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)**

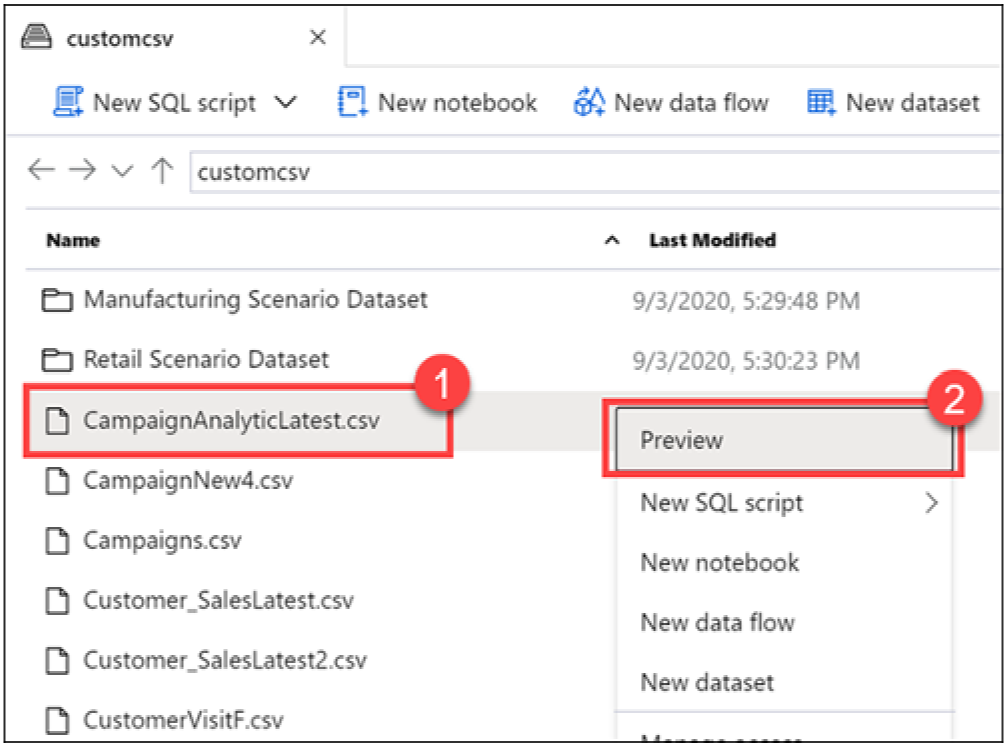
Explanation

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)

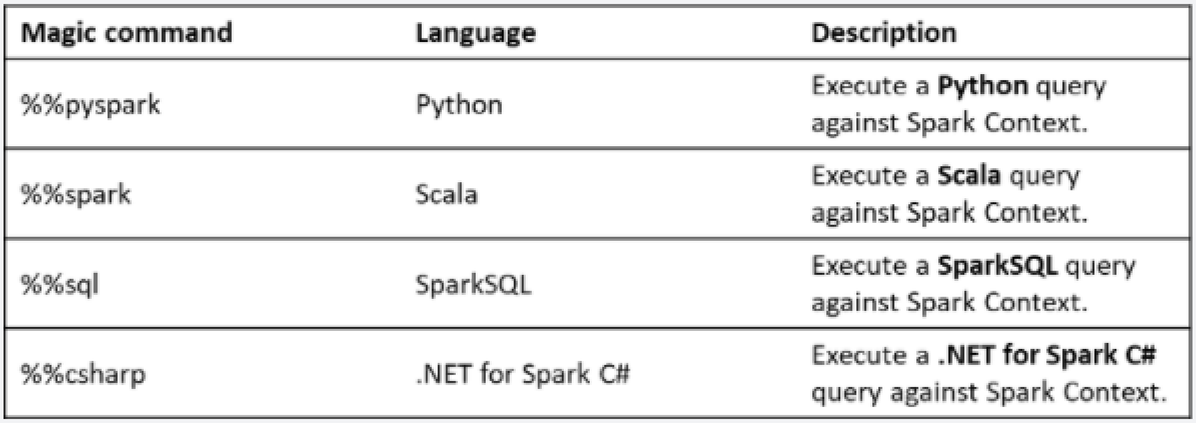
Explanation

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.

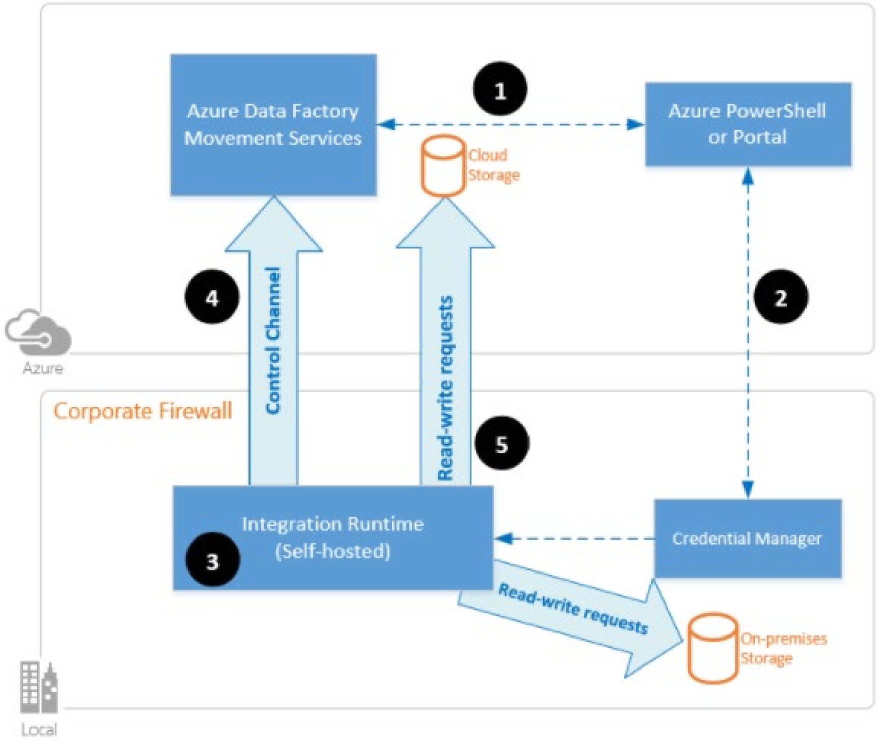
Create and configure a self-hosted integration runtime

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.

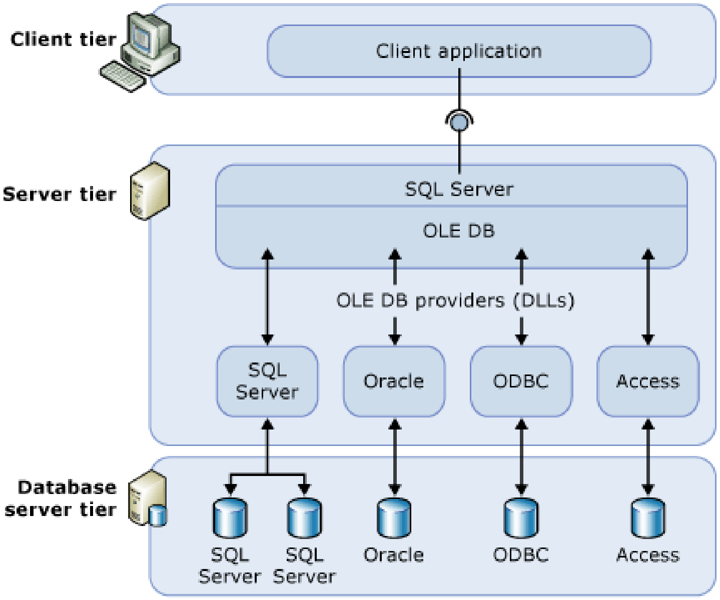
4. Azure Data Factory communicates with the self-hosted integration runtime to schedule and manage jobs. Communication is via a control channel that uses a shared Azure Relay connection. When an activity job needs to be run, Data Factory queues the request along with any credential information. It does 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.



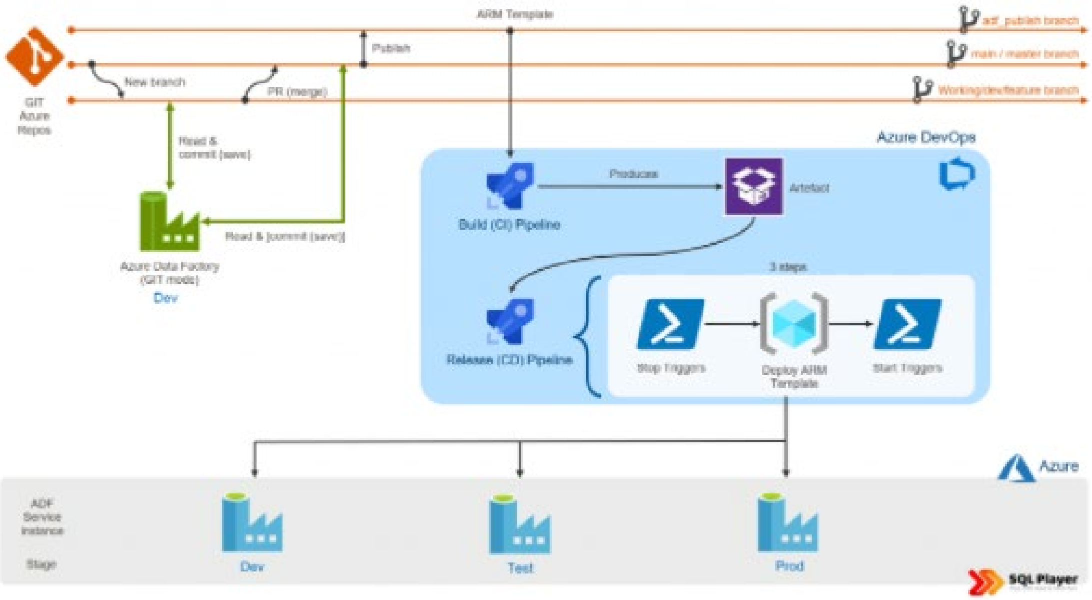
https://www.mssqltips.com/sqlservertip/3630/creating-sql-server-linked-servers-with-azure/

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.



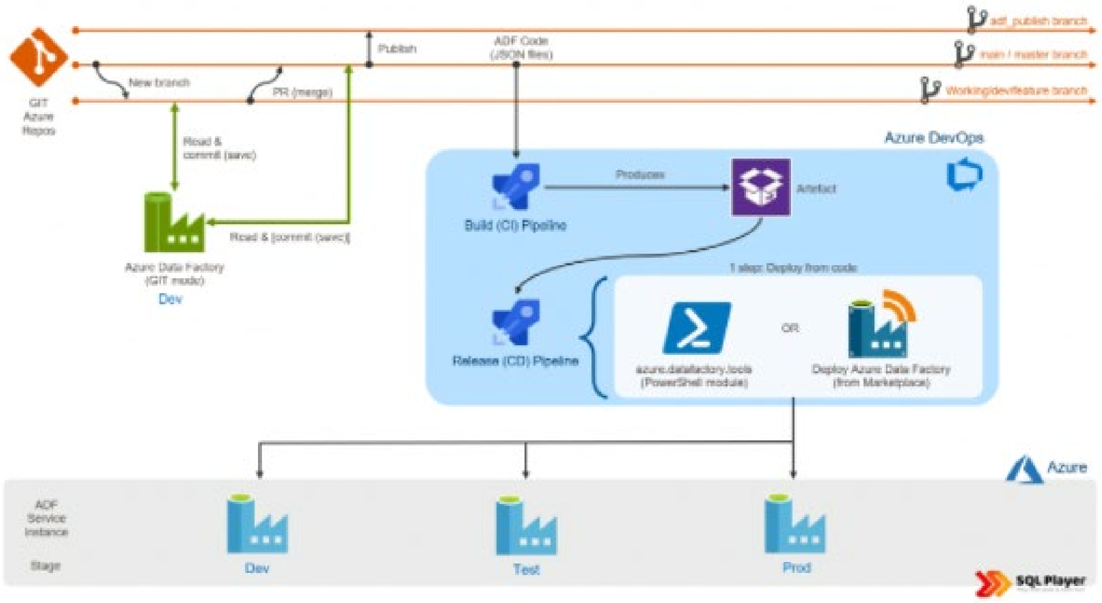
Then you can build your own CI/CD process for deployment of ADF, using Azure DevOps, for instance. I don't want to dig deeper about how to deploy ADF with this approach as I already described it in the post: Deployment of Azure Data Factory with Azure DevOps.
Why many people do not like this approach?

• 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.

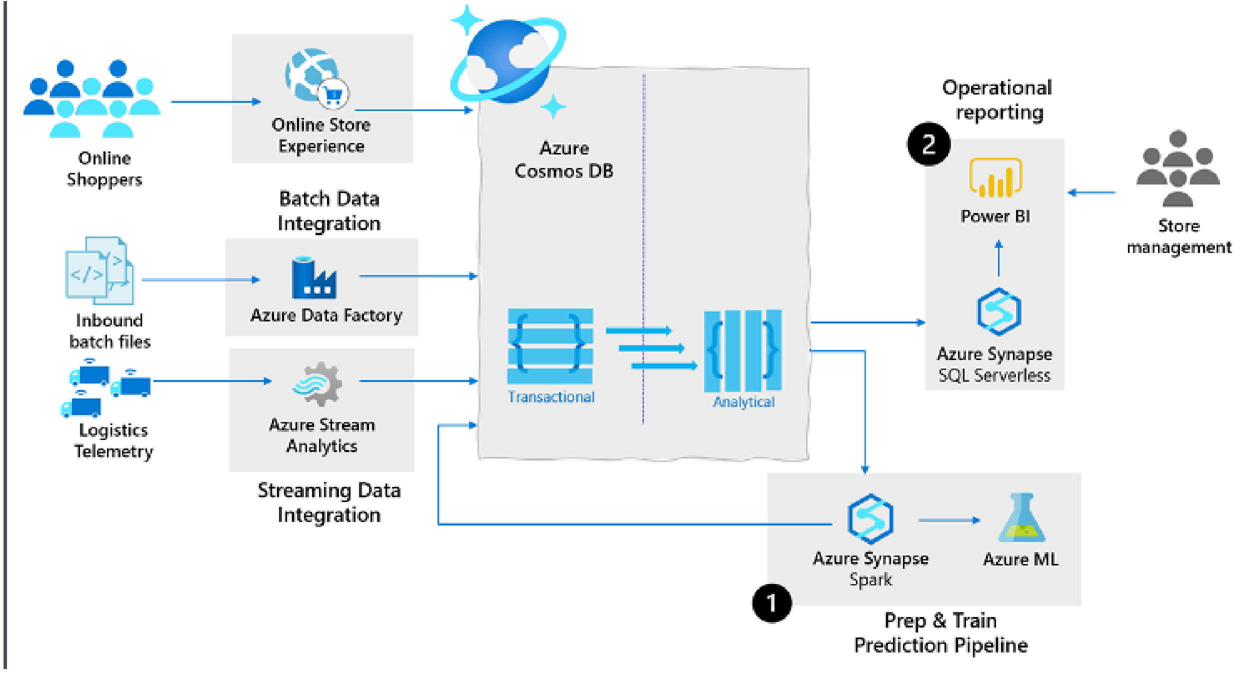
https://stackoverflow.com/questions/58813608/whats-the-difference-between-a-build-pipeline-and-a-release-pipeline-in-azure-de

Question 5: Skipped

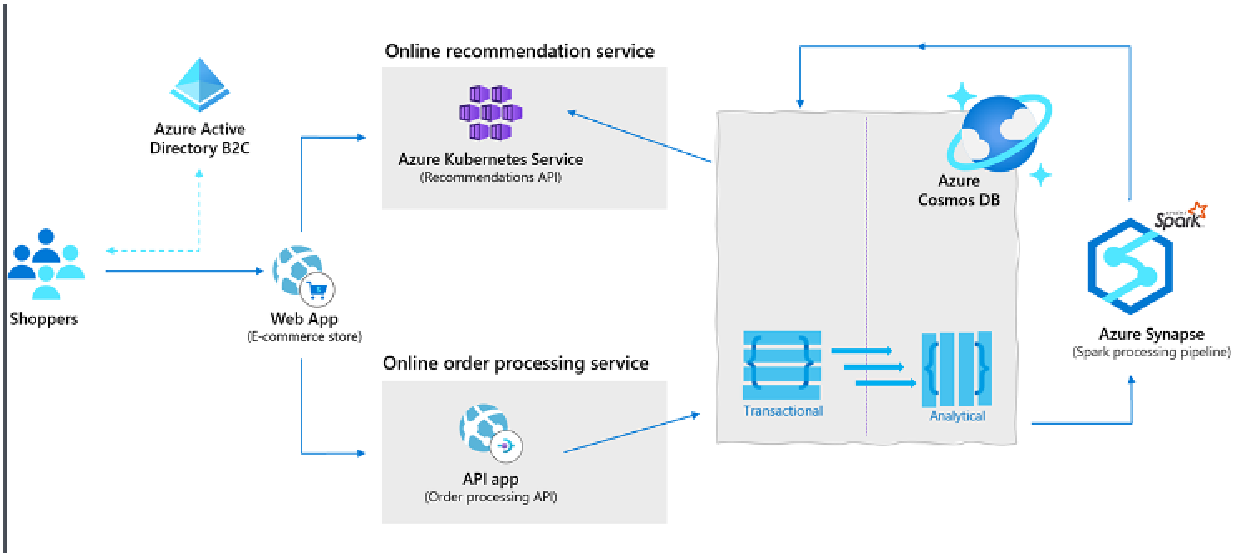
Scenario: You are working at OZcorp which is a supply chain which is generating increasing volumes of operational data every minute for orders, shipments and sales transactions, manufactures and retailers. It needs an operational database that can scale to handle the data volumes as well as an analytical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

Review the following architecture designs.

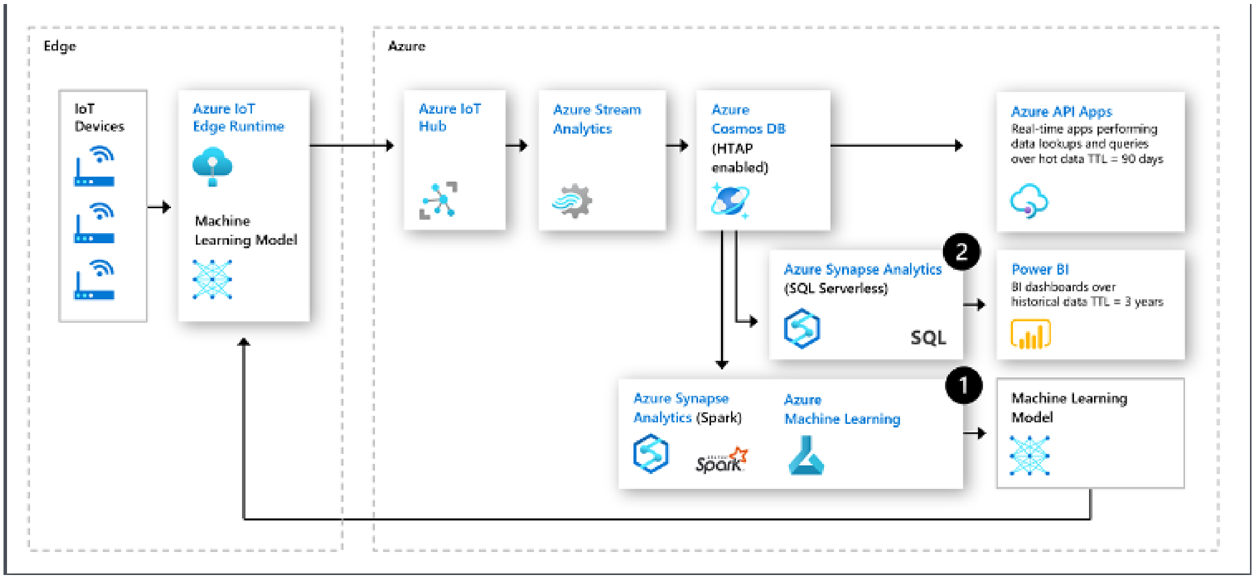
Design A:



Design B:



Design C:


Edge
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

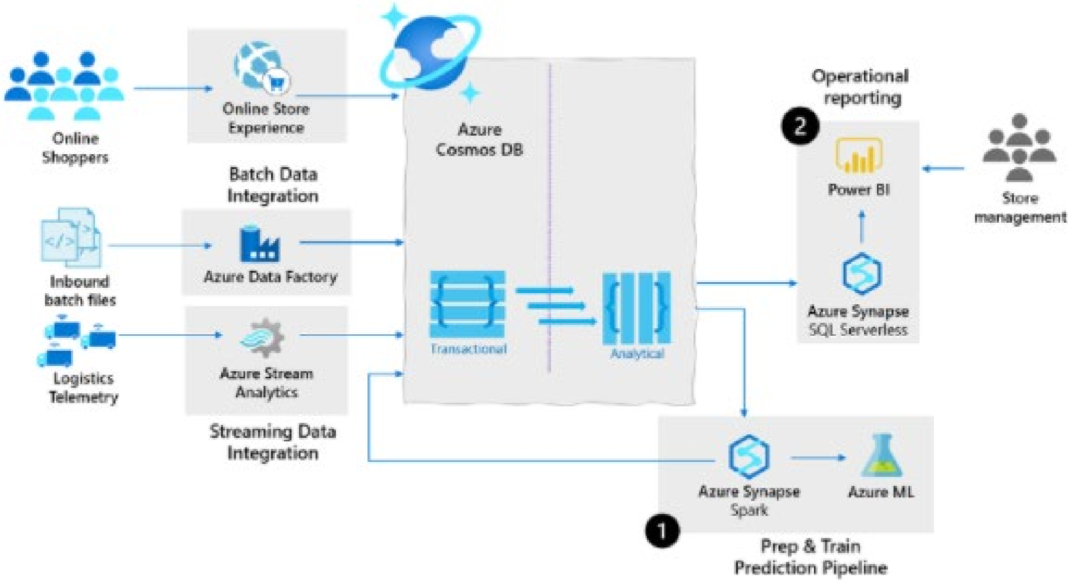
Which design would be best suited for the need?

- ○ Design B
- ○ None of the listed options
- ○ Design C
- ○ Design A (Correct)

Explanation
Supply chain analytics, forecasting and reporting.

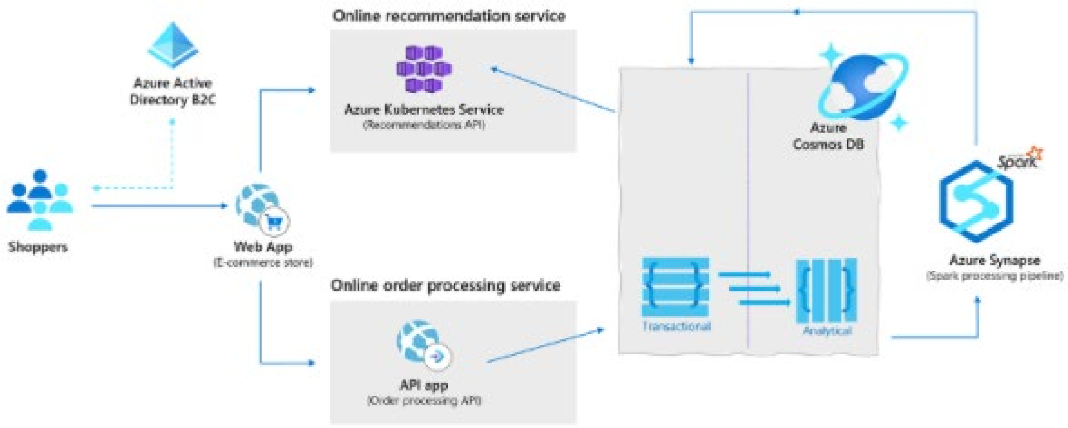
With supply chains generating increasing volumes of operational data every minute for orders, shipments and sales transactions, manufactures and retailers need an operational database that can scale to handle the data volumes as well as an analytical platform to get to a level of real-time contextual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

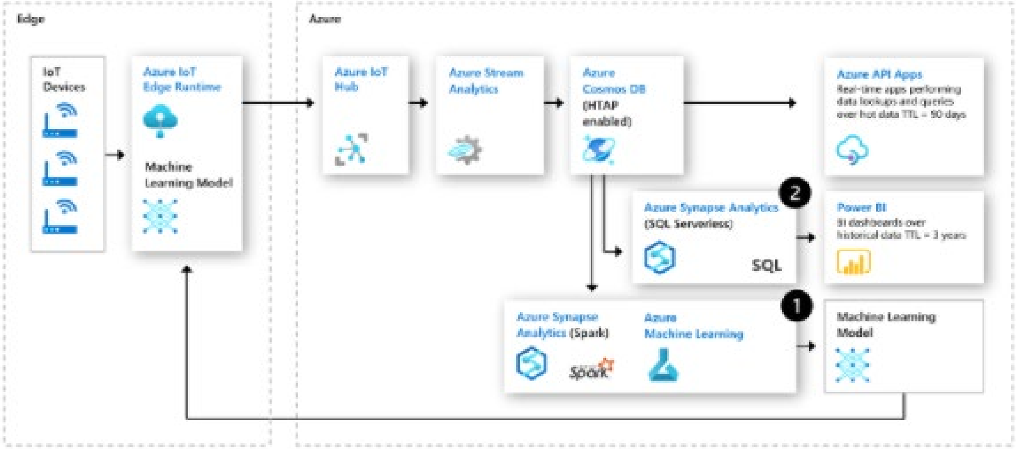
In retail, many web-based retailers will perform real-time basket analysis to make product recommendations to customers who are about to purchase products. This increased revenues for these organizations as the provided targeted suggestions at the point of sales.



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
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

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://azureopendatastorage.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

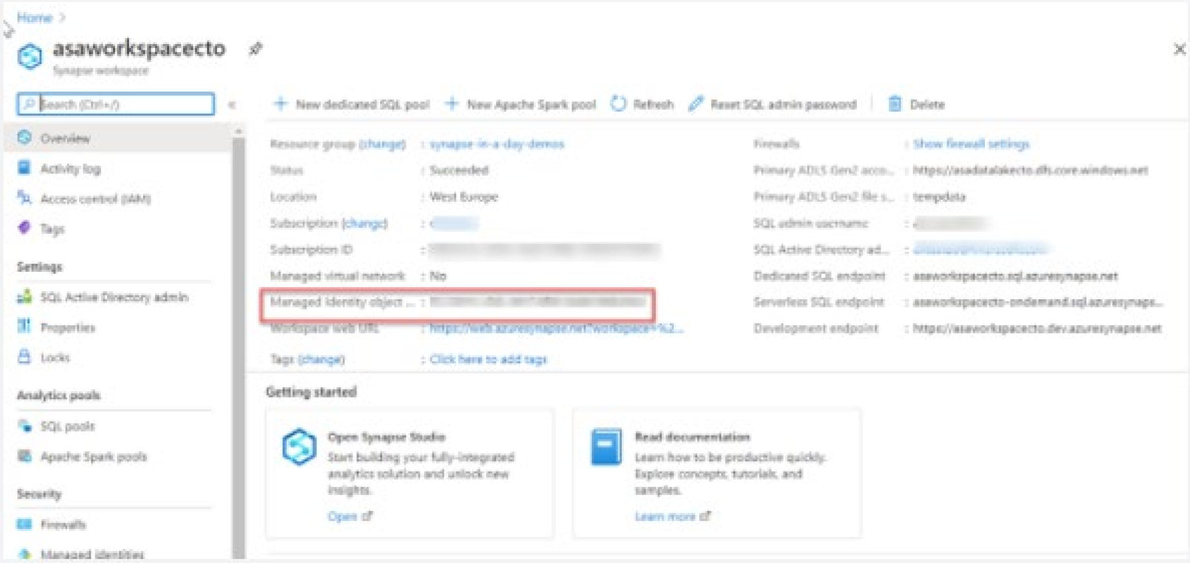
Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD. You can use the Managed Identity capability to authenticate to any service that support Azure Active Directory authentication.

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.azuresynapse...
Development endpoint : https://asaworkspacecto.dev.azuresynapse.net

Getting started

Open Synapse Studio
Start building your fully-integrated analytics solution and unlock new insights.
Open

Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more

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

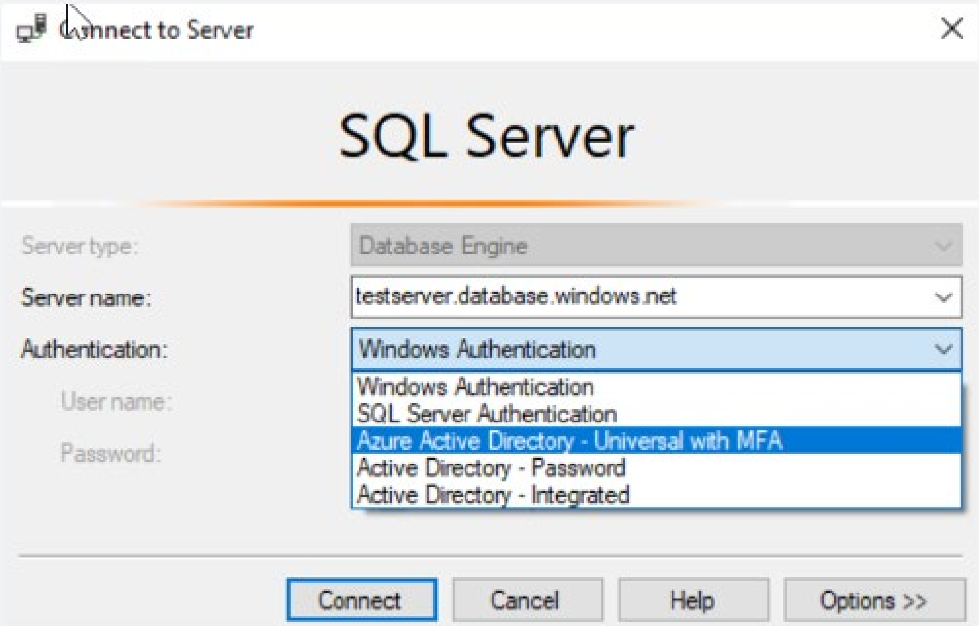
SQL Authentication

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.



Server type: Database Engine
Server name: testserver.database.windows.net
Authentication: Windows Authentication

Connect to Server
SQL Server
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

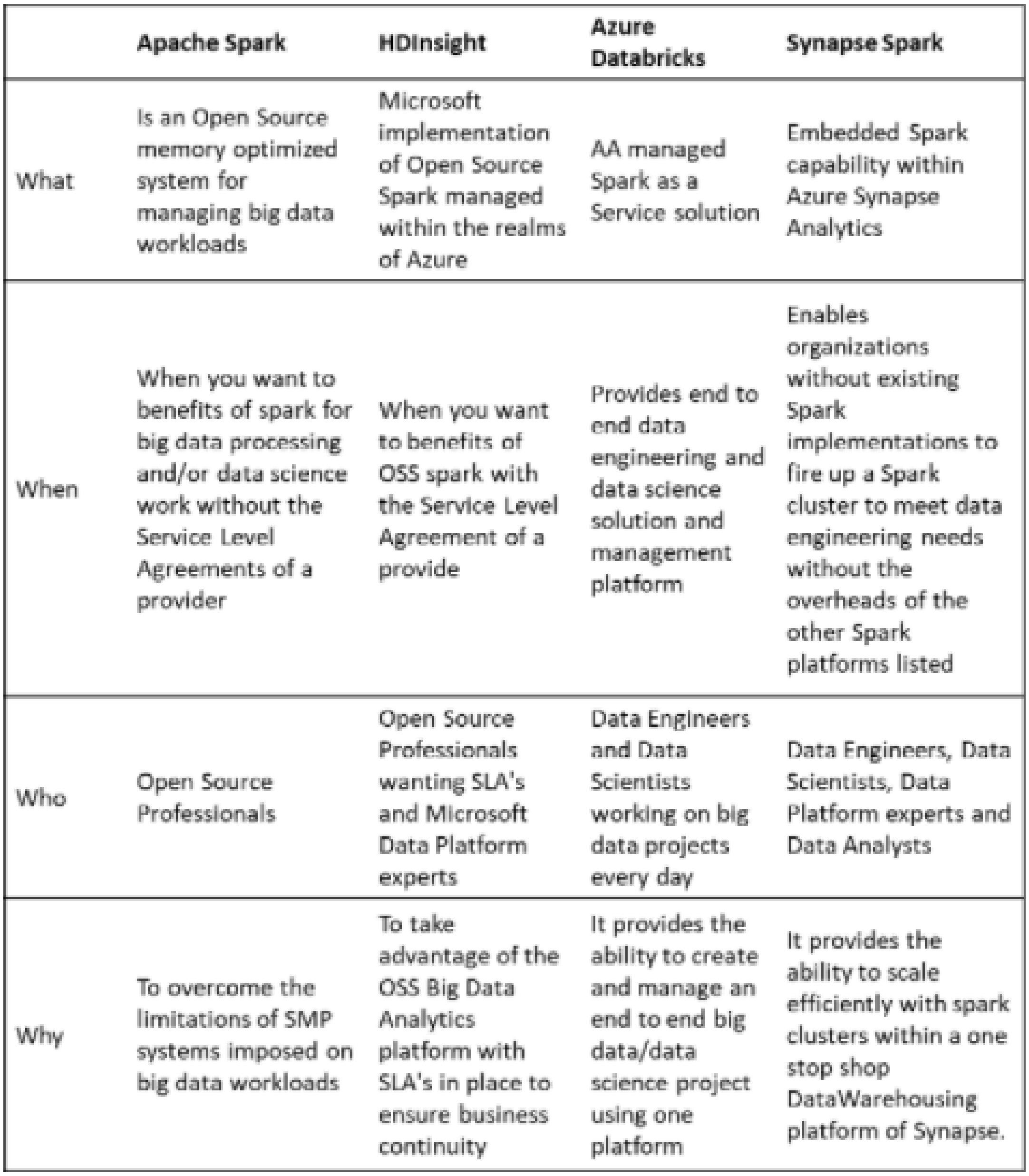
Spark Instances:

• 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)**

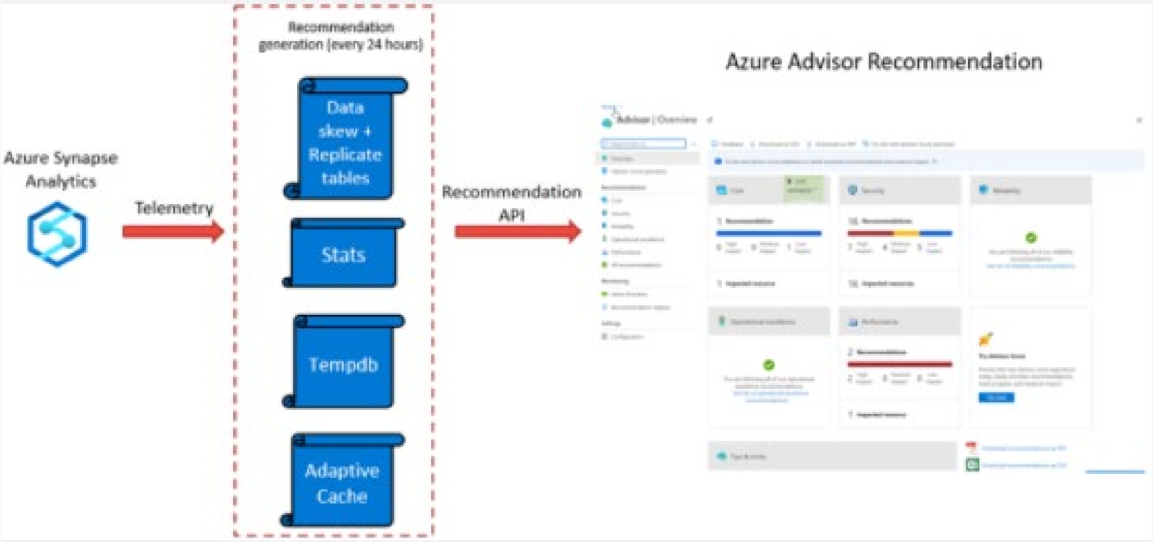
Explanation

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
Advisor | Overview

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.

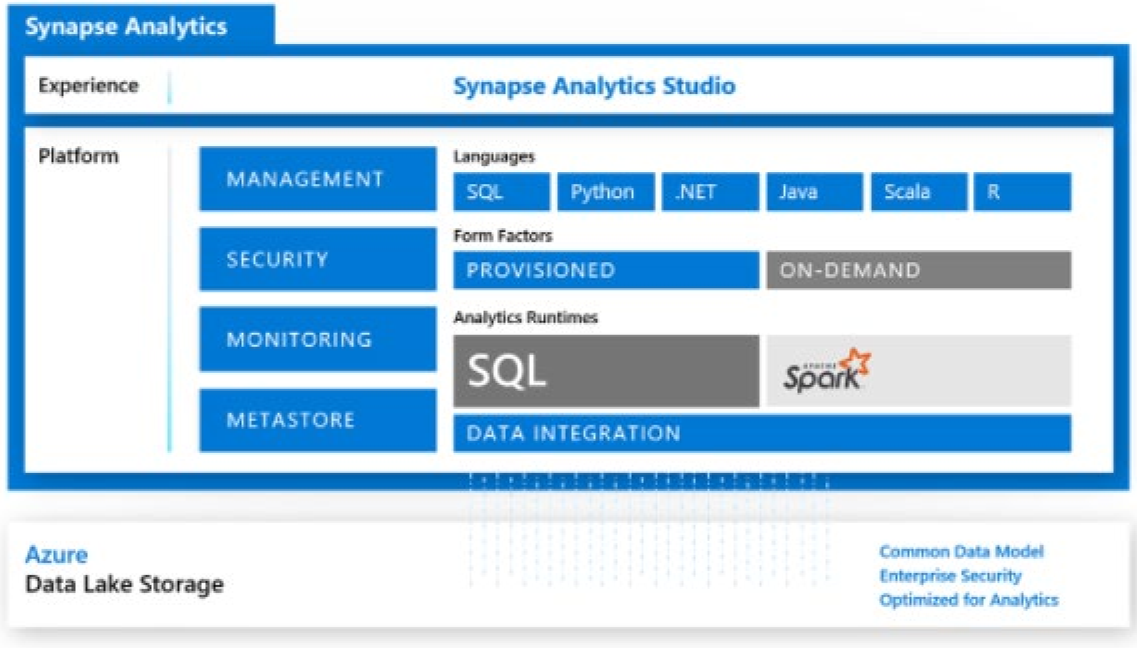
- ○ Azure Synapse Link
- ○ Azure Synapse Pipelines
- ○ Apache Spark for Azure Synapse **(Correct)**
- ○ Azure Cosmos DB
- ○ Azure Synapse SQL

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.

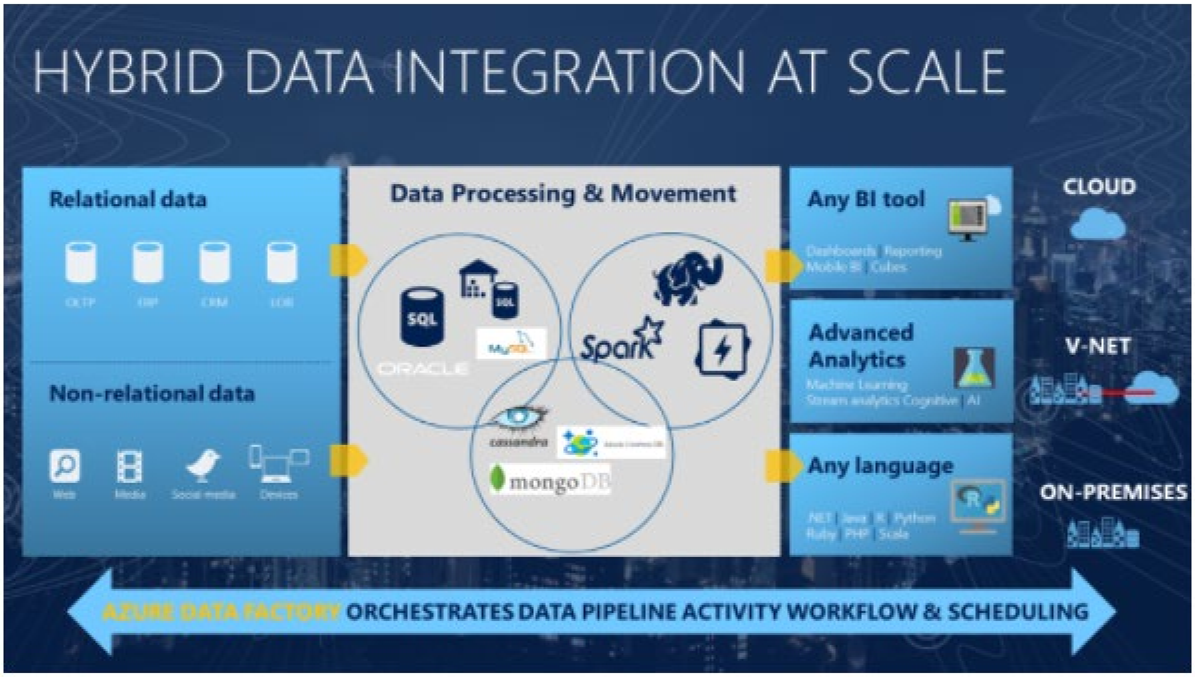
Automation and orchestration are different, but related concepts. Automation helps make your business more efficient by reducing or replacing human interaction with IT systems and instead using software to perform tasks in order to reduce cost, complexity, and errors.

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

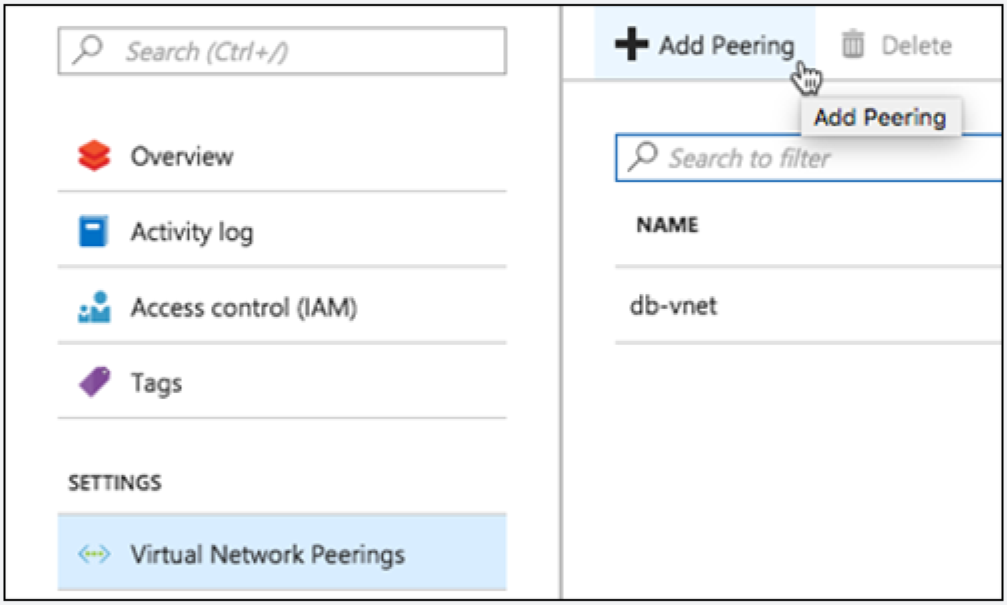
When planning and implementing your Azure Databricks deployments, you have a number of considerations about networking and network security implementation details.

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.

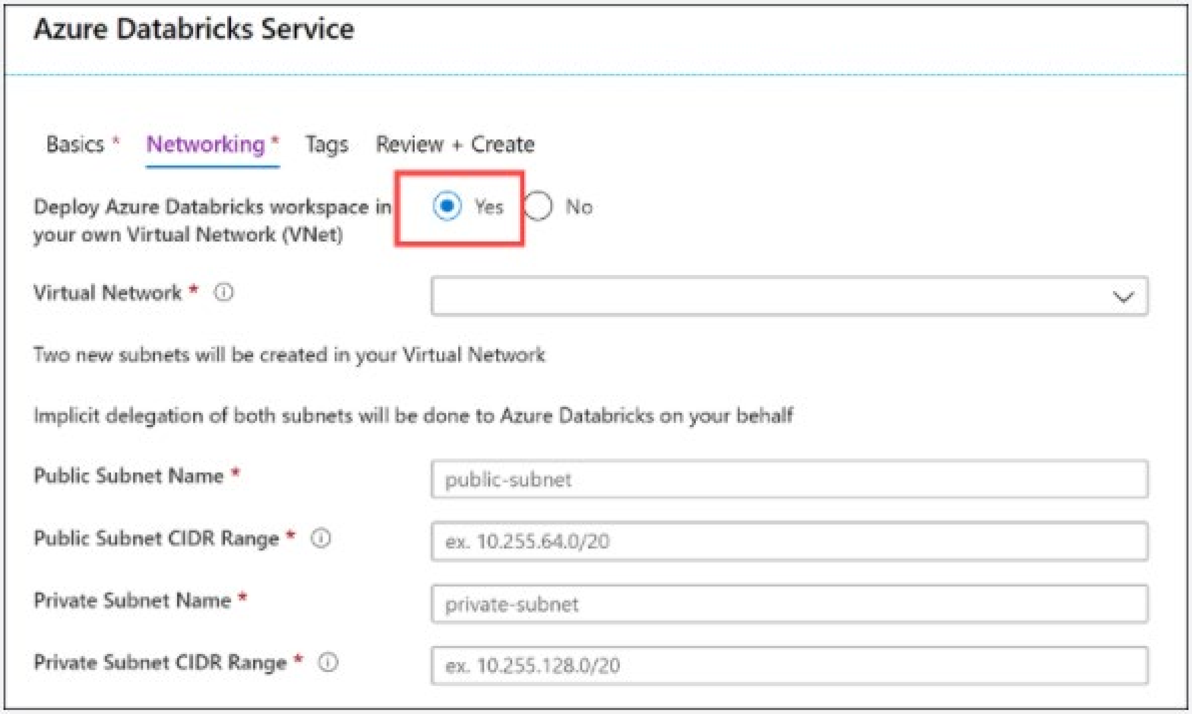
Features enabled through VNet injection include:

- 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

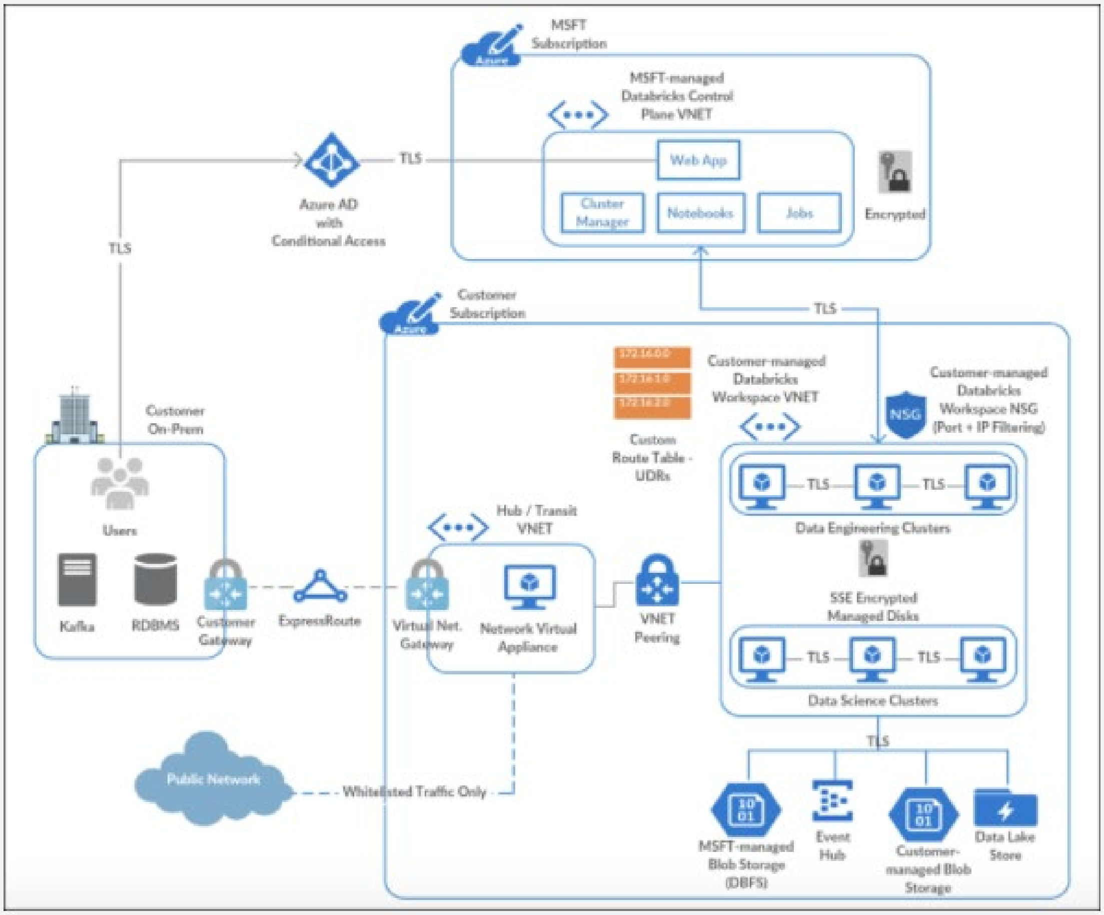
Private Subnet CIDR Range * ⓘ ex. 10.255.128.0/20

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
TLS
Azure AD with Conditional Access
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
Customer-managed Databricks Workspace NSG (Port + IP Filtering)
NSG
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.

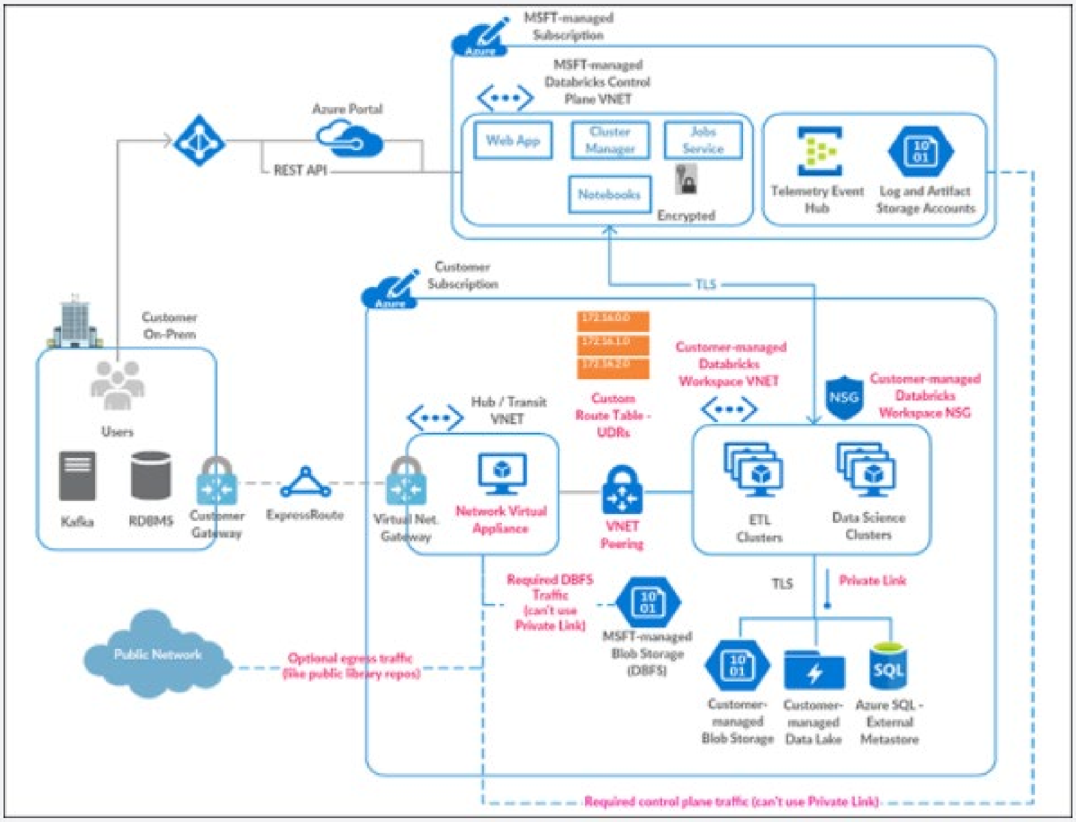
Azure VNet service endpoints

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
Azure Portal
REST API
Web App
Cluster Manager
Jobs Service
Notebooks
Encrypted
Telemetry Event Hub
10 01
Log and Artifact Storage Accounts
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
Customer On-Prem
Users
Kafka
RDBMS
Customer Gateway
ExpressRoute
Virtual Net. Gateway
Hub / Transit VNET
Network Virtual Appliance
VNET Peering
ETL Clusters
Data Science Clusters
TLS
Private Link
Required DBFS Traffic (can't use Private Link)
10 01
MSFT-managed Blob Storage (DBFS)
Public Network
Optional egress traffic (like public library repos)
10 01
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

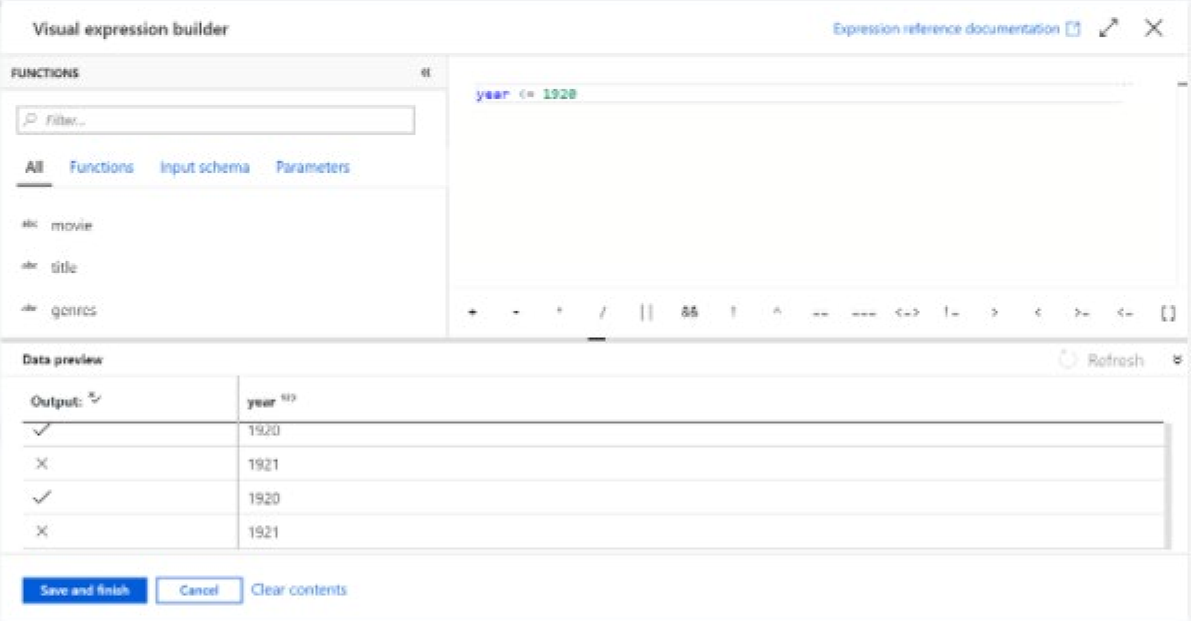
• Row modifier transformations

• 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

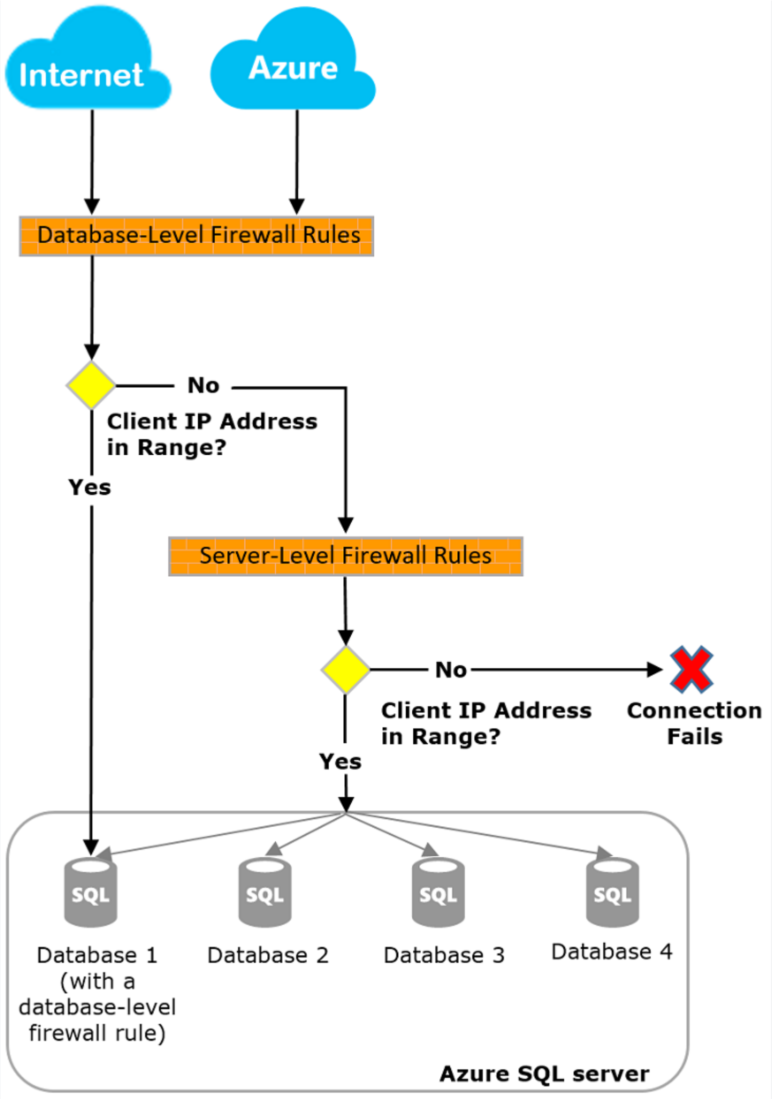
To ensure that the analytical data store is accessible only to the company;s on-premises network and Azure services with the restriction of not using a VPN, a server-level firewall IP rule should be employed. A server-level virtual network rule would be the correct answer if the VPN restriction was not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

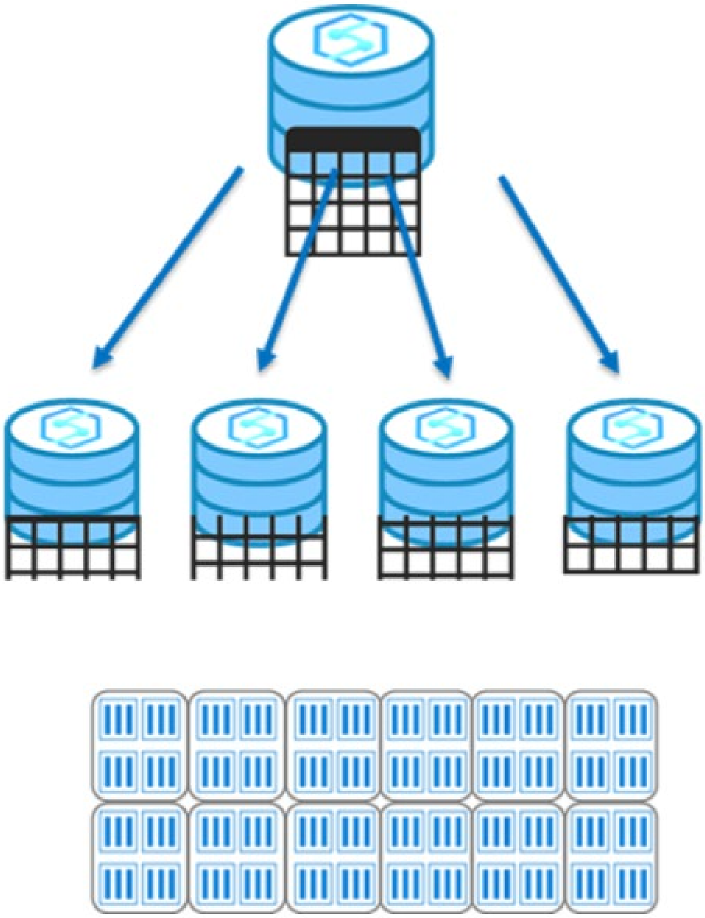
When data is loaded into Synapse Analytics dedicated SQL pools, the datasets are broken up and dispersed among the compute nodes for processing, and then written to a decoupled and scalable storage layer. This action is termed *"sharding"*.

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.

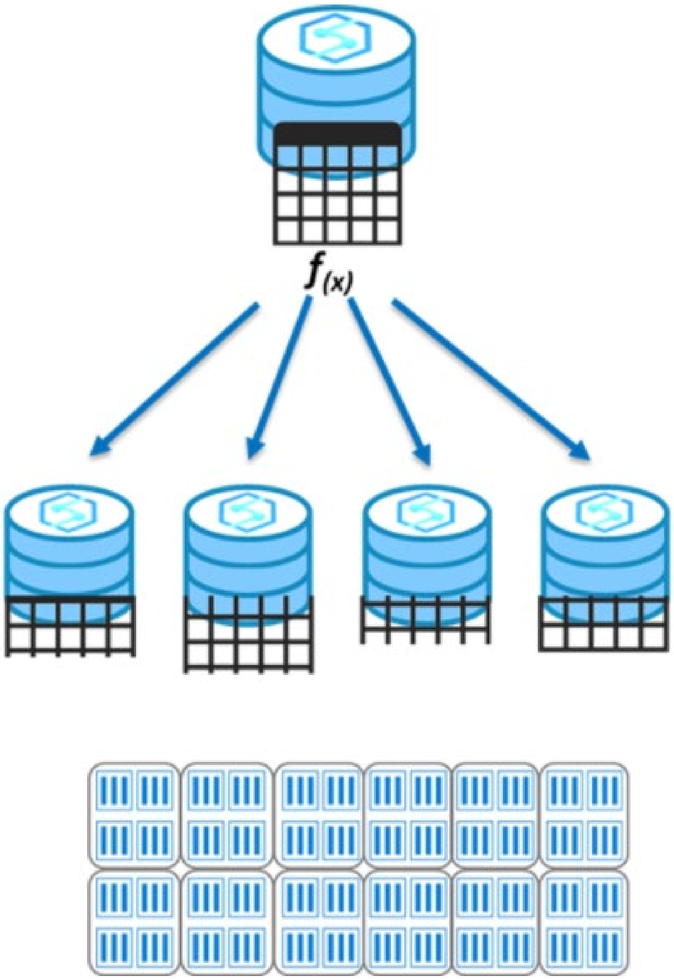
A round-robin distributed table distributes data evenly across the table but without any further optimization. A distribution is first chosen at random and then buffers of rows are assigned to distributions sequentially.

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



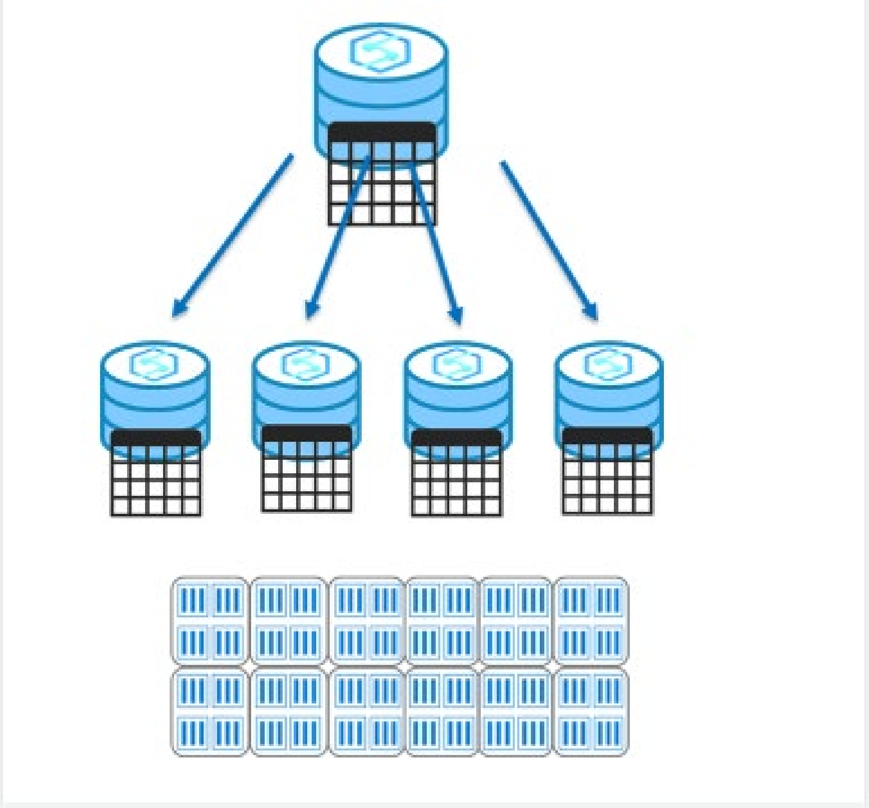
This distribution can deliver the highest query performance for joins and aggregations on large tables.

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.

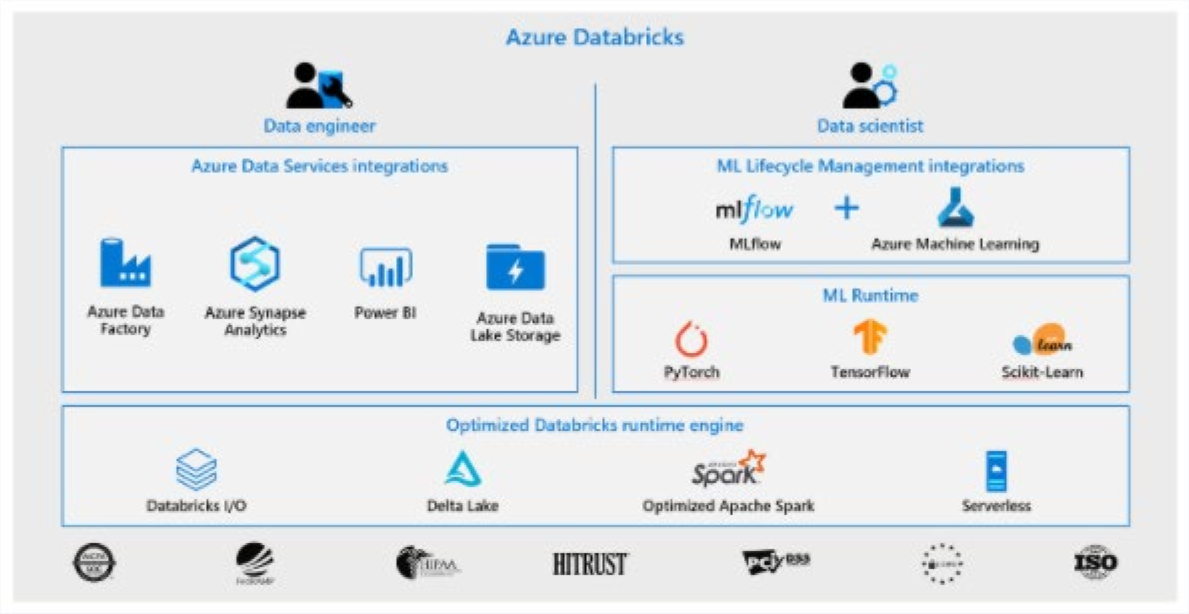
- ○ Azure Database Services
- ○ Azure VNet Peering
- ○ Azure Machine Learning Studio
- ○ Azure Kubernetes Service **(Correct)**

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.

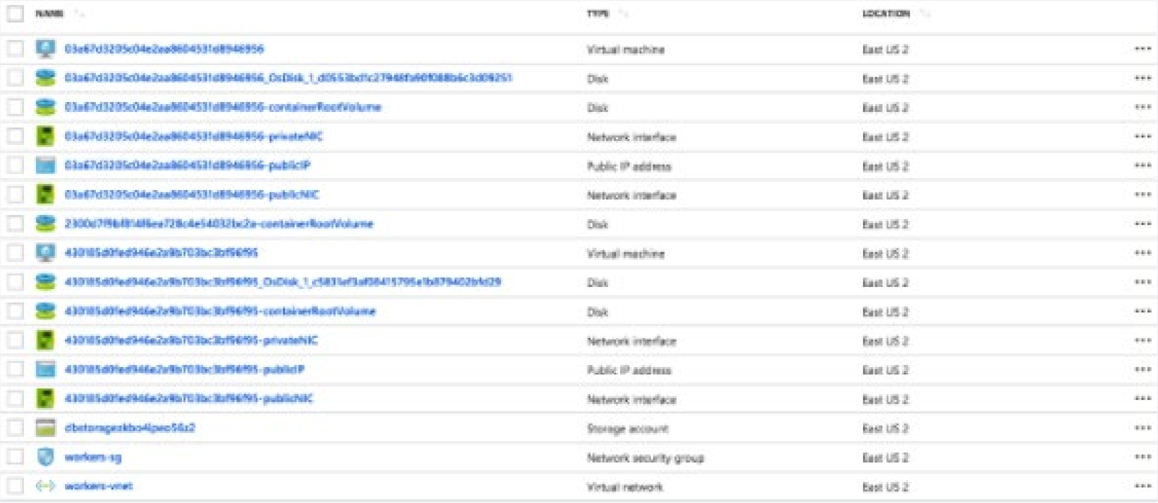
• **Real-time analytics**: Integration with IoT Hub, Azure Event Hubs, and Azure HDInsight Kafka clusters enables developers to build scalable streaming solutions for real-time analytics.

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

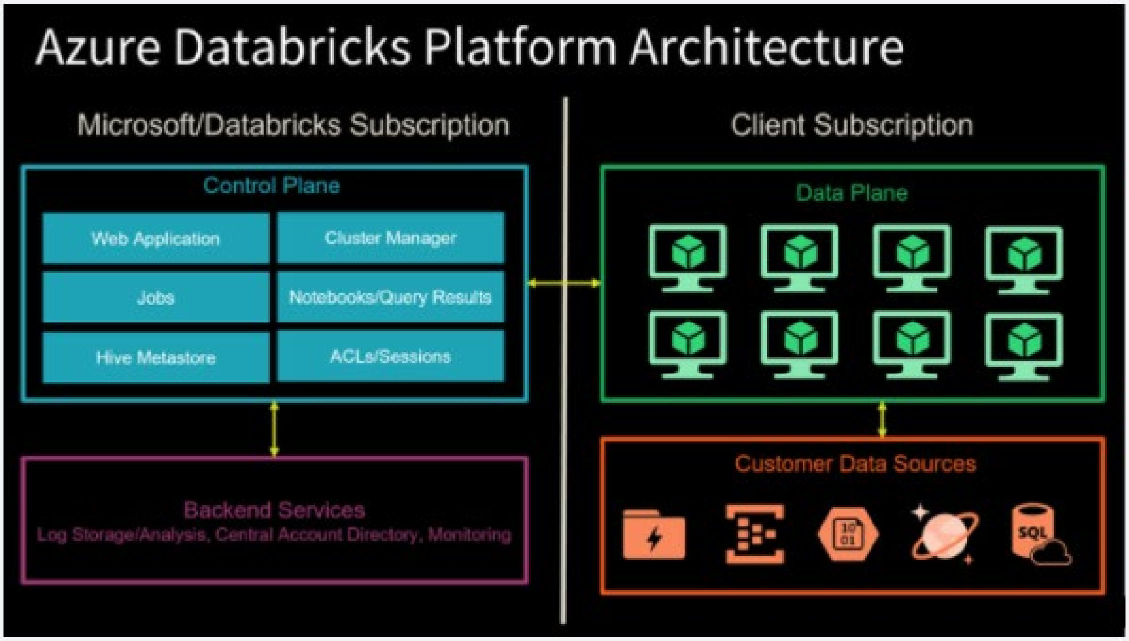
When you create an Azure Databricks service, a "Databricks appliance" is deployed as an Azure resource in your subscription. At the time of cluster creation, you specify the types and sizes of the virtual machines (VMs) to use for both the Driver and Worker nodes, but Azure Databricks manages all other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



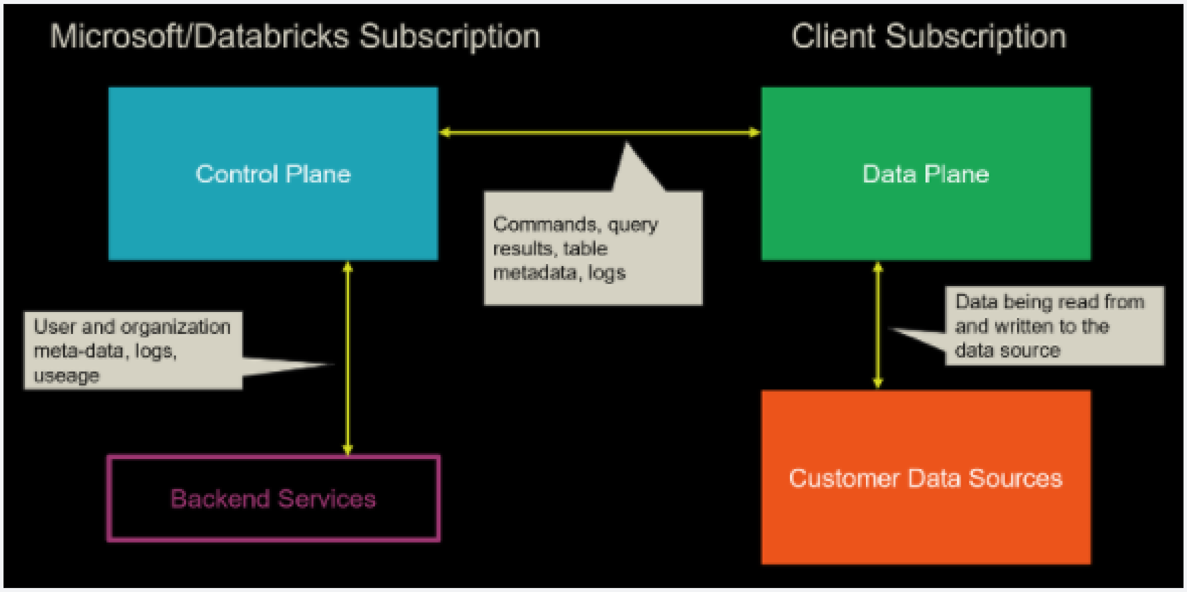
□	NAME	TYPE	LOCATION	
□	03a67d3205c04e2aa8604531d8946956	Virtual machine	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956_OsDisk_1_d0553bd1c27948fa90f088b6c3d09251	Disk	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-containerRootVolume	Disk	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-privateNIC	Network interface	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-publicIP	Public IP address	East US 2	•••
□	03a67d3205c04e2aa8604531d8946956-publicNIC	Network interface	East US 2	•••
□	2300d7f9bf8148ea728c4e54032bc2a-containerRootVolume	Disk	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95	Virtual machine	East US 2	•••
□	430185d0fed946e2a9b703bc3bf96f95_OsDisk_1_c5831ef3af08415795e1b879402bfd29	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	•••
□	dbstoragezkbo4lpeo56z2	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.

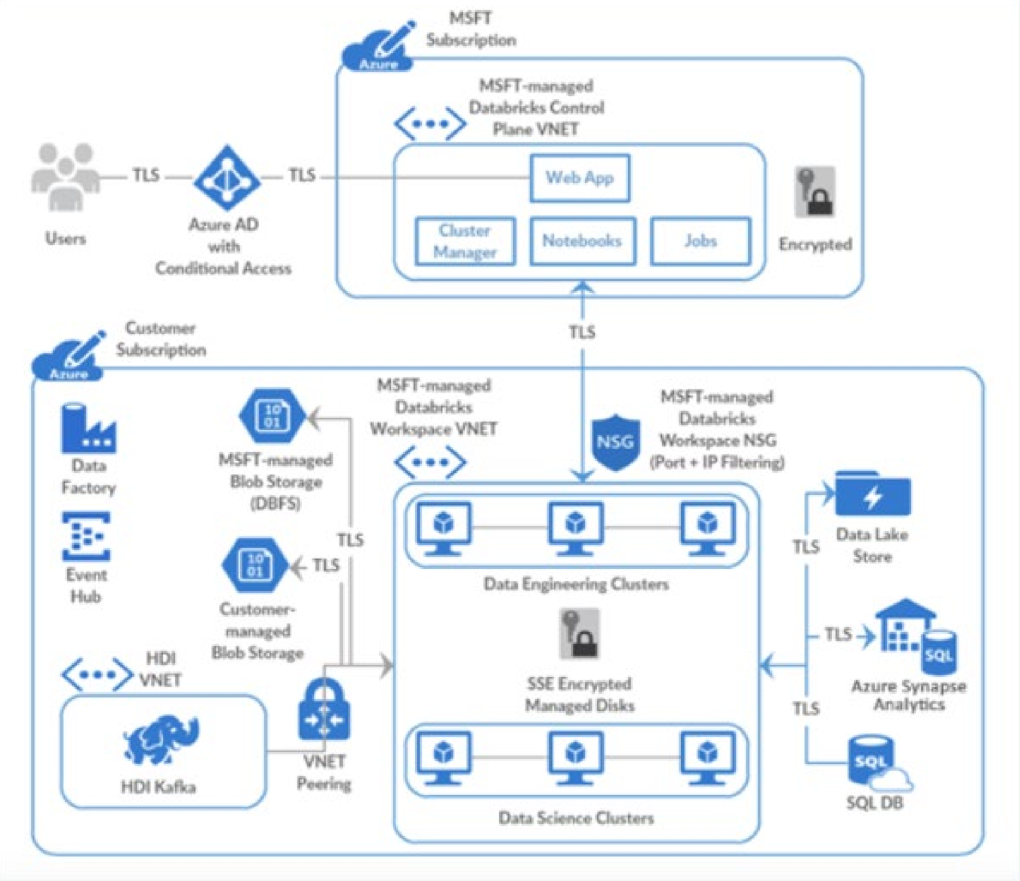


The diagram above shows a Control Plane on the left, which hosts Databricks jobs, notebooks with query results, the cluster manager, web application, Hive metastore, and security access control lists (ACLs) and user sessions. These components are managed by Microsoft in collaboration with Databricks 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
Web App
Cluster Manager
Notebooks
Jobs
Encrypted
Users
TLS
Azure AD with Conditional Access
TLS
TLS
Customer Subscription
Azure
Data Factory
Event Hub
MSFT-managed Blob Storage (DBFS)
MSFT-managed Databricks Workspace VNET
NSG
MSFT-managed Databricks Workspace NSG (Port + IP Filtering)
TLS
TLS
Customer-managed Blob Storage
Data Engineering Clusters
SSE Encrypted Managed Disks
Data Science Clusters
HDI VNET
HDI Kafka
VNET Peering
TLS
Data Lake Store
TLS
Azure Synapse Analytics
TLS
SQL
SQL DB

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:

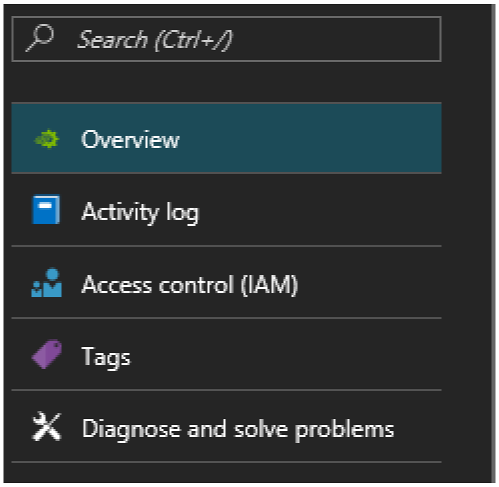
• If the backlog event metric keeps increasing, its an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

• 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.

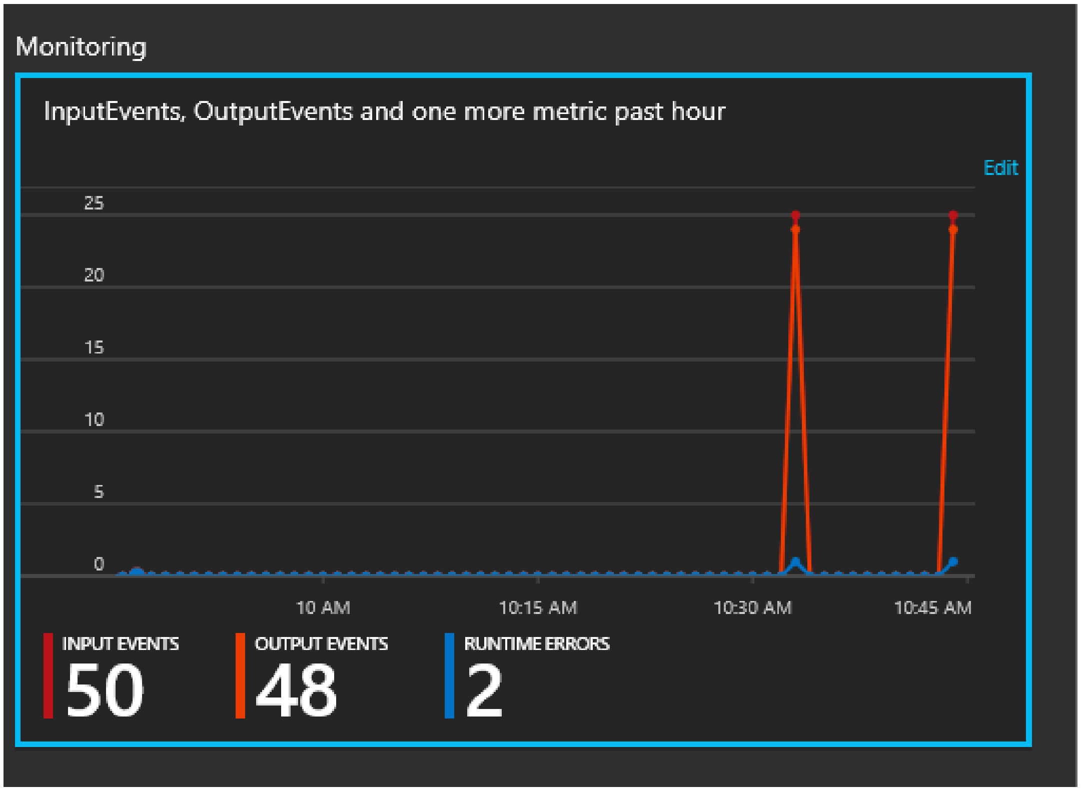
Understand Stream Analytics job monitoring and how to monitor queries

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

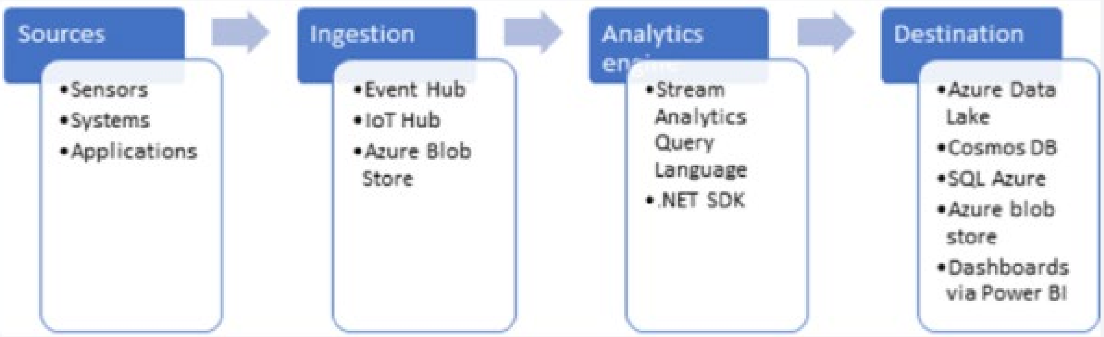
Microsoft Azure Stream Analytics is an event processing engine. It enables the consumption and analysis of high volumes of streaming data generated by sensors, devices, or applications. Stream Analytics processes the data in real time. A typical event processing pipeline built on top of Stream Analytics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

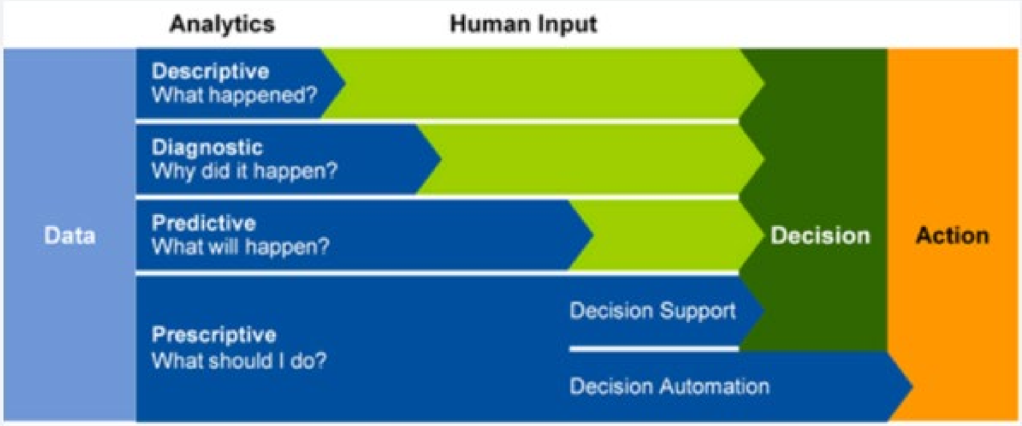
Explanation

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

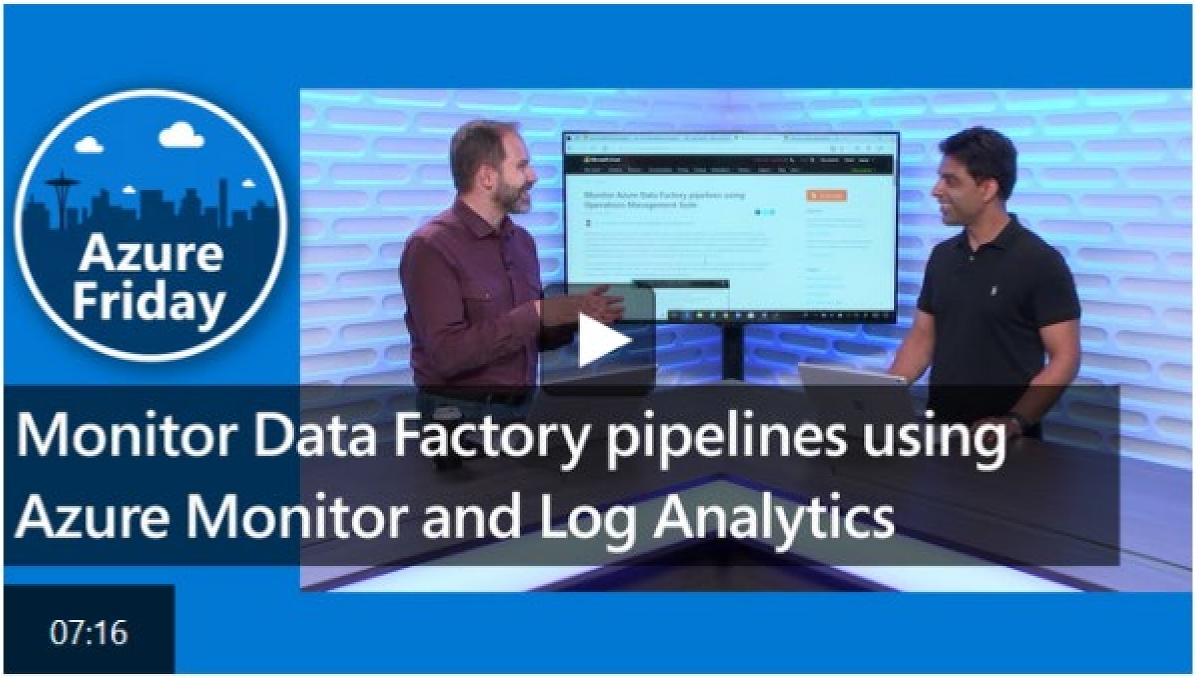
Explanation

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.

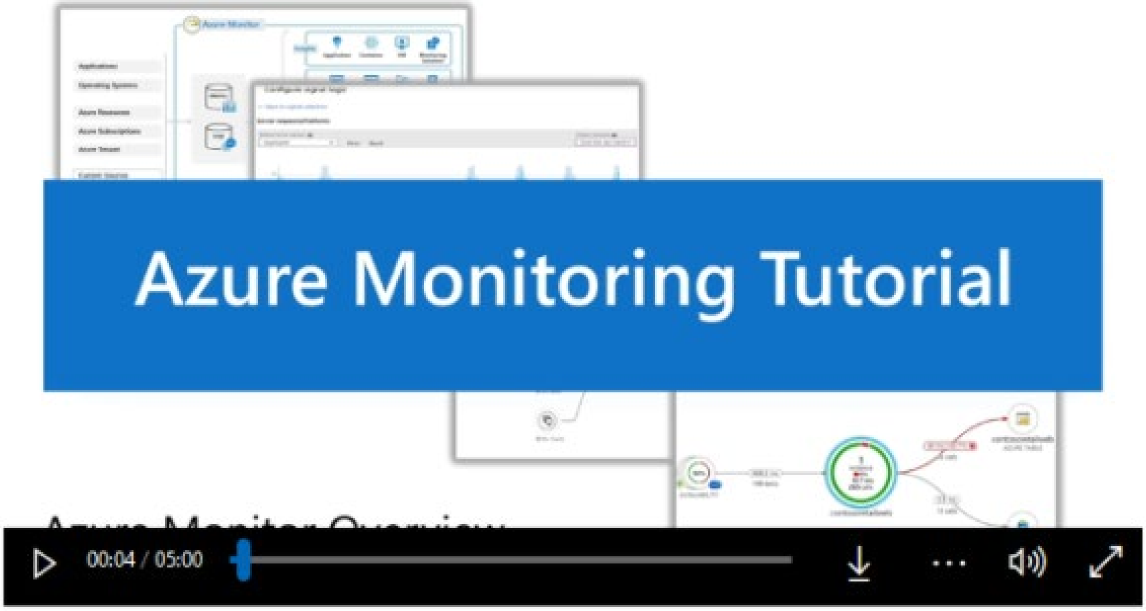
Monitor and Alert Data Factory by using Azure Monitor

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

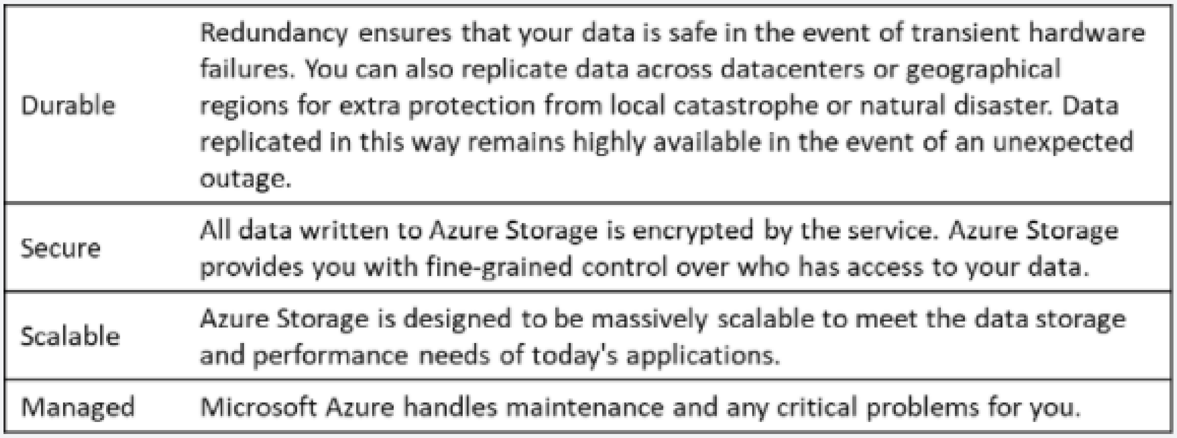
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

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

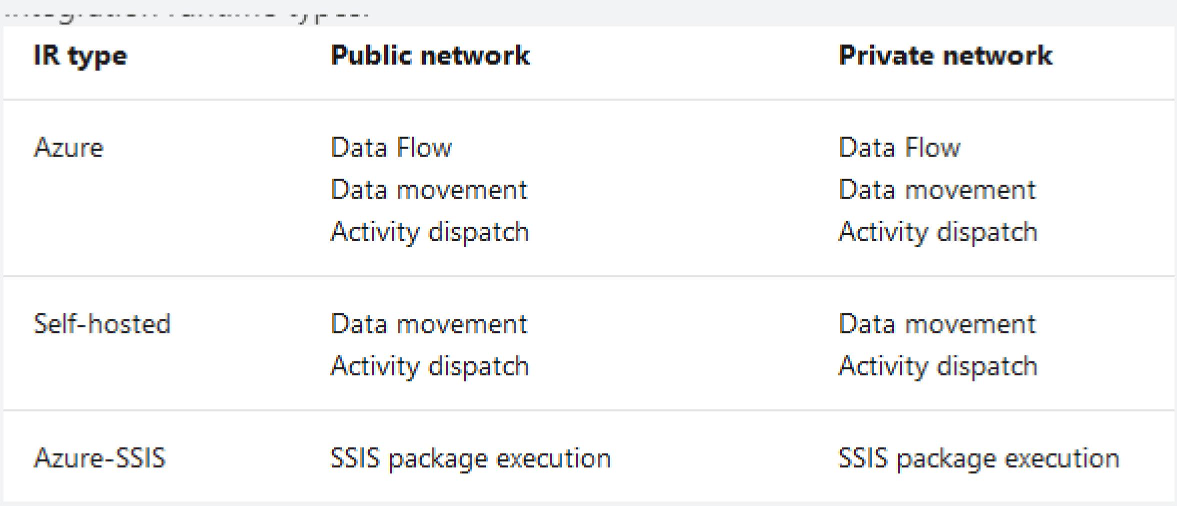
Data Factory offers three types of Integration Runtime (IR), and you should choose the type that best serve the data integration capabilities and network environment needs you're looking for. These three types are:

- 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.

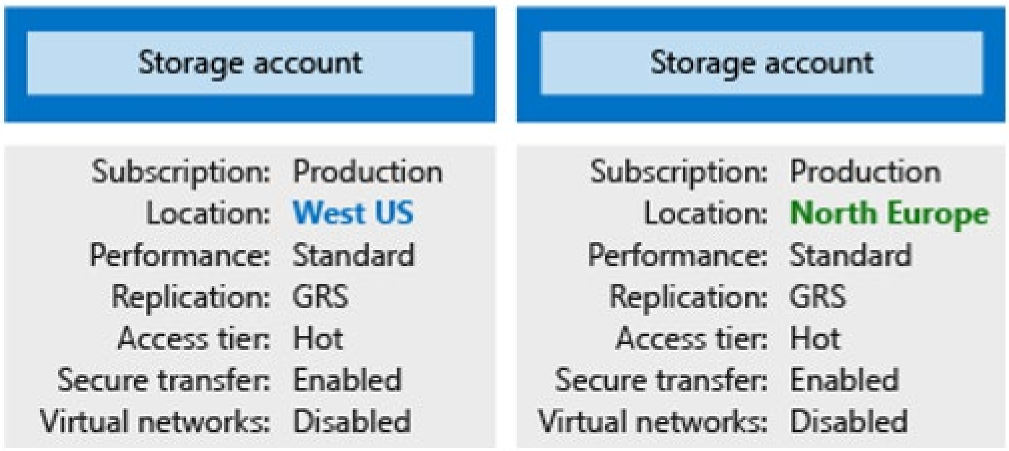
When considering Azure Storage Accounts, which option meet the data diversity requirement?

- ○ 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.