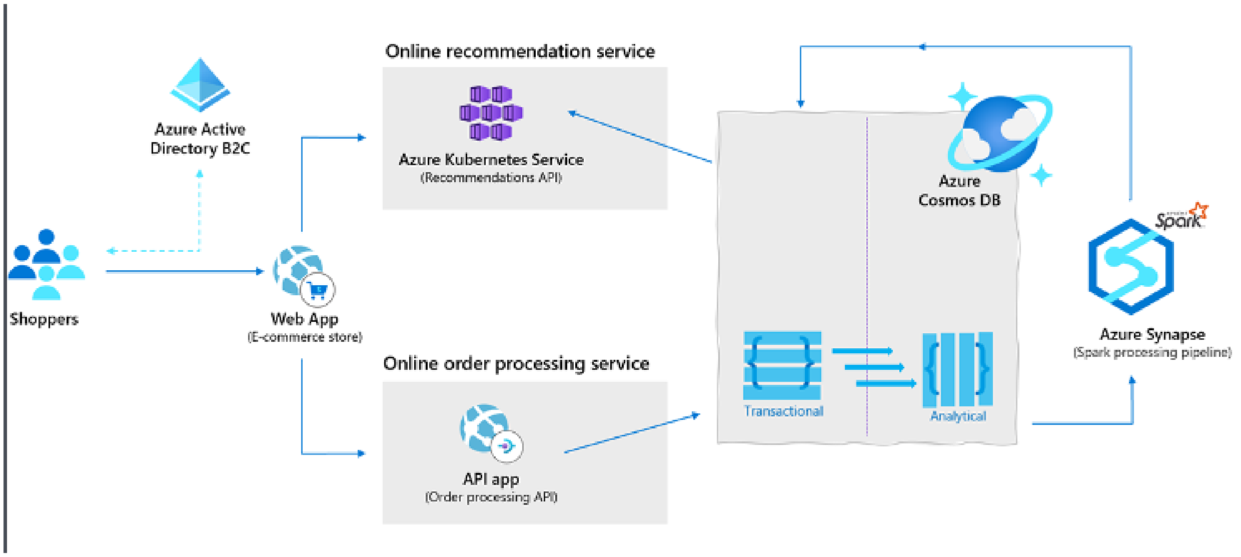
Scenario: You are working at OZcorp which is a supply chain which is generating increasing volumes of operational data every minute for orders, shipments and sales transactions, manufactures and retailers. It needs an operational database that can scale to handle the data volumes as well as an analytical platform to get to a level of real-time contextual intelligence to stay ahead of the curve.

Review the following architecture designs.

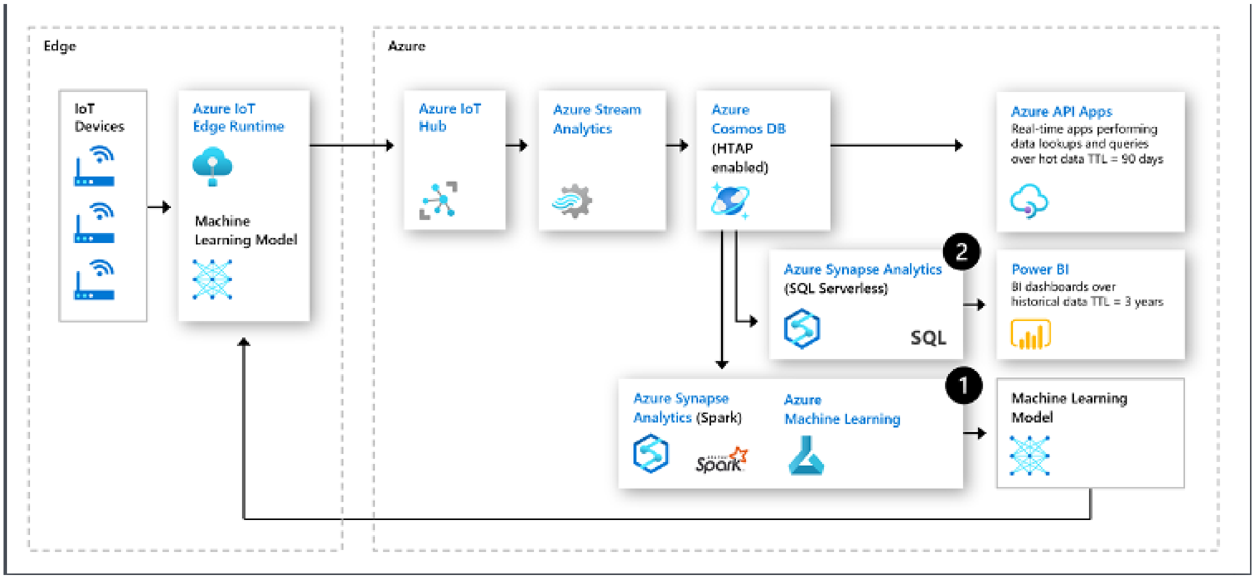
Design A:



Design B:



Design C:


Edge
Azure
IoT Devices
Azure IoT Edge Runtime
Machine Learning Model
Azure IoT Hub
Azure Stream Analytics
Azure Cosmos DB (HTAP enabled)
Azure API Apps
Real-time apps performing data lookups and queries over hot data TTL = 90 days
2
Azure Synapse Analytics (SQL Serverless)
SQL
Power BI
BI dashboards over historical data TTL = 3 years
1
Azure Synapse Analytics (Spark)
Spark
Azure Machine Learning
Machine Learning Model

Which design would be best suited for the need?

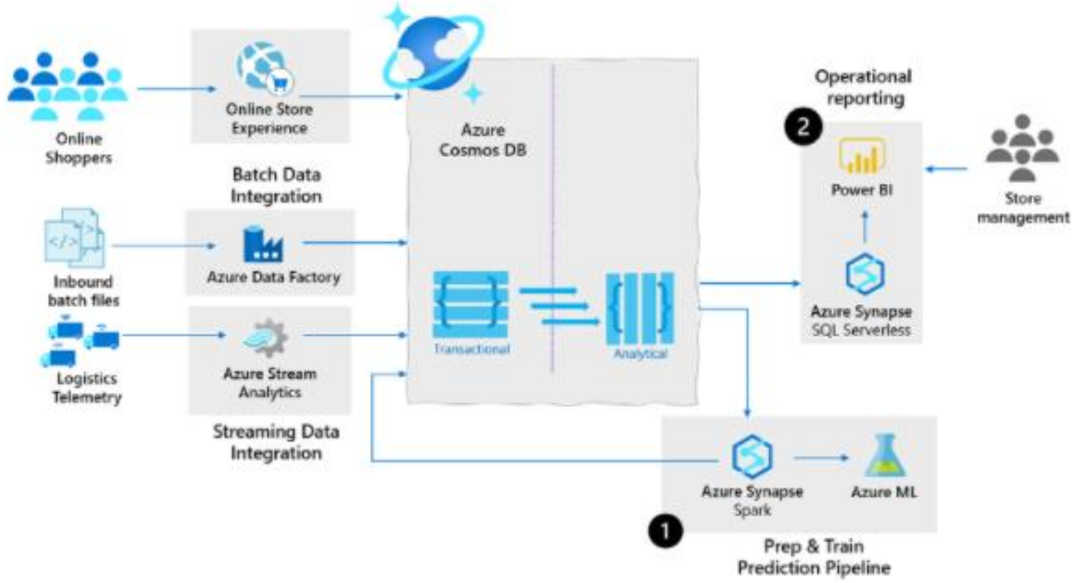
- ○ Design B
- ○ None of the listed options
- ○ Design C
- ○

Design A
(Correct)

Explanation
Supply chain analytics, forecasting and reporting.

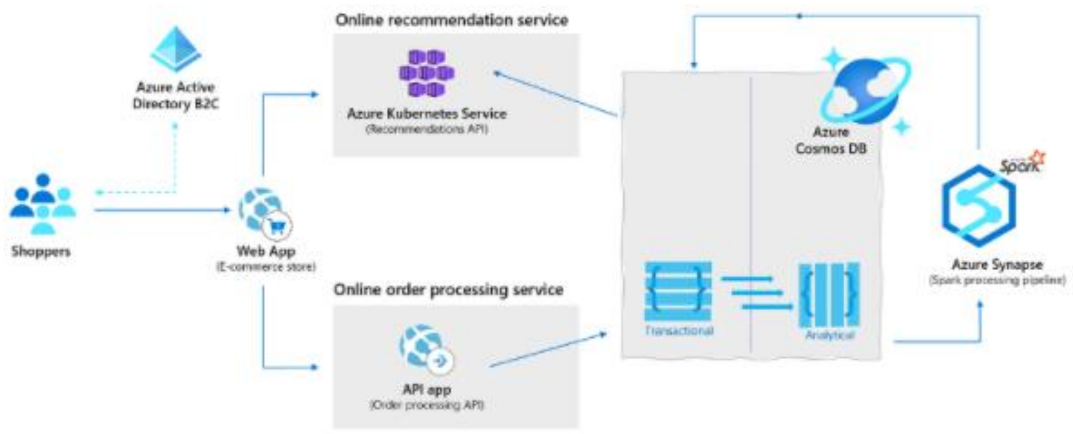
With supply chains generating increasing volumes of operational data every minute for orders, shipments and sales transactions, manufactures and retailers need an operational database that can scale to handle the data volumes as well as an analytical platform to get to a level of real-time contextual intelligence to stay ahead of the curve.

Azure Synapse Link for Cosmos DB allows these organizations to store data from their sales systems, ingest real-time telemetry data from in vehicle systems and integrate date from their ERP systems into a common operational store in Azure Cosmos DB and then leverage the data from Synapse analytics to enable both predictive analytics scenarios such as stock out monitoring and supply chain bottleneck management (1) in addition to enabling operational reporting directly on their operation data using standard reporting tools such as Power BI (2).



Retail real-time personalization.

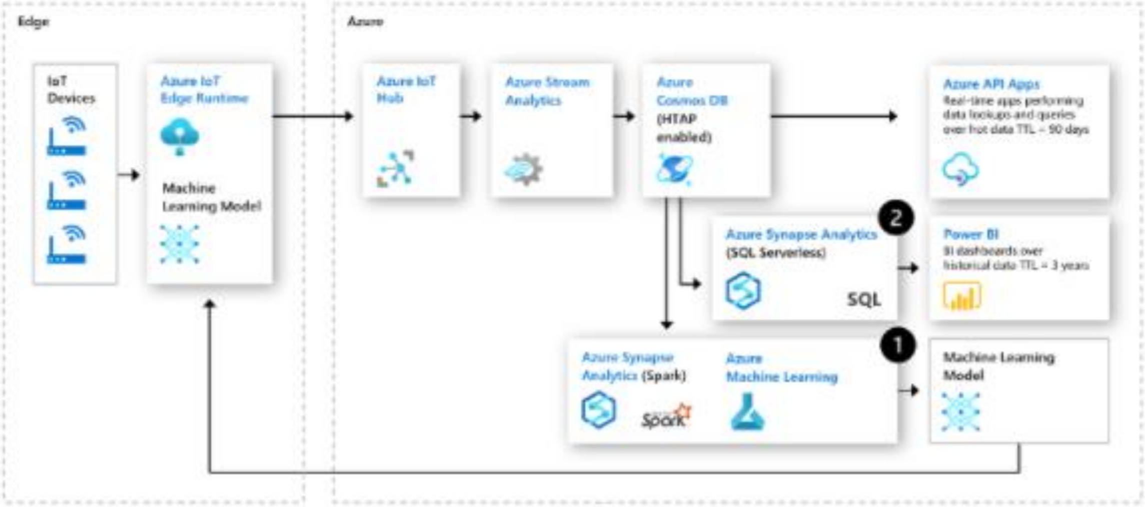
In retail, many web-based retailers will perform real-time basket analysis to make product recommendations to customers who are about to purchase products. This increased revenues for these organizations as the provided targeted suggestions at the point of sales.



Predictive maintenance using anomaly detection with IOT

Industrial IOT innovations have drastically reduced downtimes of machinery and increased overall efficiency across all fields of industry. One of such innovations is predictive maintenance analytics for machinery at the edge of the cloud.

The following architecture leverages the cloud native HTAP capabilities of Azure Synapse Link for Azure Cosmos DB in IoT predictive maintenance:


Edge
IoT Devices
Azure IoT Edge Runtime
Machine Learning Model
Azure
Azure IoT Hub
Azure Stream Analytics
Azure Cosmos DB (HTAP enabled)
Azure API Apps
Real-time apps performing data lookups and queries over hot data TTL = 90 days
2
Azure Synapse Analytics (SQL Serverless)
SQL
Power BI
BI dashboards over historical data TTL = 3 years
1
Azure Synapse Analytics (Spark)
Spark
Azure Machine Learning
Machine Learning Model

https://docs.microsoft.com/en-us/azure/cosmos-db/synapse-link-use-cases

Question 6: Skipped
Nonstructured data is stored in nonrelational systems, commonly called unstructured or NoSQL systems. Nonstructured data is stored in nonrelational systems, commonly called unstructured or NoSQL systems.

Which of the following fit this description? (Select all that apply)

- □ Document database (Correct)
- □ Key-value store (Correct)
- □ Db2
- □ CompleteDB
- □ Postgre
- □ Column database (Correct)

- □ Graph database (Correct)

Explanation
Nonstructured data

Examples of nonstructured data include binary, audio, and image files. Nonstructured data is stored in nonrelational systems, commonly called unstructured or NoSQL systems. In nonrelational systems, the data structure isn't defined at design time, and data is typically loaded in its raw format. The data structure is defined only when the data is read. The difference in the definition point gives you flexibility to use the same source data for different outputs. Nonrelational systems can also support semistructured data such as JSON file formats.

Nonstructured data is stored in nonrelational systems, commonly called unstructured or NoSQL systems. :

1. **Key-value store**: Stores key-value pairs of data in a table structure.

2. **Document database**: Stores documents that are tagged with metadata to aid document searches.

3. **Graph database**: Finds relationships between data points by using a structure that's composed of vertices and edges.

4. **Column database**: Stores data based on columns rather than rows. Columns can be defined at the query's runtime, allowing flexibility in the data that's returned performantly.

https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/non-relational-data

Question 7: Skipped
How do column statistics improve query performance?

- ○ By caching column values for queries.
- ○ By keeping track of which columns are being queried.
- ○ By caching table values for queries.
- ○ By keeping track of how much data exists between ranges in columns.

(Correct)

Explanation
Column statistics track cardinality and range density to determine which data access paths return the fewest rows for speed.

When queries are submitted, a dedicated SQL pool query optimizer tries to determine which access paths to the data will result in the least amount of effort to retrieve the data required to resolve the query. It is a cost-based optimizer, and compares the cost of various query plans, and then chooses the plan with the lowest cost.

Statistics in dedicated SQL pools

To aid this process, statistics are required that describe the amount of data that is present within ranges of values, and range of rows that may be returned to fulfill a query filter or join. Therefore, after loading data into a dedicated SQL pool, collecting statistics on your data is one of the most important things you can do for query optimization.

When you create a database in a dedicated SQL pool in Azure Synapse Analytics, the automatic creation of statistics is turned on by default. This means that statistics are created when you run the following type of Transact-SQL statements:

- `SELECT`
- `INSERT-SELECT`
- `CTAS`
- `UPDATE`
- `DELETE`
- `EXPLAIN` when containing a join or the presence of a predicate is detected

When executing the above Transact-SQL statements, that the statistics creation is performed on the fly, and as a result, there can be a slight degradation in query performance.

To avoid this, statistics are also created on any index that you create that helps aid the query optimize process. As this is an action that is performed in advance of querying the table on which the index is based, it means that the statistics are created in advance. However, you must consider that as new data is loaded into the table, the statistics may become out of date.

As such, it is important to update the statistics after you load data or update large ranges of data, so that queries can benefit from the updated statistics information.

You can check if your data warehouse has `AUTO_CREATE_STATISTICS` configured by running the following command:

```
SQL

SELECT name, is_auto_create_stats_on

FROM sys.databases
```

If your data warehouse doesn't have AUTO_CREATE_STATISTICS enabled, it is recommended that you enable this property by running the following command:

```
SQL

ALTER DATABASE <yourdatawarehousename>

SET AUTO_CREATE_STATISTICS ON
```

Statistics in serverless SQL pools

Statistics in a serverless SQL pool has the same objective of using a cost-based optimizer to choose an execution plan that will execute the fastest. How it creates its statistics is different.

Serverless SQL pool analyses incoming user queries for missing statistics. **If statistics are missing, the query optimizer creates statistics on individual columns in the query predicate or join condition to improve cardinality estimates for the query plan.** The SELECT statement will trigger automatic creation of statistics. You can also manually create statistics, this is important when working with CSV files, as automatic statistics creation is not enabled for them.

In the following example, a system stored procedure is used to specify the creation of statistics for a specific Transact-SQL statement

```
SQL

sys.sp_create_openrowset_statistics [ @stmt = ] N'statement_text'
```

To create statistics for a specific column within a csv file, you can run the following code:

```
SQL

/* make sure you have the credentials to access the storage account created

IF EXISTS (SELECT * FROM sys.credentials WHERE name = 'https://azureopendatastora
ge.blob.core.windows.net/censusdatacontainer')
```

```
DROP CREDENTIAL [https://azureopendatastorage.blob.core.windows.net/censusdatacontainer]

GO


CREATE CREDENTIAL [https://azureopendatastorage.blob.core.windows.net/censusdatacontainer]

WITH IDENTITY='SHARED ACCESS SIGNATURE',

SECRET = ''

GO

*/


/*
```

The following code will create statistics on a column named year, from a file named population.csv

```
*/


EXEC sys.sp_create_openrowset_statistics N'SELECT year

FROM OPENROWSET(

BULK ''https://sqlondemandstorage.blob.core.windows.net/csv/population/population.csv'',

FORMAT = ''CSV'',

FIELDTERMINATOR ='','',

ROWTERMINATOR = ''\n''

)

WITH (

[country_code] VARCHAR (5) COLLATE Latin1_General_BIN2,

[country_name] VARCHAR (100) COLLATE Latin1_General_BIN2,

[year] smallint,

[population] bigint

) AS [r]

‘
```

You should also update the statistics when the data in the files change. In fact, Serverless SQL pool automatically recreates statistics if data is changed significantly.

Every time statistics are automatically created, the current state of the dataset is also saved: file paths, sizes, last modification dates.

To update statistics for the year column in the dataset, which is based on the population.csv file, you need to drop and then create them, here is the drop statement:

```
SQL

EXEC sys.sp_drop_openrowset_statistics N'SELECT year

FROM OPENROWSET(

BULK ''https://sqlondemandstorage.blob.core.windows.net/csv/population/population
.csv'',

FORMAT = ''CSV'',

FIELDTERMINATOR =''',''',

ROWTERMINATOR = ''\n''

)

WITH (

[country_code] VARCHAR (5) COLLATE Latin1_General_BIN2,

[country_name] VARCHAR (100) COLLATE Latin1_General_BIN2,

[year] smallint,

[population] bigint

) AS [r]

‘
```

To update statistics for a statement, you need to drop and create statistics. The following stored procedure is used to drop statistics against a specific Transact-SQL text:

```
SQL

sys.sp_drop_openrowset_statistics [ @stmt = ] N'statement_text’
```

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

Question 8: Skipped

Scenario: You are working on an Azure Synapse Analytics Workspace as part of your project. One of the requirements is to have Azure Synapse Analytics Workspace access an Azure Data Lake Store using the benefits of the security provided by Azure Active Directory.

Which is the best authentication method to use?

- ○

  Managed identities

  (Correct)
- ○

  SQL Authentication
- ○

  Shared access signatures
- ○

  Storage account keys

Explanation

Managed identities provides Azure services with an automatically managed identity in Azure Active Directory. You can use the Managed Identity capability to authenticate to any service that support Azure Active Directory authentication.

The following are the types of authentication that you should be aware of when working with Azure Synapse Analytics.

Azure Active Directory

Azure Active Directory is a directory service that allows you to centrally maintain objects that can be secured. The objects can include user accounts and computer accounts. An employee of an organization will typically have a user account that represents them in the organizations Azure Active Directory tenant, and they then use the user account with a password to authenticate against other resources that are stored within the directory using a process known as single sign-on.

The power of Azure Active Directory is that they only have to login once, and Azure Active Directory will manage access to other resources based on the information held within it using pass through authentication. If a user and an instance of Azure Synapse Analytics are part of the same Azure Active Directory, it is possible for the user to access Azure Synapse Analytics without an apparent login. If managed correctly, this process is seamless as the administrator would have given the user authorization to access Azure Synapse Analytics dedicated SQL pool as an example.

In this situation, it is normal for an Azure Administrator to create the user accounts and assign them to the appropriate roles and groups in Azure Active Directory. The Data Engineer will then add the user, or a group to which the user belongs to access a dedicated SQL pool.

Managed identities

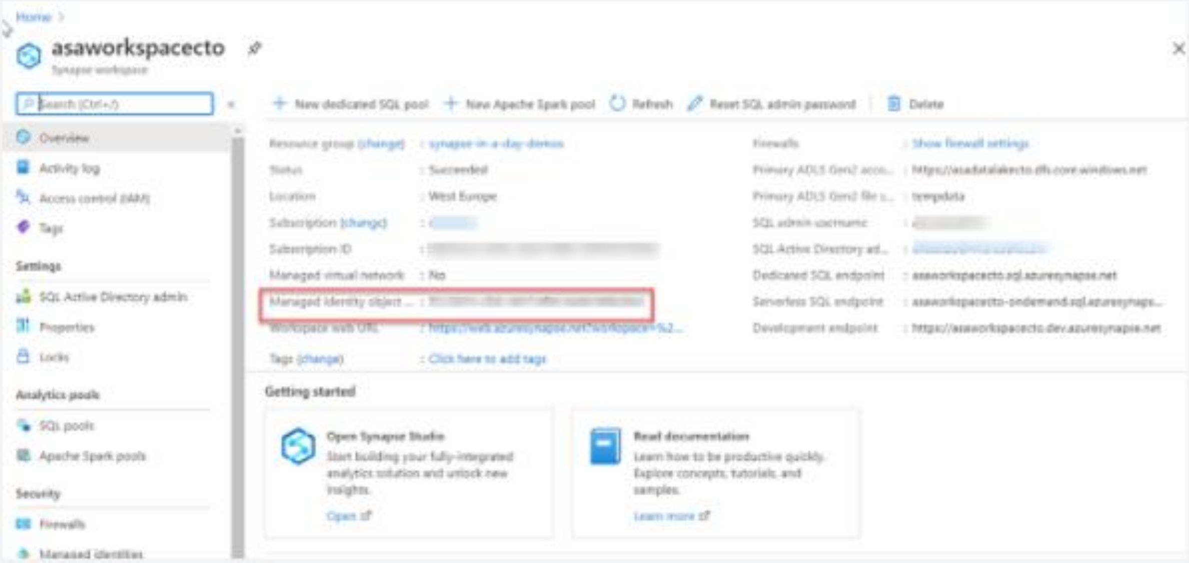
Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD. You can use the Managed Identity capability to authenticate to any service that support Azure Active Directory authentication.

Managed identities for Azure resources are the new name for the service formerly known as Managed Service Identity (MSI). A system-assigned managed identity is created for your Azure Synapse workspace when you create the workspace.

Azure Synapse also uses the managed identity to integrate pipelines. The managed identity lifecycle is directly tied to the Azure Synapse workspace. If you delete the Azure Synapse workspace, then the managed identity is also cleaned up.

The workspace managed identity needs permissions to perform operations in the pipelines. You can use the object ID or your Azure Synapse workspace name to find the managed identity when granting permissions.

You can retrieve the managed identity in the Azure portal. Open your Azure Synapse workspace in Azure portal and select **Overview** from the left navigation. The managed identity's object ID is displayed to in the main screen.



Resource group (change) : synapse-in-a-day-demos
Status : Succeeded
Location : West Europe
Subscription (change) : [redacted]
Subscription ID : [redacted]
Managed virtual network : No
Managed identity object ... : [redacted]
Workspace web URL : https://web.azuresynapse.net?workspace=%2...
Tags (change) : Click here to add tags

Firewalls : Show firewall settings
Primary ADLS Gen2 acco... : https://asadatalakecto.dfs.core.windows.net
Primary ADLS Gen2 file s... : tempdata
SQL admin username : [redacted]
SQL Active Directory ad... : [redacted]
Dedicated SQL endpoint : asaworkspacecto.sql.azuresynapse.net
Serverless SQL endpoint : asaworkspacecto-ondemand.sql.azuresynaps...
Development endpoint : https://asaworkspacecto.dev.azuresynapse.net

The managed identity information will also show up when you create a linked service that supports managed identity authentication from Azure Synapse Studio.

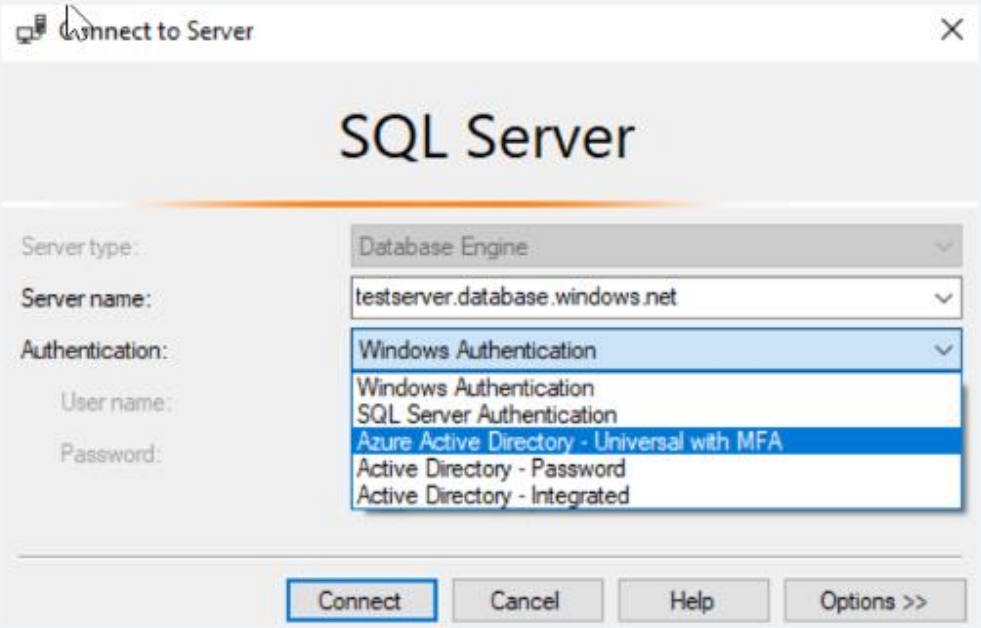
SQL Authentication

For user accounts that are not part of an Azure Active directory, then using SQL Authentication will be an alternative. In this instance, a user is created in the instance of a dedicated SQL pool. If the user in question requires administrator access, then the details of the user are held in the master database. If administrator access is not required, you can create a user in a specific database. A user then connects directly to the Azure Synapse Analytics dedicated SQL pool where they are prompted to use a username and password to access the service.

This approach is typically useful for external users who need to access the data, or if you are using third party or legacy applications against the Azure Synapse Analytics dedicated SQL pool

Multi factor authentication

Synapse SQL support connections from SQL Server Management Studio (SSMS) using Active Directory Universal Authentication.



Connect to Server

SQL Server

Server type: Database Engine

Server name: testserver.database.windows.net

Authentication: Windows Authentication

User name:
Password:
Windows Authentication
SQL Server Authentication
Azure Active Directory - Universal with MFA
Active Directory - Password
Active Directory - Integrated
Connect
Cancel
Help
Options >>

This enables you to operate in environments that use conditional access policies that enforce multi-factor authentication as part of the policy.

Keys

If you are unable to use a managed identity to access resources such as Azure Data Lake then you can use storage account keys and shared access signatures.

With storage account keys. Azure creates two of these keys (primary and secondary) for each storage account you create. The keys give access to everything in the account. You'll find the storage account keys in the Azure portal view of the storage account. Just select **Settings**, and then click **Access keys**.

As a best practice, you shouldn't share storage account keys, and you can use Azure Key Vault to manage and secure the keys.

Azure Key Vault is a secret store: a centralized cloud service for storing app secrets - configuration values like passwords and connection strings that must remain secure at all times. Key Vault helps you control your apps' secrets by keeping them in a single central location and providing secure access, permissions control, and access logging.

The main benefits of using Key Vault are:

- Separation of sensitive app information from other configuration and code, reducing risk of accidental leaks

- Restricted secret access with access policies tailored to the apps and individuals that need them

- Centralized secret storage, allowing required changes to happen in only one place

- Access logging and monitoring to help you understand how and when secrets are accessed

Secrets are stored in individual vaults, which are Azure resources used to group secrets together. Secret access and vault management is accomplished via a REST API, which is also supported by all of the Azure management tools as well as client libraries available for many popular languages. Every vault has a unique URL where its API is hosted.

Shared access signatures

If an external third-party application need access to your data, you'll need to secure their connections without using storage account keys. For untrusted clients, use a shared access signature (SAS). A shared access signature is a string that contains a security token that can be attached to a URI. Use a shared access signature to delegate access to storage objects and specify constraints, such as the permissions and the time range of access. You can give a customer a shared access signature token.

Types of shared access signatures

You can use a service-level shared access signature to allow access to specific resources in a storage account. You'd use this type of shared access signature, for example, to allow an app to retrieve a list of files in a file system or to download a file.

Use an account-level shared access signature to allow access to anything that a service-level shared access signature can allow, plus additional resources and abilities. For example, you can use an account-level shared access signature to allow the ability to create file systems.

https://docs.microsoft.com/en-us/azure/synapse-analytics/security-baseline

Question 9: Skipped

Knowing now the different concepts of spark it is imperative to understand how it fits in with the different Data services on Azure.

Which of the following is best described by:

"A managed Spark as a Service propriety Solution that provides an end to end data engineering/data science platform as a solution. This is of interest for Data Engineers and Data Scientists, working on big data projects daily because it provides the whole platform in which you have the ability to create and manage the big data/data science pipelines/projects all on one platform."

- ○ Azure Databricks **(Correct)**
- ○ Spark Pools in Azure Synapse Analytics
- ○ Apache Spark
- ○ HDI

Explanation

There are two concepts within Apache Spark Pools in Azure Synapse Analytics, namely Spark pools and Spark Instances. In short, they do the following:

Spark Pools:

• Exists as Metadata

• Creates a Spark Instance

- No costs associated with creating Pool

- Permissions can be applied

- Best practices

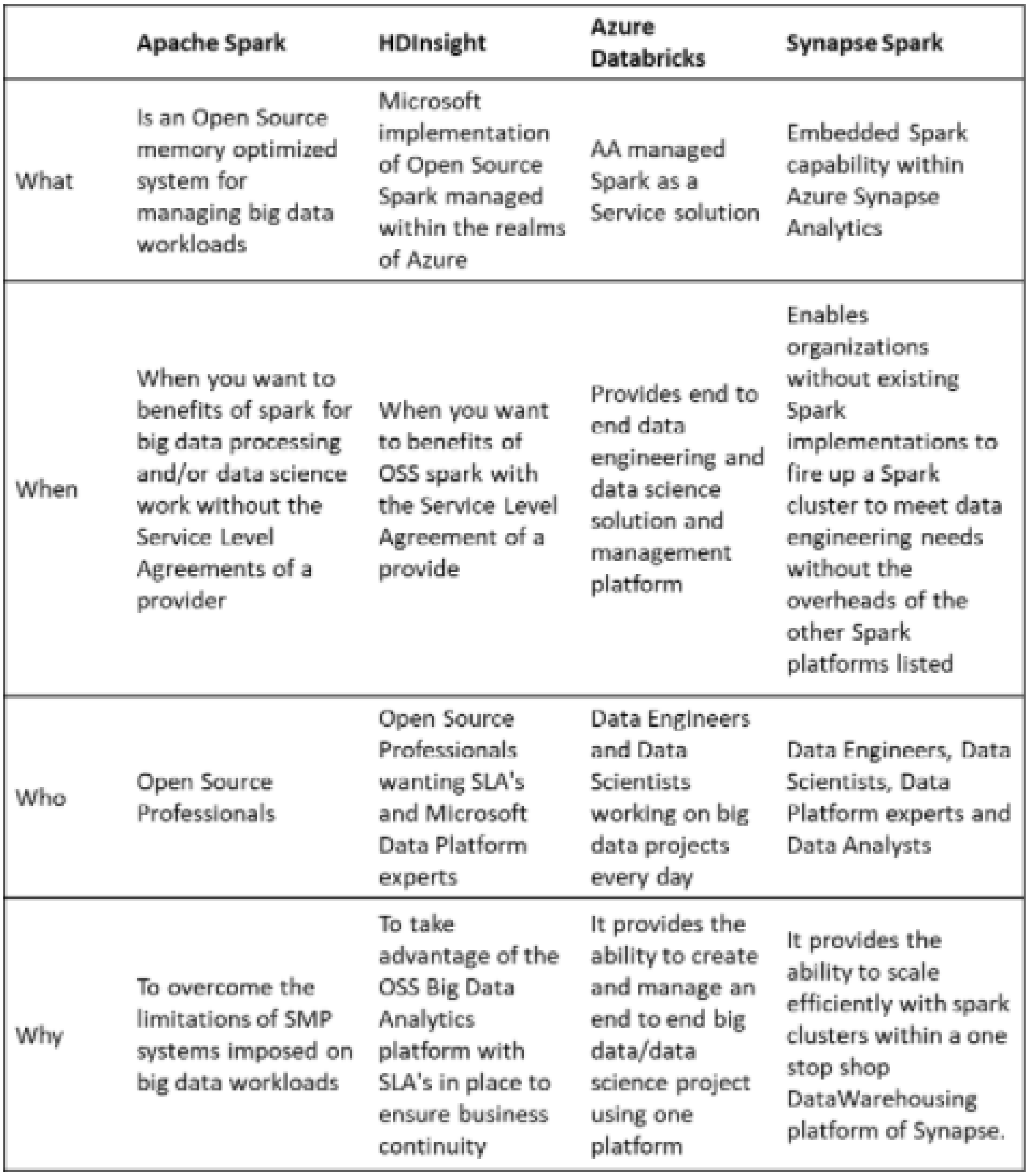
Spark Instances:

- Created when connected to Spark Pool, Session, or Job

- Multiple users can have access

- Reusable

Knowing now the different concepts of spark it is imperative to understand how it fits in with the different Data services on Azure. Below is a table where "the when to use what" is outlined:



	Apache Spark	HDInsight	Azure Databricks	Synapse Spark
What	Is an Open Source memory optimized system for managing big data workloads	Microsoft implementation of Open Source Spark managed within the realms of Azure	AA managed Spark as a Service solution	Embedded Spark capability within Azure Synapse Analytics
When	When you want to benefits of spark for big data processing and/or data science work without the Service Level Agreements of a provider	When you want to benefits of OSS spark with the Service Level Agreement of a provide	Provides end to end data engineering and data science solution and management platform	Enables organizations without existing Spark implementations to fire up a Spark cluster to meet data engineering needs without the overheads of the other Spark platforms listed
Who	Open Source Professionals	Open Source Professionals wanting SLA's and Microsoft Data Platform experts	Data Engineers and Data Scientists working on big data projects every day	Data Engineers, Data Scientists, Data Platform experts and Data Analysts
Why	To overcome the limitations of SMP systems imposed on big data workloads	To take advantage of the OSS Big Data Analytics platform with SLA's in place to ensure business continuity	It provides the ability to create and manage an end to end big data/data science project using one platform	It provides the ability to scale efficiently with spark clusters within a one stop shop DataWarehousing platform of Synapse.

***Spark Pools in Azure Synapse Analytics*:** Spark in Azure Synapse Analytics is a capability of Spark embedded in Azure Synapse Analytics in which organizations that don’t have existing spark implementations yet, get the functionality to spin up a spark cluster to meet data engineering needs without the overhead of the other Spark Platforms listed. Data Engineers, Data scientist, Data Platform Experts, and Data Analyst can come together within Synapse Analytics where the Spark cluster is spun up quickly to meet the needs. It provides scale in an efficient way for Spark Clusters and integrates with the one stop shop Data warehousing platform of Synapse.

Apache Spark: Apache Spark is an open-source memory optimized system for managing big data workloads, which is used when you want a spark engine for big data processing or data science where you don’t mind that there is no SLA provided. Usually it is of interest of Open Source Professionals and the reason for Apache spark is to overcome the limitations of what was known as SMP systems for big data workloads.

HDI: HDI is an implementation by Microsoft of Open Source Spark, managed on the Azure Platform. You can use HDI for a spark environment when you are aware of the benefits of Apache Spark in its OSS form, but you want a SLA. Usually this of interest of Open Source Professionals needing an SLA as well as Data Platform experts experienced with Microsoft.

Azure Databricks: Azure Databricks is a managed Spark as a Service propriety Solution that provides an end to end data engineering/data science platform as a solution. Azure Databricks is of interest for Data Engineers and Data Scientists, working on big data projects daily because it provides the whole platform in which you have the ability to create and manage the big data/data science pipelines/projects all on one platform.

https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-overview

Question 10: Skipped
Azure Advisor provides you with personalized messages that provide information on best practices to optimize the setup of your Azure services. Azure Advisor recommendations are free, and the recommendations are based on telemetry data that is generated by Azure Synapse Analytics. The telemetry data that is captured by Azure Synapse Analytics include which of the following? (Select all that apply)

- □ Encryption deficiencies
- □ TempDB utilization data **(Correct)**
- □ Adaptive Cache **(Correct)**
- □ Column statistics data **(Correct)**
- □ Data Skew and replicated table information **(Correct)**

Explanation

Azure Advisor provides you with personalized messages that provide information on best practices to optimize the setup of your Azure services. It analyzes your resource configuration and usage telemetry and then recommends solutions that can help you improve the cost effectiveness, performance, Reliability (formerly called High availability), and security of your Azure resources.

How Azure Synapse Analytics works with Azure Advisor

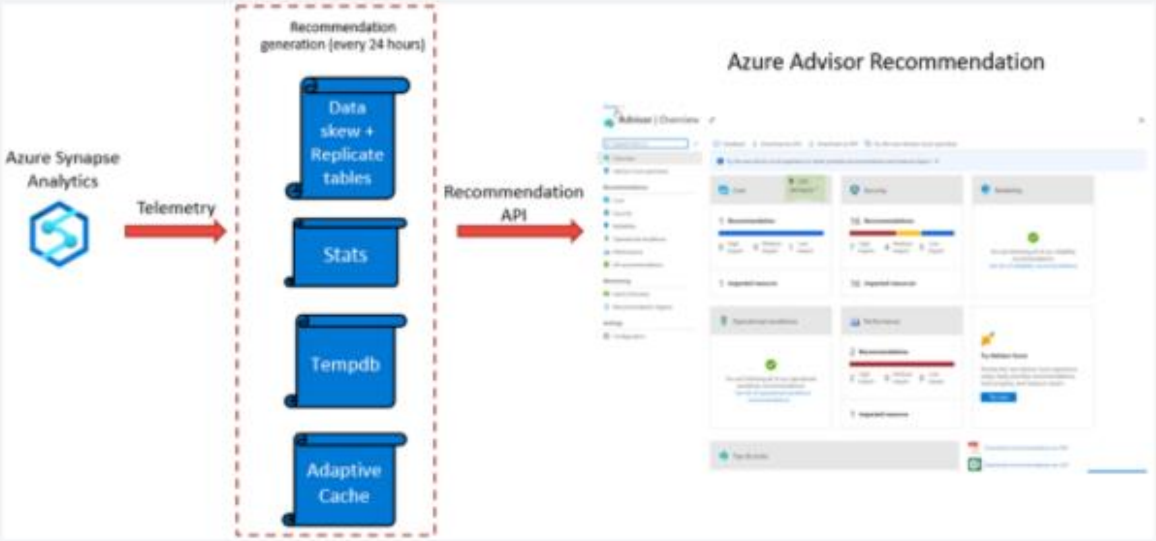
Azure Advisor recommendations are free, and the recommendations are based on telemetry data that is generated by Azure Synapse Analytics. The telemetry data that is captured by Azure Synapse Analytics include:

- Data Skew and replicated table information

- Column statistics data

- TempDB utilization data

- Adaptive Cache


Azure Synapse Analytics
Telemetry
Recommendation generation (every 24 hours)
Data skew + Replicate tables
Stats
Tempdb
Adaptive Cache
Recommendation API
Azure Advisor Recommendation

Azure Advisor recommendations are checked every 24 hours, as the recommendation API is queried against the telemetry generated from with Azure Synapse Analytics, and the recommendation dashboards are then updated to reflect the information that the telemetry has generated. This can then be viewed in the Azure Advisor dashboard.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-recommendations

Question 11: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Azure Synapse Analytics can work by acting as the one stop shop to meet all of your analytical needs in an integrated environment.

You can develop big data engineering and machine learning solutions using [?]. You can take advantage of the big data computation engine to deal with complex compute transformations that would take too long in a data warehouse.

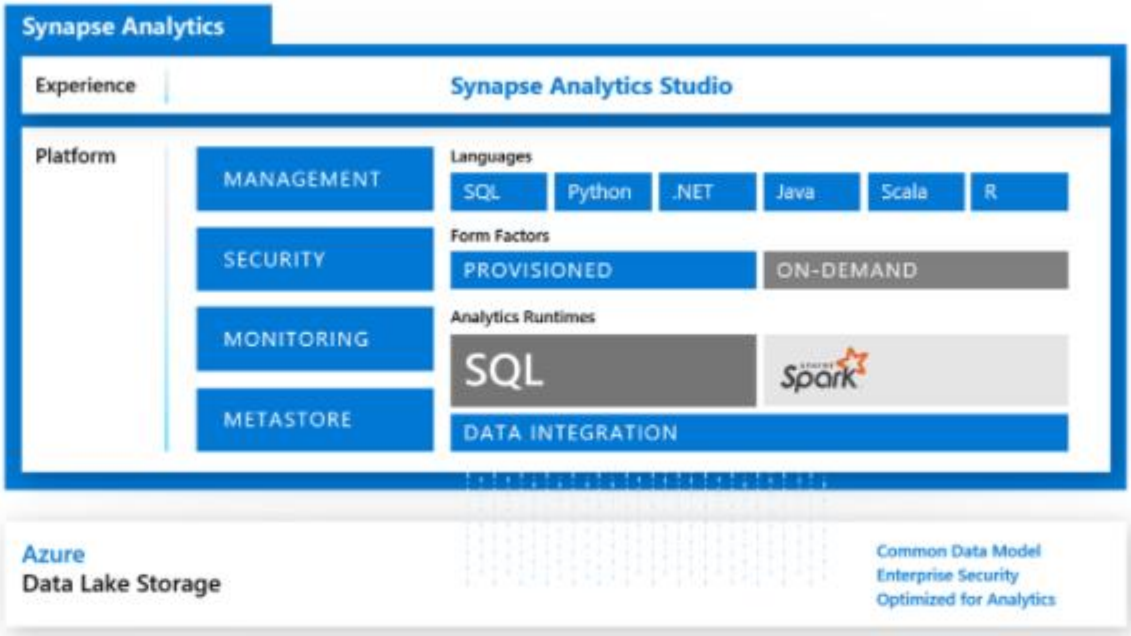
- ○ Azure Synapse Link
- ○ Azure Synapse Pipelines
- ○ Apache Spark for Azure Synapse (Correct)
- ○ Azure Cosmos DB
- ○ Azure Synapse SQL

Explanation

Azure Synapse Analytics can work by acting as the one stop shop to meet all of your analytical needs in an integrated environment. It does this by providing the following capabilities:

Analytics capabilities offered through Azure Synapse SQL through either dedicated SQL pools or SQL Serverless pools

Azure Synapse SQL is a distributed query system that enables you to implement data warehousing and data virtualization scenarios using standard T-SQL experiences familiar to data engineers. Synapse SQL offers both serverless and dedicated resource models to work with both descriptive and diagnostic analytical scenarios. For predictable performance and cost, create dedicated SQL pools to reserve processing power for data stored in SQL tables. For unplanned or ad-hoc workloads, use the always-available, serverless SQL endpoint.



Apache Spark pool with full support for Scala, Python, SparkSQL, and C#

You can develop big data engineering and machine learning solutions using Apache Spark for Azure Synapse. You can take advantage of the big data computation engine to deal with complex compute transformations that would take too long in a data warehouse. For machine learning workloads, you can use SparkML algorithms and AzureML integration for Apache Spark 2.4 with built-in support for Linux Foundation Delta Lake. There is a simple model for provisioning and scaling the Spark clusters to meet your compute needs, regardless of the operations that you are performing on the data.

Data integration to integrate your data with Azure Synapse Pipelines

Azure Synapse Pipelines leverages the capabilities of Azure Data Factory and is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale. Using Azure Synapse Pipelines, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores. You can build complex ETL processes that transform data visually with data flows or by using compute services such as Azure HDInsight Hadoop, or Azure Databricks.

Perform operational analytics with near real-time hybrid transactional and analytical processing with Azure Synapse Link

Azure Synapse Analytics enables you to reach out to operational data using Azure Synapse Link, and is achieved without impacting the performance of the transactional data store. For this to happen, you have to enable the feature within both Azure Synapse Analytics, and within the data store to which Azure Synapse Analytics will connect, such as Azure Cosmos DB. In the case of Azure Cosmos DB, this will create an analytical data store. As data changes in the transactional system, the changed data is fed to the analytical store in a Column store format from which Azure Synapse Link can query with no disruption to the source system.

https://docs.microsoft.com/en-us/azure/synapse-analytics/overview-what-is

Question 12: Skipped
To create and manage Data Factory objects including datasets, linked services, pipelines, triggers, and integration runtimes, the user account that you use to sign into Azure must be a member of which of the role groups? (Select all that apply)

- ☐ CDN Security Profile role
- ☐ Network Manager role
- ☐ Virtual Machine Contributor role
- ☐ Administrator role (Correct)
- ☐ Contributor role (Correct)
- ☐ Custom role with required rights (Correct)
- ☐ Owner role (Correct)
- ☐ DNS Admin Zone role

Explanation
To create Data Factory instances, the user account that you use to sign in to Azure must be a member of the *contributor* or *owner* role, or an *administrator* of the Azure subscription.

To create and manage Data Factory objects including datasets, linked services, pipelines, triggers, and integration runtimes, the following requirements must be met:

• To create and manage child resources in the Azure portal, you must belong to the *Data Factory Contributor* role at the resource group level or above.

• To create and manage resources with PowerShell or the SDK, the *contributor* role at the resource level or above is sufficient.

Data Factory Contributor role

When you are added as a member of this role, you have the following permissions:

• Create, edit, and delete data factories and child resources including datasets, linked services, pipelines, triggers, and integration runtimes.

• Deploy Resource Manager templates. Resource Manager deployment is the deployment method used by Data Factory in the Azure portal.

• Manage App Insights alerts for a data factory.

• At the resource group level or above, lets users deploy Resource Manager template.

• Create support tickets.

If the Data Factory Contributor role does not meet your requirement, you can create your own custom role.

https://docs.microsoft.com/en-us/azure/role-based-access-control/built-in-roles

Question 13: Skipped

What is meant by orchestration? Select the best description.

- ○

  Orchestration enables you to ingest the data from a data source to prepare it for transformation and/or analysis. In addition, Orchestration can fire up compute services on demand.

- ○

  None of the listed options.

- ○

  Orchestration typically contains the transformation logic or the analysis commands of the Azure Data Factory’s work.

- ○

Orchestration helps make your business more efficient by reducing or replacing human interaction with IT systems and instead using software to perform tasks in order to reduce cost, complexity, and errors.

- ○

  Orchestration is the automated configuration, management, and coordination of computer systems, applications, and services.
  (Correct)

Explanation
What is meant by orchestration?

Orchestration is the automated configuration, management, and coordination of computer systems, applications, and services. Orchestration helps IT to more easily manage complex tasks and workflows.

IT teams must manage many servers and applications, but doing so manually isn't a scalable strategy. The more complex an IT system, the more complex managing all the moving parts can become. The need to combine multiple automated tasks and their configurations across groups of systems or machines increases. That's where orchestration can help.

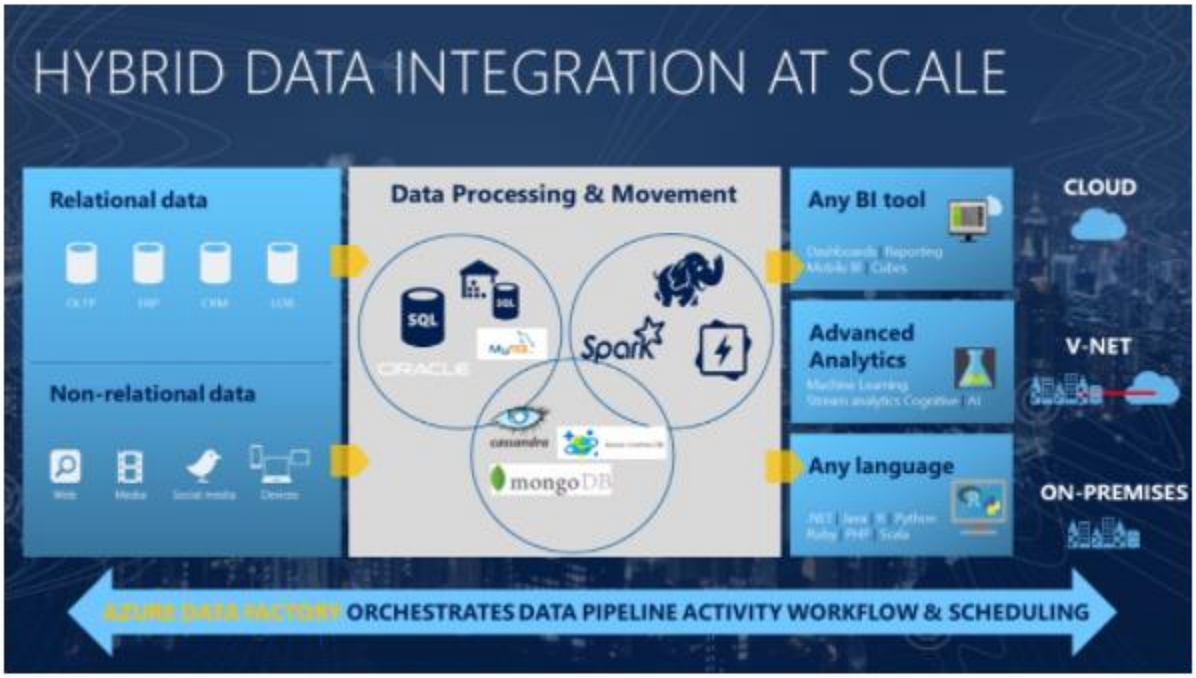
Automation and orchestration are different, but related concepts. Automation helps make your business more efficient by reducing or replacing human interaction with IT systems and instead using software to perform tasks in order to reduce cost, complexity, and errors.

https://www.redhat.com/en/topics/automation/what-is-orchestration

To use an analogy, think about a symphony orchestra. The central member of the orchestra is the conductor. The conductor does not play the instruments, they simply lead the symphony members through the entire piece of music that they perform. The musicians use their own skills to produce particular sounds at various stages of the symphony, so they may only learn certain parts of the music. The conductor orchestrates the entire piece of music, and therefore is aware of the entire score that is being performed. They will also use specific arm movements that provide instructions to the musicians how a piece of music should be played.

ADF can use a similar approach, whilst it has native functionality to ingest and transform data, sometimes it will instruct another service to perform the actual work required on its behalf, such as a Databricks to execute a transformation query. So, in this case, it would be Databricks that performs the work, not ADF. ADF merely orchestrates the execution of the query, and then provides the pipelines to move the data onto the next step or destination.

It also provides rich visualizations to display the lineage and dependencies between your data pipelines, and monitor all your data pipelines from a single unified view to easily pinpoint issues and setup monitoring alerts.



https://cloudblogs.microsoft.com/industry-blog/en-gb/technetuk/2020/08/25/data-orchestration-with-azure-data-factory/

Question 14: Skipped

When planning and implementing your Azure Databricks deployments, you have a number of considerations with respect to compliance. In many industries, it is imperative to maintain compliance through a combination of following best practices in storing and handling data, and by using services that maintain compliance certifications and attestations.

Azure Databricks has which the following compliance certifications?

- ☐ SOC 1 (SSAE 16/SSAE 18)
- ☐ PCI DSS (Correct)
- ☐ HIPAA (Correct)

- □
  AICPA
  (Correct)
- □
  SOC2, Type 2
  (Correct)
- □
  ISAE 3402
- □
  ISO 27018
  (Correct)
- □
  HITRUST
  (Correct)
- □
  SOC2, Type 1
- □
  ISO 27001
  (Correct)

Explanation

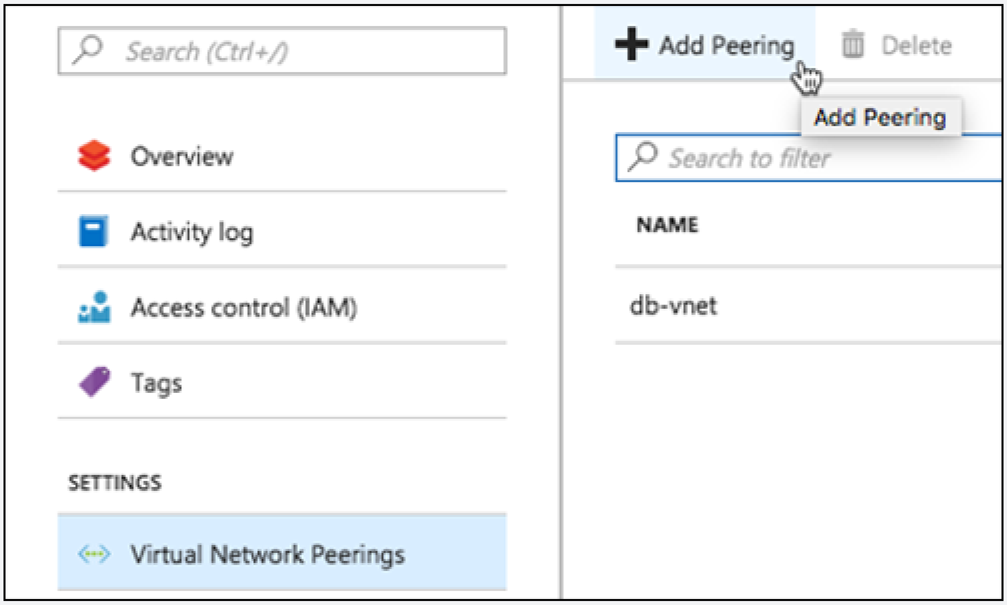
When planning and implementing your Azure Databricks deployments, you have a number of considerations about networking and network security implementation details.

Network security

VNet Peering

Virtual network (VNet) peering allows the virtual network in which your Azure Databricks resource is running to peer with another Azure virtual network. Traffic between virtual machines in the peered virtual networks is routed through the Microsoft backbone infrastructure, much like traffic is routed between virtual machines in the same virtual network, through private IP addresses only.

VNet peering is only required if using the standard deployment without VNet injection.



VNet Injection

If you're looking to do specific network customizations, you could deploy Azure Databricks data plane resources in your own VNet. In this scenario, instead of using the managed VNet, which restricts you from making changes, you "bring your own" VNet where you have full control. Azure Databricks will still create the managed VNet, but it will not use it.

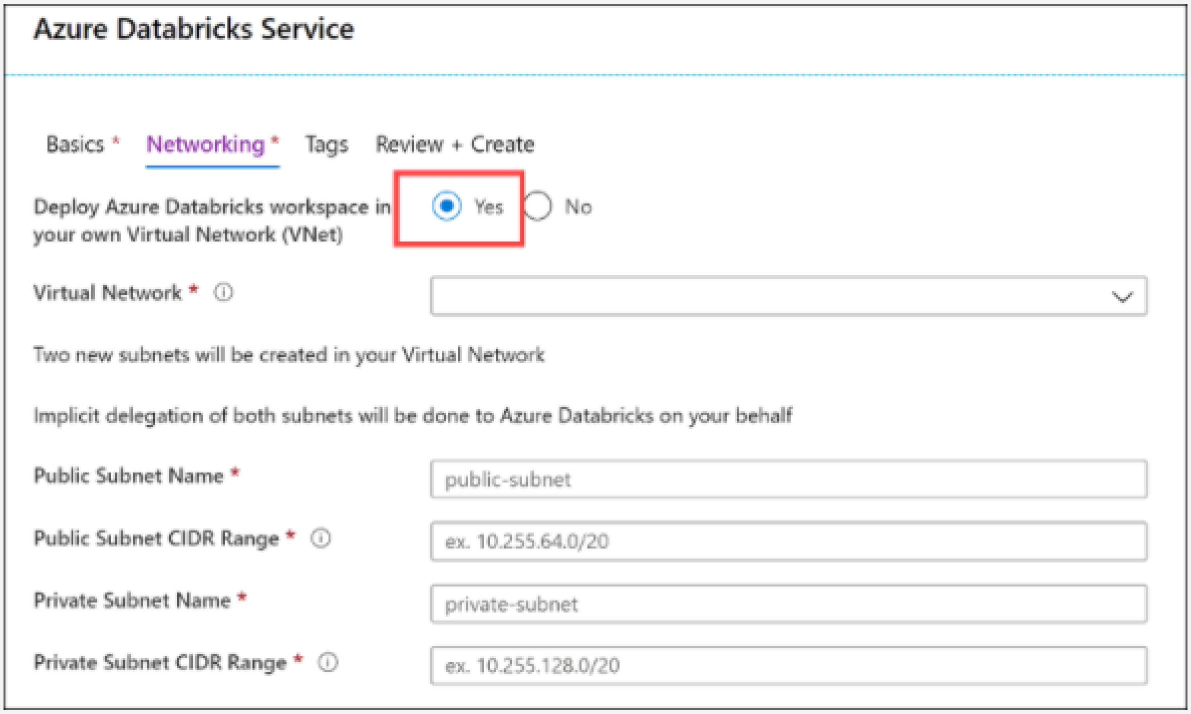
Features enabled through VNet injection include:

- On-Premises Data Access

- Single-IP SNAT and Firewall-based filtering via custom routing

- Service Endpoint

To enable VNet injection, select the **Deploy Azure Databricks workspace in your own Virtual Network** option when provisioning your Azure Databricks workspace.



Azure Databricks Service

Deploy Azure Databricks workspace in your own Virtual Network (VNet) ● Yes ○ No

Virtual Network *

Two new subnets will be created in your Virtual Network

Implicit delegation of both subnets will be done to Azure Databricks on your behalf

Public Subnet Name * public-subnet

Public Subnet CIDR Range * ex. 10.255.64.0/20

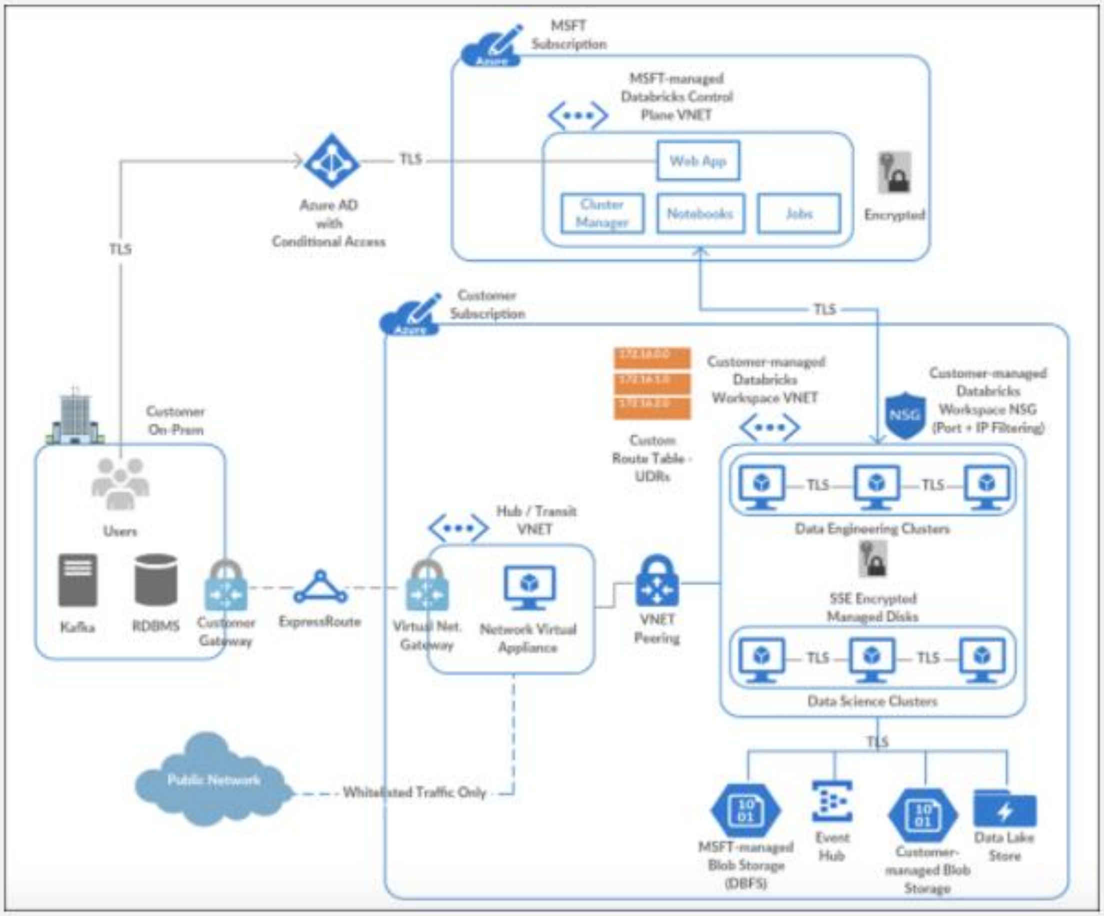
Private Subnet Name * private-subnet

Private Subnet CIDR Range * ex. 10.255.128.0/20

When you compare the deployed Azure Databricks resources in a VNet injection deployment vs. the standard deployment you saw earlier, there are some slight differences. The primary difference is that the clusters in the Data Plane are hosted within a customer-managed Azure Databricks workspace VNet instead of a Microsoft-managed one. The Control Plane is still hosted within a Microsoft-managed VNet, but the TLS connection is still created for you that routes traffic between both VNets. However, the network security groups (NSG) becomes customer-managed as well in this configuration. The only resource in the Data Plane that is still managed by Microsoft is the Blob Storage service that provides DBFS.

Also, inter-node TLS communication between the clusters in the Data Plane is enabled in this deployment. One thing to note is that, while inter-node TLS is more secure, there is a slight impact on performance vs. the non-inter-node TLS found in a basic deployment.

If your Azure Databricks workspace is deployed to your own virtual network (VNet), you can use custom routes, also known as user-defined routes (UDR), to ensure that network traffic is routed correctly for your workspace. For example, if you connect the virtual network to your on-premises network, traffic may be routed through the on-premises network and unable to reach the Azure Databricks control plane. User-defined routes can solve that problem. The diagram below shows UDRs, as well as the other components of a VNet injection deployment.


MSFT Subscription
Azure
MSFT-managed Databricks Control Plane VNET
Web App
Cluster Manager
Notebooks
Jobs
Encrypted
Azure AD with Conditional Access
TLS
TLS
Customer Subscription
Azure
TLS
172.16.0.0
172.16.1.0
172.16.2.0
Custom Route Table - UDRs
Customer-managed Databricks Workspace VNET
NSG
Customer-managed Databricks Workspace NSG (Port + IP Filtering)
Customer On-Prem
Users
Kafka
RDBMS
Customer Gateway
ExpressRoute
Hub / Transit VNET
Virtual Net. Gateway
Network Virtual Appliance
VNET Peering
TLS
TLS
Data Engineering Clusters
SSE Encrypted Managed Disks
TLS
TLS
Data Science Clusters
TLS
Public Network
Whitelisted Traffic Only
MSFT-managed Blob Storage (DBFS)
Event Hub
Customer-managed Blob Storage
Data Lake Store

You can create different Azure Databricks workspaces in the same VNet. However, you will need separate pairs of dedicated subnets per Azure Databricks workspace. As such, the VNet network range has to be fairly large to accommodate those. The VNet CIDR can be anywhere between /16 and /24, and the subnet CIDR can be anywhere between /18 and /26.

Secure connectivity to other Azure data services

Your Azure Databricks deployment likely includes other Azure data services, such as Azure Blob Storage, Azure Data Lake Storage Gen2, Azure Cosmos DB, and Azure Synapse Analytics. We recommend ensuring traffic between Azure Databricks and Azure data services such as these remains on the Azure network backbone, instead of traversing over the public internet. To do this, you should use Azure Private Link or Service Endpoints.

Azure Private Link

Using Azure Private Link is currently the most secure way to access Azure data services from Azure Databricks. Private Link enables you to access Azure PaaS Services (for example, Azure Storage, Azure Cosmos DB, and SQL Database) and Azure hosted customer/partner services over a Private Endpoint in your virtual network. Traffic between your virtual network and the service traverses over the Microsoft network backbone, eliminating exposure from the public Internet. You can also create your own Private Link Service in your virtual network (VNet) and deliver it privately to your customers.

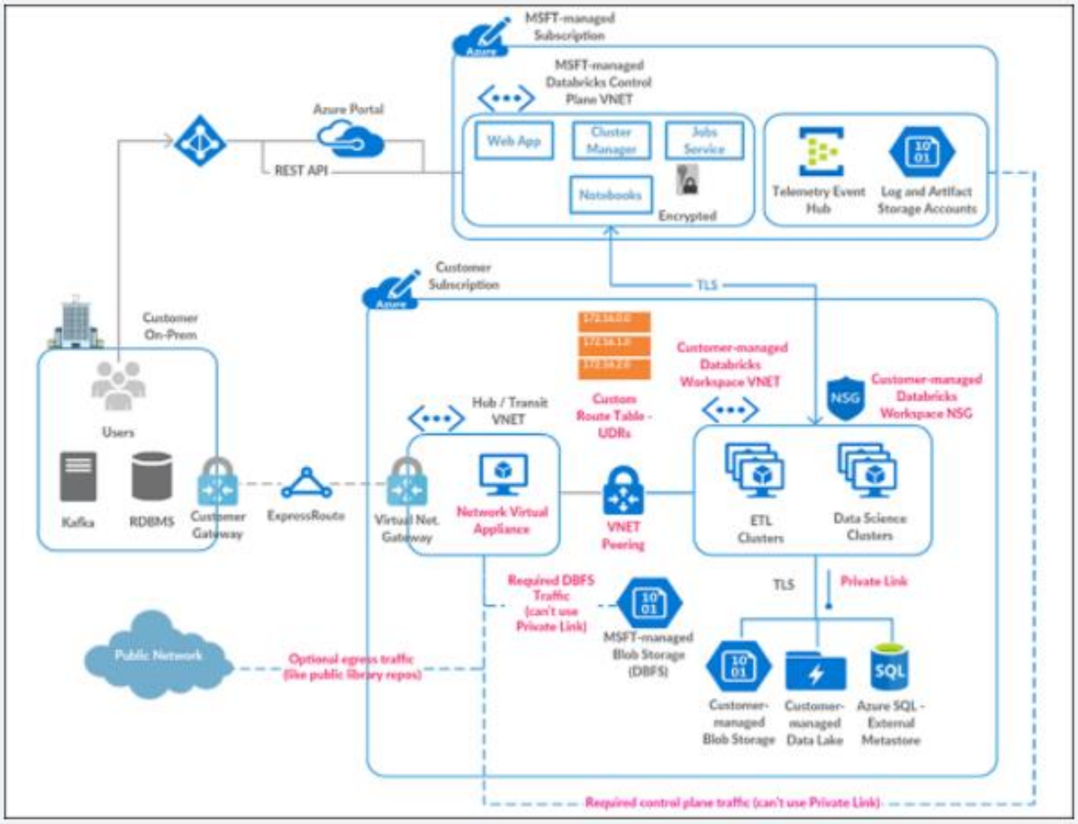
Azure VNet service endpoints

Virtual Network (VNet) service endpoints extend your virtual network private address space. The endpoints also extend the identity of your VNet to the Azure services over a direct connection. Endpoints allow you to secure your critical Azure service resources to only your virtual networks. Traffic from your VNet to the Azure service always remains on the Microsoft Azure network backbone.

Read more about securely accessing Azure data sources from Azure Databricks.

Combining VNet injection and Private Link

The following diagram shows how you may use Private Link in combination with VNet injection in a hub and spoke topology to prevent data exfiltration:


MSFT-managed Subscription
Azure
MSFT-managed Databricks Control Plane VNET
Web App
Cluster Manager
Jobs Service
Notebooks
Encrypted
Telemetry Event Hub
Log and Artifact Storage Accounts
Azure Portal
REST API
Customer Subscription
Azure
TLS
Customer On-Prem
Users
Kafka
RDBMS
Customer Gateway
ExpressRoute
Virtual Net. Gateway
Hub / Transit VNET
Network Virtual Appliance
172.16.0.0
172.16.1.0
172.16.2.0
Custom Route Table - UDRs
VNET Peering
Customer-managed Databricks Workspace VNET
NSG
Customer-managed Databricks Workspace NSG
ETL Clusters
Data Science Clusters
TLS
Private Link
Required DBFS Traffic (can't use Private Link)
MSFT-managed Blob Storage (DBFS)
Public Network
Optional egress traffic (like public library repos)
Customer-managed Blob Storage
Customer-managed Data Lake
SQL
Azure SQL - External Metastore
Required control plane traffic (can't use Private Link)

Compliance

In many industries, it is imperative to maintain compliance through a combination of following best practices in storing and handling data, and by using services that maintain compliance certifications and attestations.

Azure Databricks has the following compliance certifications:

- HITRUST
- AICPA
- PCI DSS
- ISO 27001
- ISO 27018
- HIPAA (Covered by MSFT Business Associates Agreement (BAA))

• SOC2, Type 2

Audit logs

Databricks provides comprehensive end-to-end audit logs of activities performed by Databricks users, allowing your enterprise to monitor detailed Databricks usage patterns. Azure Monitor integration enables you to capture the audit logs and make then centrally available and fully searchable.

Services / Entities included are:

• Accounts

• Clusters

• DBFS

• Genie

• Jobs

• ACLs

• SSH

• Tables

https://docs.microsoft.com/en-us/azure/security/fundamentals/network-overview

Question 15: Skipped
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

[?] is a cloud-integration service which orchestrates the movement of data between various data stores. [?] processes and transforms data by using compute services such as Azure HDInsight, Hadoop, Spark, and Azure Machine Learning. Publish output data to data stores such as Azure SQL Data Warehouse so that business intelligence applications can consume the data.

- ○
  Azure Storage Explorer
- ○
  Azure Databricks

- ○
  Azure Data Lake Storage
- ○
  Azure Data Factory
  (Correct)
- ○
  Azure Data Catalog
- ○
  Azure SQL Datawarehouse
- ○
  Azure Cosmos DB

Explanation
Azure Data Factory

Data Factory is a cloud-integration service. It orchestrates the movement of data between various data stores.

As a data engineer, you can create data-driven workflows in the cloud to orchestrate and automate data movement and data transformation. Use Data Factory to create and schedule data-driven workflows (called pipelines) that can ingest data from data stores.

Data Factory processes and transforms data by using compute services such as Azure HDInsight, Hadoop, Spark, and Azure Machine Learning. Publish output data to data stores such as Azure SQL Data Warehouse so that business intelligence applications can consume the data. Ultimately, you use Data Factory to organize raw data into meaningful data stores and data lakes so your organization can make better business decisions.

https://docs.microsoft.com/en-us/azure/data-factory/introduction

Question 16: Skipped
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Azure Data Factory provides a variety of methods for ingesting data, and also provides a range of methods to perform transformations.

These methods are:

• Mapping Data Flows

• Compute Resources

• SSIS Packages

Mapping Data Flows provides a number of different transformations types that enable you to modify data. They are broken down into the following categories:

• Schema modifier transformations

• Row modifier transformations

• Multiple inputs/outputs transformations

Some of the transformations that you can define have a(n) [?] that will enable you to customize the functionality of a transformation using columns, fields, variables, parameters, functions from your data flow in these boxes. To build the expression, use the [?], which is launched by clicking in the expression text box inside the transformation. You'll also sometimes see "Computed Column" options when selecting columns for transformation.

- ○ Data Stream Expression Builder
- ○ Data Expression Script Builder
- ○ Data Expression Orchestrator
- ○ Mapping Data Flow
- ○ Data Flow Expression Builder (Correct)
- ○ Wrangling Data Flow

Explanation
Just as Azure Data Factory provides a variety of methods for ingesting data, it also provides a range of methods to perform transformations. You can pick a method that matches the skillsets of your team or takes advantage of existing technologies that you already have in your data estate. There is also the opportunity to perform transformations without writing code at all using the Mapping Data Flow.

Transforming data using Mapping Data Flow

Mapping Data Flows provide an environment for building a wide range of data transformations visually without the need to use code. The resulting data flows that are created are subsequently executed on scaled-out Apache Spark clusters that are automatically provisioned when you execute the Mapping Data Flow. Mapping Data Flows also provides the capability to monitor the execution of the transformations so that you can view how the transformations are progressing, or to understand any errors that may occur

Mapping Data Flows provides a number of different transformations types that enable you to modify data. They are broken down into the following categories:

- Schema modifier transformations

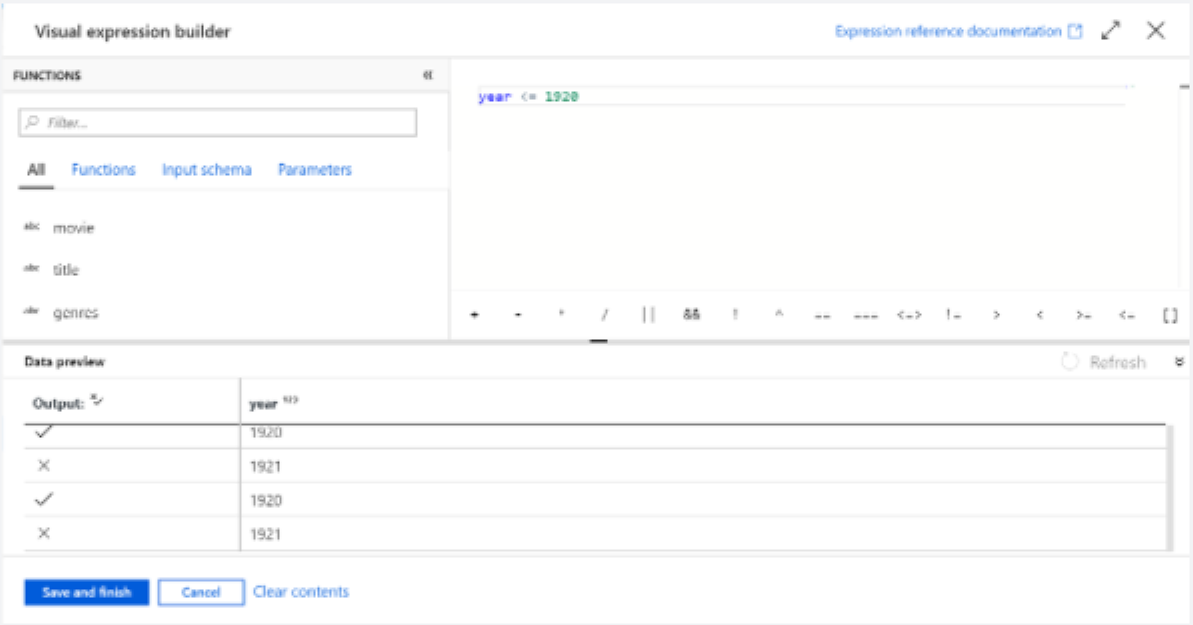
- Row modifier transformations

- Multiple inputs/outputs transformations

Data Flow Expression Builder

Some of the transformations that you can define have a **Data Flow Expression Builder** that will enable you to customize the functionality of a transformation using columns, fields, variables, parameters, functions from your data flow in these boxes.

To build the expression, use the Expression Builder, which is launched by clicking in the expression text box inside the transformation. You'll also sometimes see "Computed Column" options when selecting columns for transformation. When you click that, you'll also see the Expression Builder launched.



```
year <= 1920
```

Output:	year
✓	1920
×	1921
✓	1920
×	1921

The Expression Builder tool defaults to the text editor option. the auto-complete feature reads from the entire Azure Data Factory Data Flow object model with syntax checking and highlighting.

https://docs.microsoft.com/en-us/azure/data-factory/transform-data

Question 17: Skipped

Which language can be used to define Spark job definitions?

- ○ PowerShell
- ○ Transact-SQL
- ○ C#
- ○ Java
- ○ PySpark (Correct)

Explanation

Pyspark can be used to define spark job definitions.

https://intellipaat.com/blog/tutorial/spark-tutorial/pyspark-tutorial/

Question 18: Skipped
Which component enables you to perform code free transformations in Azure Synapse Analytics?

- ○ Flow capabilities
- ○ Studio
- ○ Mapping data flow (Correct)
- ○ Copy activity
- ○ Monitoring capabilities
- ○ Control capabilities

Explanation
You can natively perform data transformations with Azure Data Factory code free using the Mapping Data Flow task.

https://docs.microsoft.com/en-us/azure/data-factory/tutorial-data-flow

Question 19: Skipped
Whilst Azure Synapse Analytics is used for the storage of data for analytical purposes, SQL Pools do support the use of transactions and adhere to the ACID (Atomicity, Consistency, Isolation, and Durability) transaction principles associated with relational database management systems.

As such, locking, and blocking mechanisms are put in place to maintain transactional integrity while providing adequate workload concurrency. These blocking aspects may significantly delay the completion of queries.

To improve the response time, turn [?] the `READ_COMMITTED_SNAPSHOT` database option for a user database when connected to the master database.

- ○

OFF

- ○ ON (Correct)
- ○ None of the listed options.
- ○ `READ_COMMITTED_SNAPSHOT` is not the correct setting to adjust.

Explanation
Whilst Azure Synapse Analytics is used for the storage of data for analytical purposes, SQL Pools do support the use of transactions and adhere to the ACID (Atomicity, Consistency, Isolation, and Durability) transaction principles associated with relational database management systems.

As such, locking, and blocking mechanisms are put in place to maintain transactional integrity while providing adequate workload concurrency. These blocking aspects may significantly delay the completion of queries. The isolation level of the transactional support is defaulted to READ UNCOMMITTED. You can change it to READ COMMITTED SNAPSHOT ISOLATION by **turning ON the `READ_COMMITTED_SNAPSHOT` database option for a user database when connected to the master database.**

Once enabled, all transactions in this database are executed under READ COMMITTED SNAPSHOT ISOLATION and setting READ UNCOMMITTED on session level will not be honoured.

If you experience delays in the completion of queries, the Read Committed Snapshot Isolation level should be employed to alleviate this. Read Committed Snapshot, makes a copy of the rows that are being referenced in a query if it is being updated, so that the data is consistent. The version of the data being used remains only for the duration of the query and any dependant queries, which are faster for query completion at the expense of space needed to storer multiple versions of the data during workloads.

To enable `READ COMMITTED SNAPSHOT ISOLATION`, run this command when connecting to the `MASTER` database.

```
SQL

ALTER DATABASE MyDatabase

SET READ_COMMITTED_SNAPSHOT ON
```

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-develop-transactions

Question 20: Skipped

Scenario: Big Belly Foods, Inc. (BB) owns and operates 300 convenience stores across LatAm. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas. The company has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

BB employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks. You have been hired as an Azure Expert SME and you are to consult the IT team on various Azure related projects.

Business Requirements:

BB 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:

BB 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 BB.

• 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. BB 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:

BB 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.

- BB 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.

- BB 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.

- BB does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

The Ask:

The team looks to you for direction on what should be used to prevent users outside the BB on-premises network from accessing the analytical data store. Which of the following should you recommend?

- ○
  A server-level virtual network rule
- ○
  A database-level firewall IP rule
- ○
  A database-level virtual network rule
- ○
  A server-level firewall IP rule
  (Correct)

Explanation

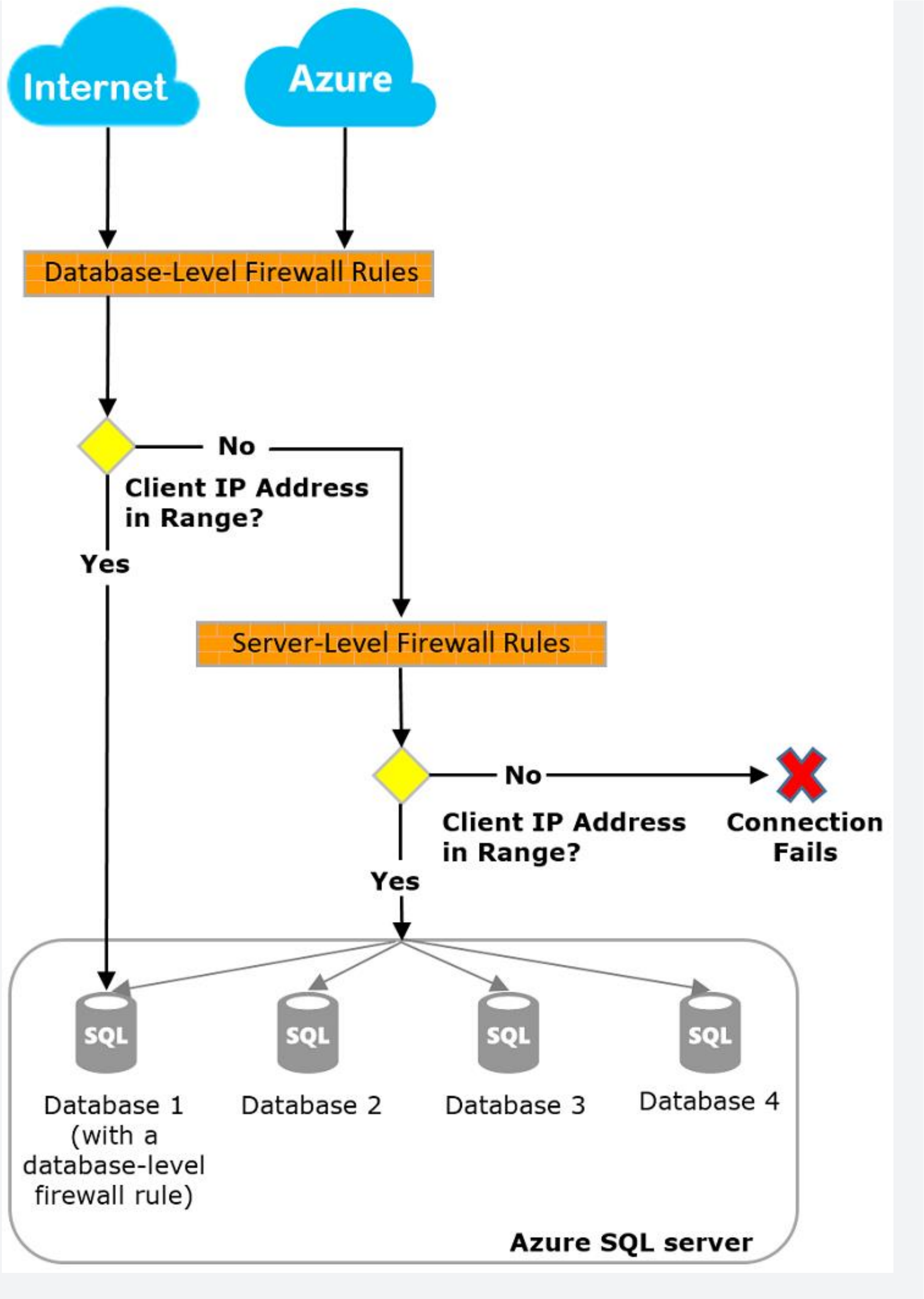
To ensure that the analytical data store is accessible only to the company;s on-premises network and Azure services with the restriction of not using a VPN, a server-level firewall IP rule should be employed. A server-level virtual network rule would be the correct answer if the VPN restriction was not indicated.

Azure SQL Database and Azure Synapse IP firewall rules

When you create a new server in Azure SQL Database or Azure Synapse Analytics named *mysqlserver*, for example, a server-level firewall blocks all access to the public endpoint for the server (which is accessible at *mysqlserver.database.windows.net*). For simplicity, *SQL Database* is used to refer to both SQL Database and Azure Synapse Analytics.

How the firewall works

Connection attempts from the internet and Azure must pass through the firewall before they reach your server or database, as the following diagram shows.


Internet
Azure
Database-Level Firewall Rules
No
Client IP Address in Range?
Yes
Server-Level Firewall Rules
No
Client IP Address in Range?
Connection Fails
Yes
SQL
SQL
SQL
SQL
Database 1 (with a database-level firewall rule)
Database 2
Database 3
Database 4
Azure SQL server

Server-level IP firewall rules

These rules enable clients to access your entire server, that is, all the databases managed by the server. The rules are stored in the *master* database. You can have a maximum of 128 server-level IP firewall rules for a server. If you have the **Allow Azure Services and resources to access this server** setting enabled, this counts as a single firewall rule for the server.

You can configure server-level IP firewall rules by using the Azure portal, PowerShell, or Transact-SQL statements.

To use the portal or PowerShell, you must be the subscription owner or a subscription contributor.

To use Transact-SQL, you must connect to the *master* database as the server-level principal login or as the Azure Active Directory administrator. (A server-level IP firewall rule must first be created by a user who has Azure-level permissions.)

*Note: By default, during creation of a new logical SQL server from the Azure portal, the **Allow Azure Services and resources to access this server** setting is set to **No**.*

Database-level IP firewall rules

Database-level IP firewall rules enable clients to access certain (secure) databases. You create the rules for each database (including the *master* database), and they're stored in the individual database.

You can only create and manage database-level IP firewall rules for master and user databases by using Transact-SQL statements and only after you configure the first server-level firewall.

If you specify an IP address range in the database-level IP firewall rule that's outside the range in the server-level IP firewall rule, only those clients that have IP addresses in the database-level range can access the database.

You can have a maximum of 128 database-level IP firewall rules for a database. For more information about configuring database-level IP firewall rules, see the example later in this article and see sp_set_database_firewall_rule (Azure SQL Database).

Recommendations for how to set firewall rules

MS recommends that you use database-level IP firewall rules whenever possible. This practice enhances security and makes your database more portable. **Use server-level IP firewall rules for administrators. Also use them when you have many databases that**

have the same access requirements, and you don't want to configure each database individually.

https://docs.microsoft.com/en-us/azure/azure-sql/database/firewall-configure

Question 21: Skipped

When data is loaded into Synapse Analytics dedicated SQL pools, the datasets are broken up and dispersed among the compute nodes for processing, and then written to a decoupled and scalable storage layer. This action is termed *"sharding"*.

The design decisions around how to split and disperse this data among the nodes and then to the storage is important to querying workloads, as the correct selection minimizes data movement that is a primary cause of performance issues in an Azure Synapse dedicated SQL Pool environment.

Which of the following are valid table distribution types available in Synapse Analytics SQL Pools?

- □ Replicated tables (Correct)
- □ Hash distribution (Correct)
- □ Round robin distribution (Correct)
- □ Centralized table distribution
- □ Distributed table schema
- □ Merkle table distribution

Explanation

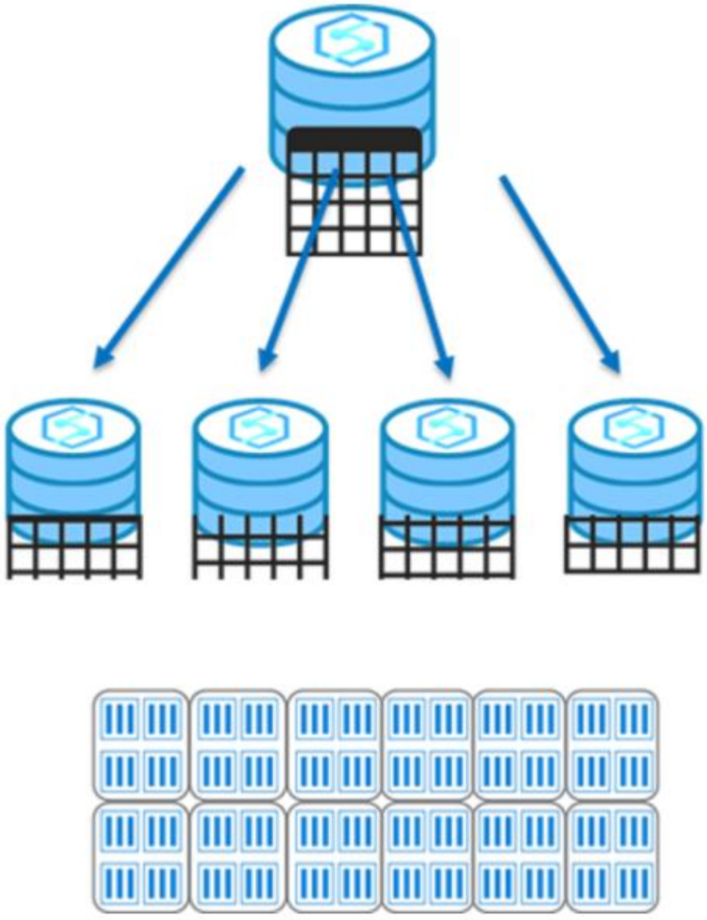
When data is loaded into Synapse Analytics dedicated SQL pools, the datasets are broken up and dispersed among the compute nodes for processing, and then written to a decoupled and scalable storage layer. This action is termed *"sharding"*.

The design decisions around how to split and disperse this data among the nodes and then to the storage is important to querying workloads, as the correct selection minimizes data movement that is a primary cause of performance issues in an Azure Synapse dedicated SQL Pool environment.

There are three main table distributions available in Synapse Analytics SQL Pools.

Selecting the correct table distribution can have an impact on the data load and query performance as follows:

Round robin distribution



This is the default distribution created for a table and delivers fast performance when used for loading data.

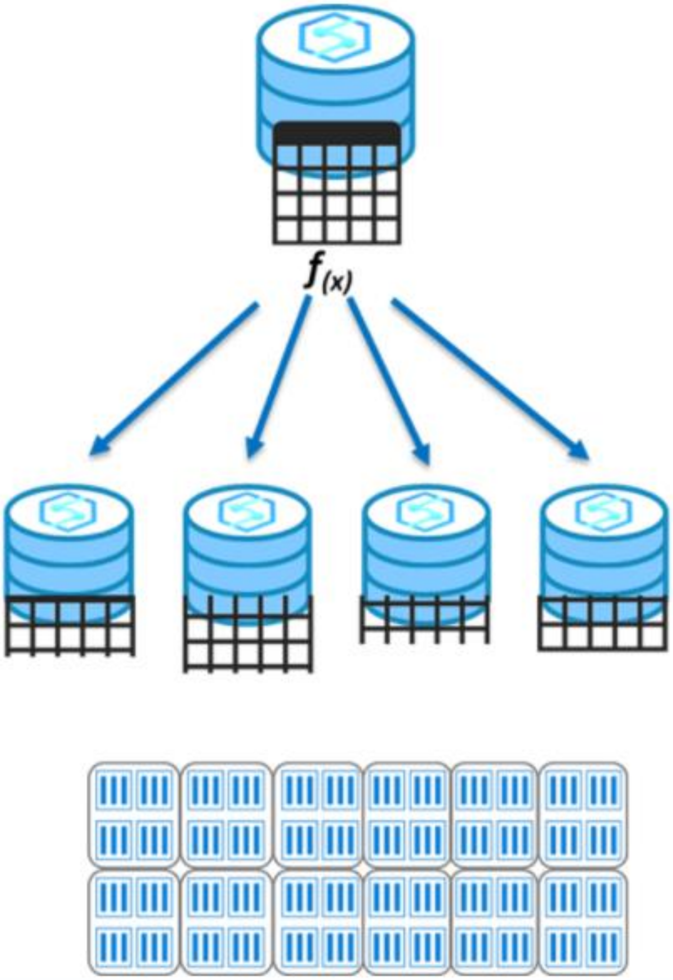
A round-robin distributed table distributes data evenly across the table but without any further optimization. A distribution is first chosen at random and then buffers of rows are assigned to distributions sequentially.

It is quick to load data into a round-robin table, but query performance can often be better with hash distributed tables for larger datasets.

Joins on round-robin tables may negatively affect query workloads, as data that is gathered for processing then has to be reshuffled to other compute nodes, which take additional time and processing.

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

Hash distribution



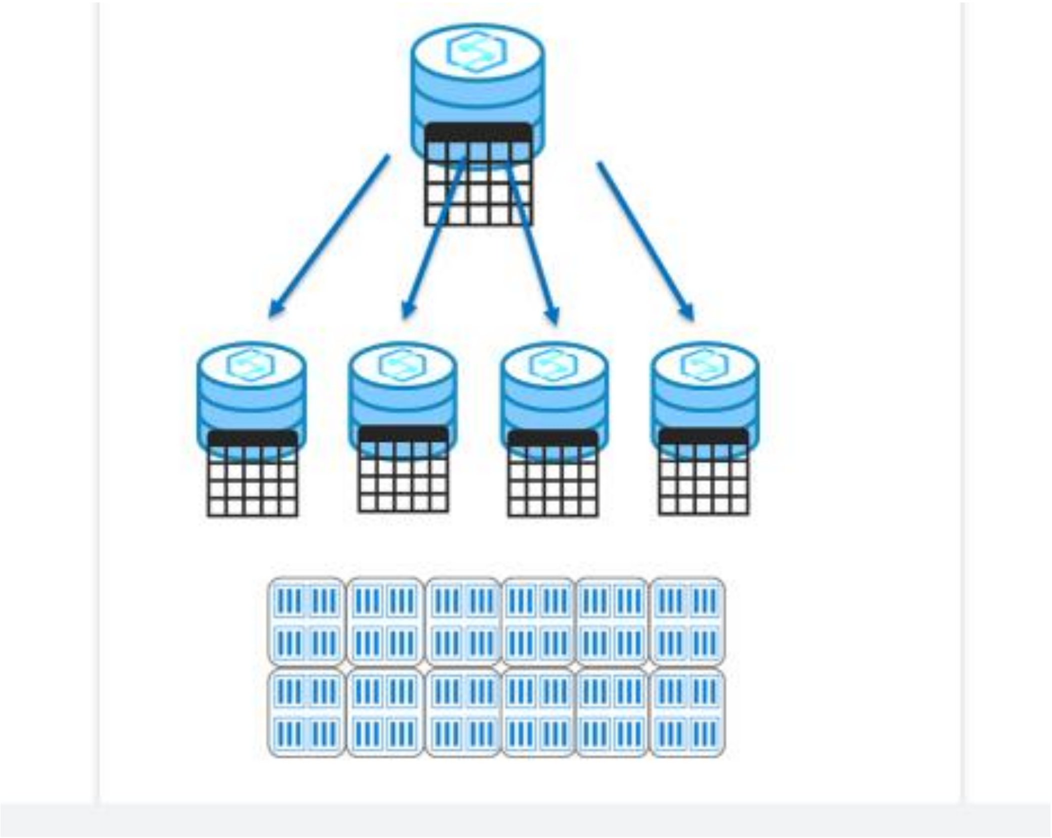
This distribution can deliver the highest query performance for joins and aggregations on large tables.

To shard data, a hash function is used to deterministically assign each row to a distribution. In the table definition, one of the columns is designated as the distribution column.

There are performance considerations for the selection of a distribution column, such as distinctness, data skew, and the types of queries that run on the system.

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

Replicated tables



A replicated table provides the fastest query performance for small tables.

A table that is replicated caches a full copy of the table on each compute node. Consequently, replicating a table removes the need to transfer data among compute nodes before a join or aggregation. As such extra storage is required and there is additional overhead that is incurred when writing data, which make large tables impractical.

Frequent data modifications will cause the cached copy to be invalidated, and require the table be recached.

Scaling the SQL Pool will also require the table be recached.

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

Question 22: Skipped

Consider: Continuous Integration/Continuous Delivery lifecycle

Which feature commits the changes of Azure Data Factory work in a custom branch created with the main branch in a Git repository?

- ○ Commit
- ○ DDL commands
- ○ TCL commands
- ○ Pull request (Correct)
- ○ Repo
- ○ DML commands

Explanation

Continuous Integration/Continuous Delivery lifecycle

Below is a sample overview of the CI/CD lifecycle in an Azure data factory that's configured with Azure Repos Git.

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 master or collaboration branch to get their changes reviewed by peers.**

4. After a pull request is approved and changes are merged in the master branch, the changes get published to the development factory.

5. When the team is ready to deploy the changes to a test or UAT (User Acceptance Testing) factory, the team goes to their Azure Pipelines release and deploys the desired version of the development factory to UAT. This deployment takes place as part of an Azure Pipelines task and uses Resource Manager template parameters to apply the appropriate configuration.

6. After the changes have been verified in the test factory, deploy to the production factory by using the next task of the pipelines release.

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

Question 23: Skipped
When creating a typical project, when would you create your storage account(s)?

- ○ At the beginning, during project setup. (Correct)
- ○ At any stage of the project, as long as it is before you need to analyze data.
- ○ At the end, during resource cleanup.
- ○ After deployment, when the project is running.

Explanation
Storage accounts are stable for the lifetime of a project. It's common to create them at the start of a project.

https://docs.microsoft.com/en-us/azure/storage/common/storage-account-create?tabs=azure-portal

Question 24: Skipped
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Azure Databricks is a fully-managed, cloud-based Big Data and Machine Learning platform, which empowers developers to accelerate AI and innovation by simplifying the process of building enterprise-grade production data applications. Built as a joint effort by Databricks and Microsoft, Azure Databricks provides data science and engineering teams with a single platform for Big Data processing and Machine Learning.

By combining the power of Databricks, an end-to-end, managed Apache Spark platform optimized for the cloud, with the enterprise scale and security of Microsoft's Azure platform, Azure Databricks makes it simple to run large-scale Spark workloads.

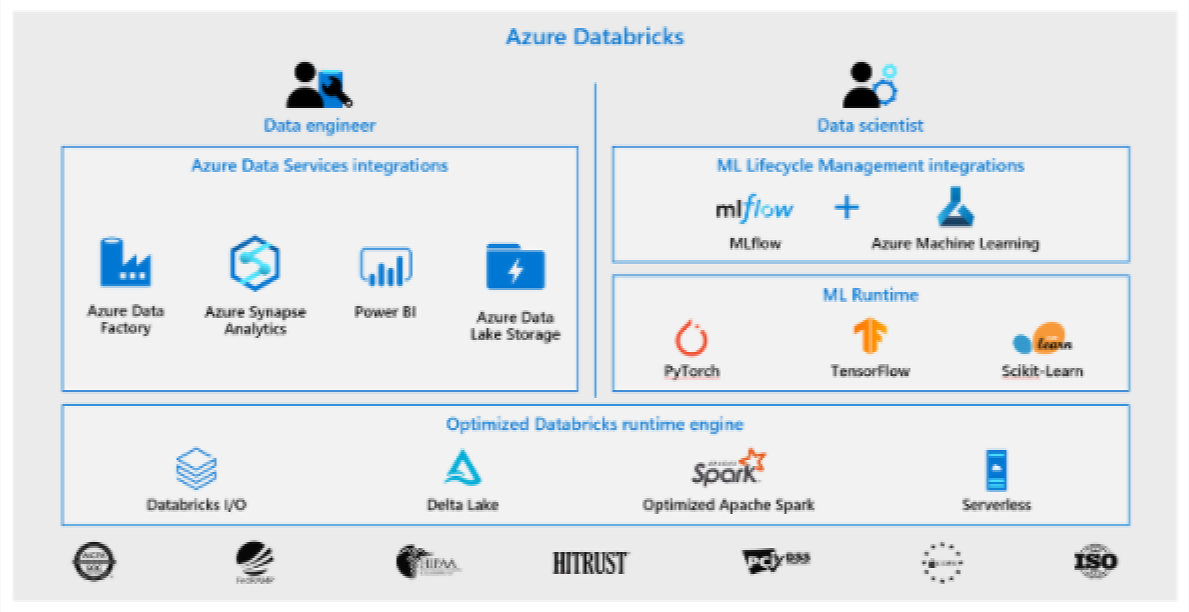
Internally, [?] is used to run the Azure Databricks control-plane and data-planes via containers running on the latest generation of Azure hardware (Dv3 VMs), with NvMe SSDs capable of blazing 100us latency on IO.

- ○ Azure Database Services
- ○ Azure VNet Peering
- ○ Azure Machine Learning Studio
- ○ Azure Kubernetes Service **(Correct)**

Explanation
Azure Databricks is a fully-managed, cloud-based Big Data and Machine Learning platform, which empowers developers to accelerate AI and innovation by simplifying the process of building enterprise-grade production data applications. Built as a joint effort by Databricks and Microsoft, Azure Databricks provides data science and engineering teams with a single platform for Big Data processing and Machine Learning.

By combining the power of Databricks, an end-to-end, managed Apache Spark platform optimized for the cloud, with the enterprise scale and security of Microsoft's Azure platform, Azure Databricks makes it simple to run large-scale Spark workloads.

Conceptual view of Azure Databricks



To provide the best platform for data engineers, data scientists, and business users, Azure Databricks is natively integrated with Microsoft Azure, providing a "first party" Microsoft service. The Azure Databricks collaborative workspace enables these teams to work together through features such as user management, git source code repository integration, and user workspace folders.

Microsoft is working to integrate Azure Databricks closely with all features of the Azure platform. Below is a list of some of the integrations completed so far:

• **VM types**: Many existing VMs can be used for clusters, including F-series for machine learning scenarios, M-series for massive memory scenarios, and D-series for general purpose.

• **Security and Privacy**: Ownership and control of data is with the customer, and Microsoft aims for Azure Databricks to adhere to all the compliance certifications that the rest of Azure provides.

• **Flexibility in network topology**: Azure Databricks supports deployments into virtual networks (VNETs), which can control which sources and sinks can be accessed and how they are accessed.

• **Orchestration**: ETL/ELT workflows (including analytics workloads in Azure Databricks) can be operationalized using Azure Data Factory pipelines.

• **Power BI**: Power BI can be connected directly to Databricks clusters using JDBC in order to query data interactively at massive scale using familiar tools.

• **Azure Active Directory**: Azure Databricks workspaces deploy into customer subscriptions, so naturally AAD can be used to control access to sources, results, and jobs.

• **Data stores**: Azure Storage and Data Lake Store services are exposed to Databricks users via Databricks File System (DBFS) to provide caching and optimized analysis over existing data. Azure Databricks easily and efficiently uploads results into Azure Synapse Analytics, Azure SQL Database, and Azure Cosmos DB for further analysis and real-time serving, making it simple to build end-to-end data architectures on Azure.

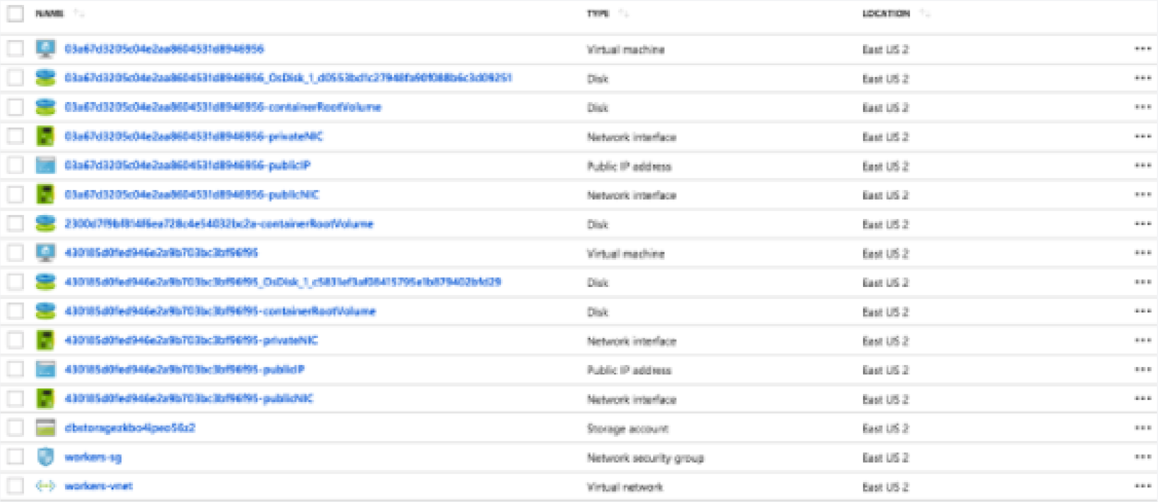
• **Real-time analytics**: Integration with IoT Hub, Azure Event Hubs, and Azure HDInsight Kafka clusters enables developers to build scalable streaming solutions for real-time analytics.

For developers, this design provides three things. First, it enables easy connection to any storage resources in their account, such as an existing Blob storage or Data Lake Store. Second, they are able to take advantage of deep integrations with other Azure services to quickly build data applications. Third, Databricks is managed centrally from the Azure control centre, requiring no additional setup, which allows developers to focus on core business value, not infrastructure management.

Azure Databricks platform architecture

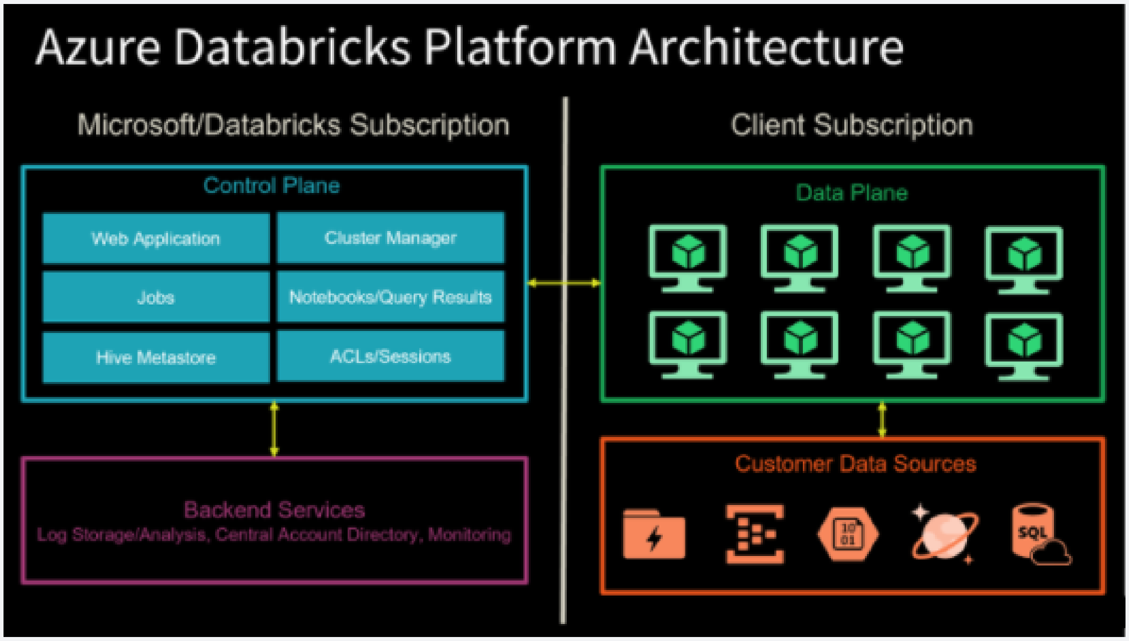
When you create an Azure Databricks service, a "Databricks appliance" is deployed as an Azure resource in your subscription. At the time of cluster creation, you specify the types and sizes of the virtual machines (VMs) to use for both the Driver and Worker nodes, but Azure Databricks manages all other aspects of the cluster.

The "Databricks appliance" is deployed into Azure as a managed resource group within your subscription. This resource group contains the Driver and Worker VMs, along with other required resources, including a virtual network, a security group, and a storage account. All metadata for your cluster, such as scheduled jobs, is stored in an Azure Database with geo-replication for fault tolerance.



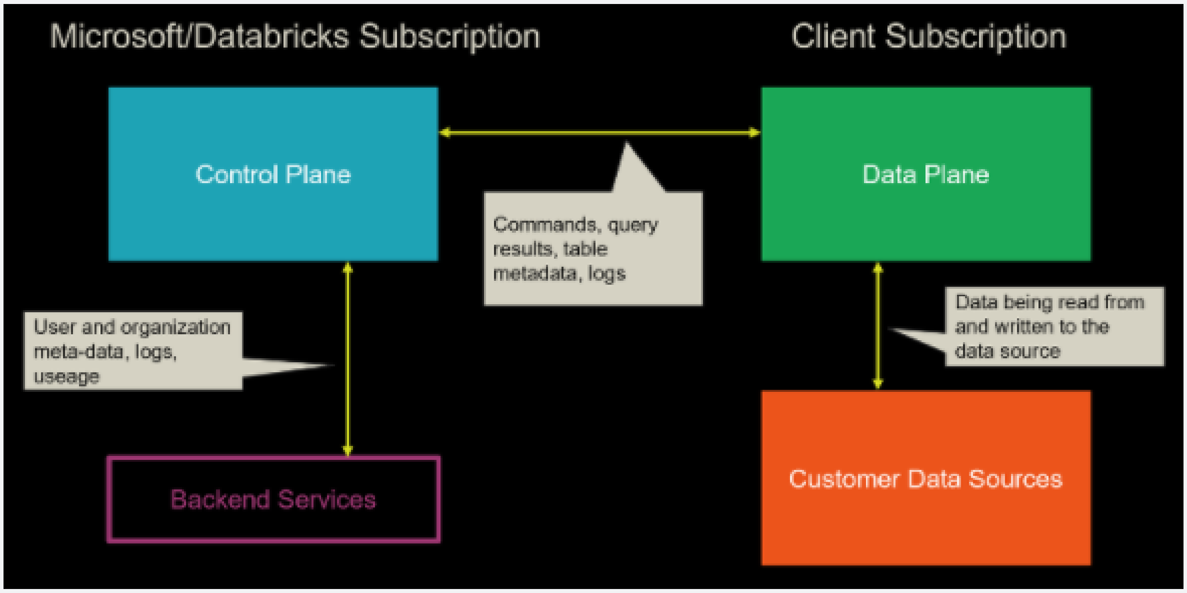
□	NAME	TYPE	LOCATION	
□	03a67d3205c04e2aa8604531d8946956	Virtual machine	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956_OsDisk_1_d0553bd1c27948fa90f088b6c3d08251	Disk	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-containerRootVolume	Disk	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-privateNIC	Network interface	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-publicIP	Public IP address	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-publicNIC	Network interface	East US 2	•••
□	2300d7f9b8f14f6ea728c4e54032bc2a-containerRootVolume	Disk	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95	Virtual machine	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95_OsDisk_1_c5831ef3af08415795e1b8794026fd29	Disk	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95-containerRootVolume	Disk	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95-privateNIC	Network interface	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95-publicIP	Public IP address	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95-publicNIC	Network interface	East US 2	•••
□	dbstoragezkbo4lpeo58z2	Storage account	East US 2	•••
□	workers-sg	Network security group	East US 2	•••
□	workers-vnet	Virtual network	East US 2	•••

Internally, Azure Kubernetes Service (AKS) is used to run the Azure Databricks control-plane and data-planes via containers running on the latest generation of Azure hardware (Dv3 VMs), with NvMe SSDs capable of blazing 100us latency on IO. These make Databricks I/O performance even better. In addition, accelerated networking provides the fastest virtualized network infrastructure in the cloud. Azure Databricks utilizes this to further improve Spark performance.

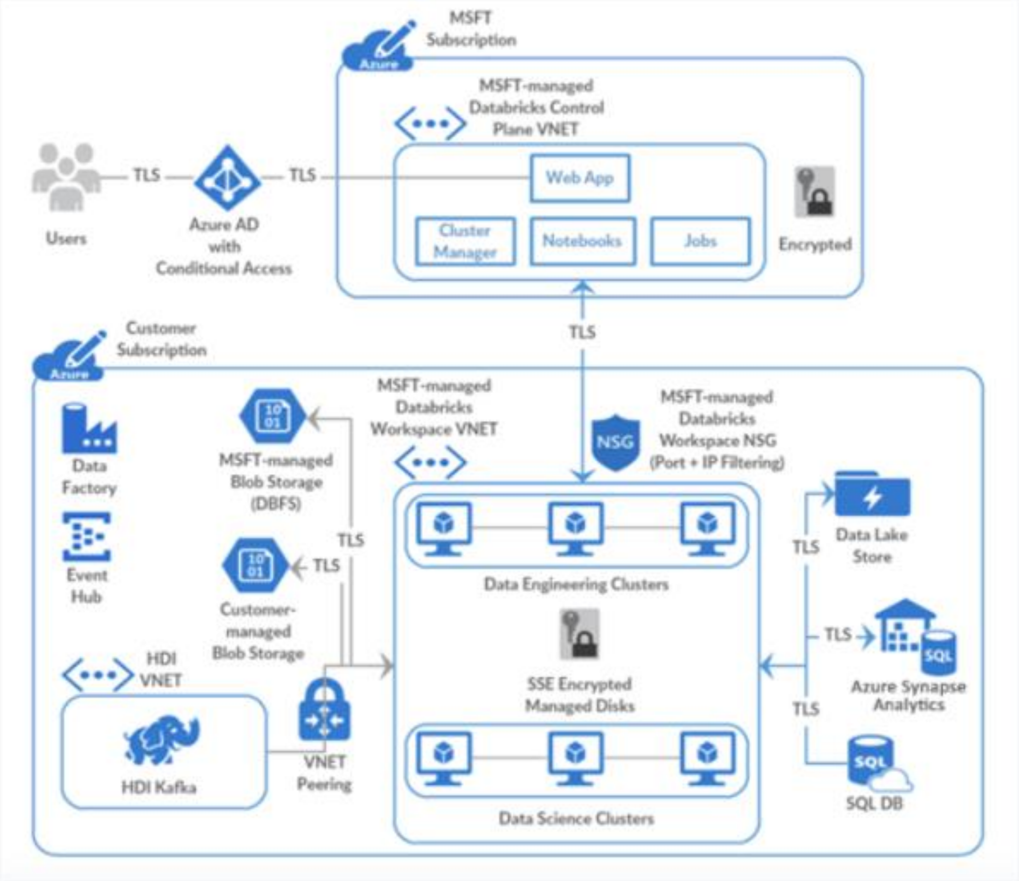


The diagram above shows a Control Plane on the left, which hosts Databricks jobs, notebooks with query results, the cluster manager, web application, Hive metastore, and security access control lists (ACLs) and user sessions. These components are managed by Microsoft in collaboration with Databricks and do not reside within your Azure subscription.

On the right-hand side is the Data Plane, which contains all the Databricks runtime clusters hosted within the workspace. All data processing and storage exists within the client subscription. This means no data processing ever takes place within the Microsoft/Databricks-managed subscription.


Microsoft/Databricks Subscription
Client Subscription
Control Plane
Data Plane
Commands, query results, table metadata, logs
User and organization meta-data, logs, useage
Data being read from and written to the data source
Backend Services
Customer Data Sources

Moving one level deeper, the diagram above shows what is being exchanged between the Azure Databricks platform components. Since the web app and cluster manager is part of the Control Plane, any commands executed in a notebook are sent from the cluster manager to the customer's clusters in the Data Plane. This is because the data processing only occurs within the customer's own subscription, as stated earlier. Any table metadata and logs are exchanged between these two high-level components. Customer data sources within the client subscription exchange data with the Data Plane through read and write activities.


MSFT Subscription
Azure
MSFT-managed Databricks Control Plane VNET
Users
TLS
Azure AD with Conditional Access
TLS
Web App
Cluster Manager
Notebooks
Jobs
Encrypted
TLS
Customer Subscription
Azure
MSFT-managed Databricks Workspace VNET
MSFT-managed Databricks Workspace NSG (Port + IP Filtering)
NSG
Data Factory
MSFT-managed Blob Storage (DBFS)
Event Hub
TLS
TLS
Customer-managed Blob Storage
Data Engineering Clusters
SSE Encrypted Managed Disks
Data Science Clusters
TLS
Data Lake Store
TLS
Azure Synapse Analytics
TLS
SQL DB
HDI VNET
HDI Kafka
VNET Peering

The diagram above shows a standard deployment that contains the boundaries between the Control Plane and the Data Plane with the Azure components deployed to each. At the top of the diagram is the Control Plane that exists within the Microsoft subscription. The customer subscription is at the bottom of the diagram, which contains the Data Plane and data sources.

A Microsoft-managed Azure Databricks workspace virtual network (VNet) exists within the customer subscription. Information exchanged between this VNet and the Microsoft-managed Azure Databricks Control Plane VNet is sent over a secure TLS connection through ports (22 and 5557) that are enabled by Network Security Groups (NSGs) and protected with port IP filtering.

The Blob Storage account provides default file storage within the workspace (databricks file system (DBFS)). This resource and all other Microsoft-managed resources are completely locked from changes made by the customer. All other resources within the customer subscription are customer-managed and can be added or modified per your Azure subscription permissions. Connectivity between these resources and the Databricks clusters that reside within the Data Plane is secured via TLS.

To clarify, you can write to the default DBFS file storage as needed, but you cannot change the Blob Storage account settings since the account is managed by the Microsoft-managed Control Plane. As a best practice, only use the default storage for temporary files and mount additional storage accounts (Blob Storage or Azure Data Lake Storage Gen2) that you create in your Azure subscription, for long-term file storage. This is because the default file storage is tied to the lifecycle of your Azure Databricks account. If you delete the Azure Databricks account, the default storage gets deleted with it.

If you need advanced network connectivity, such as custom VNet peering and VNet injection, you could deploy Azure Databricks Data Plane resources within your own VNet.

https://docs.databricks.com/getting-started/overview.html

Question 25: Skipped

Scenario: Pym Tech is a U.S. based Technology manufacturer headed by Hank Pym. Their headquarters is located at Treasure Island, San Francisco California and business is booming.

The expansion plans are underway which have presented several IT challenges which Hank has contracted you to advise his IT staff on.

At the moment, the topic is monitoring an Azure Stream Analytics job. The Backlogged Input Events count has been 20 for the last hour. Frank wants to reduce the Backlogged Input Events count.

Which of the following should you recommend Hank to do?

- ○ Drop late arriving events from the job.
- ○ Add an Azure Storage account to the job.
- ○ Increase the streaming units for the job. (Correct)
- ○ Stop the job.

Explanation

You recommend Hank to increase the streaming units for the job.

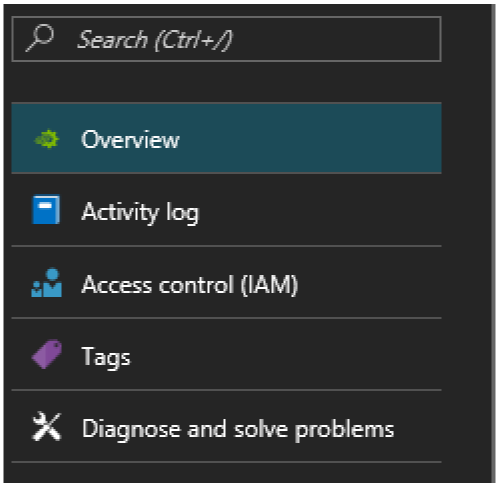
General symptoms of the job hitting system resource limits include:

• If the backlog event metric keeps increasing, its an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

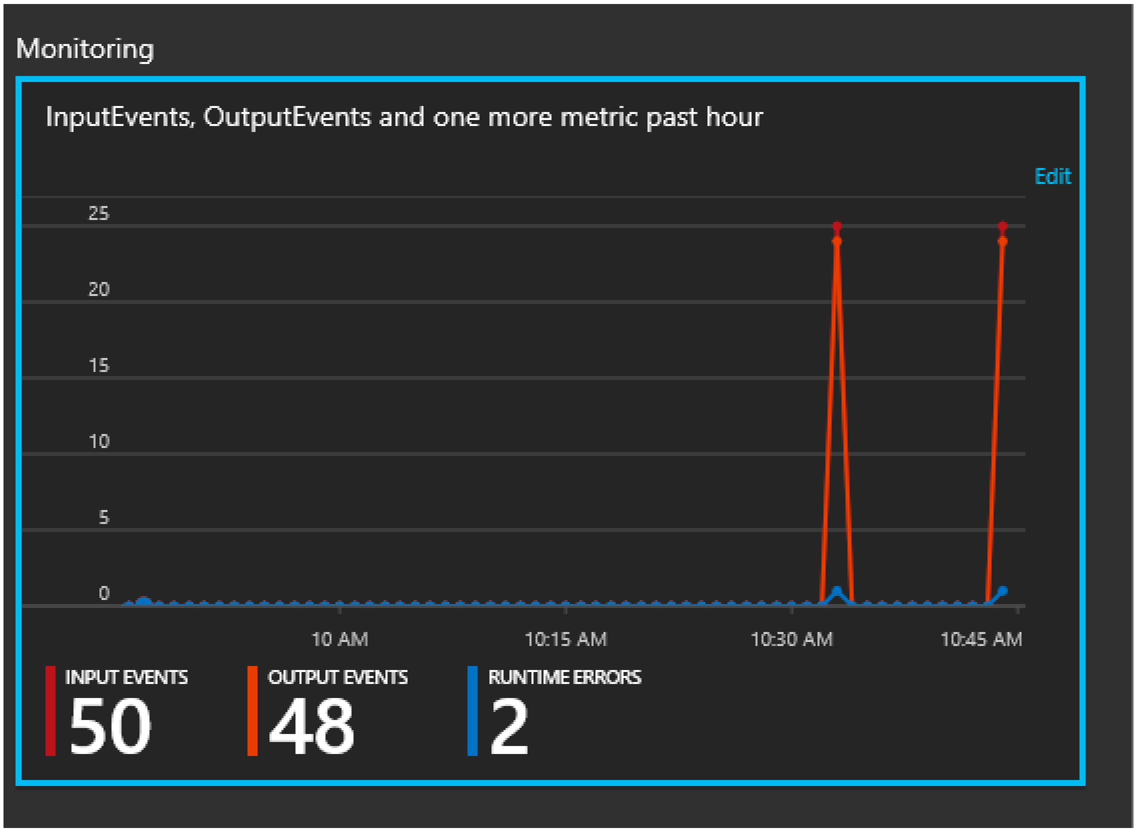
• Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.

Understand Stream Analytics job monitoring and how to monitor queries

The Azure portal surfaces key performance metrics that can be used to monitor and troubleshoot your query and job performance. To see these metrics, browse to the Stream Analytics job you are interested in seeing metrics for and view the **Monitoring** section on the Overview page.



The window will appear as shown:



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

Question 26: Skipped

Scenario: Big Belly Foods, Inc. (BB) owns and operates 300 convenience stores across LatAm. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas. The company has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

BB employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks. You have been hired as an Azure Expert SME and you are to consult the IT team on various Azure related projects.

Business Requirements:

BB 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:

BB 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 BB.

- 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. BB 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:

BB 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.

• BB 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.

• BB 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.

• BB does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

The Ask:

The team looks to you for direction on what should be used together to secure sensitive customer contact information. Which of the following should you recommend using to do this?

- ○ Transparent Data Encryption (TDE)
- ○ Column-level security
- ○ Data sensitivity labels
  (Correct)
- ○ Row-level security

Explanation

To limit the business analysts access to customer contact information, such as phone numbers, should be done with Data sensitivity labels.

Transparent Data Encryption (TDE) is incorrect; it 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.

Data Discovery & Classification

Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labelling, 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.

What is Data Discovery & Classification?

Data Discovery & Classification introduces a set of basic services and new capabilities in Azure. It forms a new information-protection paradigm for SQL Database, SQL Managed Instance, and Azure Synapse, aimed at protecting the data and not just the database. The paradigm includes:

Discovery and recommendations: The classification engine scans your database and identifies columns that contain potentially sensitive data. It then provides you with an easy way to review and apply recommended classification via the Azure portal.

Labelling: 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.

Query result-set sensitivity: The sensitivity of a query result set is calculated in real time for auditing purposes.

Visibility: You can view the database-classification state in a detailed dashboard in the Azure portal. Also, you can download a report in Excel format to use for compliance and auditing purposes and other needs.

Discover, classify, and label sensitive columns

This section describes the steps for:

Discovering, classifying, and labelling columns that contain sensitive data in your database.

Viewing the current classification state of your database and exporting reports.

The classification includes two metadata attributes:

Labels: The main classification attributes, used to define the sensitivity level of the data stored in the column.

Information types: Attributes that provide more granular information about the type of data stored in the column.

Define and customize your classification taxonomy

Data Discovery & Classification comes with a built-in set of sensitivity labels and a built-in set of information types and discovery logic. You can now customize this taxonomy and define a set and ranking of classification constructs specifically for your environment.

You define and customize of your classification taxonomy in one central place for your entire Azure organization. That location is in Azure Security Centre, as part of your security policy. Only someone with administrative rights on the organization's root management group can do this task.

As part of policy management for information protection, you can define custom labels, rank them, and associate them with a selected set of information types. You can also add your own custom information types and configure them with string patterns. The patterns are added to the discovery logic for identifying this type of data in your databases.

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

Question 27: Skipped

In Spark Structured Streaming, what method should be used to read streaming data into a `DataFrame`?

- ○

  `df.spark.stream.read`

- ○

  `df.spark.read`

- ○

`spark.readStream`
(Correct)

- ○
  `spark.stream.read`
- ○
  `df.spark.readStream`

Explanation
Use the `spark.readStream` method to start reading data from a streaming query into a `DataFrame`.

https://kontext.tech/column/streaming-analytics/475/spark-structured-streaming-read-from-and-write-into-kafka-topics

Question 28: Skipped
The Stream Analytics query language is a subset of which query language?

- ○
  T-SQL
  (Correct)
- ○
  MQL
- ○
  QUEL
- ○
  OPath
- ○
  Gremlin
- ○
  CQL

Explanation
The query language you use in Stream Analytics is based heavily on T-SQL.

https://docs.microsoft.com/en-us/stream-analytics-query/stream-analytics-query-language-reference

Question 29: Skipped
What optimization does the following command perform: `OPTIMIZE Students ZORDER BY Grade`?

- ○

  Creates an order-based index on the Grade field to improve filters against that field.
- ○

  Ensures that all data backing, for example, Grade=8 is colocated, then rewrites the sorted data into new Parquet files.

  (Correct)
- ○

  Ensures that all data backing, for example, Grade=8 is colocated, then updates a graph that routes requests to the appropriate files.
- ○

  Both creates an order-based index on the Grade field to improve filters against that field and ensures that all data backing, for example, Grade=8 is colocated, then updates a graph that routes requests to the appropriate files.

Explanation

ZOrdering colocates related information in the same set of files.

https://towardsdatascience.com/delta-lake-enables-effective-caching-mechanism-and-query-optimization-in-addition-to-acid-96c216b95134

Question 30: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Microsoft Azure Stream Analytics is an event processing engine. It enables the consumption and analysis of high volumes of streaming data generated by sensors, devices, or applications. [?] processes the data in real time.

- ○

  Azure StreamSets
- ○

  Azure Multistream Processing
- ○

  Azure EventStream
- ○

  Azure Stream Analytics

  (Correct)

Explanation

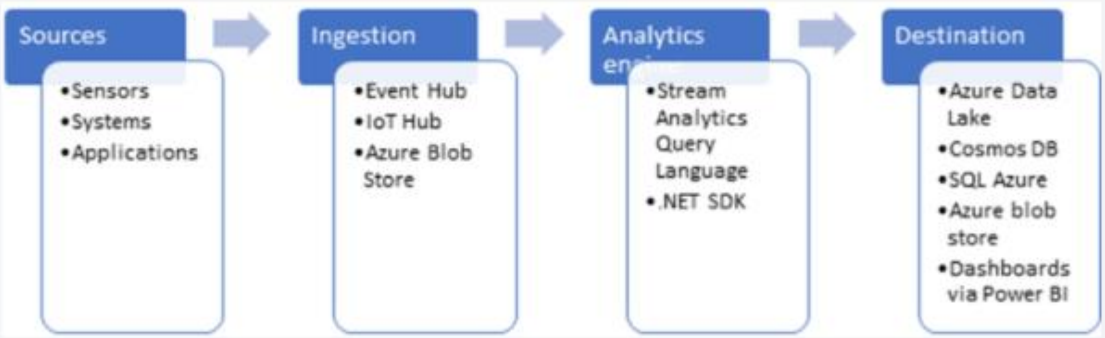
Microsoft Azure Stream Analytics is an event processing engine. It enables the consumption and analysis of high volumes of streaming data generated by sensors, devices, or applications. Stream Analytics processes the data in real time. A typical event processing pipeline built on top of Stream Analytics consists of the following four components:

• **Event producer**: Any application, system, or sensor that continuously produces event data of interest. Examples can include a sensor that tracks the flow of water in a utility pipe to an application such as Twitter that generates tweets against a single hashtag.

• **Event ingestion system**: Takes the data from the source system or application to pass onto an analytics engine. Azure Event Hubs, Azure IoT Hub, or Azure Blob storage can all serve as the ingestion system.

• **Stream analytics engine**: Where compute is run over the incoming streams of data and insights are extracted. Azure Stream Analytics exposes the Stream Analytics query language (SAQL), a subset of Transact-SQL that's tailored to perform computations over streaming data. The engine supports windowing functions that are fundamental to stream processing and are implemented by using the SAQL.

• **Event consumer**: A destination of the output from the stream analytics engine. The target can be storage, such as Azure Data Lake, Azure Cosmos DB, Azure SQL Database, or Azure Blob storage, or dashboards powered by Power BI.



Operational aspects

Stream Analytics guarantees *exactly once* event processing and *at-least-once* event delivery, so events are never lost. It has built-in recovery capabilities in case the delivery of an event fails. Also, Stream Analytics provides built-in checkpointing to maintain the state of your job and produces repeatable results.

Because Azure Stream Analytics is a PaaS service, it's fully managed and highly reliable. Its built-in integration with various sources and destinations and flexible programmability model enhance programmer productivity. The Stream Analytics engine enables in-memory compute, so it offers superior performance. All these factors contribute to low total cost of ownership (TCO) of Azure Stream Analytics.

https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

Question 31: Skipped

True or False: The self-hosted integration runtime is logically registered to the Azure Data Factory and the compute resource used to support its functionality as provided by you. Therefore there is an explicit location property for self-hosted IR.

- ○ True
- ○ False (Correct)

Explanation

In Data Factory, an activity defines the action to be performed. A linked service defines a target data store or a compute service. An integration runtime provides the bridge between the activity and linked services.

Self-hosted integration runtime

A self-hosted integration runtime is capable of:

• Running copy activity between a cloud data stores and a data store in private network.

• Dispatching the following transform activities against compute resources in on-premises or Azure Virtual Network:

  • HDInsight Hive activity (BYOC-Bring Your Own Cluster)

  • HDInsight Pig activity (BYOC)

  • HDInsight MapReduce activity (BYOC)

  • HDInsight Spark activity (BYOC)

  • HDInsight Streaming activity (BYOC)

  • Machine Learning Batch Execution activity

- Machine Learning Update Resource activities
- Stored Procedure activity
- Data Lake Analytics U-SQL activity
- Custom activity (runs on Azure Batch)
- Lookup activity
- Get Metadata activity.

The self-hosted integration runtime is logically registered to the Azure Data Factory and the compute resource used to support its functionality as provided by you. **Therefore there is no explicit location property for self-hosted IR.** When used to perform data movement, the self-hosted IR extracts data from the source and writes into the destination.

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

Question 32: Skipped

Which type of analytics answers the question *"What is likely to happen in the future based on previous trends and patterns?"*

- ○ Predictive (Correct)
- ○ Descriptive
- ○ Diagnostic
- ○ Scenario

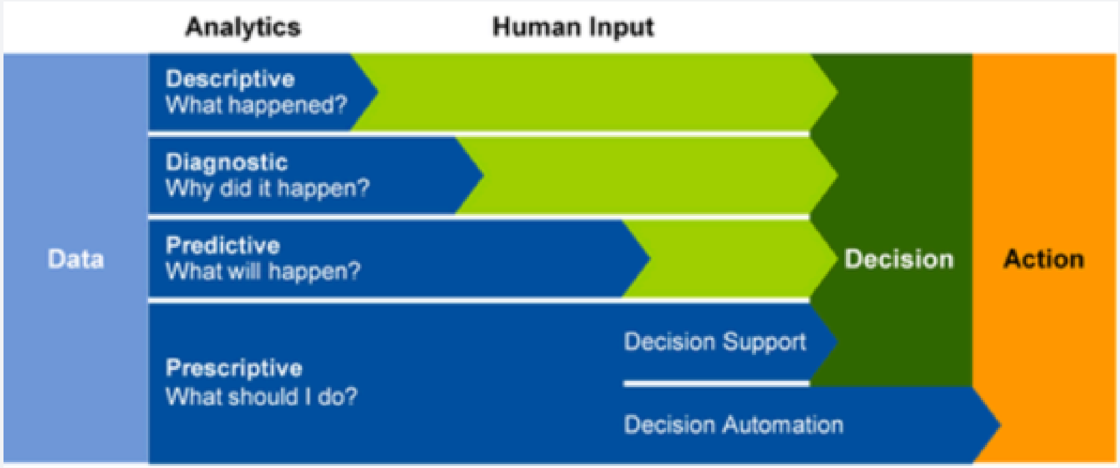
Explanation

Diagnostic analytics

Diagnostic analytics deals with answering the question "Why is it happening?" this may involve exploring information that already exists in a data warehouse, but typically involves a wider search of your data estate to find more data to support this type of analysis.

You can use the same SQL serverless capability within Azure Synapse Analytics that enables you to interactively explore data within a data lake. This can quickly enable a user to search for additional data that may help them to understand “Why is it happening?”

https://www.valamis.com/hub/descriptive-analytics



Predictive analytics

Azure Synapse Analytics also enables you to answer the question “What is likely to happen in the future based on previous trends and patterns?” by using its integrated Apache Spark engine. This can also be used in conjunction with other services such as Azure Machine Learning Services, or Azure Databricks.

https://www.ibm.com/analytics/predictive-analytics

Prescriptive analytics

This type of analytics looks at executing actions based on real-time or near real-time analysis of data, using predictive analytics. Azure Synapse Analytics provides this capability through both Apache Spark, Azure Synapse Link, and by integrating streaming technologies such as Azure Stream Analytics.

https://www.talend.com/resources/what-is-prescriptive-analytics/

Azure Synapse Analytics gives the users of the service the freedom to query data on their own terms, using either serverless or dedicated resources at scale. Azure Synapse Analytics brings these two worlds together with a unified data integration experience to ingest, prepare, manage, and serve data using Azure Synapse Pipelines. In addition, you can visualize the data in the form of dashboards and reports for immediate analysis using Power BI which is integrated into the service too.

https://docs.microsoft.com/en-us/azure/synapse-analytics/overview-what-is

Question 33: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Azure Data Lake Storage Gen2 provides a first-class data lake solution that enables enterprises to consolidate their data.

Along with role-based access control (RBAC), Azure Data Lake Storage Gen2 provides [?] that are POSIX-compliant, and that restrict access to only authorized users, groups, or service principals. It applies restrictions in a way that's flexible, fine-grained, and manageable.

- ○ Transport Layer Security (TLS)
- ○ Transmission Control Protocol (TCP)
- ○ Transparent Data Encryption (TDE)
- ○ Online Transaction Processing (OLTP)
- ○ Access Control Lists (ACLs) **(Correct)**

Explanation

Azure Data Lake Storage Gen2 provides a first-class data lake solution that enables enterprises to consolidate their data.

Along with role-based access control (RBAC), Azure Data Lake Storage Gen2 provides access control lists (ACLs) that are POSIX-compliant, and that restrict access to only authorized users, groups, or service principals. It applies restrictions in a way that's flexible, fine-grained, and manageable. Azure Data Lake Storage Gen2 authenticates through Azure Active Directory OAuth 2.0 bearer tokens. This allows for flexible authentication schemes, including federation with Azure AD Connect and multifactor authentication that provides stronger protection than just passwords.

More significantly, these authentication schemes are integrated into the main analytics services that use the data. These services include Azure Databricks, HDInsight, and Azure Synapse Analytics. Management tools, such as Azure Storage Explorer, are also included. After authentication finishes, permissions are applied at the finest granularity to ensure the right level of authorization for an enterprise's big-data assets.

The Azure Storage end-to-end encryption of data and transport layer protections complete the security shield for an enterprise data lake. The same set of analytics engines and tools can take advantage of these additional layers of protection, resulting in complete protection of your analytics pipelines.

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

Question 34: Skipped

Scenario: You are a consultant at Avengers Security which has an SaaS solution which uses Azure SQL Database with elastic pools. The solution contains a dedicated database for each customer organization where each organization has peak usage at staggered periods throughout the year.

Required: Implement an Azure SQL Database elastic pool to minimize cost.

Which option or options should you recommend to the Avengers IT team to configure?

- ○ Number of transactions only
- ○ CPU usage only
- ○ eDTUs and max data size (Correct)
- ○ Number of databases only
- ○ eDTUs per database only

Explanation

The best size for a pool depends on the aggregate resources needed for all databases in the pool. This involves determining the following:

• Maximum resources utilized by all databases in the pool (either maximum DTUs or maximum vCores depending on your choice of resourcing model).

• Maximum storage bytes utilized by all databases in the pool.

Note: Elastic pools enable the developer to purchase resources for a pool shared by multiple databases to accommodate unpredictable periods of usage by individual databases. You can configure resources for the pool based either on the DTU-based purchasing model or the vCore-based purchasing model.

References:

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-elastic-pool

Question 35: Skipped

Large data projects can be complex. The projects often involve hundreds of decisions. Multiple people are typically involved, and each person helps take the project from design to production.

Roles such as business stakeholders, business analysts, and business intelligence developers are well known and valuable.

Which of the available roles is best described by:

"Performs advanced analytics to extract value from data. Their work can vary from descriptive analytics to predictive analytics. Descriptive analytics evaluate data through a process known as exploratory data analysis (EDA). They are used in machine learning to apply modelling techniques that can detect anomalies or patterns. These are an important part of forecast models."

- ○ Project Manager
- ○ Solution Architects
- ○ AI Engineer
- ○ Data Scientist (Correct)

- ○ RPA Developers
- ○ BI Engineer
- ○ System Administrators
- ○ Data Engineer

Explanation
Data Scientist

Data scientists perform advanced analytics to extract value from data. Their work can vary from descriptive analytics to predictive analytics. Descriptive analytics evaluate data through a process known as exploratory data analysis (EDA). Predictive analytics are used in machine learning to apply modelling techniques that can detect anomalies or patterns. These are an important part of forecast models.

Descriptive and predictive analytics are just one aspect of data scientists' work. Some data scientists might even work in the realms of deep learning, iteratively experimenting to solve a complex data problem by using customized algorithms.

Anecdotal evidence suggests that most of the work in a data science project is spent on data wrangling and feature engineering. Data scientists can speed up the experimentation process when data engineers use their skills to successfully wrangle data.

https://www.whizlabs.com/blog/azure-data-engineer-roles/

Question 36: Skipped
Scenario: You are setting up database permissions for a mid-level manager in your company. This manager is only allowed to see information about their direct reports.

Which type of security would typically be best used in for this scenario?

- ○ Table-level security
- ○ Column-level security
- ○ Row-level security

(Correct)

- ○ Dynamic Data Masking

Explanation

Authentication is the process of validating credentials as you access resources in a digital infrastructure. This ensures that you can validate that an individual, or a service that wants to access a service in your environment can prove who they are. Azure Synapse Analytics provides several different methods for authentication.

Column level security in Azure Synapse Analytics

Generally speaking, column level security is simplifying a design and coding for the security in your application. It allows you to restrict column access in order to protect sensitive data. For example, if you want to ensure that a specific user 'Leo' can only access certain columns of a table because he's in a specific department. The logic for 'Leo' only to access the columns specified for the department he works in, is a logic that is located in the database tier, rather on the application level data tier. If he needs to access data from any tier, the database should apply the access restriction every time he tries to access data from another tier. The reason for doing so, is to make sure that your security is reliable and robust since we're reducing the surface area of the overall security system. Column level security will also eliminate the necessity for the introduction of view, where you would filter out columns, to impose access restrictions on 'Leo'

The way to implement column level security, is by using the `GRANT` T-SQL statement. Using this statement, SQL and Azure Active Directory (AAD) support the authentication.

The syntax to use for implementing column level security looks as follows:

```
SQL

GRANT <permission> [ ,...n ] ON

[ OBJECT :: ][ schema_name ]. object_name [ ( column [ ,...n ] ) ] // specifying
the column access

TO <database_principal> [ ,...n ]

[ WITH GRANT OPTION ]

[ AS <database_principal> ]

<permission> ::=

SELECT

| UPDATE

<database_principal> ::=

Database_user // specifying the database user

| Database_role // specifying the database role

| Database_user_mapped_to_Windows_User

| Database_user_mapped_to_Windows_Group
```

So when would you use column-level security? Let's say that you are a financial services firm, and can only have account manager allowed to have access to a customer's social security number, phone numbers or other personal identifiable information. It is imperative to distinguish the role of an account manager versus the manager of the account managers.

Another use case might be related to the Healthcare Industry. Let's say you have a specific health care provider. This healthcare provider only wants doctors and nurses to be able to access medical records. The billing department should not have access to view this data. Column-level security would typically be the option to use.

Row level security in Azure Synapse Analytics

Row-level security (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. RLS, just like column-level security, can simply help and enable your design and coding of your application security. However, compared to column-level security where it's focused on the columns (parameters), RLS helps you implement restrictions on data row access. Let's say that your employee can only access rows of data that are important of the department, you should implement RLS. If you want to restrict for example, customer's data access that is only relevant to the company, you can implement RLS. The restriction on access of the rows, is a logic that is located in the database tier, rather on the application level data tier. If 'Leo' needs to access data from any tier, the database should apply the access restriction every time he tries to access data from another tier. The reason for doing so, is to make sure that your security is reliable and robust since we're reducing the surface area of the overall security system.

The way to implement RLS is by using the `CREATE SECURITY POLICY[!INCLUDEtsql]` statement. The predicates are created as inline table-valued functions. It is imperative to understand that within Azure Synapse, it only supports filter predicates. If you need to use a block predicate, you won't be able to find support at this moment within in Azure synapse.

Description of row level security in relation to filter predicates

RLS within Azure Synapse supports one type of security predicates, which are Filter predicates, not block predicates.
What filter predicates do, are silently filtering the rows that are available for read operations such as `SELECT`, `UPDATE`, `DELETE`.

The access to row-level data in a table, is restricted as an inline table-valued function, which is a security predicate. This table-valued function will then be invoked and enforced by the security policy that you need. An application, is not aware of rows that are filtered from the result set for filter predicates. So what will happen is that if all rows are filtered, a null set is returned.

When you are using filter predicates, it will be applied when data is read from the base table. The filter predicate affects all get operations such as `SELECT`, `DELETE`, `UPDATE`. You are unable to select or delete rows that have been filtered. It is not possible for you to update a row that has been filtered. What you can do, is update rows in a way that they will be filtered afterwards.

Permissions

If you want to create, alter or drop the security policies, you would have to use the `ALTER ANY SECURITY POLICY` permission. The reason for that is when you are creating or dropping a security policy it requires `ALTER` permissions on the schema.

In addition to that, there are other permissions required for each predicate that you would add:

- `SELECT` and `REFERENCES` permissions on the inline table-valued function being used as a predicate.

- `REFERENCES` permission on the table that you target to be bound to the policy.

- `REFERENCES` permission on every column from the target table used as arguments.

Once you've set up the security policies, they will apply to all the users (including dbo users in the database) Even though DBO users can alter or drop security policies, their changes to the security policies can be audited. If you have special circumstances where highly privileged users, like a sysadmin or db_owner, need to see all rows to troubleshoot or validate data, you would still have to write the security policy in order to allow that.

If you have created a security policy where `SCHEMABINDING = OFF`, in order to query the target table, the user must have the SELECT or EXECUTE permission on the predicate function. They also need permissions to any additional tables, views, or functions used within the predicate function. If a security policy is created with `SCHEMABINDING = ON` (the default), then these permission checks are bypassed when users query the target table.

Best practices

There are some best practices to take in mind when you want to implement RLS. We recommended creating a separate schema for the RLS objects. RLS objects in this context would be the predicate functions, and security policies. Why is that a best practice? It helps to separate the permissions that are required on these special objects from the target tables. In addition to that, separation for different policies and predicate functions may be needed in multi-tenant-databases. However, it is not a standard for every case.

Another best practice to bear in mind is that the `ALTER ANY SECURITY POLICY` permission should only be intended for highly privileged users (such as a security policy manager). The security policy manager should not require `SELECT` permission on the tables they protect.

In order to avoid potential runtime errors, you should take in mind type conversions in predicate functions that you write. Also, you should try to avoid recursion in predicate functions. The reason for this is to avoid performance degradation. Even though the query optimizer will try to detect the direct recursions, there is no guarantee to find the indirect recursions. With an indirect recursion we mean where a second function call the predicate function.

It would also be recommended to avoid the use of excessive table joins in predicate functions. This would maximize performance.

Generally speaking when it comes to the logic of predicates, you should try to avoid logic that depends on session-specific SET options. Even though this is highly unlikely to be used in practical applications, predicate functions whose logic depends on certain session-specific `SET` options can leak information if users are able to execute arbitrary queries. For example, a predicate function that implicitly converts a string to **datetime** could filter different rows based on the `SET DATEFORMAT` option for the current session.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security

Question 37: Skipped

Scenario: Pym Tech is a U.S. based Technology manufacturer headed by Hank Pym. Their headquarters is located at Treasure Island, San Francisco California and business is booming.

The expansion plans are underway which have presented several IT challenges which Hank has contracted you to advise his IT staff on.

At the moment, the topic is examination of the pipeline failures in the company's Azure data factory from the last 60 days.

Which of the following should you recommend Hank to use?

- ○ The Activity log blade for the Data Factory resource
- ○ The Monitor & Manage app in Data Factory
- ○

Azure Monitor
(Correct)

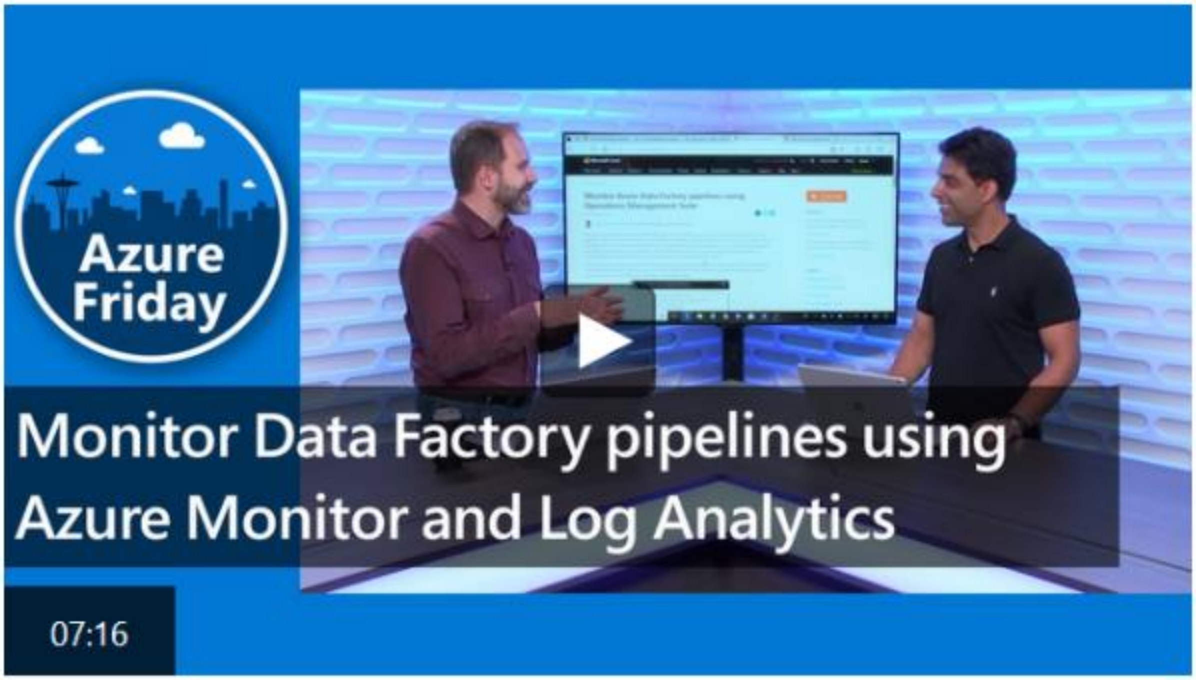
- ○
The Resource health blade for the Data Factory resource

Explanation
You should recommend Frank to use Data Factory stores pipeline-run data for only 45 days. They should use Azure Monitor if Frank wants to keep that data for a longer time.

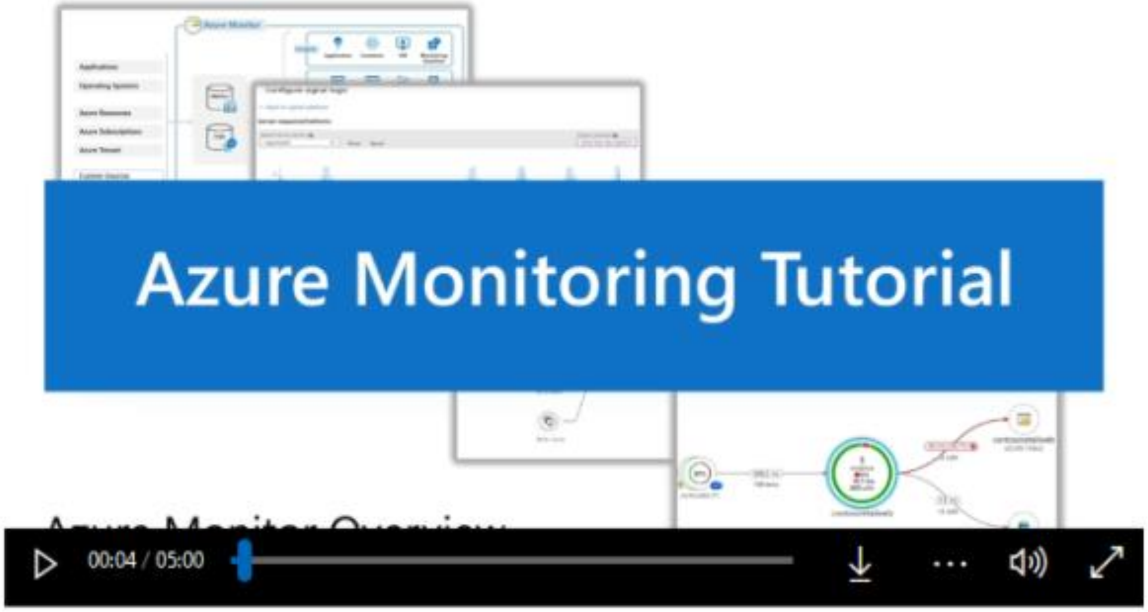
Monitor and Alert Data Factory by using Azure Monitor

Cloud applications are complex and have many moving parts. Monitors provide data to help ensure that your applications stay up and running in a healthy state. Monitors also help you avoid potential problems and troubleshoot past ones. You can use monitoring data to gain deep insights about your applications. This knowledge helps you improve application performance and maintainability. It also helps you automate actions that otherwise require manual intervention.

Azure Monitor provides base-level infrastructure metrics and logs for most Azure services. Azure diagnostic logs are emitted by a resource and provide rich, frequent data about the operation of that resource. Azure Data Factory (ADF) can write diagnostic logs in Azure Monitor.


Azure Friday
Monitor Data Factory pipelines using Azure Monitor and Log Analytics
07:16

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor



https://www.microsoft.com/en-us/videoplayer/embed/RE4qXeL

Question 38: Skipped

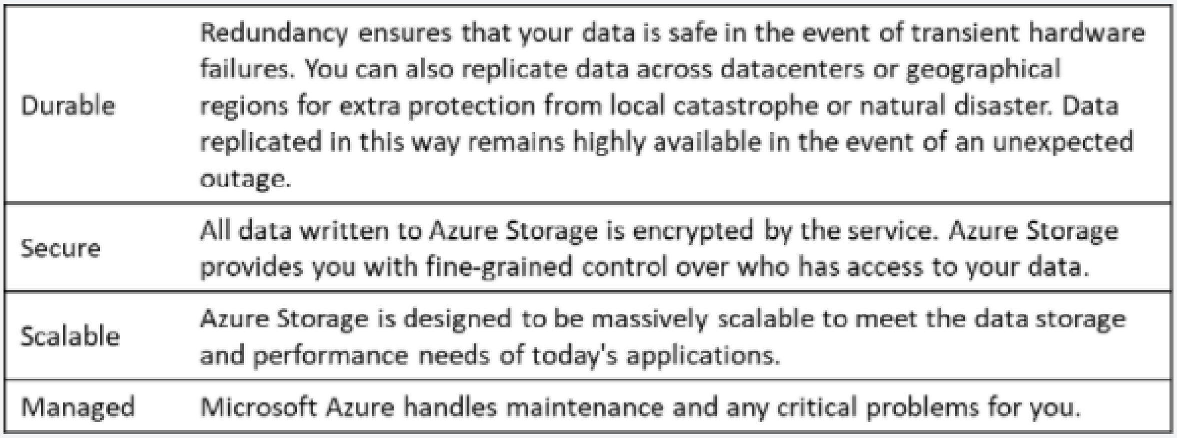
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Microsoft Azure Storage is a managed service that provides durable, secure, and scalable storage in the cloud. The Azure Queue service is used to store and retrieve messages. Queue messages can be up to [A] KB in size, and a queue can contain millions of messages. Queues are used to store lists of messages to be processed [B].

- ○ [A] 32, [B] synchronously
- ○ [A] 25, [B] sequentially
- ○ [A] 50, [B] in a time bound manner
- ○ [A] 64, [B] asynchronously (Correct)

Explanation

Microsoft Azure Storage is a managed service that provides durable, secure, and scalable storage in the cloud.



Durable	Redundancy ensures that your data is safe in the event of transient hardware failures. You can also replicate data across datacenters or geographical regions for extra protection from local catastrophe or natural disaster. Data replicated in this way remains highly available in the event of an unexpected outage.
Secure	All data written to Azure Storage is encrypted by the service. Azure Storage provides you with fine-grained control over who has access to your data.
Scalable	Azure Storage is designed to be massively scalable to meet the data storage and performance needs of today's applications.
Managed	Microsoft Azure handles maintenance and any critical problems for you.

A single Azure subscription can host up to 200 storage accounts, each of which can hold 500 TB of data.

Azure data services

Azure storage includes four types of data:

• Azure Blobs: A massively scalable object store for text and binary data. Can include support for Azure Data Lake Storage Gen2.

• **Files**: Managed file shares for cloud or on-premises deployments.

• Azure Queues: A messaging store for reliable messaging between application components.

• Azure Tables: A NoSQL store for schema-less storage of structured data. Table Storage is not covered in this module.

• Azure Disks: Block-level storage volumes for Azure VMs.

All of these data types in Azure Storage are accessible from anywhere in the world over HTTP or HTTPS. Microsoft provides SDKs for Azure Storage in various languages, and a REST API. You can also visually explore your data right in the Azure portal.

Queues

The Azure Queue service is used to store and retrieve messages. Queue messages can be up to 64 KB in size, and a queue can contain millions of messages. Queues are used to store lists of messages to be processed asynchronously.

You can use queues to loosely connect different parts of your application together. For example, we could perform image processing on the photos uploaded by our users. Perhaps we want to provide some sort of face detection or tagging capability, so people can search through all the images they have stored in our service. We could use queues to pass messages to our image-processing service to let it know that new images have been uploaded and are ready for processing. This sort of architecture would allow you to develop and update each part of the service independently.

https://docs.microsoft.com/en-us/azure/storage/common/storage-introduction

Question 39: Skipped

Before we can create an Azure Cosmos DB container with an analytical store, we must first enable Azure Synapse Link on the Azure Cosmos DB account.

True or False: You cannot disable the Synapse Link feature once it is enabled on the account.

- ○ True (Correct)
- ○ False

Explanation

Before we can create an Azure Cosmos DB container with an analytical store, we must first enable Azure Synapse Link on the Azure Cosmos DB account. You cannot disable the Synapse Link feature once it is enabled on the account. Enabling Synapse Link on the account has no billing implications until containers are created with the analytical store enabled.

https://docs.microsoft.com/en-us/azure/cosmos-db/analytical-store-introduction

If you need to turn off the Synapse Link capability, you have 2 options.

• The first one is to delete and re-create a new Azure Cosmos DB account, migrating the data if necessary.

• The second option is to open a support ticket, to get help on a data migration to another account.

Deleting the Azure Cosmos DB account with disable and remove Azure Synapse Link.

https://docs.microsoft.com/en-us/azure/cosmos-db/synapse-link-frequently-asked-questions

Question 40: Skipped

Which component of Azure Synapse analytics allows the different engines to share the databases and tables between Spark pools and SQL on-demand engine?

- ○ Azure Data Explorer
- ○ Azure Data Warehouse
- ○ Azure Synapse Link
- ○ Azure Synapse Studio
- ○ None of the listed options (Correct)
- ○ Azure Stream Analytics
- ○ Azure Synapse Spark pools
- ○ Azure Synapse Pipeline

Explanation

Azure Synapse shared metadata gives the workspace SQL engines access to databases and tables created with Spark.

Azure Synapse Analytics allows the different workspace computational engines to share databases and tables between its serverless Apache Spark pools and serverless SQL pool.

The sharing supports the so-called modern data warehouse pattern and gives the workspace SQL engines access to databases and tables created with Spark. It also allows the SQL engines to create their own objects that aren't being shared with the other engines.

Support the modern data warehouse

The shared metadata model supports the modern data warehouse pattern in the following way:

1. Data from the data lake is prepared and structured efficiently with Spark by storing the prepared data in (possibly partitioned) Parquet-backed tables contained in possibly several databases.

2. The Spark created databases and all their tables become visible in any of the Azure Synapse workspace Spark pool instances and can be used from any of the Spark jobs. This capability is subject to the permissions since all Spark pools in a workspace share the same underlying catalogue meta store.

3. The Spark created databases and their Parquet-backed tables become visible in the workspace serverless SQL pool. Databases are created automatically in the serverless SQL pool metadata, and both the external and managed tables created by a Spark job are made accessible as external tables in the serverless SQL pool metadata in the dbo schema of the corresponding database.

Object synchronization occurs asynchronously. Objects will have a slight delay of a few seconds until they appear in the SQL context. Once they appear, they can be queried, but not updated nor changed by the SQL engines that have access to them.

Shared metadata objects

Spark allows you to create databases, external tables, managed tables, and views. Since Spark views require a Spark engine to process the defining Spark SQL statement, and cannot be processed by a SQL engine, only databases and their contained external and managed tables that use the Parquet storage format are shared with the workspace SQL engine. Spark views are only shared among the Spark pool instances.

https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/overview

Question 41: Skipped
By default, the Azure Data Factory user interface experience (UX) authors directly against the data factory service.

Which of the following are the limitations of this experience? (Select all that apply)

- □
  The Data Factory service doesn't include a repository for storing the JSON entities for your changes. The only way to save changes is via the "Publish All" button and all changes are published directly to the data factory service.
  (Correct)
- □

Data Factory may be configured with GitHub to allow for easier change tracking and collaboration.

- □
  All the listed options.
- □
  The Data Factory service isn't optimized for collaboration and version control.
  (Correct)
- □
  The Azure Resource Manager template required to deploy Data Factory itself is not included.
  (Correct)

Explanation
By default, the Azure Data Factory user interface experience (UX) authors directly against the data factory service. This experience has the following limitations:

• The Data Factory service doesn't include a repository for storing the JSON entities for your changes. The only way to save changes is via the **Publish All** button and all changes are published directly to the data factory service.

• The Data Factory service isn't optimized for collaboration and version control.

• The Azure Resource Manager template required to deploy Data Factory itself is not included.

To provide a better authoring experience, Azure Data Factory allows you to configure a Git repository with either Azure Repos or GitHub. Git is a version control system that allows for easier change tracking and collaboration.

Note: Authoring directly with the Data Factory service is disabled in the Azure Data Factory UX when a Git repository is configured. Changes made via PowerShell or an SDK are published directly to the Data Factory service, and are not entered into Git.

https://docs.microsoft.com/en-us/azure/data-factory/source-control

Key word is **limitations** – *"Data Factory may be configured with GitHub to allow for easier change tracking and collaboration"* is not a limitation, it is an option.

Question 42: Skipped
Scenario: Big Belly Foods, Inc. (BB) owns and operates 300 convenience stores across LatAm. The company sells a variety of packaged foods and drinks, as well as a variety

of prepared foods, such as sandwiches and pizzas. The company has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

BB employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks. You have been hired as an Azure Expert SME and you are to consult the IT team on various Azure related projects.

Business Requirements:

BB 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:

BB 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 BB.

- 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. BB 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:

BB 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.

- BB 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.

- BB 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.

- BB does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

The Ask:

The team looks to you for direction on what should be used together to import the daily inventory data from the SQL server to Azure Data Lake Storage. Which Azure Data Factory components should you recommend for the Integration runtime type?

- ○ Azure-SSIS integration runtime
- ○ Azure-SAML integration runtime

- ○ Self-hosted integration runtime **(Correct)**
- ○ Azure integration runtime

Explanation

The following are the recommends you should present:

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

• Schedule trigger set for an 8 hour interval.

• A copy activity type

Rational:

• 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.

The Integration Runtime (IR) is the compute infrastructure used by Azure Data Factory to provide the following data integration capabilities across different network environments:

Data Flow: Execute a Data Flow in managed Azure compute environment.

Data movement: Copy data across data stores in public network and data stores in private network (on-premises or virtual private network). It provides support for built-in connectors, format conversion, column mapping, and performant and scalable data transfer.

Activity dispatch: Dispatch and monitor transformation activities running on a variety of compute services such as Azure Databricks, Azure HDInsight, Azure Machine Learning, Azure SQL Database, SQL Server, and more.

SSIS package execution: Natively execute SQL Server Integration Services (SSIS) packages in a managed Azure compute environment.

In Data Factory, an activity defines the action to be performed. A linked service defines a target data store or a compute service. An integration runtime provides the bridge between the activity and linked Services. It's referenced by the linked service or activity, and provides the compute environment where the activity either runs on or gets dispatched from. This way, the activity can be performed in the region closest possible to the target data store or compute service in the most performant way while meeting security and compliance needs.

Integration runtimes can be created in the Azure Data Factory UX via the management hub and any activities, datasets, or data flows that reference them.

Integration runtime types

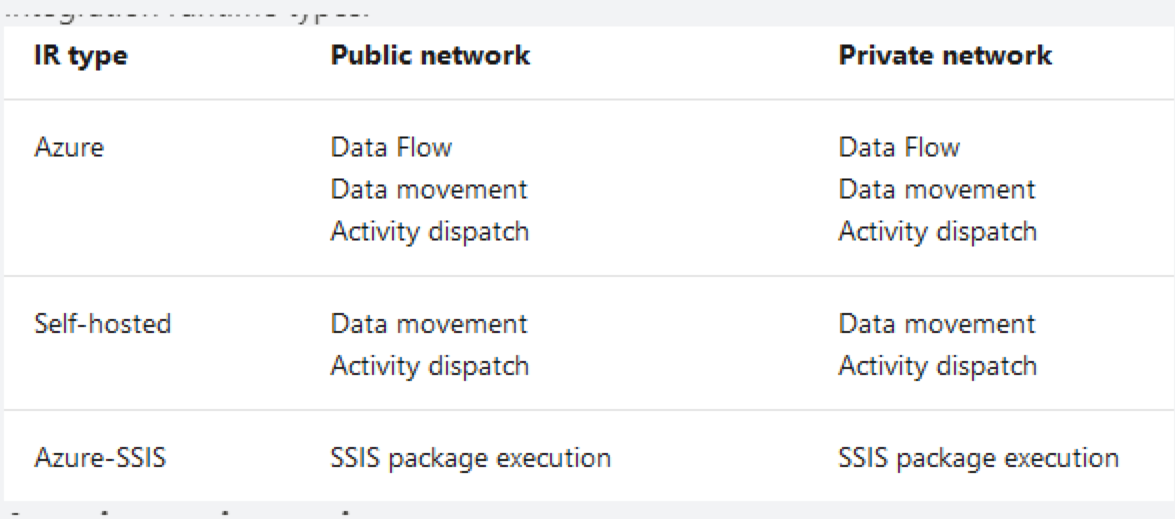
Data Factory offers three types of Integration Runtime (IR), and you should choose the type that best serve the data integration capabilities and network environment needs you're looking for. These three types are:

- Azure

- Self-hosted

- Azure-SSIS

The following table describes the capabilities and network support for each of the integration runtime types:



IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	Data Flow Data movement Activity dispatch
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Azure integration runtime

An Azure integration runtime can:

• Run Data Flows in Azure

• Run copy activity between cloud data stores

• Dispatch the following transform activities in public network: Databricks Notebook/ Jar/ Python activity, HDInsight Hive activity, HDInsight Pig activity, HDInsight MapReduce activity, HDInsight Spark activity, HDInsight Streaming activity, Azure Machine Learning Studio (classic) Batch Execution activity, Azure Machine Learning Studio (classic) Update Resource activities, Stored Procedure activity, Data Lake Analytics U-SQL activity, .NET custom activity, Web activity, Lookup activity, and Get Metadata activity.

Azure IR network environment

Azure Integration Runtime supports connecting to data stores and computes services with public accessible endpoints. Enabling Managed Virtual Network, Azure Integration Runtime supports connecting to data stores using private link service in private network environment.

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

Question 43: Skipped
Scenario: The organization you work at has two types of data:

1. Private and proprietary

2. For public consumption.

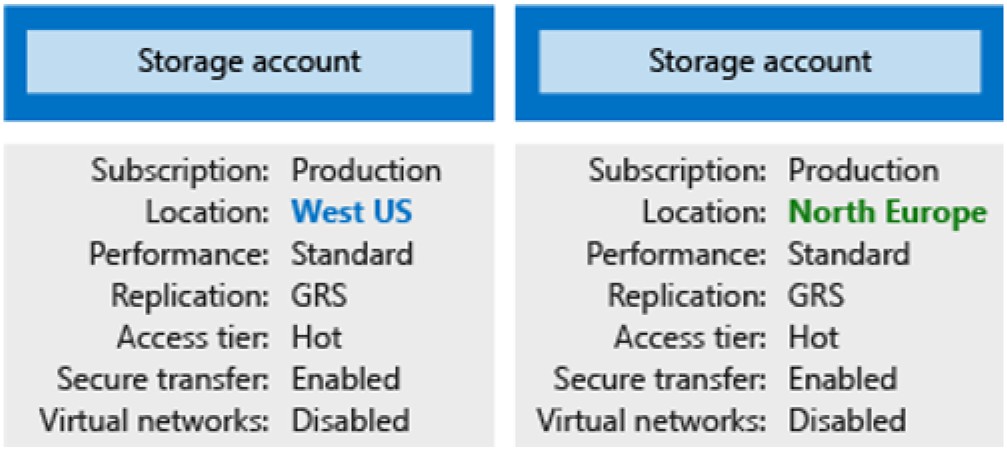
When considering Azure Storage Accounts, which option meet the data diversity requirement?

- ○ Enable virtual networks for the proprietary data and not for the public data . This will require separate storage accounts for the proprietary and public data. **(Correct)**
- ○ Locate the organization's data it in a data centre with the strictest data regulations to ensure that regulatory requirement thresholds have been met. In this way, only one storage account will be required for managing all data, which will reduce data storage costs.
- ○ Locate the organization's data it in a data centre in the required country or region with one storage account for each location.

- ○
  None of the listed options.

Explanation
How many storage accounts do you need?

A storage account represents a collection of settings like location, replication strategy, and subscription owner. You need one storage account for every group of settings that you want to apply to your data. The following illustration shows two storage accounts that differ in one setting; that one difference is enough to require separate storage accounts.



Storage account	
Subscription:	Production
Location:	West US
Performance:	Standard
Replication:	GRS
Access tier:	Hot
Secure transfer:	Enabled
Virtual networks:	Disabled

Storage account	
Subscription:	Production
Location:	North Europe
Performance:	Standard
Replication:	GRS
Access tier:	Hot
Secure transfer:	Enabled
Virtual networks:	Disabled

The number of storage accounts you need is typically determined by your data diversity, cost sensitivity, and tolerance for management overhead.

Data diversity

Organizations often generate data that differs in where it is consumed, how sensitive it is, which group pays the bills, etc. Diversity along any of these vectors can lead to multiple storage accounts. Let's consider two examples:

1. Do you have data that is specific to a country or region? If so, you might want to locate it in a data centre in that country for performance or compliance reasons. You will need one storage account for each location.

2. Do you have some data that is proprietary and some for public consumption? If so, you could enable virtual networks for the proprietary data and not for the public data. This will also require separate storage accounts.

In general, increased diversity means an increased number of storage accounts.

https://docs.microsoft.com/en-us/azure/storage/common/storage-account-overview

Question 44: Skipped

When considering Azure Data Factory, which component is able to run a data movement command or orchestrate a transformation job?

- ○ Activities (Correct)
- ○ Linked Services
- ○ Integration runtime
- ○ Datasets
- ○ SSIS

Explanation

Activities contain the transformation logic or the analysis commands of the Azure Data Factory’s work.

Activities within Azure Data Factory define the actions that will be performed on the data and there are three categories including:

• Data movement activities

• Data transformation activities

• Control activities

Data movement activities

Data movement activities simply move data from one data store to another. You can use the Copy Activity to perform data movement activities, or by using JSON. There are a wide range of data stores that are supported as a source and as a sink. This list is ever increasing, and you can find the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-movement-activities

Data transformation activities

Data transformation activities can be performed natively within the authoring tool of Azure Data Factory using the Mapping Data Flow. Alternatively, you can call a compute resource to change or enhance data through transformation, or perform analysis of the data. These include compute technologies such as Azure Databricks, Azure Batch, SQL Database and Azure Synapse Analytics, Machine Learning Services, Azure Virtual machines and HDInsight. You can make use of any existing SQL Server Integration Services (SSIS) Packages stored in a catalogue to execute in Azure

As this list is always evolving, you can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-transformation-activities

Question 45: Skipped

How does splitting source files help maintain good performance when loading into Synapse Analytics?

- ○ Compute node to storage segment alignment. **(Correct)**
- ○ Reduced possibility of data corruptions.
- ○ Optimized processing of smaller file sizes.
- ○ Having well defined “zones” established for the data coming into the Data Lake and cleansing and transformation tasks that land the data you need in a curated and optimized state.

Explanation

SQL Pools have 60 storage segments. Compute can also scale to 60 nodes and so optimizing for alignment of these 2 resources can dramatically decrease load times.

Split source files

One of the key architectural components within Azure Synapse Analytics dedicated SQL pools is the decoupled storage that is segmented into 60 parts. You should maintain alignment to multiples of this number as much as possible depending on the file sizes that you are loading, and the number of compute nodes you have provisioned. Since there are 60 storage segments and a maximum of 60 MPP compute nodes within the highest performance configuration of SQL Pools, a 1:1 file to compute node to storage segment may be viable for ultra-high workloads, reducing the load times to the minimum possible.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/data-processed

While *"Having well defined "zones" established for the data coming into the Data Lake and cleansing and transformation tasks that land the data you need in a curated and optimized state"* is in of itself is true, it has nothing to do with splitting source files.

Question 46: Skipped
What is the Python syntax for defining a DataFrame in Spark from an existing Parquet file in DBFS?

- ○
  ```
  IPGeocodeDF = read.spark.parquet("dbfs:/mnt/training/ip-geocode.parquet")
  ```
- ○
  None of the listed options
  (Correct)
- ○
  ```
  IPGeocodeDF = spark.parquet.read("dbfs:/mnt/training/ip-geocode.parquet")
  ```
- ○
  ```
  IPGeocodeDF = parquet.read("dbfs:/mnt/training/ip-geocode.parquet")
  ```
- ○
  ```
  IPGeocodeDF = spark.parquet.read("dbfs:/mnt/training/ip-geocode.parquet")
  ```

Explanation
The correct syntax is:

```
IPGeocodeDF = spark.read.parquet("dbfs:/mnt/training/ip-geocode.parquet")
```

https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html

Question 47: Skipped
Scenario: You are determining which Azure database product to use. The organization you work for needs the ability to scale up and scale down OLTP systems on demand along with Azure security and availability features.

Which of the following should be utilized?

- ○
  Azure DataNow

- ○

  Azure On-prem solution
- ○

  Azure Table Storage
- ○

  Azure Cosmos DB
- ○

  Azure SQL Database

  (Correct)

Explanation

Azure SQL Database is a managed relational database service. It supports structures such as relational data and unstructured formats such as spatial and XML data. SQL Database provides online transaction processing (OLTP) that can scale on demand. You'll also find the comprehensive security and availability that you appreciate in Azure database services.

When to use SQL Database

Use SQL Database when you need to scale up and scale down OLTP systems on demand. SQL Database is a good solution when your organization wants to take advantage of Azure security and availability features. Organizations that choose SQL Database also avoid the risks of capital expenditures and of increasing operational spending on complex on-premises systems.

SQL Database can be more flexible than an on-premises SQL Server solution because you can provision and configure it in minutes. Even more, SQL Database is backed up by the Azure service-level agreement (SLA).

Key features

SQL Database delivers predictable performance for multiple resource types, service tiers, and compute sizes. Requiring almost no administration, it provides dynamic scalability with no downtime, built-in intelligent optimization, global scalability and availability, and advanced security options. These capabilities let you focus on rapid app development and on speeding up your time to market. You no longer have to devote precious time and resources to managing virtual machines and infrastructure.

https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-database-paas-overview

Question 48: Skipped

To parallelize work, the unit of distribution is a Spark Cluster. Every Cluster has a Driver and one or more executors. Work submitted to the Cluster is split into what type of object?

- ○ Stages
- ○ Arrays
- ○ Chore
- ○ Jobs (Correct)

Explanation
Each parallelized action is referred to as a Job. The results of each Job is returned to the Driver. Depending on the work required, multiple Jobs will be required. Each Job is broken down into Stages.

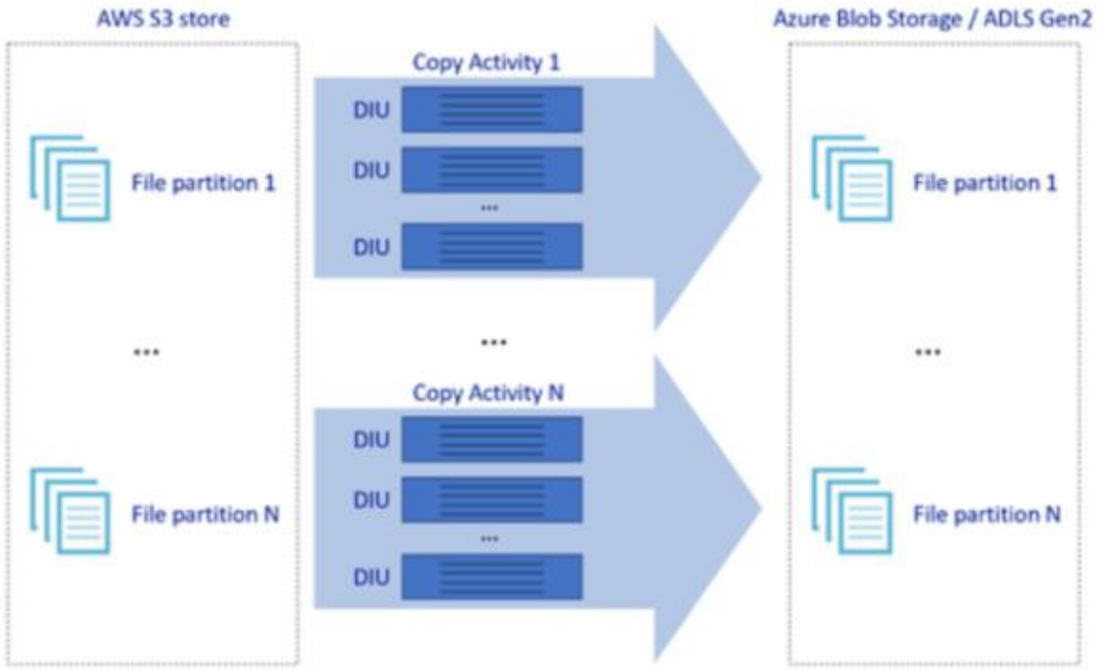
https://www.linkedin.com/pulse/catalyst-tungsten-apache-sparks-speeding-engine-deepak-rajak?articleId=6674601890514378752

Question 49: Skipped
Scenario: You have started at a new job within a company which has a Data Lake Storage Gen2 account. You have been tasked with moving of files from Amazon S3 to Azure Data Lake Storage.

Which tool should you choose?

- ○ Azure Data Factory (Correct)
- ○ Azure Data Studio
- ○ Azure Portal
- ○ Azure Storage Explorer
- ○ Azure Data Catalog

Explanation
Azure Data Factory provides a performant, robust, and cost-effective mechanism to migrate data at scale from Amazon S3 to Azure Blob Storage or Azure Data Lake Storage Gen2.



The picture above illustrates how you can achieve great data movement speeds through different levels of parallelism:

• A single copy activity can take advantage of scalable compute resources: when using Azure Integration Runtime, you can specify up to 256 DIUs for each copy activity in a serverless manner; when using self-hosted Integration Runtime, you can manually scale up the machine or scale out to multiple machines (up to 4 nodes), and a single copy activity will partition its file set across all nodes.

• A single copy activity reads from and writes to the data store using multiple threads.

• ADF control flow can start multiple copy activities in parallel, for example using For Each loop.

https://docs.microsoft.com/en-us/azure/data-factory/data-migration-guidance-s3-azure-storage

Question 50: Skipped
What is an Azure Key Vault-backed secret scope?

- ○
  It is the Key Vault Access Key used to securely connect to the vault and retrieve secrets
- ○
  A Databricks secret scope that is backed by Azure Key Vault instead of Databricks.
  (Correct)
- ○
  It is a method by which you create a secure connection to Azure Key Vault from a notebook and directly access its secrets within the Spark session
- ○
  An Azure Key Vault-backed secret scope is a private key framework managed by Microsoft.

Explanation
A secret scope is provided by Azure Databricks and can be backed by either Databricks or Azure Key Vault.

https://docs.microsoft.com/en-us/azure/databricks/security/secrets/secret-scopes

Question 51: Skipped
Which role works with Azure Cognitive Services, Cognitive Search, and the Bot Framework?

- ○
  A BI Engineer
- ○
  A Project Manager
- ○
  A Data Engineer
- ○
  An RPA Developer
- ○
  A Data Scientist
- ○
  A System Administrator
- ○
  A Solution Architect

- ○ An AI Engineer
  (Correct)

Explanation
Artificial intelligence (AI) engineers work with AI services such as Cognitive Services, Cognitive Search, and the Bot Framework.

AI Engineer

AI engineers work with AI services such as Cognitive Services, Cognitive Search, and Bot Framework. Cognitive Services includes Computer Vision, Text Analytics, Bing Search, and Language Understanding (LUIS).

Rather than creating models, AI engineers apply the prebuilt capabilities of Cognitive Services APIs. AI engineers embed these capabilities within a new or existing application or bot. AI engineers rely on the expertise of data engineers to store information that's generated from AI.

AI engineers add the intelligent capabilities of vision, voice, language, and knowledge to applications. To do this, they use the Cognitive Services offerings that are available out of the box.

When a Cognitive Services application reaches its capacity, AI engineers call on data scientists. Data scientists develop machine learning models and customize components for an AI engineer's application.

For example, an AI engineer might be working on a Computer Vision application that processes images. This AI engineer would ask a data engineer to provision an Azure Cosmos DB instance to store the metadata and tags that the Computer Vision application generates.

https://www.whizlabs.com/blog/azure-data-engineer-roles/

Question 52: Skipped
Which of the following is a good analogy for the access keys of a storage account?

- ○ IP Address
- ○ REST Endpoint
- ○ Username and password
  (Correct)

- ○ Cryptographic algorithm

Explanation

Possession of an access key identifies the account and grants you access. This is very similar to login credentials like a username and password.

https://docs.microsoft.com/en-us/azure/storage/common/storage-account-overview

Question 53: Skipped

Which correct syntax to specify the location of a checkpoint directory when defining a Delta Lake streaming query?

- ○ `.writeStream.format(“delta.parquet").option("checkpointLocation", checkpointPath) ...`
- ○ `.writeStream.format("delta").option("checkpointLocation", checkpointPath) ...` **(Correct)**
- ○ `.writeStream.format("parquet").option("checkpointLocation", checkpointPath) ...`
- ○ `.writeStream.format("delta").checkpoint("location", checkpointPath) ...`

Explanation

`.writeStream.format("delta").option("checkpointLocation", checkpointPath) ...` is the correct syntax to specify the checkpoint directory on a Delta Lake streaming query.

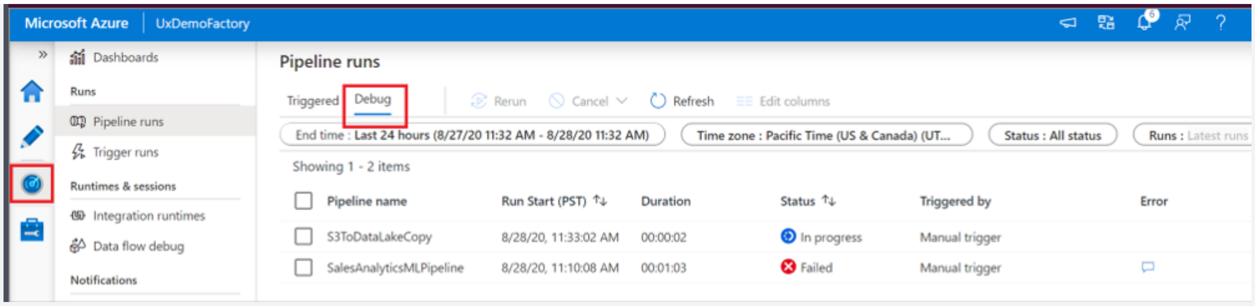
https://docs.microsoft.com/en-us/azure/databricks/delta/delta-streaming

Question 54: Skipped

You can monitor all of your pipeline runs natively in the Azure Data Factory user experience. The default monitoring view is list of triggered pipeline runs in the selected time period.

True or False: The list of pipeline and activity runs is auto refreshed every 60 seconds.

To view the results of a debug run, select the **Debug** tab.



Pipeline runs

End time : Last 24 hours (8/27/20 11:32 AM - 8/28/20 11:32 AM) | Time zone : Pacific Time (US & Canada) (UT... | Status : All status | Runs : Latest runs

Showing 1 - 2 items

Pipeline name	Run Start (PST) ↑↓	Duration	Status ↑↓	Triggered by	Error
S3ToDataLakeCopy	8/28/20, 11:33:02 AM	00:00:02	In progress	Manual trigger	
SalesAnalyticsMLPipeline	8/28/20, 11:10:08 AM	00:01:03	Failed	Manual trigger	

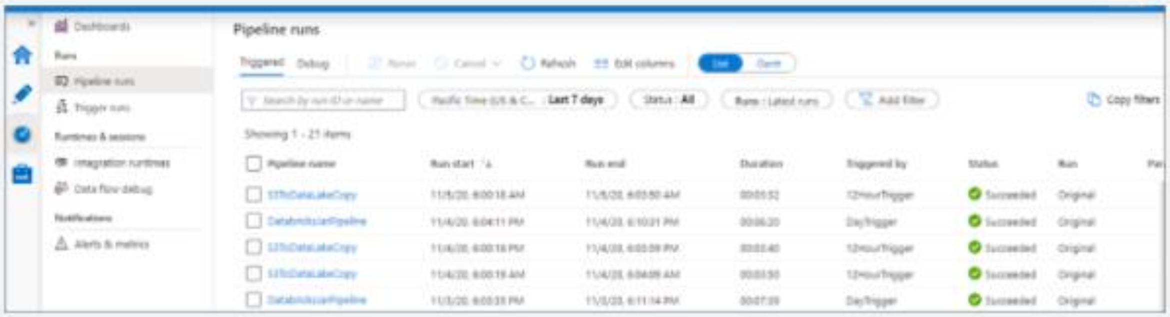
- ○ True
- ○ False (Correct)

Explanation

Once you've created and published a pipeline in Azure Data Factory, you can associate it with a trigger or manually kick off an on-demand run. You can monitor all of your pipeline runs natively in the Azure Data Factory user experience. To open the monitoring experience, select the **Monitor & Manage** tile in the data factory blade of the Azure portal. If you're already in the Azure Data Factory UX, click on the **Monitor** icon on the left sidebar.

Monitor pipeline runs

The default monitoring view is list of triggered pipeline runs in the selected time period. You can change the time range and filter by status, pipeline name, or annotation. Hover over the specific pipeline run to get run-specific actions such as rerun and the consumption report.



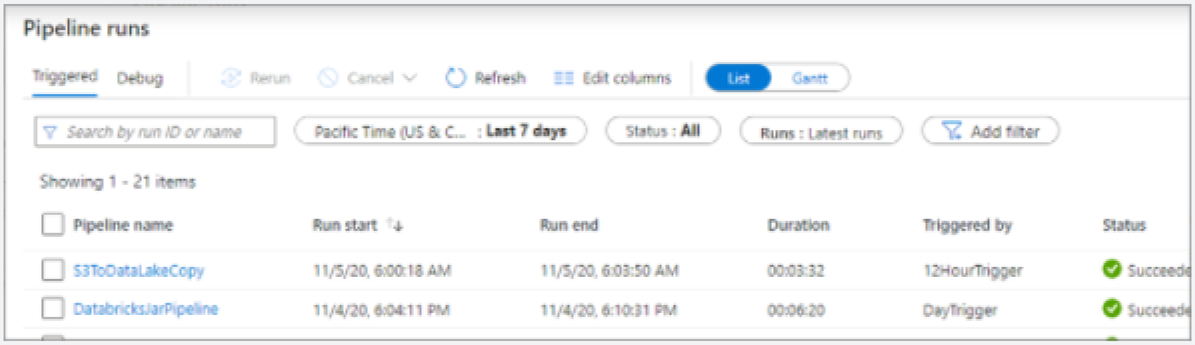
Pipeline runs

Pacific Time (US & C... Last 7 days | Status : All | Runs : Latest runs

Showing 1 - 21 items

Pipeline name	Run start ↓	Run end	Duration	Triggered by	Status	Run
S3ToDataLakeCopy	11/5/20, 6:00:18 AM	11/5/20, 6:03:50 AM	00:03:32	12HourTrigger	Succeeded	Original
DatabricksJarPipeline	11/4/20, 6:04:11 PM	11/4/20, 6:10:31 PM	00:06:20	DayTrigger	Succeeded	Original
S3ToDataLakeCopy	11/4/20, 6:00:18 PM	11/4/20, 6:03:59 PM	00:03:40	12HourTrigger	Succeeded	Original
S3ToDataLakeCopy	11/4/20, 6:00:19 AM	11/4/20, 6:04:09 AM	00:03:50	12HourTrigger	Succeeded	Original
DatabricksJarPipeline	11/3/20, 6:03:35 PM	11/3/20, 6:11:14 PM	00:07:39	DayTrigger	Succeeded	Original

You need to manually select the **Refresh** button to refresh the list of pipeline and activity runs. Autorefresh is currently not supported.



Pipeline runs

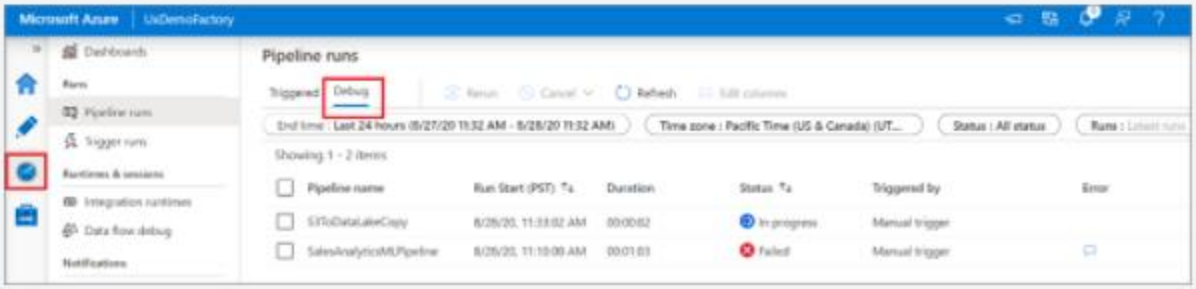
Triggered | Debug | Rerun | Cancel | Refresh | Edit columns | List | Gantt

Search by run ID or name | Pacific Time (US & C... : Last 7 days | Status : All | Runs : Latest runs | Add filter

Showing 1 - 21 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status
S3ToDataLakeCopy	11/5/20, 6:00:18 AM	11/5/20, 6:03:50 AM	00:03:32	12HourTrigger	Succeede
DatabricksJarPipeline	11/4/20, 6:04:11 PM	11/4/20, 6:10:31 PM	00:06:20	DayTrigger	Succeede

To view the results of a debug run, select the **Debug** tab.



Microsoft Azure | UsDemoFactory

Pipeline runs

Triggered | Debug | Rerun | Cancel | Refresh | Edit columns

End time : Last 24 hours (8/27/20 11:32 AM - 8/28/20 11:32 AM) | Time zone : Pacific Time (US & Canada) (UT... | Status : All status | Runs : Latest runs

Showing 1 - 2 items

Pipeline name	Run Start (PST)	Duration	Status	Triggered by	Error
S3ToDataLakeCopy	8/28/20, 11:33:02 AM	00:00:02	In progress	Manual trigger	
SalesAnalyticsMLPipeline	8/28/20, 11:10:00 AM	00:01:03	Failed	Manual trigger	

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

Question 55: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

You can use a *service-level* SAS to allow access to specific resources in a storage account. You'd use this type of SAS, for example, … [?] (Select all that apply)

- ☐
  None of the listed options.
- ☐
  All the listed options.
- ☐
  to allow an app to retrieve a list of files in a file system.
  (Correct)
- ☐

to allow the ability to create file systems.

- □
  to allow an app to download a file.
  (Correct)

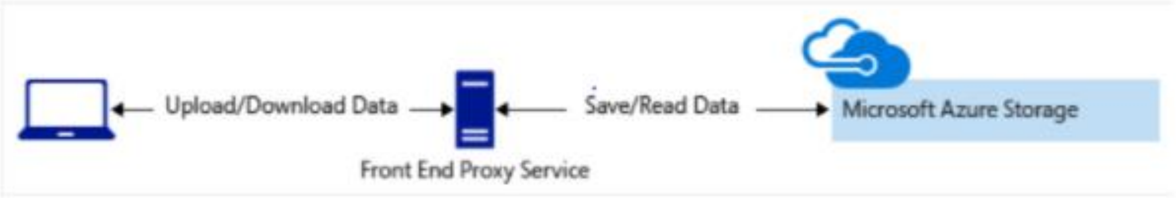
Explanation
Types of shared access signatures

You can use a *service-level* SAS to allow access to specific resources in a storage account. You'd use this type of SAS, for example, to allow an app to retrieve a list of files in a file system, or to download a file.

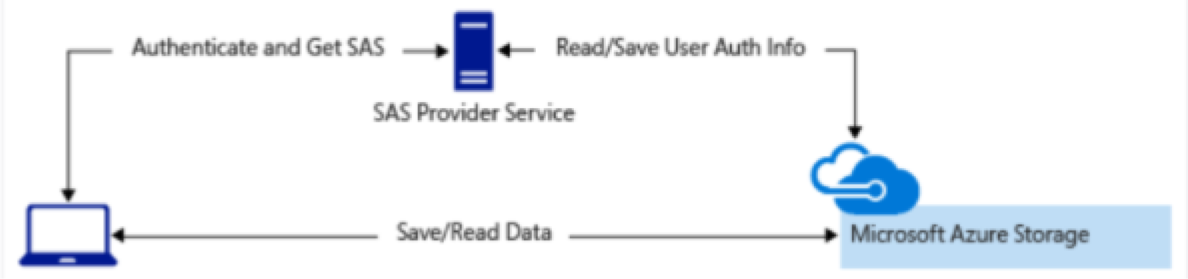
Use an *account-level* SAS to allow access to anything that a service-level SAS can allow, plus additional resources and abilities. For example, you can use an account-level SAS to allow the ability to create file systems.

You'd typically use a SAS for a service where users read and write their data to your storage account. Accounts that store user data have two typical designs:

• Clients upload and download data through a front-end proxy service, which performs authentication. This front-end proxy service has the advantage of allowing validation of business rules. But, if the service must handle large amounts of data or high-volume transactions, you might find it complicated or expensive to scale this service to match demand.



• A lightweight service authenticates the client, as needed. Next, it generates a SAS. After receiving the SAS, the client can access storage account resources directly. The SAS defines the client's permissions and access interval. It reduces the need to route all data through the front-end proxy service.



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

Question 56: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

A transactional database must adhere to the [?] properties to ensure that the database remains consistent while processing transactions.

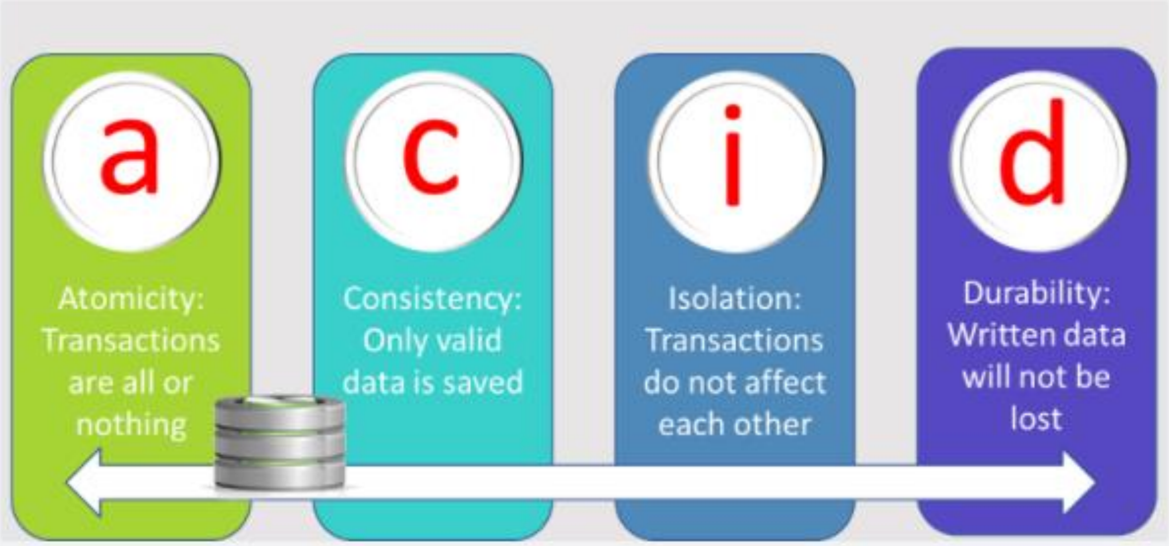
- ○ Nuclear
- ○ Atomic
- ○ ACID (Correct)
- ○ Forensic

Explanation

A transactional database must adhere to the **ACID (Atomicity, Consistency, Isolation, Durability)** properties to ensure that the database remains consistent while processing transactions.

The four letters in ACID represent the four required characteristics of database transactions:

- Atomicity
- Consistency
- Isolation
- Durability


a
Atomicity: Transactions are all or nothing
c
Consistency: Only valid data is saved
i
Isolation: Transactions do not affect each other
d
Durability: Written data will not be lost

• *Atomicity* guarantees that each transaction is treated as a single *unit*, which either succeeds completely, or fails completely. If any of the statements constituting a transaction fails to complete, the entire transaction fails and the database is left unchanged. An atomic system must guarantee atomicity in each and every situation, including power failures, errors, and crashes.

• *Consistency* ensures that a transaction can only take the data in the database from one valid state to another. A consistent database should never *lose* or *create* data in a manner that can't be accounted for. In the bank transfer example described earlier, if you add funds to an account, there must be a corresponding deduction of funds somewhere, or a record that describes where the funds have come from if they have been received externally. You can't suddenly create (or lose) money.

• *Isolation* ensures that concurrent execution of transactions leaves the database in the same state that would have been obtained if the transactions were executed sequentially. A concurrent process can't see the data in an inconsistent state (for example, the funds have been deducted from one account, but not yet credited to another.)

• *Durability* guarantees that once a transaction has been committed, it will remain committed even if there's a system failure such as a power outage or crash.

https://www.techopedia.com/definition/23949/atomicity-consistency-isolation-durability-acid-database-management-system

Question 57: Skipped

Scenario: You have been contracted by Wayne Enterprises, a company owned by Bruce Wayne with market value of over twenty seven million dollars. Bruce founded Wayne Enterprises shortly after he created the Wayne Foundation and he became the president and chairman of the company.

Bruce has come to you because his IT team needs advice on the proper type of storage to use for their files in an Azure Storage environment. Due to the various jurisdictions that Wayne Enterprises operates in, there are many compliance regulations which must be followed.

Required:

• A single storage account must be used to store all operations (includes all reads, writes and deletes)

• Retention policy dictates that an on-premises copy must exist for all historical operations

As the contracted expert on Azure, Bruce and the team look to you for direction. Which of the following actions will you recommend to them to meet the requirements?

- □ Configure the storage account to log read, write and delete operations for service type Blob (Correct)
- □ Configure the storage account to log read, write and delete operations for service-type table
- □ Use the AzCopy tool to download log data from $logs/blob (Correct)
- □ Configure the storage account to log read, write and delete operations for service type queue
- □ Use the storage client to download log data from `$logs/table`

Explanation

Storage Logging logs request data in a set of blobs in a blob container named $logs in your storage account. This container does not show up if you list all the blob containers in your account but you can see its contents if you access it directly.

Storage Analytics logs detailed information about successful and failed requests to a storage service. This information can be used to monitor individual requests and to diagnose issues with a storage service. Requests are logged on a best-effort basis. This means that most requests will result in a log record, but the completeness and timeliness of Storage Analytics logs are not guaranteed.

Storage Analytics logging is not enabled by default for your storage account. You can enable it in the Azure portal or by using PowerShell, or Azure CLI. For step-by-step guidance, see Enable and manage Azure Storage Analytics logs (classic).

You can also enable Storage Analytics logs programmatically via the REST API or the client library. Use the Get Blob Service Properties, Get Queue Service Properties, and Get Table Service Properties operations to enable Storage Analytics for each service. To see an example that enables Storage Analytics logs by using .NET, see Enable logs

Log entries are created only if there are requests made against the service endpoint. For example, if a storage account has activity in its Blob endpoint but not in its Table or Queue endpoints, only logs pertaining to the Blob service will be created.

https://docs.microsoft.com/en-us/rest/api/storageservices/enabling-storage-logging-and-accessing-log-data

To view and analyze your log data, you should download the blobs that contain the log data you are interested in to a local machine. Many storage-browsing tools enable you to download blobs from your storage account; you can also use the Azure Storage team provided command-line Azure Copy Tool (AzCopy) to download your log data.

AzCopy is a command-line utility that you can use to copy blobs or files to or from a storage account. This article helps you download AzCopy, connect to your storage account, and then transfer files.

https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10

https://www.youtube.com/watch?v=GJYAgi5eYYE

Question 58: Skipped

How can all notebooks in Synapse studio be saved?

- ○ Notebooks are synced to the Synapse Studio cloud automatically upon changes being made to a file.
- ○ Select the Publish button on the notebook command bar.
- ○ Select the Publish all button on the workspace command bar. (Correct)
- ○ Using CTRL + S

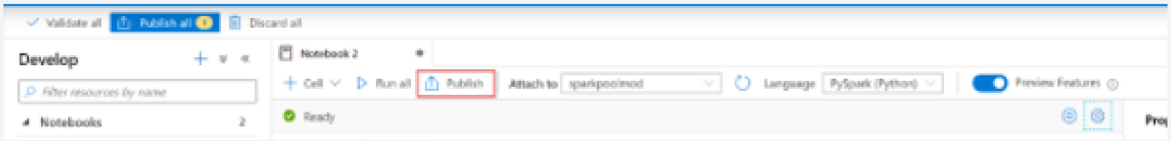
Explanation

To save all notebooks in your workspace, select the Publish all button on the workspace command bar.

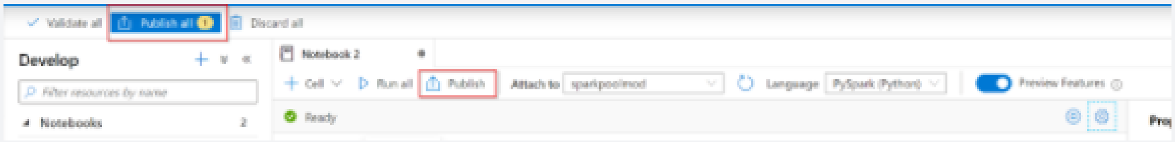
It is possible to save a single or all notebooks that you've created with Azure Synapse Studio notebooks.

You have the possibility to save a single notebook or all notebooks in your workspace.

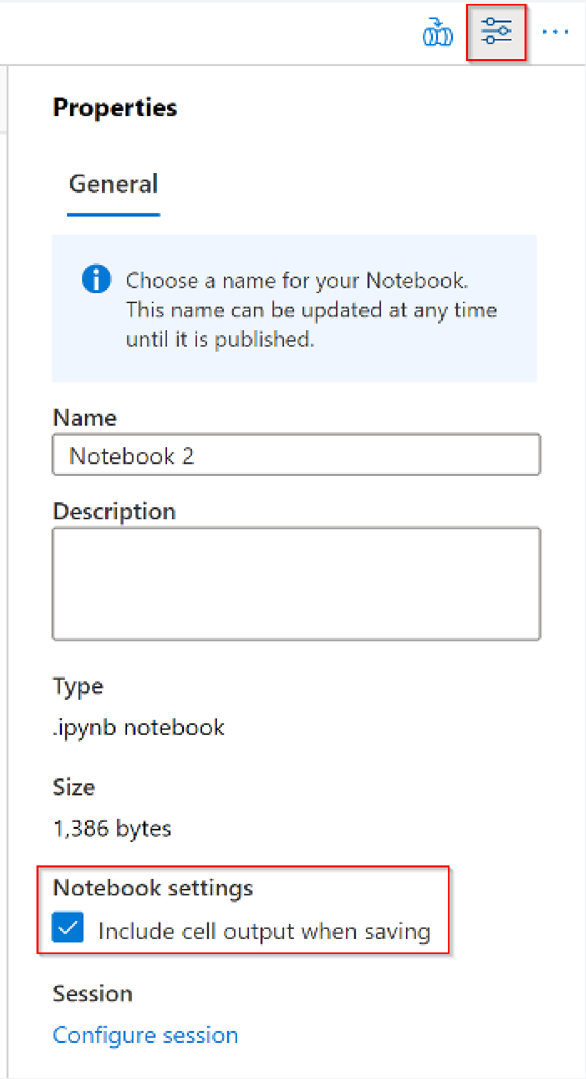
To save changes you made to a single notebook, select the **Publish** button on the notebook command bar.



To save all notebooks in your workspace, select the **Publish all** button on the workspace command bar.



ude the cell output when



Properties

General

Choose a name for your Notebook. This name can be updated at any time until it is published.

Name

Notebook 2

Description

Type

.ipynb notebook

Size

1,386 bytes

Notebook settings

☑ Include cell output when saving

Session

Configure session

saving.

https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-development-using-notebooks?tabs=classical

Question 59: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

From a high level, the Azure Databricks service launches and manages Apache Spark clusters within your Azure subscription. Apache Spark clusters are groups of computers that are treated as a single computer and handle the execution of commands issued from notebooks.

Internally, Azure Kubernetes Service (AKS) is used to ... [?]

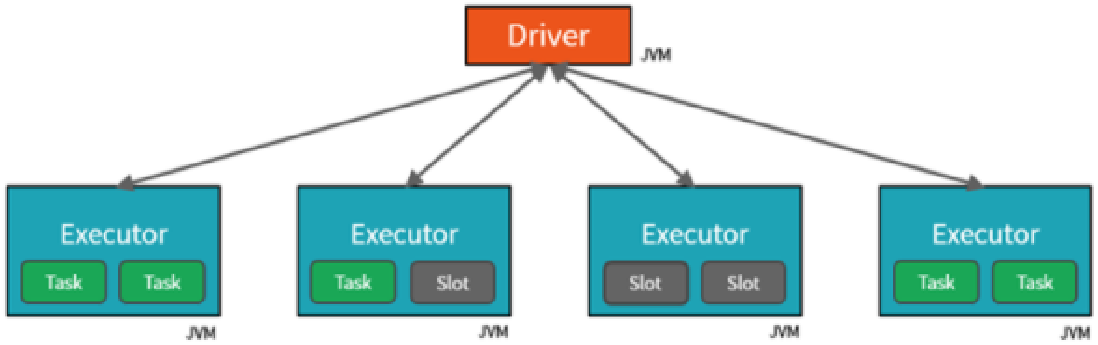
- ○
  specify the types and sizes of the virtual machines.
- ○
  provide the fastest virtualized network infrastructure in the cloud.
- ○
  run the Azure Databricks control-plane and data-planes via containers running on the latest generation of Azure hardware.
  (Correct)
- ○
  auto-scale as needed based on your usage and the setting used when configuring the cluster.
- ○
  pulls data from a specified data source.

Explanation
To gain a better understanding of how to develop with Azure Databricks, it is important to understand the underlying architecture. We will look at two aspects of the Databricks architecture: the Azure Databricks service and Apache Spark clusters.

High-level overview

From a high level, the Azure Databricks service launches and manages Apache Spark clusters within your Azure subscription. Apache Spark clusters are groups of computers that are treated as a single computer and handle the execution of commands issued from notebooks. Using a master-worker type architecture, clusters allow processing of data to be parallelized across many computers to improve scale and performance. They consist of a Spark Driver (master) and worker nodes. The driver node sends work to the worker nodes and instructs them to pull data from a specified data source.

In Databricks, the notebook interface is the driver program. This driver program contains the main loop for the program and creates distributed datasets on the cluster, then applies operations (transformations & actions) to those datasets. Driver programs access Apache Spark through a SparkSession object regardless of deployment location.



Microsoft Azure manages the cluster, and auto-scales it as needed based on your usage and the setting used when configuring the **cluster.** Auto-termination can also be enabled, which allows Azure to terminate the cluster after a specified number of minutes of inactivity.

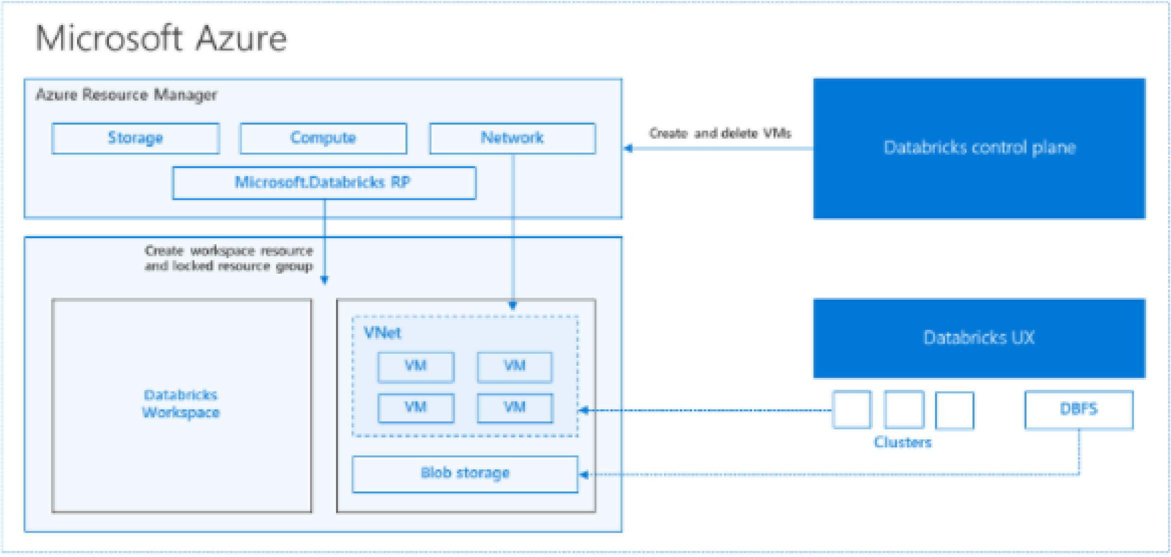
Under the covers

Now let's take a deeper look under the covers. When you create an Azure Databricks service, a "Databricks appliance" is deployed as an Azure resource in your subscription. At the time of cluster creation, you specify the types and sizes of the virtual machines (VMs) to use for both the Driver and Worker nodes, but Azure Databricks manages all other aspects of the cluster.

You also have the option of using a Serverless Pool. A Serverless Pool is self-managed pool of cloud resources that is auto-configured for interactive Spark workloads. You provide the minimum and maximum number of workers and the worker type, and Azure Databricks provisions the compute and local storage based on your usage.

The "Databricks appliance" is deployed into Azure as a managed resource group within your subscription. This resource group contains the Driver and Worker VMs, along with other required resources, including a virtual network, a security group, and a storage account. All metadata for your cluster, such as scheduled jobs, is stored in an Azure Database with geo-replication for fault tolerance.

Internally, Azure Kubernetes Service (AKS) is used to run the Azure Databricks control-plane and data-planes via containers running on the latest generation of Azure hardware (Dv3 VMs), with NvMe SSDs capable of blazing 100us latency on IO. These make Databricks I/O performance even better. In addition, accelerated networking provides the fastest virtualized network infrastructure in the cloud. Azure Databricks utilizes these features to further improve Spark performance. Once the services within this managed resource group are ready, you will be able to manage the Databricks cluster through the Azure Databricks UI and through features such as auto-scaling and auto-termination.



https://databricks.com/blog/2017/11/15/a-technical-overview-of-azure-databricks.html

Question 60: Skipped

Scenario: You are working as a consultant at Avengers Security. At the moment, you are consulting with Tony, the lead of the IT team and the topic of discussion is access provisioning for an Azure Data Lake Storage Gen2 account.

Quentin Beck is a team member who has contributor access to the storage account, as well as the application ID access key. One of Quentin's tasks on his to-do list is to use PolyBase to load data into Azure SQL data warehouse.

Required: Configure PolyBase to connect the data warehouse to the storage account.

Tony has listed out a few items that he thinks Quentin should create to perform the task, but is not sure if they are correct and is not sure of the order of operations needed to complete the requirement successfully.

a. A database encryption key

b. An asymmetric key

c. An external data source

d. An external file format

e. A database scoped credential

Since you are an Azure SME, he looks to you for advice to identify the correct items to create and for you to arrange them in the correct order.

Which of the following identifies the correct items needed in the correct order to fulfill the requirement?

- ○ c → e → a → d
- ○ a → d → c
- ○ c → d → a → e
- ○ e → c → d (Correct)
- ○ a → d → c → b → e

Explanation

Step 1: A database scoped credential

To access your Data Lake Storage account, you will need to create a Database Master Key to encrypt your credential secret used in the next step. You then create a database scoped credential.

Step 2: An external data source

Create the external data source. Use the CREATE EXTERNAL DATA SOURCE command to store the location of the data. Provide the credential created in the previous step.

Step 3: An external file format

Configure data format: To import the data from Data Lake Storage, you need to specify the External File Format. This object defines how the files are written in Data Lake Storage.

Load data from Azure Data Lake Storage into dedicated SQL pools in Azure Synapse Analytics

Create the target table

Connect to your dedicated SQL pool and create the target table you will to load to. In this example, we are creating a product dimension table.

```
SQL
-- A: Create the target table
-- DimProduct
CREATE TABLE [dbo].[DimProduct]
(
    [ProductKey] [int] NOT NULL,
    [ProductLabel] [nvarchar](255) NULL,
    [ProductName] [nvarchar](500) NULL
)
WITH
(
    DISTRIBUTION = HASH([ProductKey]),
    CLUSTERED COLUMNSTORE INDEX
    --HEAP
);
```

Create the COPY statement

Connect to your SQL dedicated pool and run the COPY statement. For a complete list of examples, visit the following documentation: Securely load data using dedicated SQL pools.

```
SQL
-- B: Create and execute the COPY statement


COPY INTO [dbo].[DimProduct]
--The column list allows you map, omit, or reorder input file columns to target table columns.
--You can also specify the default value when there is a NULL value in the file.
--When the column list is not specified, columns will be mapped based on source and target ordinality
(
        ProductKey default -1 1,
        ProductLabel default 'myStringDefaultWhenNull' 2,
        ProductName default 'myStringDefaultWhenNull' 3
)
--The storage account location where you data is staged
FROM 'https://storageaccount.blob.core.windows.net/container/directory/'
WITH
(
   --CREDENTIAL: Specifies the authentication method and credential access your storage account
   CREDENTIAL = (IDENTITY = '', SECRET = ''),
   --FILE_TYPE: Specifies the file type in your storage account location
   FILE_TYPE = 'CSV',
   --FIELD_TERMINATOR: Marks the end of each field (column) in a delimited text (CSV) file
   FIELDTERMINATOR = '|',
   --ROWTERMINATOR: Marks the end of a record in the file
   ROWTERMINATOR = '0x0A',
   --FIELDQUOTE: Specifies the delimiter for data of type string in a delimited text (CSV) file
   FIELDQUOTE = '',
   ENCODING = 'UTF8',
   DATEFORMAT = 'ymd',
   --MAXERRORS: Maximum number of reject rows allowed in the load before the COPY operation is canceled
   MAXERRORS = 10,
   --ERRORFILE: Specifies the directory where the rejected rows and the corresponding error reason should be written
   ERRORFILE = '/errorsfolder',
) OPTION (LABEL = 'COPY: ADLS tutorial');
```

Optimize columnstore compression

By default, tables are defined as a clustered columnstore index. After a load completes, some of the data rows might not be compressed into the columnstore. There's a variety of reasons why this can happen. To learn more, see manage columnstore indexes.

To optimize query performance and columnstore compression after a load, rebuild the table to force the columnstore index to compress all the rows.

```
SQL

ALTER INDEX ALL ON [dbo].[DimProduct] REBUILD;
```

Optimize statistics

It is best to create single-column statistics immediately after a load. There are some choices for statistics. For example, if you create single-column statistics on every column it might take a long time to rebuild all the statistics. If you know certain columns are not going to be in query predicates, you can skip creating statistics on those columns.

If you decide to create single-column statistics on every column of every table, you can use the stored procedure code sample `prc_sqldw_create_stats` in the statistics article.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store#create-the-target-table

Question 61: Skipped

Once Azure Synapse Link is configured on Cosmos DB, what is the first step to perform to use Azure Synapse Analytics serverless SQL pools to query the Azure Cosmos DB data?

- ○ None of the listed options
- ○ Use the `OPENROWSET` function
- ○ `CREATE` database (Correct)
- ○ Use a `SELECT` clause

Explanation

Before being able to issue any queries using Azure Synapse Analytics serverless SQL pools, you first must create a database.

https://docs.microsoft.com/en-us/azure/azure-monitor/insights/azure-sql

Question 62: Skipped

Scenario: You are working as a consultant at Avengers Security and advising the IT team on the design of a hybrid solution to synchronize data and on-premises Microsoft SQL Server database to Azure SQL Database.

Required: An assessment of databases must be done in order to determine whether or not data will move without compatibility issues.

The Avengers IT team has many different tools at their disposal and it is your responsibility to advise them on which tool to use. Which of the following is the best for the application?

- ○ Microsoft Assessment and Planning Toolkit
- ○ SQL Vulnerability Assessment (VA)
- ○ SQL Server Migration Assistant (SSMA)
- ○ Data Migration Assistant (DMA) **(Correct)**

Explanation

The Data Migration Assistant (DMA) helps you upgrade to a modern data platform by detecting compatibility issues that can impact database functionality in your new version of SQL Server or Azure SQL Database. DMA recommends performance and reliability improvements for your target environment and allows you to move your schema, data, and uncontained objects from your source server to your target server.

Data Migration Assistant is a client-side tool that you can install on a Windows-compatible workstation or server. It has two major functions in the migration of the social database to the Azure SQL Database platform in this module.

• First, it assesses your existing database and identifies any incompatibilities between that database and Azure SQL Database.

• It then generates a report of the things you need to fix before you can migrate.

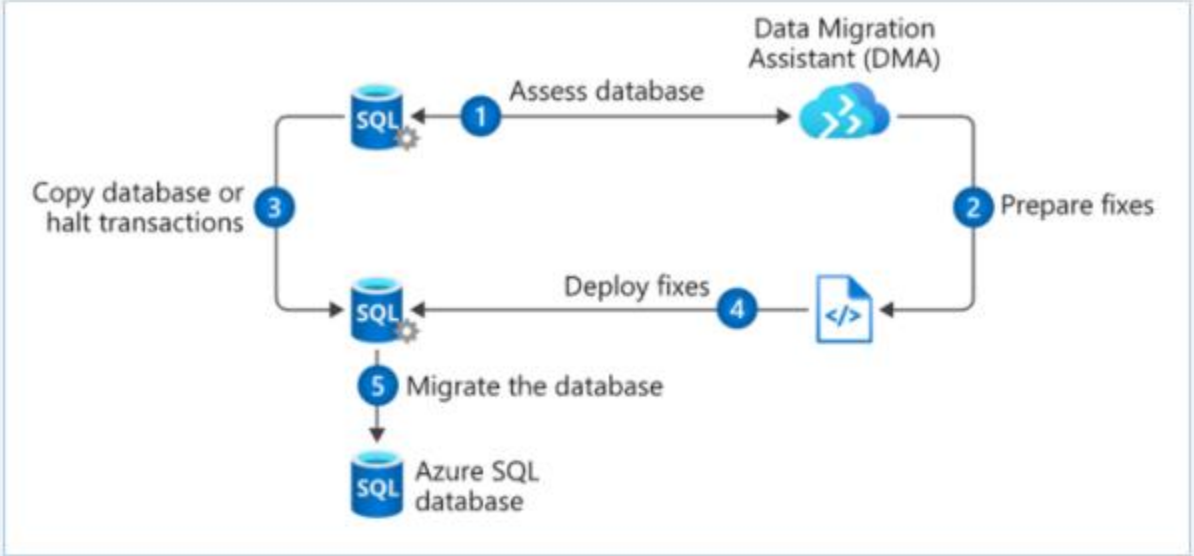
As you make changes, you can rerun Data Migration Assistant to generate an updated report of changes that you need to make. This capability helps you to not only track your progress, but also catch any new issues that might have been introduced during your coding phase.

Migration process overview

Migrating your company's social media database is a multi-step process. The workflow begins with a *pre-migration* phase, in which you determine which databases need to be migrated. You also look for any compatibility issues between your existing database and Azure SQL Database.

After you resolve any incompatibility issues, you're ready for the *migration* phase. First, you migrate the schema to the Azure SQL Database Service. Then, you're ready to migrate the data itself by using Azure Database Migration Service.

The last step in your workflow is the *post-migration* phase. During this phase, you do any required testing. Then you update apps, reports, and other tools that will need to use the new database for their data.



Pre-migration

The pre-migration phase begins with *discovery*, or taking inventory of your existing databases and the tools and apps that rely on them. For this simple exercise, we're concerned with only a single social database. In practice, it can be a much more complex step.

You need to identify everything that uses your existing database. Apps, SQL Server Report Services reports, Power BI reports, and export jobs written in PowerShell are all examples of things to note so you can update them, after the migration, to point to the new Azure SQL Database.

The second step in the pre-migration phase is the *assessment*. During the assessment, you examine the database for any incompatibilities between it and the Azure SQL Database platform. Because this can be a difficult task to perform manually, Microsoft has provided Data Migration Assistant. You can use Data Migration Assistant to automatically examine your source database for any compatibility issues with Azure SQL Database.

Data Migration Assistant provides a report that you can use as a guide to update your database. As you make changes, you can rerun Data Migration Assistant to track your progress and to uncover any new issues that might arise as you make changes. The assessment phase is covered in steps 1 and 2 of the migration workflow previously illustrated.

The final stage in the pre-migration is *convert*. In the convert phase, you make any changes for compatibility that Data Migration Assistant has recommended. Then, you create the SQL scripts for deploying to the Azure SQL Database. Data Migration Assistant can be of help to you here as well. It generates all of the SQL scripts needed to deploy your schema to the target Azure SQL Database.

Migration

The migration phase involves migrating two elements: *schema* and *data*. In the convert phase of pre-migration, the Data Migration Assistant tool generated all of the code. Data Migration Assistant can run these scripts for you. Or, you can save these scripts, and run them on your own by using a tool such as SQL Server Management Studio, Azure Data Studio, or the sqlcmd utility. The schema migration can be found in step 4 of the migration workflow.

After your database schema has been migrated, you're ready to migrate your data (steps 3 and 5 in the workflow). For this step, you'll use Azure Database Migration Service to move your data up to the Azure SQL Database Service.

Database Migration Service can be run in two modes, online and offline. When it's running in online mode, there are two additional steps. The first is *sync*, in which any changes made to the data in the source system after the migration are brought into the target database. The other is *cutover*, in which the source database is taken offline, and the new Azure SQL Database becomes available.

Post-migration

The post-migration phase is a process that consists of several steps. First, you need to remediate any applications, updating any affected by the database changes. For example, you might need to update the connection strings to point to the new Azure SQL Database.

In addition, make sure there's thorough and complete testing. Validation testing will ensure that your application did not break because of changes at the database level. Construct tests to return data from both the source and target. Compare the data to ensure that queries are returning from the Azure SQL Database as they would with the original source database. Next, create performance tests that will:

- Validate that your application returns data in the times required by your organization.

- Enable you to do further optimizations, if necessary.

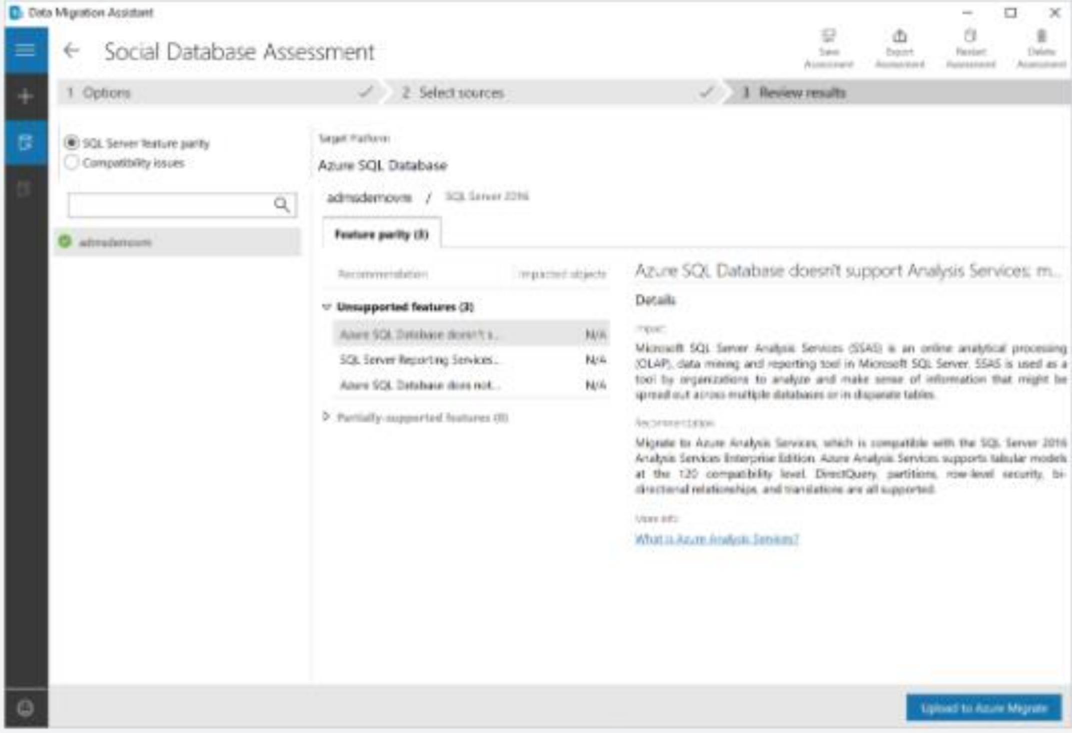
The post-migration phase is critical because it ensures that your data is both accurate and complete. In addition, it alerts you to any performance issues that might arise with the workload in the new environment.

Data migration tools in Azure

The core of data migration in Azure is the Azure Database Migration Service. You can use this service to move bulk amounts of data in a timely way. As part of Database Migration Service, Microsoft provides Data Migration Assistant. Just as its name implies, Data Migration Assistant *assists* the service by preparing the target database.

Data Migration Assistant

Data Migration Assistant is a client-side tool that you can install on a Windows-compatible workstation or server. It has two major functions in the migration of the social database to the Azure SQL Database platform in this module.



● SQL Server feature parity
○ Compatibility issues

admsdemovm

Target Platform
Azure SQL Database

admsdemovm / SQL Server 2016

Feature parity (3)

Recommendation	Impacted objects
Unsupported features (3)	
Azure SQL Database doesn't s...	N/A
SQL Server Reporting Services...	N/A
Azure SQL Database does not...	N/A
Partially-supported features (0)	

Azure SQL Database doesn't support Analysis Services: m...

Details

Impact

Microsoft SQL Server Analysis Services (SSAS) is an online analytical processing (OLAP), data mining and reporting tool in Microsoft SQL Server. SSAS is used as a tool by organizations to analyze and make sense of information that might be spread out across multiple databases or in disparate tables.

Recommendation

Migrate to Azure Analysis Services, which is compatible with the SQL Server 2016 Analysis Services Enterprise Edition. Azure Analysis Services supports tabular models at the 120 compatibility level. DirectQuery, partitions, row-level security, bi-directional relationships, and translations are all supported.

More info:

What is Azure Analysis Services?

First, it assesses your existing database and identifies any incompatibilities between that database and Azure SQL Database. It then generates a report of the things you need to fix before you can migrate. As you make changes, you can rerun Data Migration Assistant to generate an updated report of changes that you need to make. This capability helps you to not only track your progress, but also catch any new issues that might have been introduced during your coding phase.

After Data Migration Assistant completes the assessment and you've made any changes, you need to migrate the database schema to Azure SQL Database. Data Migration Assistant can help with this as well. It generates the required SQL, and then gives you the option of running the code, or saving it so you can run it yourself later.

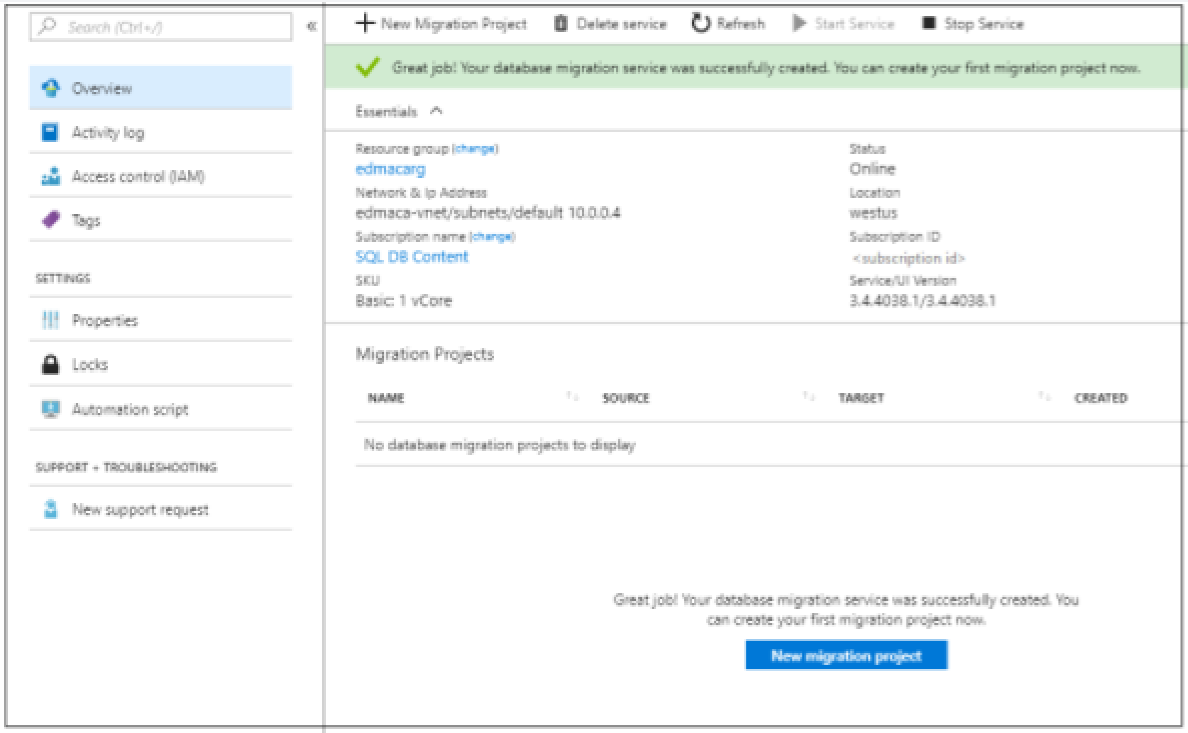
Using Data Migration Assistant is not a requirement to use Azure Database Migration Service. You have the option of coding your new database in the Azure SQL Database service manually without trying to convert an existing database.

As an example, let's say you're creating a staging database in Azure SQL Database that will later feed data into Azure Synapse Analytics. The staging database will be sourced from multiple systems, but it will migrate only small portions of the source data. In this situation, you might be better off manually crafting the new database directly on the Azure SQL Database service rather than trying to automate the job.

Azure Database Migration Service

After you've migrated your database schema by using Data Migration Assistant, or created a target database manually, you're ready to move your data. To do that, you'll use Azure Database Migration Service.

Azure Database Migration Service is a fully-managed Azure service that provides automated, seamless data migrations from multiple sources into the Azure data platforms.



✓ Great job! Your database migration service was successfully created. You can create your first migration project now.

Resource group (change)
edmacarg
Network & Ip Address
edmaca-vnet/subnets/default 10.0.0.4
Subscription name (change)
SQL DB Content
SKU
Basic: 1 vCore

Status
Online
Location
westus
Subscription ID
<subscription id>
Service/UI Version
3.4.4038.1/3.4.4038.1

Migration Projects

NAME	SOURCE	TARGET	CREATED
No database migration projects to display			

Great job! Your database migration service was successfully created. You can create your first migration project now.

Database Migration Service runs on the Azure platform, as opposed to being a client application like Data Migration Assistant. It's capable of moving large amounts of data quickly and is not dependent upon installation of a client application. Database Migration Service can operate in two modes, offline and online.

In offline mode, no more changes can be made to your source database. Data is migrated, and then your applications can begin using the new Azure SQL Database.

In online mode, your source database can remain in use while the bulk of the data is migrated. At the end of the migration, you'll take the source system offline momentarily while any final changes to the source are synced to the new Azure SQL Database. At this point, your applications can cut over to use the SQL database.

https://docs.microsoft.com/en-us/sql/dma/dma-overview

Question 63: Skipped
A pipeline in Azure Data Factory represents a logical grouping of activities where the activities together perform a certain task.

Which of the following are valid dependency conditions? (Select four)

- □
  Completed
  (Correct)
- □
  Pending
- □
  Working
- □
  Queue
- □
  Succeeded
  (Correct)
- □
  Skipped
  (Correct)
- □
  Failed
  (Correct)
- □
  Running

Explanation
Azure Data Factory

In order to work with data factory pipelines, it is imperative to understand what a pipeline in Azure Data Factory is.

A pipeline in Azure Data Factory represents a logical grouping of activities where the activities together perform a certain task.

An example of a combination of activities in one pipeline can be, ingesting and cleaning log data in combination with a mapping data flow that analyzes the log data that has been cleaned.

A pipeline enables you to manage the separate individual activities as a set, which would otherwise be managed individually. It enables you to deploy and schedule the activities efficiently, through the use of a single pipeline, versus managing each activity independently.

Activities in a pipeline are referred to as actions that you perform on your data. An activity can take zero or more input datasets and produce one or more output datasets.

An example of an action can be the use of a copy activity, where you copy data from an Azure SQL Database to an Azure DataLake Storage Gen2. To build on this example, you can use a data flow activity or an Azure Databricks Notebook activity for processing and transforming the data that was copied to your Azure Data Lake Storage Gen2 account, in order to have the data ready for business intelligence reporting solutions like in Azure Synapse Analytics.

Since there are many activities that are possible in a pipeline in Azure Data Factory, we have grouped the activities in three categories:

- *Data movement activities*: the Copy Activity in Data Factory copies data from a source data store to a sink data store.

- *Data transformation activities*: Azure Data Factory supports transformation activities such as Data Flow, Azure Function, Spark, and others that can be added to pipelines either individually or chained with another activity.

- *Control activities*: Examples of control flow activities are 'get metadata', 'For Each', and 'Execute Pipeline'.

Activities can depend on each other. What we mean, is that the activity dependency defines how subsequent activities depend on previous activities. The dependency itself can be based on a condition of whether to continue in the execution of previous defined activities in order to complete a task. An activity that depends on one or more previous activities, can have different dependency conditions.

The four dependency conditions are:

- Succeeded

- Failed

- Skipped

- Completed

For example, if a pipeline has an Activity A, followed by an Activity B and Activity B has as a dependency condition on Activity A 'Succeeded', then Activity B will only run if Activity A has the status of succeeded.

If you have multiple activities in a pipeline and subsequent activities are not dependent on previous activities, the activities may run in parallel.

https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/

Question 64: Skipped

Scenario: You are determining the type of Azure service needed to fit the following specifications and requirements:

Data classification: Structured

Operations: Read-only, complex analytical queries across multiple databases

Latency & throughput: Some latency in the results is expected based on the complex nature of the queries.

Transactional support: Not required

- ○ Azure Route Table
- ○ Azure Cosmos DB
- ○ Azure Queue Storage
- ○ Azure Blob Storage
- ○ Azure SQL Database (Correct)

Explanation

Recommended service: Azure SQL Database

Business data will most likely be queried by business analysts, who are more likely to know SQL than any other query language. Azure SQL Database could be used as the solution by itself, but pairing it with Azure Analysis Services enables data analysts to create a semantic model over the data in SQL Database. The data analysts can then share it with business users, so that they only need to connect to the model from any business intelligence (BI) tool to immediately explore the data and gain insights.

Why not other Azure services?

Azure Synapse supports OLAP solutions and SQL queries. But your business analysts will need to perform cross-database queries, which Azure Synapse does not support.

Azure Analysis Services could be used in addition to Azure SQL Database. But your business analysts are more well-versed in SQL than in working with Power BI. So they'd like a database that supports SQL queries, which Azure Analysis Services does not. In addition, the financial data you're storing in your business data set is relational and multidimensional in nature. Azure Analysis Services supports tabular data stored on the service itself, but not multidimensional data. To analyze multidimensional data with Azure Analysis Services, you can use a direct query to the SQL Database.

Azure Stream Analytics is a great way to analyze data and transform it into actionable insights, but its focus is on real-time data that is streaming in. In this scenario, the business analysts are looking at historical data only.

https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-database-paas-overview

Question 65: Skipped

Scenario: You are working in an Azure Databricks workspace and you want to filter by a productType column where the value is equal to book.

Which command meets the requirement by specifing a column value in a DataFrame's filter?

- ○ `df.filter("productType == 'book'")`
- ○ `df.filter(col("productType") == "book")` **(Correct)**
- ○ `df.col("productType").filter("book")`
- ○ `df.filter("productType = 'book'")`

Explanation

The `df.filter(col("productType") == "book")` approach is the correct way to apply the filter, by using the Column Class.

https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html

Question 66: Skipped

Scenario: You are working on a new project and creating storage accounts and blob containers for your application.

Which of the below describes a good strategy for doing this?

- ○ Create Azure Storage accounts before deploying your app. Create containers in your application as needed. (Correct)
- ○ Create Azure Storage accounts in your application as needed. Create the containers before deploying the application.
- ○ Create both your Azure Storage accounts and containers before deploying your application.
- ○ None of the listed options.
- ○ All the listed options.

Explanation

Creating an Azure Storage account is an administrative activity and can be done prior to deploying an application. Container creation is lightweight and is often driven by run-time data which makes it a good activity to do in your application.

https://docs.microsoft.com/en-us/azure/storage/common/storage-account-create?tabs=azure-portal

Question 67: Skipped

Scenario: You work in an organization where much of the transformation logic is currently held in existing SSIS packages that have been created on SQL Server. Since your boss is not familiar with Azure as well as you are, he tells you he has heard that Azure has the ability to lift and shift SSIS package so to execute them within Azure Data

Factory to leverage existing work. He asks you *“What do we need to setup in order to do this?”*

Which of the below is the correct response?

- ○
  None of the listed options.
- ○
  In order to do this you must set up an Azure-SSIS integration runtime.
  (Correct)
- ○
  In order to do this you must set up a Self-hosted solution and then upload the data.
- ○
  Your boss is mistaken, Azure does not have the ability to lift and shift SSIS package so to execute them within Azure Data Factory, it must be converted to AZ format and then ingested via Azure Storage.
- ○
  In order to do this you must set up an Azure Stored procedure to execute the lift and shift.

Explanation
You may work in an organization where much of the transformation logic is currently held in existing SSIS packages that have been created on SQL Server. You have the ability to lift and shift SSIS package so you can execute them within Azure Data Factory, so you can make use in existing work. In order to do this you must set up an Azure-SSIS integration runtime.

Azure-SSIS integration runtime

In order to make use of the Azure-SSIS integration runtime, it is assumed that there is SSIS Catalog (SSISDD) deployed on a SQL Server SSIS instance. With that prerequisite met, the Azure-SSIS integration runtime is capable of:

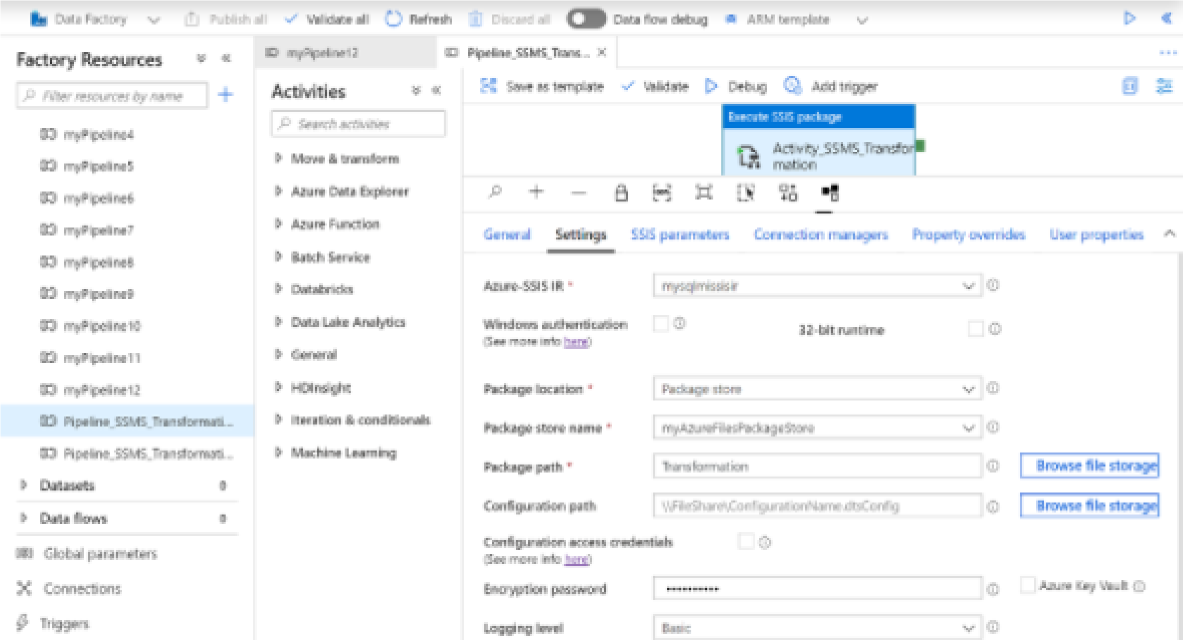
• Lift and shift existing SSIS workloads

During the provisioning of the Azure-SSIS integration runtime, you specify the following options:

• The node size (including the number of cores) and the number of nodes in the cluster.

• The existing instance of Azure SQL Database to host the SSIS Catalog Database (SSISDB), and the service tier for the database.

• The maximum parallel executions per node.



Azure-SSIS IR * mysqlmissisir

Windows authentication (See more info here) □

32-bit runtime □

Package location * Package store

Package store name * myAzureFilesPackageStore

Package path * Transformation

Configuration path \\FileShare\ConfigurationName.dtsConfig

Configuration access credentials (See more info here) □

Encryption password ••••••••••

□ Azure Key Vault

Logging level Basic

With the Azure-SSIS integration runtime enabled, you are able to manage, monitor and schedule SSIS packages using tools such as SQL Server Management Studio (SSMS) or SQL Server Data Tools (SSDT).

https://docs.microsoft.com/en-us/azure/data-factory/azure-ssis-integration-runtime-package-store

Question 68: Skipped
When is it possible to add or remove datasets if created with Azure Data Share?

- ○ It is not possible to add or remove datasets if created with Azure Data Share.
- ○ It is only possible to remove or add datasets before it's sent within Azure Data Share.
- ○ None of the listed options.
- ○

It is possible to add or remove datasets within Azure Data Share after it has been created.
(Correct)

Explanation
It is possible to add or remove datasets after it has been created in Azure Data Share.

https://docs.microsoft.com/en-us/azure/data-factory/lab-data-flow-data-share

Question 69: Skipped
Azure Cosmos DB is a globally distributed, multimodel database. Which of the following can be used to deploy it?

- □
  Cassandra API
  (Correct)
- □
  Gremlin API
  (Correct)
- □
  T-SQL API
- □
  Table API
  (Correct)
- □
  SQL API
  (Correct)
- □
  ABS API
- □
  ADLS API
- □
  U-SQL API
- □
  MongoDB API
  (Correct)

Explanation
Azure Cosmos DB is a globally distributed, multimodel database. You can deploy it by using several API models:

• SQL API

• MongoDB API

• Cassandra API

• Gremlin API

• Table API

Because of the multimodel architecture of Azure Cosmos DB, you benefit from each model's inherent capabilities. For example, you can use MongoDB for semistructured data, Cassandra for wide columns, or Gremlin for graph databases. When you move your data from SQL, MongoDB, or Cassandra to Azure Cosmos DB, applications that are built using the SQL, MongoDB, or Cassandra APIs will continue to operate.

https://docs.microsoft.com/en-us/azure/cosmos-db/faq

Question 70: Skipped
In Data Factory, an Activity defines the action to be performed. A Linked Service defines a target data store or a compute service. An Integration Runtime (IR) provides the bridge between the Activity and Linked Services.

In order to make use of the Azure-SSIS Integration Runtime, it is assumed that there is SSIS Catalog (SSISDB) deployed on a SQL Server SSIS instance. With that prerequisite met, the Azure-SSIS Integration Runtime is capable of lifting and shifting existing SSIS workloads

During the provisioning of the Azure-SSIS Integration Runtime, which are the options that must be specified? (Select all that apply)

- ☐ All the listed options
- ☐ Node size
  (Correct)
- ☐ IP address(es) of the nodes
- ☐ VM regions
- ☐ Database (SSISDB) along with the service tier for the database

(Correct)

- □
  Private Link parameters
- □
  Maximum parallel executions per node
  (Correct)
- □
  Existing instance of Azure SQL Database to host the SSIS Catalog
  (Correct)

Explanation
Integration Runtime

In Data Factory, an Activity defines the action to be performed. A Linked Service defines a target data store or a compute service. An Integration Runtime (IR) provides the bridge between the Activity and Linked Services.

Azure-SSIS Integration Runtime

To lift and shift existing SSIS workload, you can create an Azure-SSIS IR to natively execute SSIS packages. Selecting the right location for your Azure-SSIS IR is essential to achieve high performance in your extract-transform-load (ETL) workflows.

• The location of your Azure-SSIS IR does not need to be the same as the location of your data factory, but it should be the same as the location of your own Azure SQL Database or Azure SQL Database managed instance server where SSISDB is to be hosted. This way, your Azure-SSIS Integration Runtime can easily access SSISDB without incurring excessive traffics between different locations.

• If you do not have an existing Azure SQL Database or Azure SQL Database managed instance server to host SSISDB, but you have on-premises data sources/destinations, you should create a new Azure SQL Database or Azure SQL Database managed instance server in the same location of a virtual network connected to your on-premises network. This way, you can create your Azure-SSIS IR using the new Azure SQL Database or Azure SQL Database managed instance server and joining that virtual network, all in the same location, effectively minimizing data movements across different locations.

• If the location of your existing Azure SQL Database or Azure SQL Database managed instance server where SSISDB is hosted is not the same as the location of a virtual network connected to your on-premises network, first create your Azure-SSIS IR using an existing Azure SQL Database or Azure SQL Database managed instance server and joining another virtual network in the same location, and then configure a virtual network to virtual network connection between different locations.

In order to make use of the Azure-SSIS Integration Runtime, it is assumed that there is SSIS Catalog (SSISDB) deployed on a SQL Server SSIS instance. With that prerequisite met, the Azure-SSIS Integration Runtime is capable of lifting and shifting existing SSIS workloads

During the provisioning of the Azure-SSIS Integration Runtime, you specify the following options:

• The node size (including the number of cores) and the number of nodes in the cluster.

• The existing instance of Azure SQL Database to host the SSIS Catalog Database (SSISDB), and the service tier for the database.

• The maximum parallel executions per node.

With the Azure-SSIS Integration Runtime enabled, you are able to manage, monitor, and schedule SSIS packages using tools such as SQL Server Management Studio (SSMS) or SQL Server Data Tools (SSDT).

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

Question 71: Skipped

True or False: Azure Storage encrypts all data that's written to it. It is not necessary to enable encryption within your subscription.

- ○ True (Correct)
- ○ False

Explanation

Azure Storage Data security

Azure Storage encrypts all data that's written to it. Azure Storage also provides you with fine-grained control over who has access to your data. You'll secure the data by using keys or shared access signatures.

Azure Resource Manager provides a permissions model that uses role-based access control (RBAC).

https://docs.microsoft.com/en-us/azure/storage/common/storage-introduction

Question 72: Skipped

Data engineers use Azure Stream Analytics to process streaming data and respond to data anomalies in real time. You can use Stream Analytics for Internet of Things (IoT) monitoring, web logs, remote patient monitoring, and point of sale (POS) systems.

Stream Analytics can route job output to which of the following storage systems? (Select all that apply)

- □ Azure SQL Datawarehouse
- □ Azure Storage Explorer
- □ Azure SQL Database (Correct)
- □ Azure Data Lake Storage (Correct)
- □ Azure Cosmos DB (Correct)
- □ Azure Table Storage
- □ Azure Blob Storage (Correct)

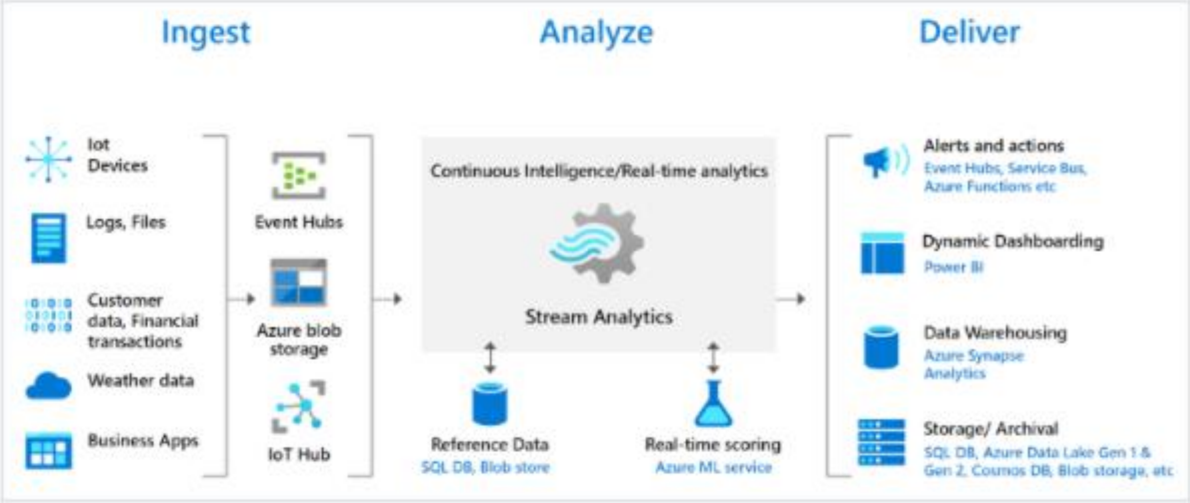
Explanation

Applications, sensors, monitoring devices, and gateways broadcast continuous event data known as *data streams*. Streaming data is high volume and has a lighter payload than nonstreaming systems.

Data engineers use Azure Stream Analytics to process streaming data and respond to data anomalies in real time. You can use Stream Analytics for Internet of Things (IoT) monitoring, web logs, remote patient monitoring, and point of sale (POS) systems.

Data processing

To process streaming data, set up Stream Analytics jobs with input and output pipelines. Inputs are provided by Event Hubs, IoT Hubs, or Azure Storage. Stream Analytics can route job output to many storage systems. These systems include Azure Blob, Azure SQL Database, Azure Data Lake Storage, and Azure Cosmos DB.



After storing the data, run batch analytics in Azure HDInsight. Or send the output to a service like Event Hubs for consumption. Or use the Power BI streaming API to send the output to Power BI for real-time visualization.

Queries

To define job transformations, use a simple, declarative Stream Analytics query language. The language should let you use simple SQL constructs to write complex temporal queries and analytics.

The Stream Analytics query language is consistent with the SQL language. If you're familiar with the SQL language, you can start creating jobs.

Data security

Stream Analytics handles security at the transport layer between the device and Azure IoT Hub. Streaming data is generally discarded after the windowing operations finish. Event Hubs uses a shared key to secure the data transfer. If you want to store the data, your storage device will provide security.

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

Question 73: Skipped
What is a step in flattening a nested schema?

- ○
  `LOAD` CSV file
- ○
  `CREATE` parquet file
- ○
  Explode Arrays
  (Correct)
- ○
  `COPY` data

Explanation
Explode Arrays is a third step in flattening nested schema's. It is necessary to transform the array in the data frame into a new dataframe where the column that you want to select is defined.

A DataFrame creates a data structure and it's one of the core data structures in Spark. In Spark, it is seen as a distributed collection of data that is organized into columns that have names.

Some use cases for transforming complex data types are as follows:

• Complex data types are increasingly common and represent a challenge for data engineers as analyzing nested schema and arrays tend to include time-consuming and complex SQL queries.

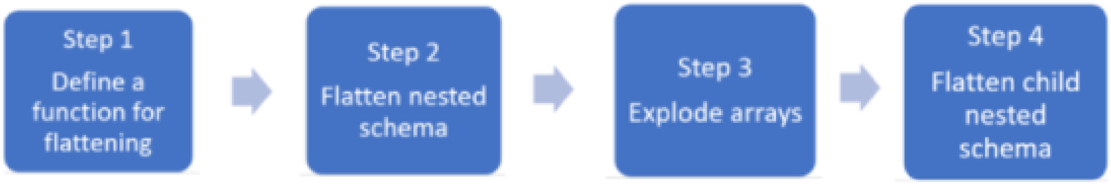
• It can be difficult to rename or cast the nested columns data type.

• Performance issues arise when working with deeply nested objects.

• Data Engineers need to understand how to efficiently process complex data types and make them easily accessible to everyone.

Synapse Spark can be used to read and transform objects into a flat structure through data frames. Synapse SQL serverless can be used to query such objects directly and return those results as a regular table. With Synapse Spark, it's easy to transform nested structures into columns and array elements into multiple rows.

In the overview below, the steps show the techniques involved to deal with complex data types:



• Step 1: Define a function for flattening We define a function to flatten the nested schema.

• Step 2: Flatten nested schema Use the function to flatten the nested schema of the data frame (df) into a new data frame.

• Step 3: Explode Arrays Transform the array in the data frame into a new dataframe where you also define the column that you want to select.

• Step 4: Flatten child nested Schema Use the function you create to flatten the nested schema of the data frame into a new data frame.

https://medium.com/@saikrishna_55717/flattening-nested-data-json-xml-using-apache-spark-75fa4c8ea2a7

Question 74: Skipped
Azure Data Lake Storage Gen2 plays a fundamental role in a wide range of big data architectures. There are stages for processing big data solutions that are common to all architectures.

Which are they? (Select four)

- □ Model and serve (Correct)
- □ Ingestion (Correct)
- □ Streamed

- □
  Store
  (Correct)
- □
  Relational
- □
  Prep and train
  (Correct)
- □
  Clusters

Explanation

Azure Data Lake Storage Gen2 plays a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:

• A modern data warehouse.

• Advanced analytics against big data.

• A real-time analytical solution.

There are four stages for processing big data solutions that are common to all architectures:

• **Ingestion** - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the Data Lake Store. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate technology to use.

• **Store** - The store phase identifies where the ingested data should be placed. In this case, we're using Azure Data Lake Storage Gen2.

• **Prep and train** - The prep and train phase identifies the technologies that are used to perform data preparation and model training and scoring for data science solutions. The common technologies that are used in this phase are Azure Databricks, Azure HDInsight or Azure Machine Learning Services.

• **Model and serve** - Finally, the model and serve phase involves the technologies that will present the data to users. These can include visualization tools such as Power BI, or other data stores such as Azure Synapse Analytics, Azure Cosmos DB, Azure SQL

Database, or Azure Analysis Services. Often, a combination of these technologies will be used depending on the business requirements.

https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-data-scenarios

Question 75: Skipped

Which Dynamic Management View enables the view the active connections against a dedicated SQL pool?

- ○ `sys.dm_pdw_nodes_os_connection_counters`
- ○ `sys.dm_pdw_dms_workers`
- ○ `DBCC PDW_SHOWEXECUTIONPLAN`
- ○ `sys.dm_pdw_exec_sessions`
- ○ `sys.dm_pdw_nodes_exec_connection`
- ○ `sys.dm_pdw_dms_workers`
- ○ `sys.dm_pdw_request_steps`
- ○ `sys.dm_pdw_exec_requests` (Correct)

Explanation

`sys.dm_pdw_exec_requests` enables you to view the active connections against a dedicated SQL pool

https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-pdw-exec-requests-transact-sql?view=aps-pdw-2016-au7

Question 76: Skipped

Which is one of the possible ways to optimize a Spark Job?

- ○

Remove all nodes

- ○
  Remove the Spark Pool
- ○
  Use the local cache option
- ○
  Use bucketing
  (Correct)
- ○
  None of the listed options

Explanation

The way bucketed tables are optimized is because it's because the metadata about how it was bucketed and sorted are stored.

Once you have checked the monitor tab within the Azure Synapse Studio environment, and feel that you could improve the performance of the run, you have several things to take in mind:

• Choose the data abstraction

• Use the optimal data format

• Use the cache option

• Check the memory efficiency

• Use Bucketing

• Optimize Joins and Shuffles if appropriate

• Optimize Job Execution

In order to optimize the Apache Spark Jobs in Azure Synapse Analytics, you need to take into account the cluster configuration for the workload you're running on that cluster. You might run into challenges where memory pressure (if not configured well, like not choosing the right size of executors), long running operations and tasks that might result in Cartesian operations. If you want to speed up the jobs, you'd have to configure the appropriate caching for that task, as well as checking joins and shuffles in relation to data skew. Therefore, it is so imperative that you monitor and review Spark Job executions that are long running or resource-consuming.

Some recommendations in order for you to optimize the Spark Job might include the following:

Choosing the data abstraction

Some of the earlier Spark versions use RDDs to abstract the data. Spark 1.3 and 1.6 introduced the use of DataFrames and Datasets. The following relative merits might help you to optimize in relation to your data abstraction:

DataFrames Using DataFrames would be a great place to start. DataFrames provide query optimization through Catalyst. It also includes a whole-stage code generation with direct memory access. One thing to take in mind is that when you want to have the best-developer-friendly experience it might be better to use DataSets, since there are no compile-time checks or domain object programming.

On that note, let's look into DataSets: *DataSets are good in complex ETL pipelines optimization where the performance impact is acceptable. Just be cautious when using DataSets in aggregations, since it might impact the performance. However, it will provide query optimization through Catalyst and is developer-friendly by providing object programming and compile-time checks. DataSets do add serialization/deserialization overhead and has a high GC overhead.

Looking at RDDs we would advise as follows: It is not necessary to use RDDs unless you want or need to build a new custom RDD. However, there is no query optimization through Catalyst as well as no whole-stage code generation and would still have a high GC overhead. The only way to use RDDs is that it needs SPark 1.x legacy APIs.

When looking at your data format, spark provides many. Formats that you can use are csv, json, xml, parquet etc. It can also be extended by other formats with external data sources. A tip that might be useful is using parquet with snappy compression (which also happen to be the default in Spar 2.x.) Why Parquet? It stores data in a columnar format, is compressed and highly optimized in Spark, as well as, splittable in order to decompress.

When it comes to the caching, there is a native built in Spark caching mechanism. It can be used through different methods like: `.persist()`, `.cache(`), and `CACHE TABL`E. When using small datasets, it might be effective. In ETL pipelines where caching of intermediate results is necessary this might come in handy too. Just take in mind that is you need to do partitioning, the spark native caching mechanism might have some downsides. The reason for that is that a cached table won't keep the partitioning data.

It is also imperative to understand how to use the memory efficiently. What you have to understand is that Spark operates by placing data in memory. Therefore, managing memory resources is an aspect of optimizing Spark jobs executions. The way to manage it, might be to check smaller data partitions and checking data size, types and distributions when you formulate a partitioning strategy. Another way to optimize is to consider Kryo data serialization: Kryo data serialization, versus the default Java serialization. Always bear in mind though, to keep monitoring and tuning the Spark configuration settings.

Another thing to look at might be bucketing.

Use bucketing

Bucketing is almost the same as data partitioning. The way it differs is that a bucket holds a set of column values instead of one. It might work well when you partition on large (millions or more) values like product identifiers. A bucket is determined by hashing the bucket key of a row. The way bucketed tables are optimized is because it's because the metadata about how it was bucketed and sorted are stored.

Some advanced bucketing features are:

- Query optimization based on bucketing meta-information.

- Optimized aggregations.

- Optimized joins.

However, bucketing doesn't exclude partitioning. They go hand in had. You can use partitioning and bucketing at the same time.

Optimize joins and shuffles

Sometimes, when you have a slower performance on join or shuffle jobs, it can be caused by data skew. What is data skew? It's asymmetry in your job data. An example might be that a job only takes 20 sec regularly, however running the same job where data is joined and shuffled can take up hours. In order to fix that data skew, you can salt the entire key, or use an isolated salt for only some subset of keys. Another option to look into might be the introduction of a bucket column and pre-aggregate in buckets first. However, there's more to causing slow joins, since it might be the join type. Spark uses the SortMerge join type. This type of join is best suited for large data sets, but is otherwise computationally expensive because it must first sort the left and right sides of data before merging them. Therefore, a Broadcast join might be better suited for smaller data sets, or where one side of the join is much smaller than the other side.

You can change the join type in your configuration by setting spark.sql.autoBroadcastJoinThreshold, or you can set a join hint using the DataFrame APIs (`dataframe.join(broadcast(df2))`).

```
Scala
// Option 1
spark.conf.set("spark.sql.autoBroadcastJoinThreshold", 1*1024*1024*1024)


// Option 2
val df1 = spark.table("FactTableA")
val df2 = spark.table("dimMP")
df1.join(broadcast(df2), Seq("PK")).
createOrReplaceTempView("V_JOIN")


sql("SELECT col1, col2 FROM V_JOIN")
```

If you did decide to use bucketed tables, you will have a third join type, the Merge join. A correctly pre-partitioned and pre-sorted dataset will skip the expensive sort phase from a SortMerge join. Another thing to take in mind is that the order of the different type of joins does matter, especially in complex queries. Therefore, it's advised to start with the most selective joins. In addition, try to move joins that increase the number of rows after aggregations when possible.

Looking at the sizing of executors in order to increase performance in your spark job, you could consider the Java garbage Collection Overhead (GC) overhead.

- Factors to reduce executor size:
    - Reduce heap size below 32 GB to keep GC overhead < 10%.
    - Reduce the number of cores to keep GC overhead < 10%.
- Factors to increase executor size:
    - Reduce communication overhead between executors.
    - Reduce the number of open connections between executors (N2) on larger clusters (>100 executors).
    - Increase heap size to accommodate for memory-intensive tasks.
    - Optional: Reduce per-executor memory overhead.
    - Optional: Increase utilization and concurrency by oversubscribing CPU.

As a general rule of thumb when selecting the executor size:

• Start with 30 GB per executor and distribute available machine cores.

• Increase the number of executor cores for larger clusters (> 100 executors).

• Modify size based both on trial runs and on the preceding factors such as GC overhead.

When running concurrent queries, consider as follows:

• Start with 30 GB per executor and all machine cores.

• Create multiple parallel Spark applications by oversubscribing CPU (around 30% latency improvement).

• Distribute queries across parallel applications.

• Modify size based both on trial runs and on the preceding factors such as GC overhead.

As stated before, it's important to keep monitoring the performance, especially outliers, using the timeline view, SQL graph, job statistics etc. It might be a case where one of the executors is slower than the other, which most frequently happens on large clusters (30+ nodes). What you then might consider is to divide the work into more tasks such that the scheduler can compensate for the slower tasks.

If there is an optimization necessary in relation to the optimization of a job execution, make sure you keep in mind the caching (an example might be using the data twice, but cache it). IF you broadcast variables on all the executors you set up, due to the variables only being serialized once, you'll have faster lookups. In another case you might use the thread pool that runs on the driver, which could result in faster operations for many tasks.

https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-performance

Question 77: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

A window function enables you to perform a mathematical equation on a set of data that is defined within a window. The mathematical equation is typically an aggregate function; however, instead of applying the aggregate function to all the rows in a table, it is applied to a set of rows that are defined by the window function, and then the aggregate is applied to it.

One of the key components of window functions is the [?] clause. This clause determines the partitioning and ordering of a `rowset` before the associated window function is applied. That is, the [?] clause defines a window or user-specified set of rows within a query result set.

- ○ `UNDER`
- ○ `OVER` (Correct)
- ○ `HAVING`
- ○ `WHERE`

Explanation
A window function enables you to perform a mathematical equation on a set of data that is defined within a window. The mathematical equation is typically an aggregate function; however, instead of applying the aggregate function to all the rows in a table, it is applied to a set of rows that are defined by the window function, and then the aggregate is applied to it.

It is used to either perform calculations against a range of data, but it can also be used to programmatically define a deduplication of data technique, or paginate results.

One of the key components of window functions is the **`OVER`** clause. This clause determines the partitioning and ordering of a `rowset` before the associated window function is applied. That is, the OVER clause defines a window or user-specified set of rows within a query result set. A window function then computes a value for each row in the window. You can use the `OVER` clause with functions to compute aggregated values such as moving averages, cumulative aggregates, running totals, or a top N per group results.

```
SQL

-- Syntax for SQL Server, Azure SQL Database, and Azure Synapse Analytics
```

```
OVER (
[ <PARTITION BY clause> ]
[ <ORDER BY clause> ]
[ <ROW or RANGE clause> ]
)

<PARTITION BY clause> ::=
PARTITION BY value_expression , ... [ n ]

<ORDER BY clause> ::=
ORDER BY order_by_expression
[ COLLATE collation_name ]
[ ASC | DESC ]
[ ,...n ]

<ROW or RANGE clause> ::=
{ ROWS | RANGE } <window frame extent>

<window frame extent> ::=
{ <window frame preceding>
| <window frame between>
}
<window frame between> ::=
BETWEEN <window frame bound> AND <window frame bound>

<window frame bound> ::=
{ <window frame preceding>
| <window frame following>
}

<window frame preceding> ::=
{
```

```
UNBOUNDED PRECEDING
| <unsigned_value_specification> PRECEDING
| CURRENT ROW
}

<window frame following> ::=
{
UNBOUNDED FOLLOWING
| <unsigned_value_specification> FOLLOWING
| CURRENT ROW
}

<unsigned value specification> ::=
{ <unsigned integer literal> }
```

https://docs.microsoft.com/en-us/sql/t-sql/queries/select-over-clause-transact-sql?view=sql-server-ver15

You can then use aggregate functions with our window by expanding on our query that uses the OVER clause. The following aggregate functions are supported including `COUNT`, `MAX`, `AVG`, `SUM`, `APPROX_COUNT`, `DISTINCT`, `MIN`, `STDEV`, `STDEVP`, `STRING_AGG`, `VAR`, `VARP`, `GROUPING`, `GROUPING_ID`, `COUNT_BIG`, `CHECKSUM_AGG`

Alternatively, you can use analytical functions, which calculate an aggregate value based on a group of rows. Unlike aggregate functions, however, analytic functions can return multiple rows for each group. Use analytic functions to compute moving averages, running totals, percentages, or top-N results within a group. Supports `LAG`, `LEAD`, `FIRST_VALUE`, `LAST_VALUE`, `CUME_DIST`, `PERCENTILE_CONT`, `PERCENTILE_DISC`, `PERCENT_RANK`

You may want to use the `ROWS` and `RANGE` clauses to further limit the rows within the partition by specifying start and end points within the partition. This is done by specifying a range of rows with respect to the current row either by logical association or physical association. Physical association is achieved by using the `ROWS` clause. Supports `PRECEDING`, `UNBOUNDING PRECEDING`, `CURRENT ROW`, `BETWEEN`, `FOLLOWING`, `UNBOUNDED FOLLOWING`

Finally, window functions support Ranking functions like `RANK`, `NTILE`, `DENSE_RANK`, `ROW_NUMBER`.

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

Question 78: Skipped

Azure Data Factory provides a variety of methods for ingesting data, and also provides a range of methods to perform transformations.

Which of the following are valid options for transforming data within Azure Data Factory? (Select three)

- □ Data Storage Activities
- □ Analytic Flows
- □ Compute Resources (Correct)
- □ Test Lab Packages
- □ Control Resources
- □ Mapping Data Flows (Correct)
- □ SSIS Packages (Correct)
- □ Data Movement Flows

Explanation

Just as Azure Data Factory provides a variety of methods for ingesting data, it also provides a range of methods to perform transformations. You can pick a method that matches the skillsets of your team or takes advantage of existing technologies that you already have in your data estate. There is also the opportunity to perform transformations without writing code at all using the Mapping Data Flow.

Transforming data using Mapping Data Flow

Mapping Data Flows provide an environment for building a wide range of data transformations visually without the need to use code. The resulting data flows that are created are subsequently executed on scaled-out Apache Spark clusters that are automatically provisioned when you execute the Mapping Data Flow. Mapping Data Flows also provides the capability to monitor the execution of the transformations so that you can view how the transformations are progressing, or to understand any errors that may occur

Transforming data using compute resources

Azure Data Factory can also call on compute resources to transform data by a data platform service that may be better suited to the job. A great example of this is that Azure Data Factory can create a pipeline to an analytical data platform such as Spark pools in an Azure Synapse Analytics instance to perform a complex calculation using python. Another example could be to send data to an Azure SQL Database instance to execute a stored procedure using Transact-SQL. There is a wide range of compute resource, and the associated activities that they can perform as shown in the following table:

Compute environment: On-demand HDInsight cluster or your own HDInsight cluster

Activities: Hive, Pig, Spark, MapReduce, Hadoop Streaming

Compute environment: Azure Batch

Activities: Custom activities

Compute environment: Azure Machine Learning Studio Machine

Activities: Learning activities: Batch Execution and Update Resource

Compute environment: Azure Machine Learning

Activities: Azure Machine Learning Execute Pipeline

Compute environment: Azure Data Lake Analytics

Activities: Data Lake Analytics U-SQL

Compute environment: Azure SQL, Azure SQL Data Warehouse, SQL Server

Activities: Stored Procedure

Compute environment: Azure Databricks

Activities: Notebook, Jar, Python

Compute environment: Azure Function

Activities: Azure Function activity

Transforming data using SQL Server Integration Services (SSIS) packages

Many organizations have decades of development investment in SSIS packages that contain both ingestion and transformation logic from on-premises and cloud data stores. Azure Data Factory provides the ability to lift and shift existing SSIS workload, by creating an Azure-SSIS Integration Runtime to natively execute SSIS packages. Using Azure-SSIS Integration Runtime will enable you to deploy and manage your existing SSIS packages with little to no change using familiar tools such as SQL Server Data Tools (SSDT) and SQL Server Management Studio (SSMS), just like using SSIS on premises.

https://docs.microsoft.com/en-us/azure/data-factory/transform-data

Question 79: Skipped
What is Apache Spark notebook?

- ○
  A cloud-based Big Data and Machine Learning platform, which empowers developers to accelerate AI and innovation by simplifying the process of building enterprise-grade production data applications.
- ○

The default Time to Live (TTL) property for records stored in an analytical store can manage the lifecycle of data and define how long it will be retained for.

- ○

  The logical Azure Databricks environment in which clusters are created, data is stored (via DBFS), and in which the server resources are housed.
- ○

  A notebook is a collection of cells. These cells are run to execute code, to render formatted text, or to display graphical visualizations.

  (Correct)

Explanation
What is Apache Spark notebook?

A notebook is a collection of cells. These cells are run to execute code, to render formatted text, or to display graphical visualizations.

What is a cluster?

The notebooks are backed by clusters, or networked computers, that work together to process your data. The first step is to create a cluster.

https://azure-ramitgridhar.blogspot.com/2019/07/azure-databricks-create-cluster-and.html

Question 80: Skipped
When loading data into Azure Synapse Analytics on a scheduled basis, it's important to try to reduce the time taken to not perform the data load, and minimize the resources needed as much as possible to maintain good performance cost-effectively.

Which of the following are valid Strategies for managing source data files? (Select all that apply)

- □

  When loading large datasets, it's best to use the compression capabilities of the file format.

  (Correct)
- □

  Maintaining a well-engineered Data Lake structure

  (Correct)
- □

  Consolidate source files
- □

Having well defined “zones” established for the data coming into the Data Lake and cleansing and transformation tasks that land the data you need in a curated and optimized state.

(Correct)

Explanation

When loading data into Azure Synapse Analytics on a scheduled basis, it’s important to try to reduce the time taken to not perform the data load, and minimize the resources needed as much as possible to maintain good performance cost-effectively.

Strategies for managing source data files include:

Maintain a well-engineered data lake structure

Maintaining a well-engineered Data Lake structure allows you to know that the data your loading regularly is consistent with the data requirements for your system. It is less important if your load is a once-off or exploratory rather than analytical. Some strategies include folder hierarchies based on the source system, and date/time or file format and focus.

In general, having well defined “zones” established for the data coming into the Data Lake and cleansing and transformation tasks that land the data you need in a curated and optimized state.

Compress and optimize files

When loading large datasets, it’s best to use the compression capabilities of the file format. It ensures that less time is spent on the process of data transfers, using instead the power of Azure Synapse’ Massively Parallel Processing (MPP) compute capabilities for decompression.

It is fairly standard to maintain curated source files in columnar compressed file formats such as RC, Gzip, Parquet, and ORC, which are all supported import formats.

Split source files

One of the key architectural components within Azure Synapse Analytics dedicated SQL pools is the decoupled storage that is segmented into 60 parts. You should maintain alignment to multiples of this number as much as possible depending on the file sizes that you are loading, and the number of compute nodes you have provisioned. Since there are 60 storage segments and a maximum of 60 MPP compute nodes within the highest performance configuration of SQL Pools, a 1:1 file to compute node to storage segment may be viable for ultra-high workloads, reducing the load times to the minimum possible.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/data-processed

Question 81: Skipped

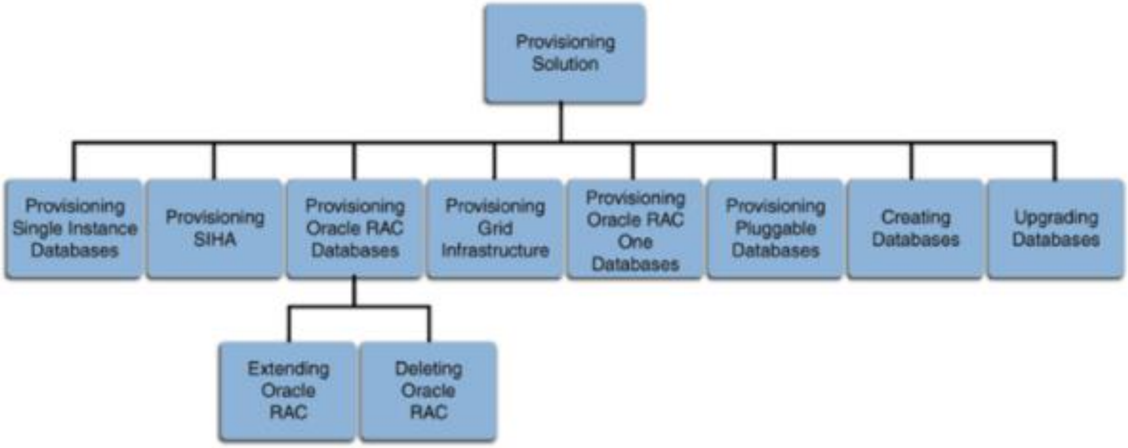
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

The act of setting up the database server is called [?].

- ○ Provisioning (Correct)
- ○ Distribution
- ○ Running up
- ○ Population

Explanation

The act of setting up the database server is called *provisioning*.


Provisioning Solution
Provisioning Single Instance Databases
Provisioning SIHA
Provisioning Oracle RAC Databases
Provisioning Grid Infrastructure
Provisioning Oracle RAC One Databases
Provisioning Pluggable Databases
Creating Databases
Upgrading Databases
Extending Oracle RAC
Deleting Oracle RAC

https://docs.oracle.com/cd/E24628_01/em.121/e27046/prov_db_overview.htm#EMLCM11094

Question 82: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Azure Storage provides a REST API to work with the containers and data stored in each account. The simplest way to handle access keys and endpoint URLs within applications is to use [?].

- ○ The account subscription key
- ○ The private access key
- ○ The instance key
- ○ The REST API endpoint
- ○ A public access key
- ○ Storage account connection strings **(Correct)**

Explanation

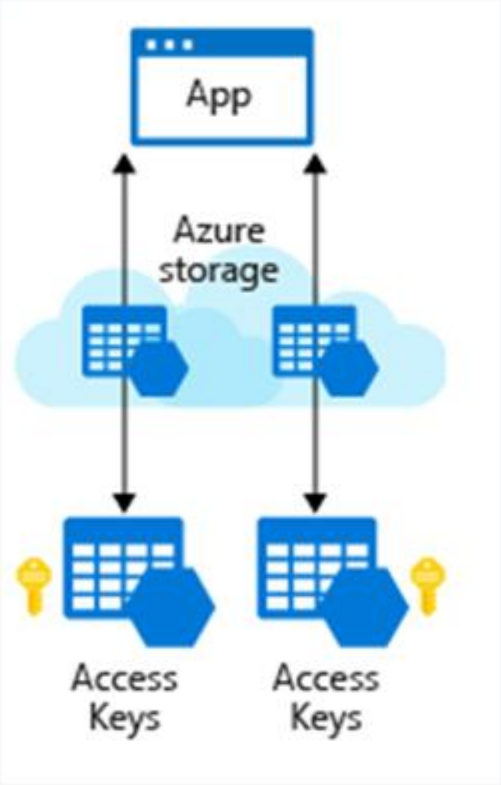
Azure Storage provides a REST API to work with the containers and data stored in each account. To work with data in a storage account, your app will need two pieces of data:

• Access key

• REST API endpoint

Security access keys

Each storage account has two unique *access keys* that are used to secure the storage account. If your app needs to connect to multiple storage accounts, your app will require an access key for each storage account.



Connection strings

The simplest way to handle access keys and endpoint URLs within applications is to use **storage account connection strings**. A connection string provides all needed connectivity information in a single text string.

Azure Storage connection strings look similar to the following example, but with the access key and account name of your specific storage account:

```
DefaultEndpointsProtocol=https;AccountName={your-storage};

AccountKey={your-access-key};

EndpointSuffix=core.windows.net
```

https://docs.microsoft.com/en-us/rest/api/storageservices/blob-service-rest-api

Question 83: Skipped

Activities within Azure Data Factory define the actions that will be performed on the data. Which are valid activity categories? (Select three)

- □ Control activities

(Correct)

- □

  Data movement activities

  (Correct)
- □

  Analytic activities
- □

  Data transformation activities

  (Correct)
- □

  Test Lab activities
- □

  Data storage activities

Explanation

Activities within Azure Data Factory define the actions that will be performed on the data and there are three categories including:

• Data movement activities

• Data transformation activities

• Control activities

Data movement activities

Data movement activities simply move data from one data store to another. You can use the Copy Activity to perform data movement activities, or by using JSON. There are a wide range of data stores that are supported as a source and as a sink. This list is ever increasing, and you can find the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-movement-activities

Data transformation activities

Data transformation activities can be performed natively within the authoring tool of Azure Data Factory using the Mapping Data Flow. Alternatively, you can call a compute resource to change or enhance data through transformation, or perform analysis of the data. These include compute technologies such as Azure Databricks, Azure Batch, SQL Database and Azure Synapse Analytics, Machine Learning Services, Azure Virtual machines and HDInsight. You can make use of any existing SQL Server Integration Services (SSIS) Packages stored in a catalogue to execute in Azure

As this list is always evolving, you can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-transformation-activities

Control activities

When graphically authoring ADF solutions, you can use the control flow within the designed to orchestrate pipeline activities that include chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger. The current capabilities include:

- Execute Pipeline Activity

Execute Pipeline activity allows a Data Factory pipeline to invoke another pipeline.

- ForEachActivity

ForEach Activity defines a repeating control flow in your pipeline. This activity is used to iterate over a collection and executes specified activities in a loop. The loop implementation of this activity is similar to Foreach looping structure in programming languages.

- WebActivity

Web Activity can be used to call a custom REST endpoint from a Data Factory pipeline. You can pass datasets and linked services to be consumed and accessed by the activity.

- Lookup Activity

Lookup Activity can be used to read or look up a record/ table name/ value from any external source. This output can further be referenced by succeeding activities.

- Get Metadata Activity

GetMetadata activity can be used to retrieve metadata of any data in Azure Data Factory.

- Until Activity

Implements Do-Until loop that is similar to Do-Until looping structure in programming languages. It executes a set of activities in a loop until the condition associated with the activity evaluates to true. You can specify a timeout value for the until activity in Data Factory.

• If Condition Activity

The If Condition can be used to branch based on condition that evaluates to true or false. The If Condition activity provides the same functionality that an if statement provides in programming languages. It evaluates a set of activities when the condition evaluates to true and another set of activities when the condition evaluates to false.

• Wait Activity

When you use a Wait activity in a pipeline, the pipeline waits for the specified period of time before continuing with execution of subsequent activities.

You can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#control-activities

Question 84: Skipped

In Azure Synapse Studio, manage integration pipelines within the Integrate hub.

When you expand Pipelines you will see which of the following? (Select three)

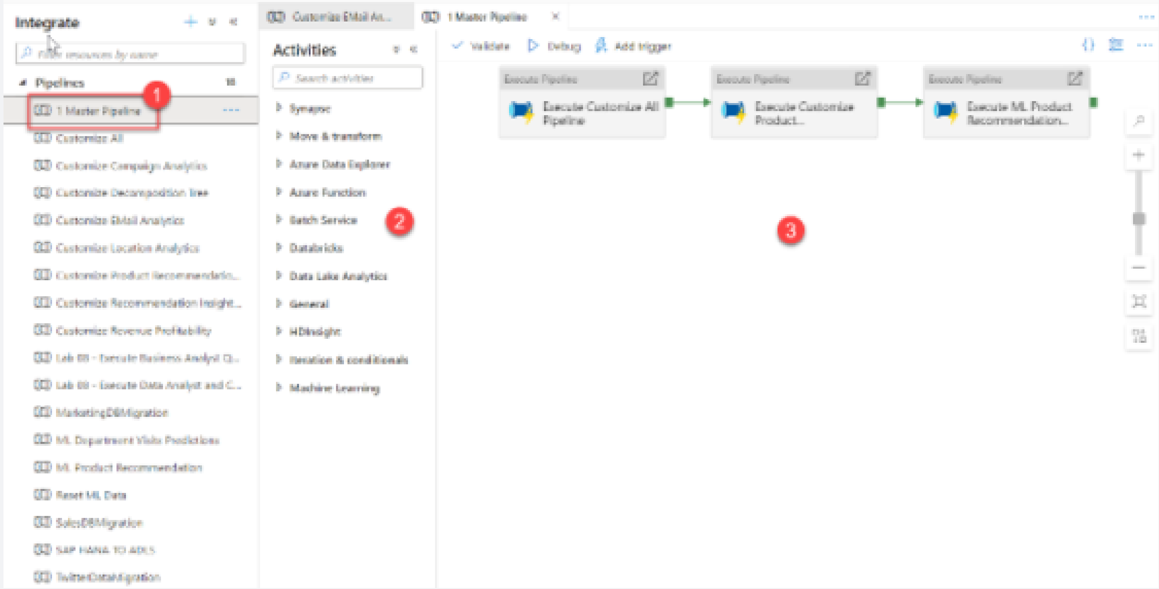
- ☐ Data flows
- ☐ Notebooks
- ☐ Pipeline canvas **(Correct)**
- ☐ Master Pipeline **(Correct)**
- ☐ SQL serverless databases
- ☐ Activities **(Correct)**

- □ SQL scripts
- □ Power BI
- □ Provisioned SQL pool databases
- □ External data sources

Explanation

In Azure Synapse Studio, manage integration pipelines within the Integrate hub. If you are familiar with Azure Data Factory, then you will feel at home in this hub. The pipeline creation experience is the same as in ADF, which gives you another powerful integration built into Synapse Analytics, removing the need to use Azure Data Factory for data movement and transformation pipelines.

When you expand Pipelines you will see **Master Pipeline (1)**. Point out the **Activities (2)** that can be added to the pipeline, and show the **pipeline canvas (3)** on the right.



This Synapse workspace contains 16 pipelines that enable us to orchestrate data movement and transformation steps over data from several sources.

The **Activities** list contains many activities that you can drag and drop onto the pipeline canvas on the right.

Expand a few activity categories to show what's available, such as Notebook, Spark, and SQL pool stored procedure activities under Synapse.

https://techcommunity.microsoft.com/t5/azure-synapse-analytics/quickly-get-started-with-azure-synapse-studio/ba-p/1961116

Question 85: Skipped

Azure Synapse Pipelines is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale.

Azure Synapse Pipelines enables you to integrate data pipelines between which of the following? (Select all that apply)

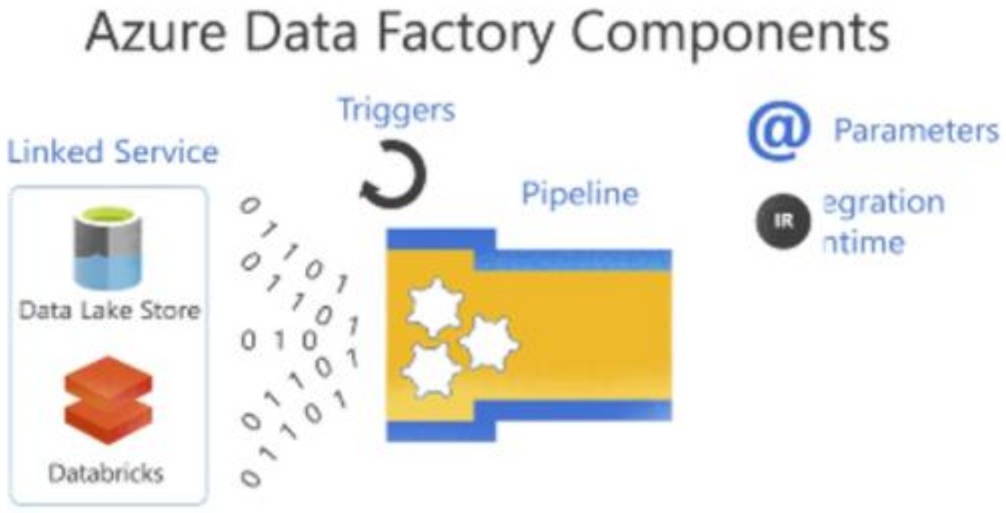
- □ SQL Serverless **(Correct)**
- □ SQL Pools **(Correct)**
- □ Hadoop Pools
- □ Spark Pools **(Correct)**
- □ Cosmos Pools
- □ Cosmos Serverless

Explanation

Azure Synapse Pipelines is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale. Using Azure Synapse Pipelines, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores. You can build complex ETL, or ELT processes that transform data visually with data flows or by using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure Synapse Analytics.

Much of the functionality of Azure Synapse Pipelines come from the Azure Data Factory features and are commonly referred to as Pipelines. Azure Synapse Pipelines enables you to integrate data pipelines between SQL Pools, Spark Pools and SQL Serverless, providing a one stop shop for all your analytical needs.

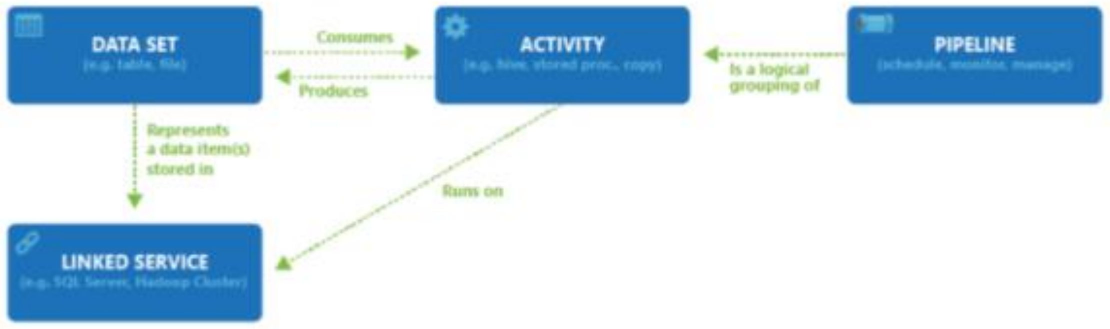
Like Azure Data Factory, Azure Synapse Pipelines is composed of four core components. These components work together to provide the platform on which you can compose data-driven workflows with steps to move and transform data.



Data Factory supports a wide variety of data sources that you can connect to through the creation of an object known as a **Linked Service**, which enables you to ingest the data from a data source in readiness to prepare the data for transformation and/or analysis. In addition, Linked Services can fire up compute services on demand. For example, you may have a requirement to start an on-demand HDInsight cluster for the purpose of just processing data through a Hive query. So Linked Services enables you to define data sources, or compute resource that is required to ingest and prepare data.

With the linked service defined, Azure Data Factory is made aware of the datasets that it should use through the creation of a **Datasets** object. Datasets represent data structures within the data store that is being referenced by the Linked Service object. Datasets can also be used by an ADF object known as an Activity.

Activities typically contain the transformation logic or the analysis commands of the Azure Data Factory’s work. Activities includes the Copy Activity that can be used to ingest data from a variety of data sources. It can also include the Mapping Data Flow to perform code-free data transformations. It can also include the execution of a stored procedure, Hive Query, or Pig script to transform the data. You can push data into a Machine Learning model to perform analysis. It is not uncommon for multiple activities to take place that may include transforming data using a SQL stored procedure and then perform analytics with Databricks. In this case, multiple activities can be logically grouped together with an object referred to as a **Pipeline**, and these can be *scheduled* to execute, or a *trigger* can be defined that determines when a pipeline execution needs to be kicked off. There are different types of triggers for different types of events.


DATA SET
(e.g. table, file)
Consumes
Produces
ACTIVITY
(e.g. hive, stored proc., copy)
Is a logical grouping of
PIPELINE
(schedule, monitor, manage)
Represents a data item(s) stored in
Runs on
LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Control flow is an orchestration of pipeline activities that includes chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger. It also includes custom-state passing and looping containers, and For-each iterators.

Parameters are key-value pairs of read-only configuration. Parameters are defined in the pipeline. The arguments for the defined parameters are passed during execution from the run context that was created by a trigger or a pipeline that was executed manually. Activities within the pipeline consume the parameter values.

Azure Synapse pipelines has an *integration runtime* that enables it to bridge between the activity and linked Services objects. It is referenced by the linked service, and provides the compute environment where the activity either runs on or gets dispatched from. This way, the activity can be performed in the region closest possible. There are three types of Integration Runtime, including Azure, Self-hosted, and Azure-SSIS.

Once all the work is complete, you can then use Data Factory to publish the final dataset to another linked service that can then be consumed by technologies such as Power BI or Machine Learning.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-partner-data-integration

Question 86: Skipped

Setting Global parameters in an Azure Data Factory pipeline, allows you to use constants for consumption in pipeline expressions.

If you have created a data flow in which you have set parameters, it is possible to execute it from a pipeline using the Execute Data Flow Activity. Once you have added the activity to the pipeline canvas, you'll find the data flow parameters in the activity's Parameters tab.

Is it possible to combine the pipeline and data flow expression parameters while mapping dataflow?

- ○ No
- ○ Yes (Correct)

Explanation

Global parameters in Azure Data Factory

Setting Global parameters in an Azure Data Factory pipeline, allows you to use constants for consumption in pipeline expressions. A use-case for setting global parameters is when you have multiple pipelines where the parameters names and values are identical. If you use the continuous integration and deployment process with Azure Data Factory, the global parameters can be overridden if you wish so, for each and every environment that you have created.

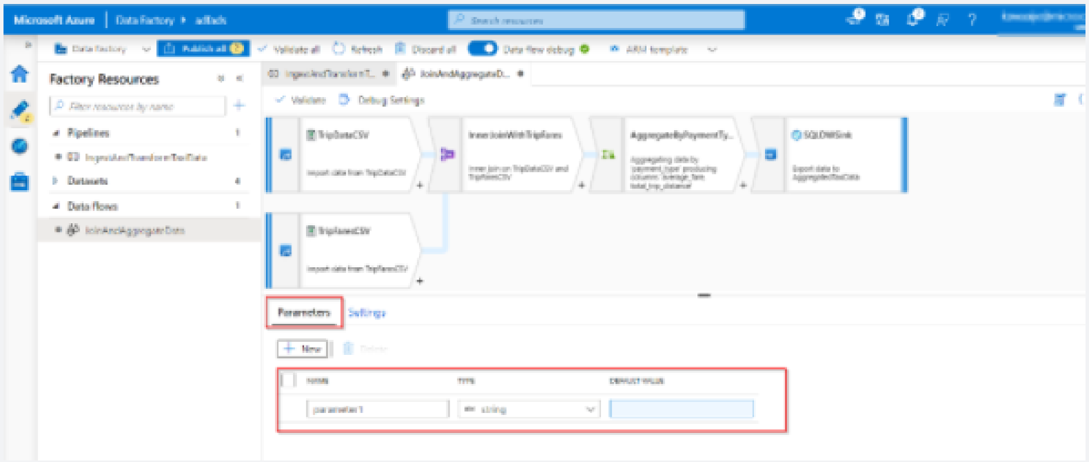
Using global parameters in a pipeline

When using global parameters in a pipeline in Azure Data Factory, it is mostly referenced in pipeline expressions. For example, if a pipeline references to a resource like a dataset or data flow, you can pass down the global parameter value through the resource parameter. The command or reference of global parameters in Azure Data Factory flows as follows: pipeline().globalParameters.

Create parameters in dataflow

To add parameters to your data flow, click on the blank portion of the data flow canvas to see the general properties. In the settings pane, you will see a tab called Parameter.

Select New to generate a new parameter. For each parameter, you must assign a name, select a type, and optionally set a default value.



Assign parameters from a pipeline in mapping dataflow

If you have created a data flow in which you have set parameters, it is possible to execute it from a pipeline using the Execute Data Flow Activity. Once you have added the activity to the pipeline canvas, you'll find the data flow parameters in the activity's Parameters tab. Assigning parameter values, ensures that you are able to use the parameters in a pipeline expression language or data flow expression language based on spark types. **You can also combine the two, that is, pipeline and data flow expression parameters.**

https://docs.microsoft.com/en-us/azure/data-factory/parameterize-linked-services

Question 87: Skipped

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

All data within an Azure Cosmos DB container is partitioned based on the [?], and applies to both the transactional store and the analytical store. Boundaries for parallelizing workloads are based on this [?].

- ○ Index key

- ○ Partition key (Correct)
- ○ Primary key
- ○ Foreign key

Explanation

Mixed entity types per container

You may want to mix different document entity types (entities) in the same container, which is useful to efficiently retrieve data for both entities using a single query. For example, you could put both customer profile and sales order data in the same container and partition it by customerId. In such a situation, you would usually add a field to your documents that identifies the entity type of each document to differentiate between them at query time. In the following sample documents, you will see that the type is added for this purpose in the following example documents:

```
JSON
{
"id": "54AB87A7-BDB9-4FAE-A668-AA9F43E26628",
"type": "customer",
"name": "Franklin Ye",
"customerId": "54AB87A7-BDB9-4FAE-A668-AA9F43E26628",
"address": {
"streetNo": 15850,
"streetName": "NE 40th St.",
"postcode": 98052
}
}



{
"_id": "000C23D8-B8BC-432E-9213-6473DFDA2BC5",
"type": "salesOrder",
"customerId": "54AB87A7-BDB9-4FAE-A668-AA9F43E26628",
"orderDate": "2014-02-16T00:00:00",

"shipDate": "2014-02-23T00:00:00",

"details": [

{

"sku": "BK-R64Y-42",

"name": "Road-550-W Yellow, 42",

"price": 1120.49,

"quantity": 1

}

]

}
```

The following query on against the transactional store would return the customer details and all orders associated with this one customer.

```
SQL

SELECT * FROM c WHERE c.customerID = "54AB87A7-BDB9-4FAE-A668-AA9F43E26628“
```

Whilst this approach to modelling is potentially useful for your Cosmos DB transactional store queries. All documents within a single container are mapped to a single analytical store, leading to sparsely populated column stores with the different data types needing to be further separated at the time of running an analytical query.

Recommendation: As with many design decisions, there is a trade-off between the efficiency of querying the transactional store and the ease of querying the analytical store. Carefully evaluate the usefulness of storing a mix of different document entity types in the same container to your transactional workloads. If you choose to do so, you will be required to filter by the property entity type property you selected.

Embedding entity arrays

When optimizing transactional data models, we choose to embed entities within an array in a document, especially for read heavy workloads where:

- There are contained relationships between entities.

- There are one-to-few relationships between entities.

- There is embedded data that changes infrequently.

- There is embedded data that will not grow without bound.

• There is embedded data that is queried frequently together.

Due to the fact that there are one to few relationships between the embedded entities that are represented within a single document, and that these are mapped to a single column within a single row within the analytical store. The entire embedded entity array will reside within a single column value, and need to be translated from its JSON representation at the time of querying in order to retrieve embedded entity values, irrespective of which of the two modes of schema representation being used.

Recommendation: Again, a balance needs to be struck between the usefulness of the entity embedding within the transactional application and the added complexity of writing queries against embedded JSON documents for your application.

Partitioning of data

All data within an Azure Cosmos DB container is partitioned based on the partition key, and applies to both the transactional store and the analytical store. Boundaries for parallelizing workloads are based on this partition key.

The orderliness associated when data appears in the analytical store for a query is only guaranteed within a partition. As an example, when documents (1) (2) (3) are inserted in the transactional store into a single partition, they are guaranteed to be present in the analytical store in the order in which they were inserted.

https://docs.microsoft.com/en-us/azure/cosmos-db/modeling-data

Question 88: Skipped
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

Security and infrastructure configuration go hand-in-hand. When you set up your Azure Databricks workspace(s) and related services, you need to make sure that security considerations do not take a back seat during the architecture design.

When enabled, authentication automatically takes place in Azure Data Lake Storage (ADLS) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that one uses to log into Azure Databricks. Commands running on a configured cluster will be able to read and write data in ADLS without needing to configure service principal credentials. Any ACLs applied at the folder or file level in ADLS are enforced based on the user's identity.

ADLS Passthrough is configured when you create a cluster in the Azure Databricks workspace. On a *standard cluster*, when you enable this setting … [?]

- ○

you will inherit user access from the Azure Active Directory (AAD) users to the Azure Databricks workspace.

- ○
  you must set two user accesses to one of the Azure Active Directory (AAD) users in the Azure Databricks workspace. The second is required as a backup or secondary user.
- ○
  you may set multiple user accesses to one of the Azure Active Directory (AAD) users in the Azure Databricks workspace. The additional access are required as a backup or auxiliary users.
- ○
  you must set single user access to one of the Azure Active Directory (AAD) users in the Azure Databricks workspace.
  (Correct)

Explanation

Security and infrastructure configuration go hand-in-hand. When you set up your Azure Databricks workspace(s) and related services, you need to make sure that security considerations do not take a back seat during the architecture design.

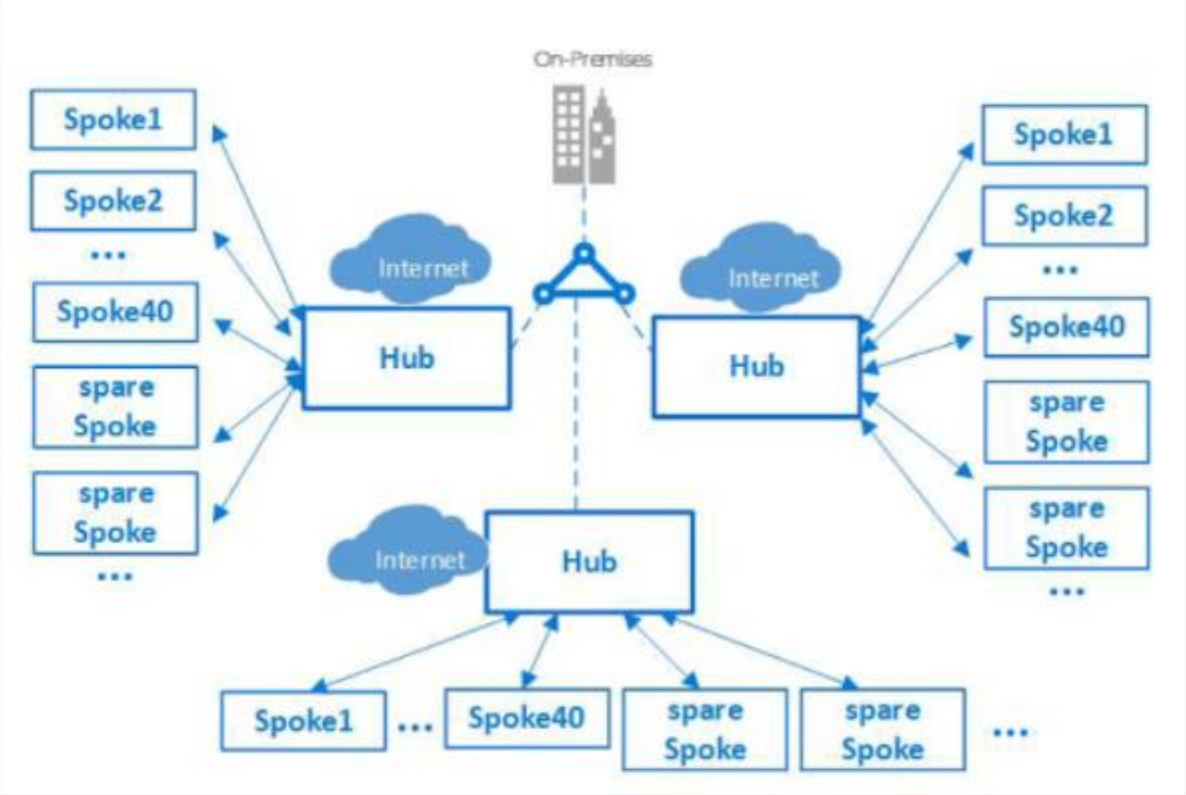
Consider isolating each workspace in its own VNet

While you can deploy more than one Workspace in a VNet by keeping the associated subnet pairs separate from other workspaces, MS recommends that you should only deploy one workspace in any VNet. Doing this perfectly aligns with the ADB's Workspace level isolation model. Most often organizations consider putting multiple workspaces in the same VNet so that they all can share some common networking resource, like DNS, also placed in the same VNet because the private address space in a VNet is shared by all resources. You can easily achieve the same while keeping the Workspaces separate by following the hub and spoke model and using VNet Peering to extend the private IP space of the workspace VNet. Here are the steps:

1. Deploy each Workspace in its own spoke VNet.

2. Put all the common networking resources in a central hub VNet, such as your custom DNS server.

3. Join the Workspace spokes with the central networking hub using VNet Peering



Do not store any production data in Default Databricks Filesystem (DBFS) Folders

This recommendation is driven by security and data availability concerns. Every Workspace comes with a default Databricks File System (DBFS), primarily designed to store libraries and other system-level configuration artifacts such as initialization scripts. You should not store any production data in it, because:

1. The lifecycle of default DBFS is tied to the Workspace. Deleting the workspace will also delete the default DBFS and permanently remove its contents.

2. One can't restrict access to this default folder and its contents.

Important: This recommendation doesn't apply to Blob or ADLS folders explicitly mounted as DBFS by the end user.

Always hide secrets in a key vault

It is a significant security risk to expose sensitive data such as access credentials openly in Notebooks or other places such as job configs, initialization scripts, etc. You should always use a vault to securely store and access them. You can either use ADB's internal Key Vault for this purpose or use Azure's Key Vault (AKV) service.

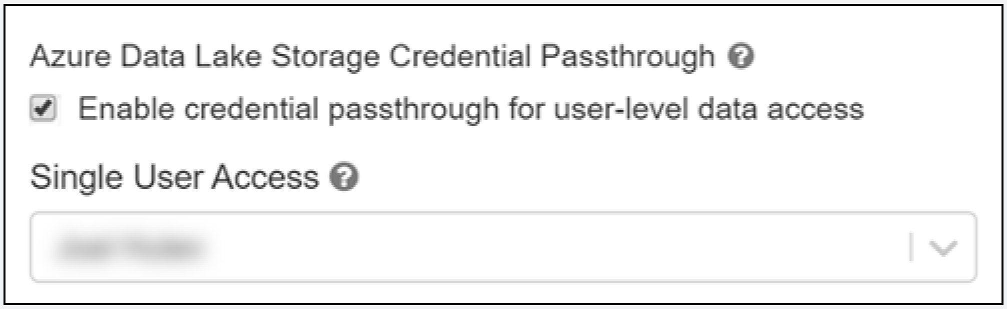
If using Azure Key Vault, create separate AKV-backed secret scopes and corresponding AKVs to store credentials pertaining to different data stores. This will help prevent users from accessing credentials that they might not have access to. Since access controls are applicable to the entire secret scope, users with access to the scope will see all secrets for the AKV associated with that scope.

Access control - Azure Data Lake Storage (ADLS) passthrough

When enabled, authentication automatically takes place in Azure Data Lake Storage (ADLS) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that one uses to log into Azure Databricks. Commands running on a configured cluster will be able to read and write data in ADLS without needing to configure service principal credentials. Any ACLs applied at the folder or file level in ADLS are enforced based on the user's identity.

ADLS Passthrough is configured when you create a cluster in the Azure Databricks workspace. ADLS Gen1 requires Databricks Runtime 5.1+. ADLS Gen2 requires 5.3+.

On a *standard cluster*, when you enable this setting you must set single user access to one of the Azure Active Directory (AAD) users in the Azure Databricks workspace. Only one user is allowed to run commands on this cluster when Credential Passthrough is enabled.



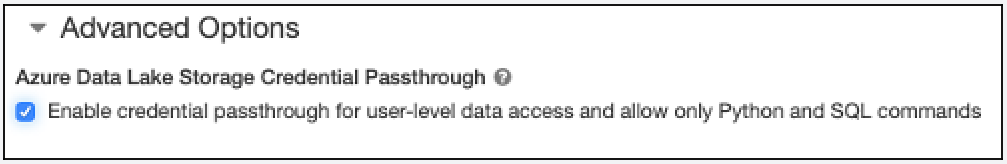
Azure Data Lake Storage Credential Passthrough

☑ Enable credential passthrough for user-level data access

Single User Access

[redacted]

High-concurrency clusters can be shared by multiple users. When you enable ADLS Passthrough on this type of cluster, it does not require you to select a single user.



▾ Advanced Options

Azure Data Lake Storage Credential Passthrough ?

☑ Enable credential passthrough for user-level data access and allow only Python and SQL commands

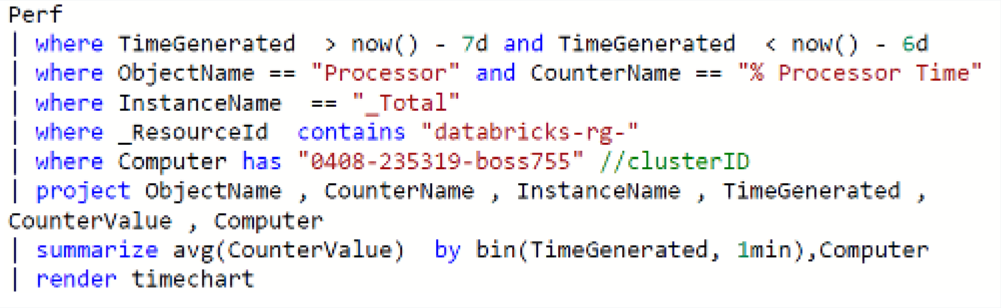
Configure audit logs and resource utilization metrics to monitor activity

An important facet of monitoring is understanding the resource utilization in Azure Databricks clusters. You can also extend this to understanding utilization across all clusters in a workspace. This information is useful in arriving at the correct cluster and VM sizes. Each VM does have a set of limits (cores/disk throughput/network throughput) which play an important role in determining the performance profile of an Azure Databricks job.

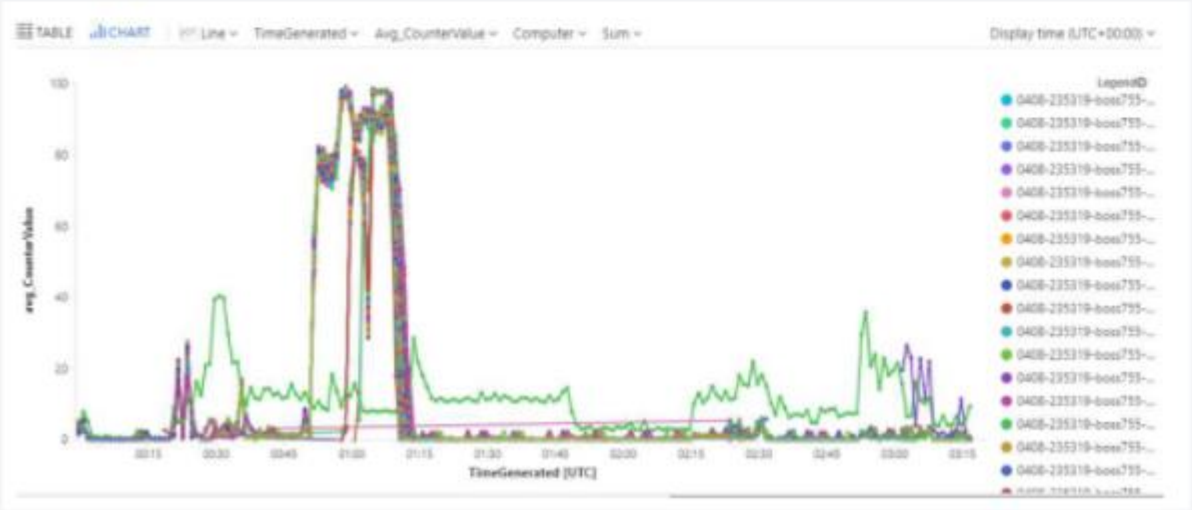
In order to get utilization metrics of an Azure Databricks cluster, you can stream the VM's metrics to an Azure Log Analytics Workspace (see Appendix A) by installing the Log Analytics Agent on each cluster node.

Querying VM metrics in Log Analytics once you have started the collection using the above document

You can use Log analytics directly to query the Perf data. Here is an example of a query which charts out CPU for the VMs in question for a specific cluster ID. See log analytics overview for further documentation on log analytics and query syntax.



```
Perf
| where TimeGenerated  > now() - 7d and TimeGenerated  < now() - 6d
| where ObjectName == "Processor" and CounterName == "% Processor Time"
| where InstanceName  == "_Total"
| where _ResourceId  contains "databricks-rg-"
| where Computer has "0408-235319-boss755" //clusterID
| project ObjectName , CounterName , InstanceName , TimeGenerated ,
CounterValue , Computer
| summarize avg(CounterValue)  by bin(TimeGenerated, 1min),Computer
| render timechart
```



1. https://docs.microsoft.com/azure/azure-monitor/learn/quick-collect-linux-computer

2. https://github.com/Microsoft/OMS-Agent-for-Linux/blob/master/docs/OMS-Agent-for-Linux.md

3. https://github.com/Microsoft/OMS-Agent-for-Linux/blob/master/docs/Troubleshooting.md

Question 89: Skipped
Within the context of Azure Databricks, sharing data from one worker to another can be a costly operation.

Spark has optimized this operation by using a format called Tungsten which prevents the need for expensive serialization and de-serialization of objects in order to get data from one JVM to another.

The data that is "shuffled" is in a format known as `UnsafeRow`, or more commonly, the Tungsten Binary Format.

When we shuffle data, it creates what is known as [?].

- ○

  A Stage
- ○

  A Lineage

- ○

  A Stage boundary

  (Correct)
- ○

  A Pipeline

Explanation

As opposed to narrow transformations, wide transformations cause data to shuffle between executors. This is because a wide transformation requires sharing data across workers. **Pipelining** helps us optimize our operations based on the differences between the two types of transformations.

Pipelining

• Pipelining is the idea of executing as many operations as possible on a single partition of data.

• Once a single partition of data is read into RAM, Spark will combine as many narrow operations as it can into a single **Task**

• Wide operations force a shuffle, conclude a stage, and end a pipeline.

Shuffles

A shuffle operation is triggered when data needs to move between executors.

To carry out the shuffle operation Spark needs to:

• Convert the data to the UnsafeRow, commonly referred to as **Tungsten Binary Format**.

• Write that data to disk on the local node - at this point the slot is free for the next task.

• Send that data across the wire to another executor

• Technically the Driver decides which executor gets which piece of data.

• Then the executor pulls the data it needs from the other executor's shuffle files.

• Copy the data back into RAM on the new executor

• The concept, if not the action, is just like the initial read "every“ `DataFrame` starts with.

• The main difference being it's the 2nd+ stage.

As we will see in a moment, this amounts to a free cache from what is effectively temp files.

Some actions induce in a shuffle. Good examples would include the operations `count()` and `reduce(..)`.

UnsafeRow (also known as Tungsten Binary Format)

Sharing data from one worker to another can be a costly operation.

Spark has optimized this operation by using a format called **Tungsten**.

Tungsten prevents the need for expensive serialization and de-serialization of objects in order to get data from one JVM to another.

The data that is "shuffled" is in a format known as `UnsafeRow`, or more commonly, the Tungsten Binary Format.

`UnsafeRow` is the in-memory storage format for Spark SQL, DataFrames & Datasets.

Advantages include:

• Compactness:

• Column values are encoded using custom encoders, not as JVM objects (as with RDDs).

• The benefit of using Spark 2.x's custom encoders is that you get almost the same compactness as Java serialization, but significantly faster encoding/decoding speeds.

• Also, for custom data types, it is possible to write custom encoders from scratch.

• Efficiency: Spark can operate *directly out of Tungsten*, without first deserializing Tungsten data into JVM objects.

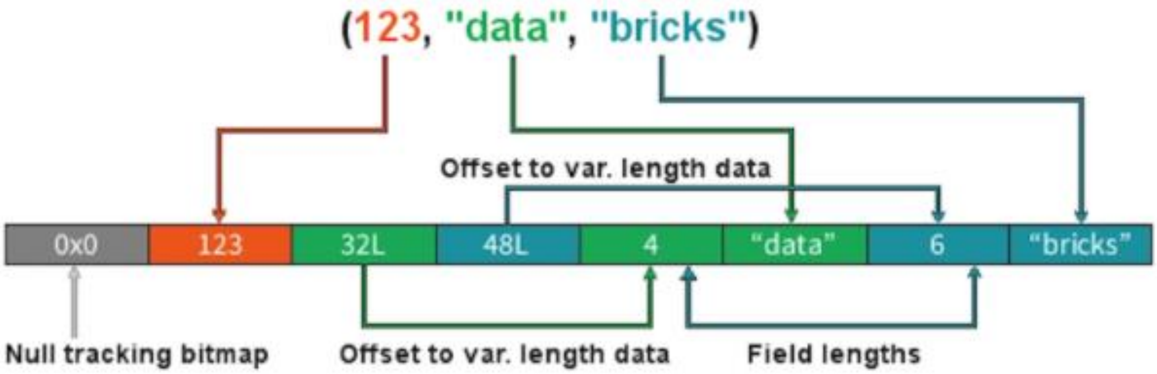
How UnsafeRow works

• The first field, "123", is stored in place as its primitive.

• The next 2 fields, "data" and "bricks", are strings and are of variable length.

• An offset for these two strings is stored in place (32L and 48L respectively shown in the picture below).

- The data stored in these two offset's are of format "length + data".
- At offset 32L, we store 4 + "data" and likewise at offset 48L we store 6 + "bricks".



Stages

- **When we shuffle data, it creates what is known as a stage boundary.**
- Stage boundaries represent a process bottleneck.

Take for example the following transformations:

Step	**Transformation**
1	Read
2	Select
3	Filter
4	GroupBy
5	Select
6	Filter
7	Write

Spark will break this one job into two stages (steps 1-4b and steps 4c-7):

Stage #1

Step Transformation

1 Read

2 Select

3 Filter

4a GroupBy 1/2

4b shuffle write

Stage #1

Step Transformation

4c shuffle read

4d GroupBy 2/2

5 Select

6 Filter

7 Write

In **Stage #1**, Spark will create a pipeline of transformations in which the data is read into RAM (Step #1), and then perform steps #2, #3, #4a & #4b

All partitions must complete **Stage #1** before continuing to **Stage #2**

- It's not possible to group all records across all partitions until every task is completed.

- This is the point at which all the tasks must synchronize.

- This creates our bottleneck.

- Besides the bottleneck, this is also a significant performance hit: disk IO, network IO and more disk IO.

Once the data is shuffled, we can resume execution...

For **Stage #2**, Spark will again create a pipeline of transformations in which the shuffle data is read into RAM (Step #4c) and then perform transformations #4d, #5, #6 and finally the write action, step #7.

Lineage

From the developer's perspective, we start with a read and conclude (in this case) with a write:

Step	**Transformation**
1	Read
2	Select
3	Filter
4	GroupBy
5	Select
6	Filter
7	Write

However, Spark starts with the action (write(..) in this case).

Next, it asks the question, what do I need to do first?

It then proceeds to determine which transformation precedes this step until it identifies the first transformation.

Step	**Transformation**	
7	Write	Depends on #6
6	Filter	Depends on #5
5	Select	Depends on #4
4	GroupBy	Depends on #3
3	Filter	Depends on #2

2 Select Depends on #1

1 Read First

Why Work Backwards?

Question: So what is the benefit of working backward through your action's lineage?

Answer: It allows Spark to determine if it is necessary to execute every transformation.

Take another look at our example:

• Say we've executed this once already

• On the first execution, step #4 resulted in a shuffle

• Those shuffle files are on the various executors (src & dst)

• Because the transformations are immutable, no aspect of our lineage can change.

• That means the results of our last shuffle (if still available) can be reused.

Why Work Backwards?

Step Transformation

7 Write Depends on #6

6 Filter Depends on #5

5 Select Depends on #4

4 GroupBy <<< shuffle

3 Filter don't care

2 Select don't care

1 Read don't care

In this case, what we end up executing is only the operations from **Stage #2**.

This saves us the initial network read and all the transformations in **Stage #1**

Step Transformation

1 Read skipped

2 Select skipped

3 Filter skipped

4a GroupBy 1/2 skipped

4b shuffle write skipped

4c shuffle read -

4d GroupBy 2/2 -

5 Select -

6 Filter -

7 Write

And Caching...

The reuse of shuffle files (also known as our temp files) is just one example of Spark optimizing queries anywhere it can.

We cannot assume this will be available to us.

Shuffle files are by definition temporary files and will eventually be removed.

However, we cache data to explicitly accomplish the same thing that happens inadvertently with shuffle files.

In this case, the lineage plays the same role. Take for example:

Step Transformation

7 Write Depends on #6

6 Filter Depends on #5

5 Select <<< cache

4 GroupBy <<< shuffle files

3 Filter ?

2 Select ?

1 Read ?

In this case we cached the result of the select(..).

We never even get to the part of the lineage that involves the shuffle, let alone Stage #1.

Instead, we pick up with the cache and resume execution from there:

Step Transformation

1 Read skipped

2 Select skipped

3 Filter skipped

4a GroupBy 1/2 skipped

4b shuffle write skipped

4c shuffle read skipped

4d GroupBy 2/2 skipped

5a cache read -

5b Select -

6 Filter -

7 Write

https://databricks.com/blog/2015/04/28/project-tungsten-bringing-spark-closer-to-bare-metal.html

Question 90: Skipped

What happens to Databricks activities (notebook, JAR, Python) in Azure Data Factory if the target cluster in Azure Databricks isn't running when the cluster is called by Data Factory?

- ○ If the target cluster is stopped, Databricks will start the cluster before attempting to execute.

  (Correct)

- ○
  The Databricks activity will fail in Azure Data Factory – you must always have the cluster running.
- ○
  Whenever a cluster is paused or shut down, ADF will recover from the last operational PiT.
- ○
  Simply add a Databricks cluster start activity before the notebook, JAR, or Python Databricks activity.

Explanation

This situation will result in a longer execution time because the cluster must start, but the activity will still execute as expected.

https://docs.microsoft.com/en-us/azure/data-factory/transform-data-databricks-python