Q. 1 Which Azure data platform is commonly used to process data in an ELT framework?

- ○ Azure Data Lake Storage
- ◉ Azure Data Factory
- ○ Azure Data Catalog
- ○ Azure Stream Analytics
- ○ Azure Databricks

Report Error

◉ Azure Data Factory

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

Report Error

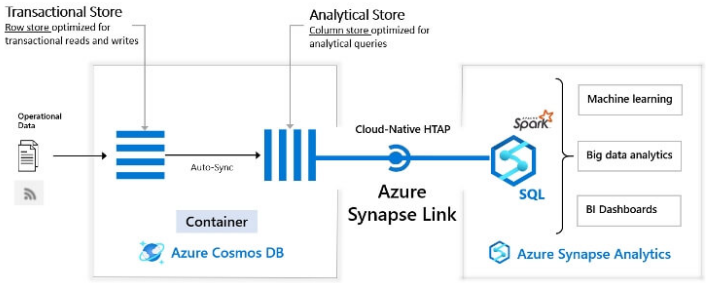
Q. 2 Which Azure Synapse Analytics component enables you to perform Hybrid Transactional and Analytical Processing?

- ○ None of the listed options
- ○ Azure Data Warehouse
- ○ Azure Synapse Spark pools
- ◉ Azure Synapse Studio
- ○ Azure Stream Analytics
- ○ Azure Synapse Link

Report Error

Azure Synapse Link

Explanation:- Azure Synapse Link is the component that enables Hybrid Transactional and Analytical Processing.
Azure Synapse Link for Azure Cosmos DB is a cloud-native hybrid transactional and analytical processing (HTAP) capability that enables you to run near real-time analytics over operational data in Azure Cosmos DB. Azure Synapse Link creates a tight seamless integration between Azure Cosmos DB and Azure Synapse Analytics.
Using Azure Cosmos DB analytical store, a fully isolated column store, Azure Synapse Link enables no Extract-Transform-Load (ETL) analytics in Azure Synapse Analytics against your operational data at scale. Business analysts, data engineers and data scientists can now use Synapse Spark or Synapse SQL interchangeably to run near real-time business intelligence, analytics, and machine learning pipelines. You can achieve this without impacting the performance of your transactional workloads on Azure Cosmos DB.
The following image shows the Azure Synapse Link integration with Azure Cosmos DB and Azure Synapse Analytics:
https://docs.microsoft.com/en-us/azure/cosmos-db/synapse-link



Q. 3

Azure Data Factory is a 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 Data Factory, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.

Correct or incorrect: Each data factory has a single dedicated pipeline. When additional pipelines are needed for workloads, additional data factory deployments can be used to create an unlimited number of pipelines.

- Correct
- Incorrect

Report Error

Incorrect

Explanation:- Azure Data Factory is a 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 Data Factory, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.
A data factory can have one or more pipelines. A pipeline is a logical grouping of activities that together perform a task. For example, a pipeline could contain a set of activities that ingest and clean log data, and then kick off a mapping data flow to analyze the log data. The pipeline allows you to manage the activities as a set instead of each one individually. You deploy and schedule the pipeline instead of the activities independently.
https://docs.microsoft.com/en-us/azure/data-factory/introduction

Report Error

Q. 4

Scenario: Honest Eddie's Car Dealership is an establishment in South Carolina USA, which is dedicated to the purchase and sale of cars and light trucks. Eddie has hired you as an expert consultant for Azure projects and you are holding a workgroup session with the team discussing inputs for Azure Stream Analytics jobs. The team has collectively created a list of key benefits of using Azure Stream Analytics to process streaming data. Eddie thinks that some of the responses are not correct and has asked you to confirm which are not correct.

Which of the following statements are not primary benefits of using Azure Stream Analytics to process streaming data?

- ○ Being able to rapidly deploy queries into production by creating and starting an Azure Stream Analytics job
- ◉ The ability to write and test transformation queries in the Azure portal
- ○ All of the listed options are primary benefits
- ○ Integration with Azure Blob storage

Report Error

- ○ Being able to rapidly deploy queries into production by creating and starting an Azure Stream Analytics job
- ○ The ability to write and test transformation queries in the Azure portal
- ○ All of the listed options are primary benefits
- ◉ Integration with Azure Blob storage

Explanation:-

Integration with Azure Blob storage is not one of the primary benefits of using Azure Stream Analytics to process streaming data. The integration with Blob storage can be used to process static data.

Azure Stream Analytics, the recommended service for stream analytics on Azure, easily integrates with your applications and connected devices and sensors to transform streaming data. The process of consuming data streams, analyzing them, and deriving actionable insights is called stream processing. You can transform streaming data using the SQL-like Stream Analytics Query Language to perform temporal and other aggregations against a data stream to gain insights. Streaming data inputs include Azure Event Hubs and IoT Hub. And static data held in Blob storage can also be processed using Stream Analytics jobs.

Stream processing refers to the continuous ingestion, transformation, and analysis of data streams generated by applications, IoT devices and sensors, and other sources to derive actionable insights in near-real-time. Data stream analysis frequently involves using temporal operations, such as windowed aggregates, temporal joins, and temporal analytic functions to measure changes or differences over time. The intent being to:

- Continuously monitor data using time-boxes windows to understand better how specific areas of interest change or fluctuate over time
- Identify and react to anomalies or irregularities within data in real-time
- Perpetually analyze new data to identify and respond to issues in real-time
- Trigger specific actions when certain thresholds are identified

The exponential propagation of connected applications, devices, and sensors has fuelled the necessity for organizations to analyze streaming data as it arrives and use the latent knowledge contained within the data to make business decisions in near-real-time. Some example use cases of streaming data analysis include:

- Anomaly detection to identify potentially fraudulent transactions in finance industries
- Making product recommendations to online customers in real-time
- Monitoring pipelines and distribution systems by oil companies
- Generating predictive maintenance schedules for industrial and manufacturing equipment
- Sentiment analysis of social media posts

Approaches to data stream processing

The primary approach to stream processing is to analyze new data continuously, transforming incoming data as it arrives to facilitate near-real-time insights. Computations and aggregations can be executed against the data using temporal analysis and sent to a Power BI dashboard for real-time visualization and analysis. This approach typically involves persisting the streaming data into a data store, such as Azure Data Lake Storage (ADLS) Gen2, for further examination or more advanced analytics workloads.

An alternative approach for processing streaming data is to persist incoming data in a data store, such as Azure Data Lake Storage (ADLS) Gen2. You can then process the static data in batches at a later time. This approach is frequently used to take advantage of lower compute costs when processing large sets of existing data.

Azure Stream Analytics is the recommended service for stream analytics on Azure. Stream Analytics provides you with the ability to ingest, process, and analyze streaming data from Azure Event Hubs (including Azure Event Hubs from Apache Kafka) and Azure IoT Hub. You can also configure static data ingestion from Azure Blob Storage. This integration allows you to quickly create hot-path analytics pipelines to generate powerful insights to drive real-time actions. Azure Stream Analytics is meant for a wide range of scenarios that include, but aren't limited to:

• Dashboards for data visualization
•Real-time alerts from temporal and spatial patterns or anomalies
• Extract, Transform, Load (ETL)
• Event Sourcing pattern
• IoT Edge

Contoso collects vehicle telemetry and wants to use Event Hubs to ingest and store the data in its raw form. They want to perform several aggregations on the telemetry data in near-real-time. In the end, they would like to visualize the aggregated data on a dashboard that automatically updates with new data as it arrives. They want the dashboard to contain various visualizations of detected anomalies, like engines overheating, abnormal oil pressure, and aggressive driving. They are interested in utilizing a mapping component to show irregularities related to locations, as well as various charts and graphs depicting this information. The CIO asked you to assist them in setting up a near-real-time analytics pipeline built on Event Hubs, Azure Stream Analytics, and Power BI.

The fastest way to get streaming analytics running in Azure is to add an Azure Stream Analytics job to your application. Your Stream Analytics job would ingest the streaming data from one of the supported inputs and run real-time analytics queries against the streams. The built-in integration with Azure Event Hubs and IoT Hub provides a rapid mechanism to create these streaming analytics pipelines. Stream Analytics also supports various inputs and outputs. It also provides the capability to use Azure Machine Learning functions to make it a robust tool for analyzing data streams. The primary benefits of processing streaming data with Azure Stream Analytics include:

• The ability to preview and visualize incoming data directly in the Azure portal.
• Using the Azure portal to write and test your transformation queries using the SQL-like Stream Analytics Query Language (SAQL). You can use the built-in functions of SAQL to find interesting patterns from the incoming stream of data.
• Rapid deployment of your queries into production by creating and starting an Azure Stream Analytics job.

With Azure Stream Analytics, you can quickly stand up real-time dashboards and alerts. An example of a simple solution, as depicted in the diagram below, includes ingesting streaming data from Event Hubs or IoT Hub. The streaming data can then be transformed using Stream Analytics windowing queries. The aggregated data are then sent to a Power BI dashboard with a streaming data set.

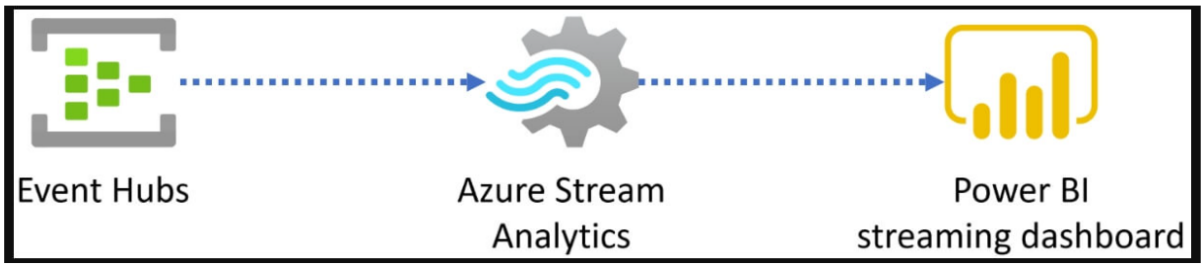
The rich out-of-the-box functionality of Azure Stream Analytics allows you to immediately take advantage of the following features without any additional setup:

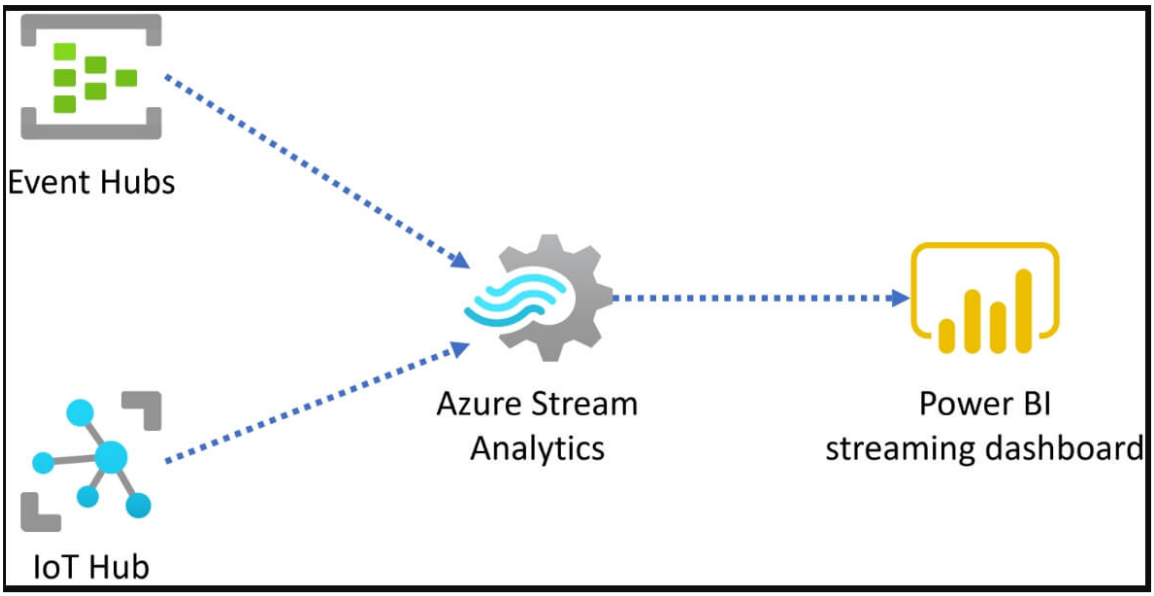
• Built-in temporal operators, such as windowed aggregates, temporal joins, and temporal analytic functions
• Native Azure input and output adapters
• Support for slow-changing reference data (also known as lookup tables), including joining with geospatial reference data for geofencing
• Integrated solutions, such as Anomaly Detection
• Multiple time windows in the same query
• Ability to compose multiple temporal operators in arbitrary sequences

Operational aspects

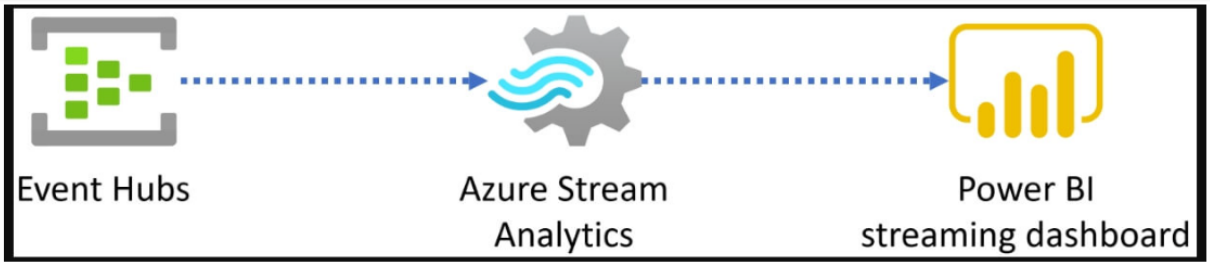
Azure Stream Analytics is a PaaS job service, so it is fully managed and highly reliable. You do not have to spend time managing clusters or worrying about downtime. Job-level billing ensures low startup costs (three Streaming Units, by default). And, jobs are scalable up to 192 Streaming Units to provide the performance necessary to run even the most demanding jobs effectively. It's much more cost-effective to run a few Stream Analytics jobs than run and maintain a Spark cluster.

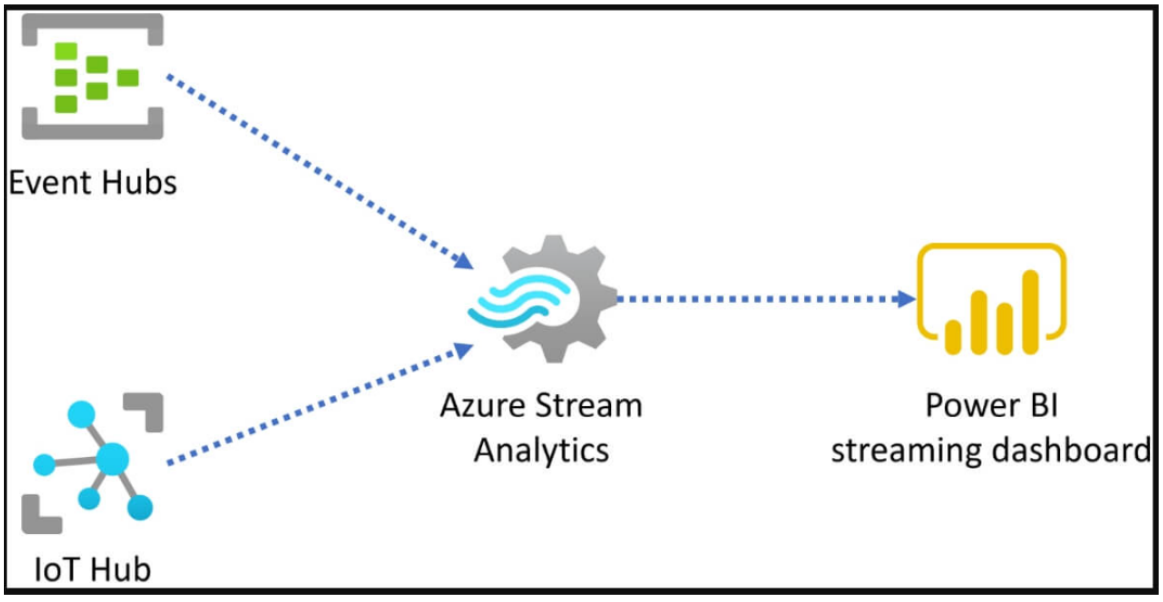
Streaming Units (SUs) represents the computing resources designated to execute Stream Analytics jobs. Increasing the number of SUs means more CPU and memory resources are allocated to the job. Azure Stream Analytics jobs perform all processing in memory to achieve the low latency required for efficient stream processing. Handling compute capacity in this manner allows you to focus on writing queries and leaves hardware management to Microsoft.





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





Q. 5 Which Workload Management capability manages minimum and maximum resource allocations during peak periods?

- ○ Workload Isolation
- ○ Workload Importance
- ◉ Workload Classification
- ○ Workload Containment

Report Error

Q. 5 Which Workload Management capability manages minimum and maximum resource allocations during peak periods?

- ○ Workload Containment
- ○ Workload Classification
- ○ Workload Importance
- ◉ Workload Isolation

Explanation:- Workload Isolation assigns maximum and minimum usage values for varying resources under load. These adjustments can be done live without having to take the SQL Pool offline.
Dedicated SQL pool workload management in Azure Synapse consists of three high-level concepts:
• Workload Classification
• Workload Importance
• Workload Isolation
Workload isolation
Workload isolation reserves resources for a workload group. Resources reserved in a workload group are held exclusively for that workload group to ensure execution. Workload groups also allow you to define the amount of resources that are assigned per request, much like resource classes do. Workload groups give you the ability to reserve or cap the amount of resources a set of requests can consume. Finally, workload groups are a mechanism to apply rules, such as query timeout, to requests.
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-workload-management

Report Error

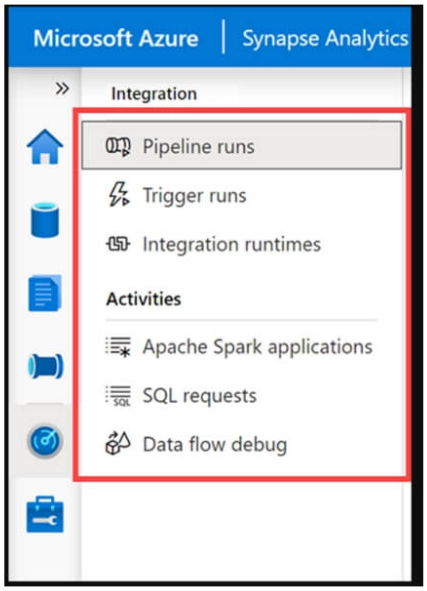
Q. 6 In Azure Synapse Studio, use the Monitor hub is where you access which of the following?

- ○ Integration runtimes
- ◉ All of these
- ○ Trigger runs
- ○ Apache Spark jobs
- ○ SQL requests
- ○ Pipeline runs

Report Error

All of these

Explanation:- In Azure Synapse Studio, use the Monitor hub to view pipeline and trigger runs, view the status of the various integration runtimes that are running, view Apache Spark jobs, SQL requests, and data flow debug activities.
The Monitor hub is your first stop for debugging issues and gaining insight on resource usage. You can see a history of all the activities taking place in the workspace and which ones are active now.
https://techcommunity.microsoft.com/t5/azure-synapse-analytics/explore-the-monitor-hub-in-synapse-studio-to-keep-track-of-all/ba-p/1987405



Q. 7

Scenario: Honest Eddie's Car Dealership is an establishment in South Carolina USA, which is dedicated to the purchase and sale of cars and light trucks. Currently the IT team is planning to use applications which publish messages to Azure Event Hub very frequently.

By default, how many partitions will a new Event Hub have?

- 2
- 3
- 1
- 4

Report Error

Q. 7

Scenario: Honest Eddie's Car Dealership is an establishment in South Carolina USA, which is dedicated to the purchase and sale of cars and light trucks. Currently the IT team is planning to use applications which publish messages to Azure Event Hub very frequently.

By default, how many partitions will a new Event Hub have?

◉ 4

Explanation:-

Event Hubs default to 4 partitions. Partitions are the buckets within an Event Hub. Each publication will go into only one partition. Each consumer group may read from one or more than one partition.

Big data apps must be able to process increased throughput by scaling out to meet increased transaction volumes.

Suppose you work in the credit card department of a bank. You're part of a team that manages the system responsible for fraud testing to determine whether to approve or decline each transaction. Your system receives a stream of transactions and needs to process them in real time.

The load on your system can spike during weekends and holidays. The system must handle the increased throughput efficiently and accurately. Given the sensitive nature of the transactions, even the slightest error can have a considerable impact.

Azure Event Hubs can receive and process a large number of transactions. It can also be configured to scale dynamically, when required, to handle increased throughput.

What is an Azure Event Hub?

Azure Event Hubs is a cloud-based, event-processing service that can receive and process millions of events per second. Event Hubs acts as a front door for an event pipeline, to receive incoming data and stores this data until processing resources are available.

An entity that sends data to the Event Hubs is called a publisher, and an entity that reads data from the Event Hubs is called a consumer or a subscriber. Azure Event Hubs sits between these two entities to divide the production (from the publisher) and consumption (to a subscriber) of an event stream. This decoupling helps to manage scenarios where the rate of event production is much higher than the consumption. The following illustration shows the role of an Event Hub.

Events

An event is a small packet of information (a datagram) that contains a notification. Events can be published individually, or in batches, but a single publication (individual or batch) can't exceed 1 MB.

Publishers and subscribers

Event publishers are any app or device that can send out events using either HTTPS or Advanced Message Queuing Protocol (AMQP) 1.0.

For publishers that send data frequently, AMQP has better performance. However, it has a higher initial session overhead, because a persistent bidirectional socket and transport-level security (TLS) or SSL/TLS has to be set up first.

For more intermittent publishing, HTTPS is the better option. Though HTTPS requires additional overhead for each request, there isn't the session initialization overhead.

Note: Existing Kafka-based clients, using Apache Kafka 1.0 and newer client versions, can also act as Event Hubs publishers.

Event subscribers are apps that use one of two supported programmatic methods to receive and process events from an Event Hub.

• EventHubReceiver - A simple method that provides limited management options.

• EventProcessorHost - An efficient method that we'll use later in this module.

Consumer groups

An Event Hub consumer group represents a specific view of an Event Hub data stream. By using separate consumer groups, multiple subscriber apps can process an event stream independently, and without affecting other apps. However, the use of many consumer groups isn't a requirement, and for many apps, the single default consumer group is sufficient.

Pricing

There are three pricing tiers for Azure Event Hubs: Basic, Standard, and Dedicated. The tiers differ in terms of supported connections, the number of available Consumer groups, and throughput. When using Azure CLI to create an Event Hubs namespace, if you don't specify a pricing tier, the default of Standard (20 Consumer groups, 1000 Brokered connections) is assigned.

Create and configure new Azure Event Hubs

There are two main steps when creating and configuring new Azure Event Hubs. The first step is to define the Event Hubs namespace. The second step is to create an Event Hub in that namespace.

Define an Event Hubs namespace

An Event Hubs namespace is a containing entity for managing one or more Event Hubs. Creating an Event Hubs namespace typically involves the following configuration:

Define namespace-level settings

Certain settings such as namespace capacity (configured using throughput units), pricing tier, and performance metrics are defined at the namespace level. These settings apply to all the Event Hubs within that namespace. If you don't define these settings, a default value is used: 1 for capacity and Standard for pricing tier.

Keep the following aspects in mind:

• You must balance your configuration against your Azure budget expectations.

• You might consider configuring different Event Hubs for different throughput requirements. For example, if you have a sales data app, and you're planning for two Event Hubs, it would make sense to use a separate namespace for each hub.

• You'll configure one namespace for high throughput collection of real-time sales data telemetry and one namespace for infrequent event log collection. This way, you only need to configure (and pay for) high throughput capacity on the telemetry hub.

1. Select a unique name for the namespace. The namespace is accessible through this URL: namespace.servicebus.windows.net

2. Define the following optional properties:

• Enable Kafka. This option enables Kafka apps to publish events to the Event Hub.

• Make this namespace zone redundant. Zone-redundancy replicates data across separate data centers with their independent power, networking, and cooling infrastructures.

• Enable Auto-Inflate and Auto-Inflate Maximum Throughput Units. Auto-Inflate provides an automatic scale-up option by increasing the number of throughput units up to a maximum value. This option is useful to avoid throttling in situations when incoming or outgoing data rates exceed the currently set number of throughput units.

Azure CLI commands to create an Event Hubs namespace

To create a new Event Hubs namespace, use the az eventhubs namespace commands. Here's a brief description of the subcommands you'll use in the exercise.

Configure a new Event Hub

After you create the Event Hubs namespace, you can create an Event Hub. When creating a new Event Hub, there are several mandatory parameters.

The following parameters are required to create an Event Hub:

• Event Hub name - Event Hub name that is unique within your subscription and:

• Is between 1 and 50 characters long.

• Contains only letters, numbers, periods, hyphens, and underscores.

• Starts and ends with a letter or number.

• Partition Count - The number of partitions required in an Event Hub (between 2 and 32). The partition count should be directly related to the expected number of concurrent consumers and can't be changed after the hub has been created. The partition separates the message stream so that consumer or receiver apps only need to read a specific subset of the data stream. If not defined, this value defaults to 4.

• Message Retention - The number of days (between 1 and 7) that messages will remain available if the data stream needs to be replayed for any reason. If not defined, this value defaults to 7.

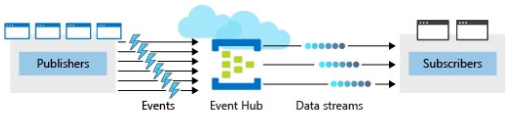
You can also optionally configure an Event Hub to stream data to an Azure Blob storage or Azure Data Lake Store account.

Azure CLI commands to create an Event Hub

To create a new Event Hub with the Azure CLI, you'll run the az eventhubs eventhub command set. Here's a brief description of the subcommands we'll be using.

Summary

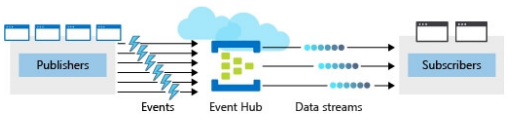
To deploy Azure Event Hubs, you must configure an Event Hubs namespace, and then configure the Event Hub itself. In the next unit, you'll go through the detailed configuration steps to create a new namespace and Event Hub.



Command	Description
create	Create the Event Hubs namespace.
authorization-rule	All Event Hubs within the same Event Hubs namespace share common connection credentials. You'll need these credentials when you configure apps to send and receive messages using the Event Hub. This command returns the connection string for your Event Hubs namespace.

Command	Description
create	Creates the Event Hub in a specified namespace.
show	Displays the details of your Event Hub.

https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-quickstart-cli



Command	Description
create	Create the Event Hubs namespace.
authorization-rule	All Event Hubs within the same Event Hubs namespace share common connection credentials. You'll need these credentials when you configure apps to send and receive messages using the Event Hub. This command returns the connection string for your Event Hubs namespace.

Command	Description
create	Creates the Event Hub in a specified namespace.
show	Displays the details of your Event Hub.

- ○ 3
- ○ 2
- ○ 1

Report Error

Q. 8 What is a dataframe?

- ○ A parquet file
- ○ An Array
- ◉ A creation of a data structure
- ○ A CSV file

Report Error

- [x] A creation of a data structure

Explanation:- A DataFrame creates a data structure and it's one of the core data structures in Spark.
What are dataframes?
Basically you could view DataFrames as you might see in excel. It's like a box with squares in it, that organizes data, which we could also refer to as a table of data.
What does a table of data mean?
It is a single set of two-dimensional data that can have multiple rows and columns in the data. Each row, is a sample of data. Each column is a variable or parameter that is able to describe the row that contains the sample of data.
A DataFrame creates a data structure and it's one of the core data structures in Spark. In Spark, it is seen as a distributed collection of data that is organized into columns that have names.
What you see in Data Engineering is that you start with reading or loading data that can be unstructured, semi-structured, or structured, which is stored in a DataFrame and start transforming that data in order to get insights. You can use different functionalities in order to do so, like using Spark SQL, PySpark, and others.
Usually when you see 'df' in some code it refers to a dataframe.
You can either create your own dataframe as this example shows:
Python
new_rows = [('CA',22, 45000),("WA",35,65000) ,("WA",50,85000)]
demo_df = spark.createDataFrame(new_rows, ['state', 'age', 'salary'])
demo_df.show()
Or load a file that contains data into a dataframe like in the below example where the open taxi dataset is used:
Python
from azureml.opendatasets import NycTlcYellow
data = NycTlcYellow()
data_df = data.to_spark_dataframe()
display(data_df.limit(10))
Once you're at the stage where you'd like to manipulate the data that is stored in a DataFrame, you can use User-Defined Functions (UDFs) that are column-based and help you transform and manipulate the data stored in a DataFrame.
https://www.tutorialspoint.com/spark_sql/spark_sql_dataframes.htm

Q. 9 Which Dynamic Management View enables the view the active connections against a dedicated SQL pool?

- [ ] sys.dm_pdw_dms_workers
- [x] sys.dm_pdw_exec_sessions
- [ ] DBCC PDW_SHOWEXECUTIONPLAN
- [ ] sys.dm_pdw_exec_requests
- [ ] sys.dm_pdw_request_steps
- [ ] sys.dm_pdw_nodes_exec_connection

Report Error

Q. 9 Which Dynamic Management View enables the view the active connections against a dedicated SQL pool?

- ○ sys.dm_pdw_exec_sessions
- ○ sys.dm_pdw_nodes_exec_connection
- ○ DBCC PDW_SHOWEXECUTIONPLAN
- ○ sys.dm_pdw_dms_workers
- ◉ sys.dm_pdw_exec_requests

Explanation:- sys.dm_pdw_exec_requests enables you to view the active connections against a dedicated SQL pool
https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-pdw-exec-requests-transact-sql?view=aps-pdw-2016-au7

- ○ sys.dm_pdw_request_steps

Report Error

Q. 10 How does splitting source files help maintain good performance when loading into Synapse Analytics?

- ○ Reduced possibility of data corruptions.
- ◉ Compute node to storage segment alignment.
- ○ 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.
- ○ Optimized processing of smaller file sizes.

Report Error

Q. 10 How does splitting source files help maintain good performance when loading into Synapse Analytics?

- ○ Optimized processing of smaller file sizes.
- ◉ Compute node to storage segment alignment.

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.

- ○ 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.
- ○ Reduced possibility of data corruptions.

Report Error

Q. 11 Which type of analytics answers the question "What is likely to happen in the future based on previous trends and patterns?"

- (●) Predictive
- ( ) Diagnostic
- ( ) Descriptive
- ( ) Scenario

Report Error

(●) Predictive

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

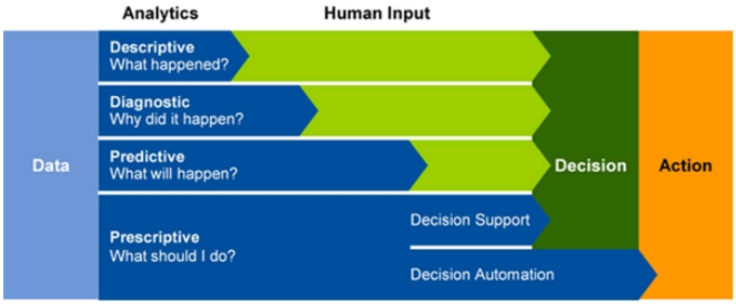
Prescriptive analytics

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



Q. 12

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
- ( ) HDI
- ( ) Spark Pools in Azure Synapse Analytics
- ( ) Apache Spark

Report Error

(●) Azure Databricks

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
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

	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.

Q. 13 Correct or Incorrect : Azure Storage encrypts all data that's written to it. It is not necessary to enable encryption within your subscription.

- (●) Incorrect
- ( ) Correct

Report Error

Q. 13 Correct or Incorrect : Azure Storage encrypts all data that's written to it. It is not necessary to enable encryption within your subscription.

- ( ) Incorrect
- (●) Correct

Explanation:- Azure Storage Data security
Azure Storage encrypts all data that's written to it. Azure Storage also provides you with fine-grained control over who has access to your data. You'll secure the data by using keys or shared access signatures.
Azure Resource Manager provides a permissions model that uses role-based access control (RBAC).
https://docs.microsoft.com/en-us/azure/storage/common/storage-introduction

Report Error

Q. 14 What is the Python syntax for defining a DataFrame in Spark from an existing Parquet file in DBFS?

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

Report Error

Q. 14 What is the Python syntax for defining a DataFrame in Spark from an existing Parquet file in DBFS?

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

Explanation:- The correct syntax is:
IPGeocodeDF = spark.read.parquet("dbfs:/mnt/training/ip-geocode.parquet")
https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html

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

Report Error

Q. 15

A time dimension table is one of the most consistently used dimension tables. This type of table enables consistent granularity for temporal analysis and reporting and usually contains temporal hierarchies, such as Year ? Quarter ? Month ? Day. In

Correct or Incorrect : It is more performant to filter on stored attributes in a large dimension table than always calculating the time attributes at query time.

- Incorrect
- Correct

Report Error

Incorrect

Explanation:- It is more performant to filter on stored attributes in a small dimension table than always calculating the time attributes at query time.
A time dimension table is one of the most consistently used dimension tables. This type of table enables consistent granularity for temporal analysis and reporting and usually contains temporal hierarchies, such as Year ? Quarter ? Month ? Day. In addition to consistency in time attributes, this design will also help query performance. It is more performant to filter on stored attributes in a small dimension table than always calculating the time attributes at query time.
Time dimension tables can contain business-specific attributes that are useful references for reporting and filters, such as fiscal periods and public holidays.
This is the schema of the time dimension table that you will create:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview

Column	Data Type
DateKey	int
DateAltKey	datetime
CalendarYear	int
CalendarQuarter	int
MonthOfYear	int
MonthName	nvarchar(15)
DayOfMonth	int
DayOfWeek	int
DayName	nvarchar(15)
FiscalYear	int
FiscalQuarter	int

Report Error

Q. 16 Which component enables you to perform code free transformations in Azure Synapse Analytics?

- Monitoring capabilities
- Flow capabilities
- Studio
- Mapping data flow
- Copy activity
- Control capabilities

Report Error

Q. 16 Which component enables you to perform code free transformations in Azure Synapse Analytics?

- Control capabilities
- Flow capabilities
- Mapping data flow

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

- Studio
- Monitoring capabilities
- Copy activity

Report Error

Q. 17

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

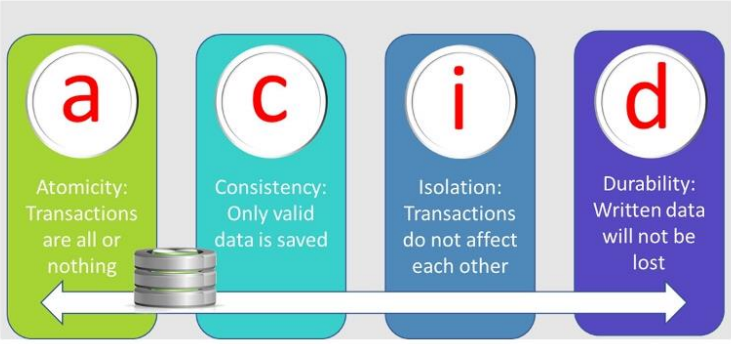
A transactional database must adhere to the [?] properties to ensure that the database remains consistent while processing transactions.

- Nuclear
- Forensic
- ACID
- Atomic

Report Error

- ACID

Explanation:- A transactional database must adhere to the ACID (Atomicity, Consistency, Isolation, Durability) properties to ensure that the database remains consistent while processing transactions.
The four letters in ACID represent the four required characteristics of database transactions:
• Atomicity
• Consistency
• Isolation
• Durability
• Atomicity guarantees that each transaction is treated as a single unit, which either succeeds completely, or fails completely. If any of the statements constituting a transaction fails to complete, the entire transaction fails and the database is left unchanged. An atomic system must guarantee atomicity in each and every situation, including power failures, errors, and crashes.
• Consistency ensures that a transaction can only take the data in the database from one valid state to another. A consistent database should never lose or create data in a manner that can't be accounted for. In the bank transfer example described earlier, if you add funds to an account, there must be a corresponding deduction of funds somewhere, or a record that describes where the funds have come from if they have been received externally. You can't suddenly create (or lose) money.
• Isolation ensures that concurrent execution of transactions leaves the database in the same state that would have been obtained if the transactions were executed sequentially. A concurrent process can't see the data in an inconsistent state (for example, the funds have been deducted from one account, but not yet credited to another.)
• Durability guarantees that once a transaction has been committed, it will remain committed even if there's a system failure such as a power outage or crash.
https://www.techopedia.com/definition/23949/atomicity-consistency-isolation-durability-acid-database-management-system



Q. 18

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

The act of setting up the database server is called [?].

- (●) Provisioning
- ( ) Distribution
- ( ) Population
- ( ) Running up

Report Error

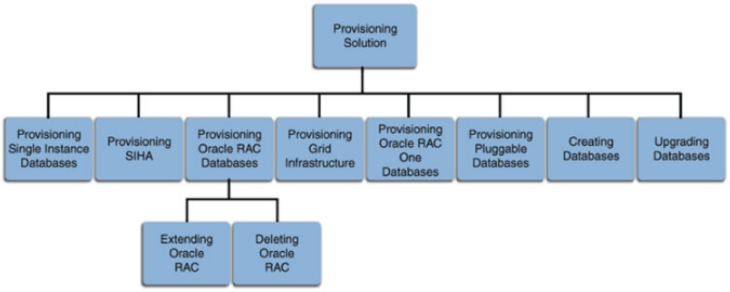
Q. 18

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

The act of setting up the database server is called [?].

- ○ Population
- ○ Distribution
- ○ Running up
- ◉ Provisioning

Explanation:- The act of setting up the database server is called provisioning.
https://docs.oracle.com/cd/E24628_01/em.121/e27046/prov_db_overview.htm#EMLCM11094



Report Error

Q. 19

Scenario: You are working on a new project and creating storage accounts and blob containers for your application.

Which of the below describes a good strategy for doing this?

- ○ Create Azure Storage accounts in your application as needed. Create the containers before deploying the application.
- ○ All the listed options.
- ○ None of the listed options.
- ◉ Create Azure Storage accounts before deploying your app. Create containers in your application as needed.
- ○ Create both your Azure Storage accounts and containers before deploying your application.

Report Error

Q. 19

Scenario: You are working on a new project and creating storage accounts and blob containers for your application.

Which of the below describes a good strategy for doing this?

- ○ All the listed options.
- ○ None of the listed options.
- ○ Create Azure Storage accounts in your application as needed. Create the containers before deploying the application.
- ○ Create both your Azure Storage accounts and containers before deploying your application.
- ◉ Create Azure Storage accounts before deploying your app. Create containers in your application as needed.

Explanation:- Creating an Azure Storage account is an administrative activity and can be done prior to deploying an application. Container creation is lightweight and is often driven by run-time data which makes it a good activity to do in your application.
https://docs.microsoft.com/en-us/azure/storage/common/storage-account-create?tabs=azure-portal

Report Error

Q. 20

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.

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

Report Error

◉ Access Control Lists (ACLs)

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

- ○ Transparent Data Encryption (TDE)

Q. 21

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?

- [ ] Merkle table distribution
- [ ] Centralized table distribution
- [x] Distributed table schema
- [x] Replicated tables
- [x] Round robin distribution
- [ ] Hash distribution

Report Error

- [x] Replicated tables

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

- [x] Distributed table schema
- [x] Merkle table distribution
- [x] Centralized table distribution
- [x] Round robin distribution

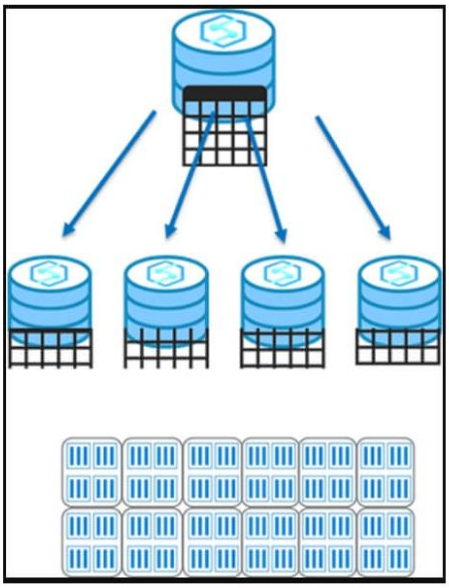
Explanation:- When data is loaded into Synapse Analytics dedicated SQL pools, the datasets are broken up and dispersed among the compute nodes for processing, and then written to a decoupled and scalable storage layer. This action is termed "sharding".
The design decisions around how to split and disperse 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

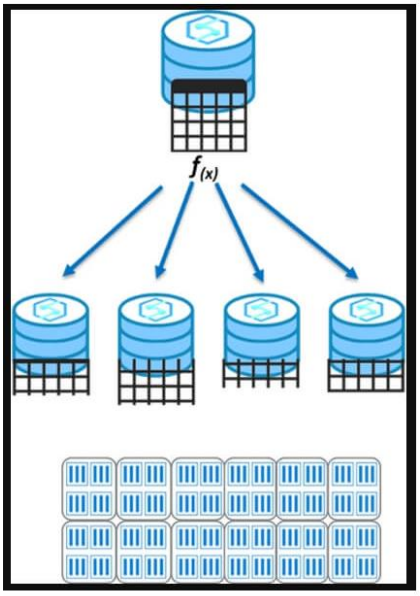
Replicated tables
A replicated table provides the fastest query performance for small tables.
A table that is replicated caches a full copy of the table on each compute node. Consequently, replicating a table removes the need to transfer data among compute nodes before a join or aggregation. As such 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

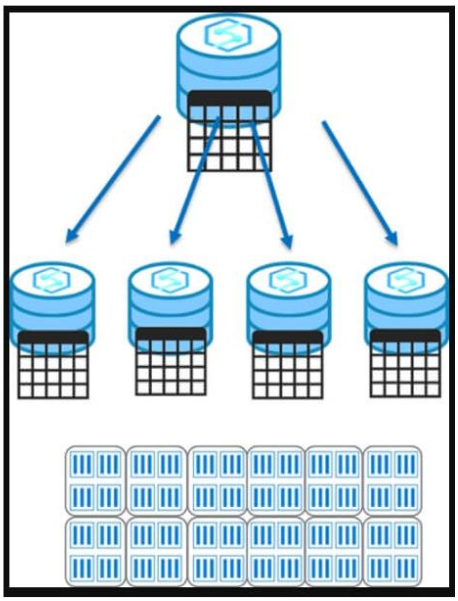
- [x] Hash distribution

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

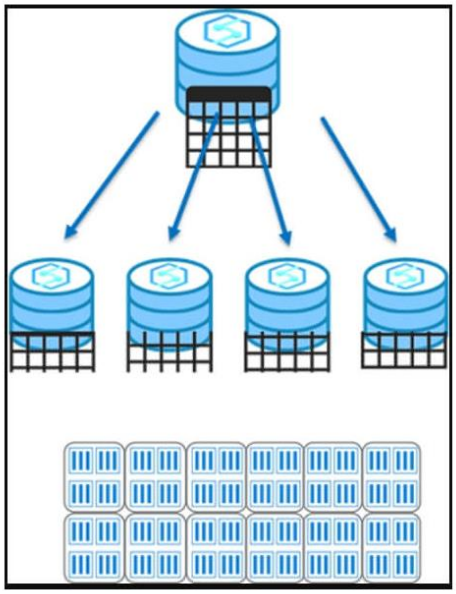
Replicated tables
A replicated table provides the fastest query performance for small tables.
A table that is replicated caches a full copy of the table on each compute node. Consequently, replicating a table removes the need to transfer data among compute nodes before a join or aggregation. As such 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.

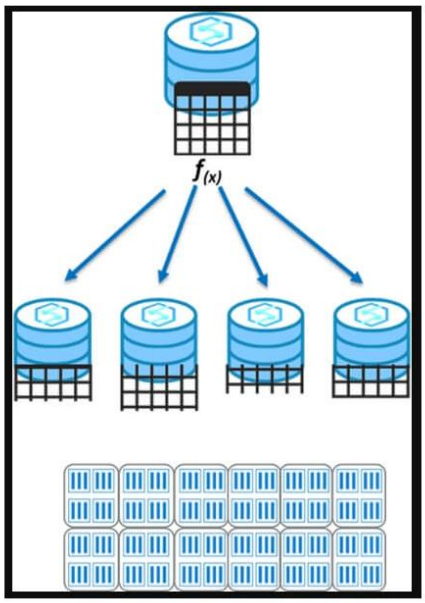


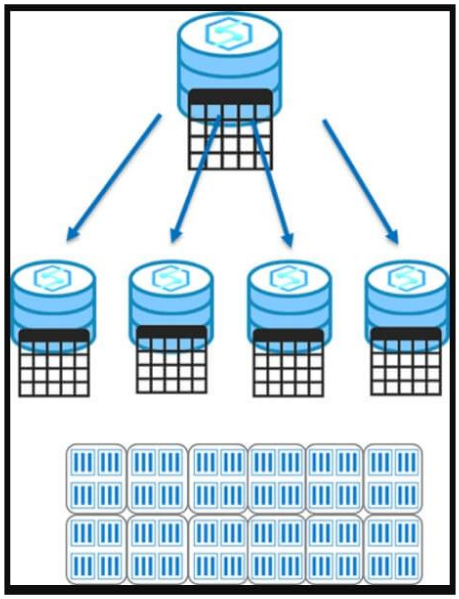




https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables







Report Error

Q. 22 Once Azure Synapse Link is configured on Cosmos DB, what is the first step to perform to use Azure Synapse Analytics serverless SQL pools to query the Azure Cosmos DB data?

- ○ Use the OPENROWSET function
- ○ CREATE database
- ○ None of the listed options
- ◉ Use a SELECT clause

Report Error

Q. 22 Once Azure Synapse Link is configured on Cosmos DB, what is the first step to perform to use Azure Synapse Analytics serverless SQL pools to query the Azure Cosmos DB data?

- ◉ CREATE database

Explanation:- Before being able to issue any queries using Azure Synapse Analytics serverless SQL pools, you first must create a database.
https://docs.microsoft.com/en-us/azure/azure-monitor/insights/azure-sql

- ○ Use the OPENROWSET function
- ○ None of the listed options
- ○ Use a SELECT clause

Report Error

Q. 23 When is it possible to add or remove datasets if created with Azure Data Share?

- ○ None of the listed options.
- ○ It is not possible to add or remove datasets if created with Azure Data Share.
- ◉ It is possible to add or remove datasets within Azure Data Share after it has been created.
- ○ It is only possible to remove or add datasets before it's sent within Azure Data Share.

Report Error

Q. 23 When is it possible to add or remove datasets if created with Azure Data Share?

- ○ None of the listed options.
- ◉ It is possible to add or remove datasets within Azure Data Share after it has been created.

Explanation:- It is possible to add or remove datasets after it has been created in Azure Data Share.
https://docs.microsoft.com/en-us/azure/data-factory/lab-data-flow-data-share

- ○ It is only possible to remove or add datasets before it's sent within Azure Data Share.
- ○ It is not possible to add or remove datasets if created with Azure Data Share.

Report Error

Q. 24 When creating a typical project, when would you create your storage account(s)?

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

Report Error

Q. 24 When creating a typical project, when would you create your storage account(s)?

- ( ) At the end, during resource cleanup.
- ( ) At any stage of the project, as long as it is before you need to analyze data.
- (●) At the beginning, during project setup.

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

- ( ) After deployment, when the project is running.

Report Error

Q. 25

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

A window function enables you to perform a mathematical equation on a set of data that is defined within a window. The mathematical equation is typically an aggregate function; however, instead of applying the aggregate function to all the rows in a table, it is applied to a set of rows that are defined by the window function, and then the aggregate is applied to it.

One of the key components of window functions is the [?] clause. This clause determines the partitioning and ordering of a rowset before the associated window function is applied.

That is, the [?] clause defines a window or user-specified set of rows within a query result set.

- ( ) UNDER
- ( ) OVER
- ( ) WHERE
- (●) HAVING

Report Error

○ HAVING

◉ OVER

Explanation:- A window function enables you to perform a mathematical equation on a set of data that is defined within a window. The mathematical equation is typically an aggregate function; however, instead of applying the aggregate function to all the rows in a table, it is applied to a set of rows that are defined by the window function, and then the aggregate is applied to it.
It is used to either perform calculations against a range of data, but it can also be used to programmatically define a deduplication of data technique, or paginate results.
One of the key components of window functions is the OVER clause. This clause determines the partitioning and ordering of a rowset before the associated window function is applied. That is, the OVER clause defines a window or user-specified set of rows within a query result set. A window function then computes a value for each row in the window. You can use the OVER clause with functions to compute aggregated values such as moving averages, cumulative aggregates, running totals, or a top N per group results.
SQL
-- Syntax for SQL Server, Azure SQL Database, and Azure Synapse Analytics

OVER (
[ ]
[ ]
[ ]
)

::=
PARTITION BY value_expression , ... [n]

::=

ORDER BY order_by_expression
[COLLATE collation_name]
[ASC | DESC]
[,...n]

::=
{ ROWS | RANGE }

::=
{
|
}
::=
BETWEEN AND

::=
{
|
}

::=
{
UNBOUNDED PRECEDING
| PRECEDING
| CURRENT ROW
}

::=

{
UNBOUNDED FOLLOWING
| FOLLOWING
| CURRENT ROW
}

::=
{ }
https://docs.microsoft.com/en-us/sql/t-sql/queries/select-over-clause-transact-sql?view=sql-server-ver15
You can then use aggregate functions with our window by expanding on our query that uses the OVER clause. The following aggregate functions are supported including COUNT, MAX, AVG, SUM, APPROX_COUNT, DISTINCT, MIN, STDEV, STDEVP, STRING_AGG, VAR, VARP, GROUPING, GROUPING_ID, COUNT_BIG, CHECKSUM_AGG
Alternatively, you can use analytical functions, which calculate an aggregate value based on a group of rows. Unlike aggregate functions, however, analytic functions can return multiple rows for each group. Use analytic functions to compute moving averages, running totals, percentages, or top-N results within a group. Supports LAG, LEAD, FIRST_VALUE, LAST_VALUE, CUME_DIST, PERCENTILE_CONT, PERCENTILE_DISC, PERCENT_RANK
You may want to use the ROWS and RANGE clauses to further limit the rows within the partition by specifying start and end points within the partition. This is done by specifying a range of rows with respect to the current row either by logical association or physical association. Physical association is achieved by using the ROWS clause. Supports PRECEDING, UNBOUNDING PRECEDING, CURRENT ROW, BETWEEN, FOLLOWING, UNBOUNDED FOLLOWING
Finally, window functions support Ranking functions like RANK, NTILE, DENSE_RANK, ROW_NUMBER.
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

- ○ WHERE
- ○ UNDER

Report Error

Q. 26

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
- ○ Storage account keys
- ○ SQL Authentication
- ◉ Shared access signatures

Report Error

○ SQL Authentication

◉ Managed identities

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.

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.

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

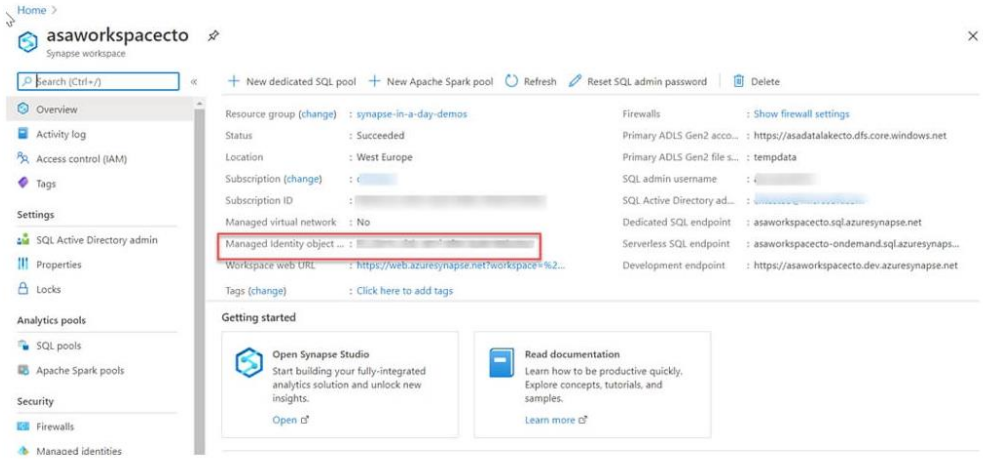
Secrets are stored in individual vaults, which are Azure resources used to group secrets together. Secret access and vault management is accomplished via a REST API, which is also supported by all of the Azure management tools as well as client libraries available for many popular languages. Every vault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

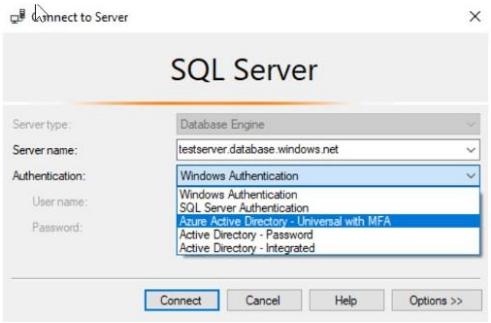
Shared access signatures

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.

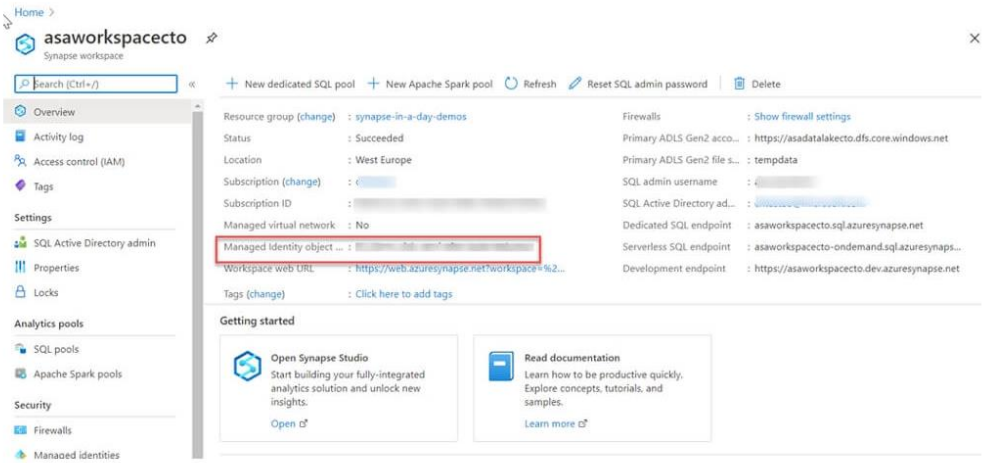
Types of shared access signatures

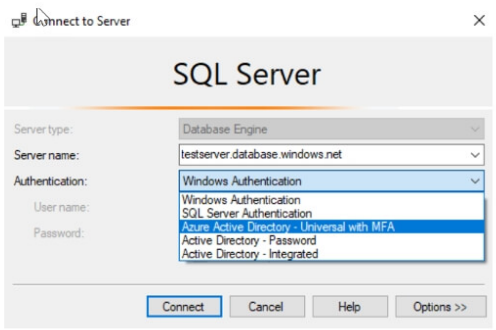
You can use a service-level shared access signature to allow access to specific resources in a storage account. You'd use this type of shared access signature, for example, to allow an app to retrieve a list of files in a file system or to download a file.
Use an account-level shared access signature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





- ○ Shared access signatures
- ○ Storage account keys

Report Error

Q. 27

Azure Data Lake Storage Gen2 plays a fundamental role in a wide range of big data architectures. There are stages for processing big data solutions that are common to all architectures.

Which are they? (Select four)

- ☑ Ingestion
- ☑ Model and serve
- ☑ Store
- ☐ Clusters
- ☐ Streamed
- ☑ Prep and train

Report Error

☑ Prep and train

Explanation:- Azure Data Lake Storage Gen2 plays a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:
• A modern data warehouse.
• Advanced analytics against big data.
• A real-time analytical solution.
There are four stages for processing big data solutions that are common to all architectures:
• Ingestion - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the Data Lake Store. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate technology to use.
• Store - The store phase identifies where the ingested data should be placed. In this case, we're using Azure Data Lake Storage Gen2.
• Prep and train - The prep and train phase identifies the technologies that are used to perform data preparation and model training and scoring for data science solutions. The common technologies that are used in this phase are Azure Databricks, Azure HDInsight or Azure Machine Learning Services.
• Model and serve - Finally, the model and serve phase involves the technologies that will present the data to users. These can include visualization tools such as Power BI, or other data stores such as Azure Synapse Analytics, Azure Cosmos DB, Azure SQL Database, or Azure Analysis Services. Often, a combination of these technologies will be used depending on the business requirements.
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-data-scenarios

☑ Model and serve

Explanation:- Azure Data Lake Storage Gen2 plays a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:
• A modern data warehouse.
• Advanced analytics against big data.
• A real-time analytical solution.
There are four stages for processing big data solutions that are common to all architectures:
• Ingestion - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the Data Lake Store. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate technology to use.
• Store - The store phase identifies where the ingested data should be placed. In this case, we're using Azure Data Lake Storage Gen2.
• Prep and train - The prep and train phase identifies the technologies that are used to perform data preparation and model training and scoring for data science solutions. The common technologies that are used in this phase are Azure Databricks, Azure HDInsight or Azure Machine Learning Services.
• Model and serve - Finally, the model and serve phase involves the technologies that will present the data to users. These can include visualization tools such as Power BI, or other data stores such as Azure Synapse Analytics, Azure Cosmos DB, Azure SQL Database, or Azure Analysis Services. Often, a combination of these technologies will be used depending on the business requirements.
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-data-scenarios

☑ Clusters

☑ Ingestion

Explanation:- Azure Data Lake Storage Gen2 plays a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:
• A modern data warehouse.
• Advanced analytics against big data.
• A real-time analytical solution.
There are four stages for processing big data solutions that are common to all architectures:
• Ingestion - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the Data Lake Store. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate technology to use.
• Store - The store phase identifies where the ingested data should be placed. In this case, we're using Azure Data Lake Storage Gen2.
• Prep and train - The prep and train phase identifies the technologies that are used to perform data preparation and model training and scoring for data science solutions. The common technologies that are used in this phase are Azure Databricks, Azure HDInsight or Azure Machine Learning Services.
• Model and serve - Finally, the model and serve phase involves the technologies that will present the data to users. These can include visualization tools such as Power BI, or other data stores such as Azure Synapse Analytics, Azure Cosmos DB, Azure SQL Database, or Azure Analysis Services. Often, a combination of these technologies will be used depending on the business requirements.
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-data-scenarios

Store

Explanation:- Azure Data Lake Storage Gen2 plays a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:
• A modern data warehouse.
• Advanced analytics against big data.
• A real-time analytical solution.
There are four stages for processing big data solutions that are common to all architectures:
• Ingestion - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the Data Lake Store. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate technology to use.
• Store - The store phase identifies where the ingested data should be placed. In this case, we're using Azure Data Lake Storage Gen2.
• Prep and train - The prep and train phase identifies the technologies that are used to perform data preparation and model training and scoring for data science solutions. The common technologies that are used in this phase are Azure Databricks, Azure HDInsight or Azure Machine Learning Services.
• Model and serve - Finally, the model and serve phase involves the technologies that will present the data to users. These can include visualization tools such as Power BI, or other data stores such as Azure Synapse Analytics, Azure Cosmos DB, Azure SQL Database, or Azure Analysis Services. Often, a combination of these technologies will be used depending on the business requirements.
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-data-scenarios

Streamed

Report Error

Q. 28 Which of the following is a good analogy for the access keys of a storage account?

- IP Address
- Cryptographic algorithm
- Username and password
- REST Endpoint

Report Error

Q. 28 Which of the following is a good analogy for the access keys of a storage account?

Username and password

Explanation:- Possession of an access key identifies the account and grants you access. This is very similar to login credentials like a username and password.
https://docs.microsoft.com/en-us/azure/storage/common/storage-account-overview

- Cryptographic algorithm
- IP Address
- REST Endpoint

Report Error

Q. 29 Which role works with Azure Cognitive Services, Cognitive Search, and the Bot Framework?

- ○ A System Administrator
- ○ A Project Manager
- ○ A Solution Architect
- ○ An RPA Developer
- ◉ An AI Engineer
- ○ A Data Engineer

Report Error

Q. 29 Which role works with Azure Cognitive Services, Cognitive Search, and the Bot Framework?

- ◉ An AI Engineer

Explanation:- Artificial intelligence (AI) engineers work with AI services such as Cognitive Services, Cognitive Search, and the Bot Framework.
AI Engineer
AI engineers work with AI services such as Cognitive Services, Cognitive Search, and Bot Framework. Cognitive Services includes Computer Vision, Text Analytics, Bing Search, and Language Understanding (LUIS).
Rather than creating models, AI engineers apply the prebuilt capabilities of Cognitive Services APIs. AI engineers embed these capabilities within a new or existing application or bot. AI engineers rely on the expertise of data engineers to store information that's generated from AI.
AI engineers add the intelligent capabilities of vision, voice, language, and knowledge to applications. To do this, they use the Cognitive Services offerings that are available out of the box.
When a Cognitive Services application reaches its capacity, AI engineers call on data scientists. Data scientists develop machine learning models and customize components for an AI engineer's application.
For example, an AI engineer might be working on a Computer Vision application that processes images. This AI engineer would ask a data engineer to provision an Azure Cosmos DB instance to store the metadata and tags that the Computer Vision application generates.
https://www.whizlabs.com/blog/azure-data-engineer-roles/

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

Report Error

Q. 30

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

All data within an Azure Cosmos DB container is partitioned based on the [?], and applies to both the transactional store and the analytical store. Boundaries for parallelizing workloads are based on this [?].

- ○ Index key
- ◉ Partition key
- ○ Foreign key
- ○ Primary key

Report Error

- ○ Index key
- ◉ Partition key

Explanation:- Mixed entity types per container
You may want to mix different document entity types (entities) in the same container, which is useful to efficiently retrieve data for both entities using a single query. For example, you could put both customer profile and sales order data in the same container and partition it by customerId. In such a situation, you would usually add a field to your documents that identifies the entity type of each document to differentiate between them at query time. In the following sample documents, you will see that the type is added for this purpose in the following example documents:
JSON

```
{
"id": "54AB87A7-BDB9-4FAE-A668-AA9F43E26628",
"type": "customer",
"name": "Franklin Ye",
"customerId": "54AB87A7-BDB9-4FAE-A668-AA9F43E26628",
"address": {
"streetNo": 15850,
"streetName": "NE 40th St.",
"postcode": 98052
}
}
```

{
"_id": "000C23D8-B8BC-432E-9213-6473DFDA2BC5",
"type": "salesOrder",
"customerId": "54AB87A7-BDB9-4FAE-A668-AA9F43E26628",
"orderDate": "2014-02-16T00:00:00",
"shipDate": "2014-02-23T00:00:00",
"details": [
{
"sku": "BK-R64Y-42",
"name": "Road-550-W Yellow, 42",
"price": 1120.49,
"quantity": 1
}
]
}

The following query on against the transactional store would return the customer details and all orders associated with this one customer.

SQL

SELECT * FROM c WHERE c.customerID = "54AB87A7-BDB9-4FAE-A668-AA9F43E26628"

Whilst this approach to modelling is potentially useful for your Cosmos DB transactional store queries. All documents within a single container are mapped to a single analytical store, leading to sparsely populated column stores with the different data types needing to be further separated at the time of running an analytical query.

Recommendation: As with many design decisions, there is a trade-off between the efficiency of querying the transactional store and the ease of querying the analytical store. Carefully evaluate the usefulness of storing a mix of different document entity types in the same container to your transactional workloads. If you choose to do so, you will be required to filter by the property entity type property you selected.

Embedding entity arrays

When optimizing transactional data models, we choose to embed entities within an array in a document, especially for read heavy workloads where:

• There are contained relationships between entities.
• There are one-to-few relationships between entities.
• There is embedded data that changes infrequently.
• There is embedded data that will not grow without bound.
• There is embedded data that is queried frequently together.

Due to the fact that there are one to few relationships between the embedded entities that are represented within a single document, and that these are mapped to a single column within a single row within the analytical store. The entire embedded entity array will reside within a single column value, and need to be translated from its JSON representation at the time of querying in order to retrieve embedded entity values, irrespective of which of the two modes of schema representation being used.

Recommendation: Again, a balance needs to be struck between the usefulness of the entity embedding within the transactional application and the added complexity of writing queries against embedded JSON documents for your application.

Partitioning of data

All data within an Azure Cosmos DB container is partitioned based on the partition key, and applies to both the transactional store and the analytical store. Boundaries for parallelizing workloads are based on this partition key.

The orderliness associated when data appears in the analytical store for a query is only guaranteed within a partition. As an example, when documents (1) (2) (3) are inserted in the transactional store into a single partition, they are guaranteed to be present in the analytical store in the order in which they were inserted.

https://docs.microsoft.com/en-us/azure/cosmos-db/modeling-data

- ○ Primary key
- ○ Foreign key

Report Error

Q. 31

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

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

The Avengers IT team has many different tools at their disposal and it is your responsibility to advise them on which tool to use.

Which of the following is the best for the application?

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

Report Error

◉ Data Migration Assistant (DMA)

Explanation:-

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

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

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

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

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

Migration process overview

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

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

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

Pre-migration

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

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

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

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

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

Migration

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

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

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

Post-migration

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

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

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

• Enable you to do further optimizations, if necessary.

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

Data migration tools in Azure

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

Data Migration Assistant

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

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

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

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

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

Azure Database Migration Service

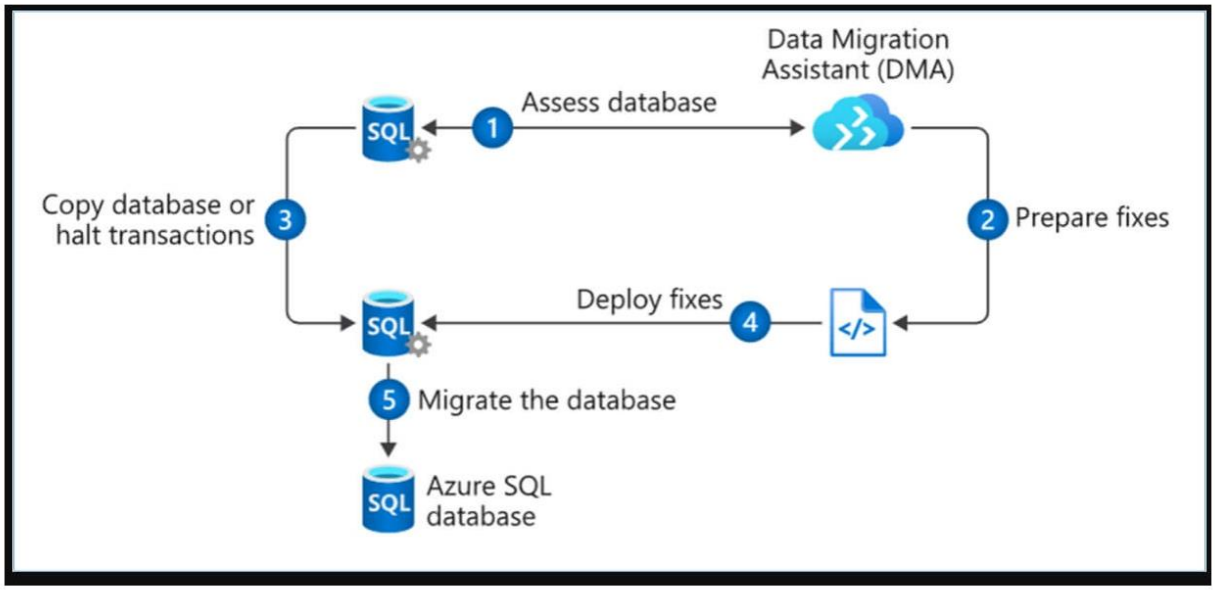
After you've migrated your database schema by using Data Migration Assistant, or created a target database manually, you're ready to move your data. To do that, you'll use Azure Database Migration Service.

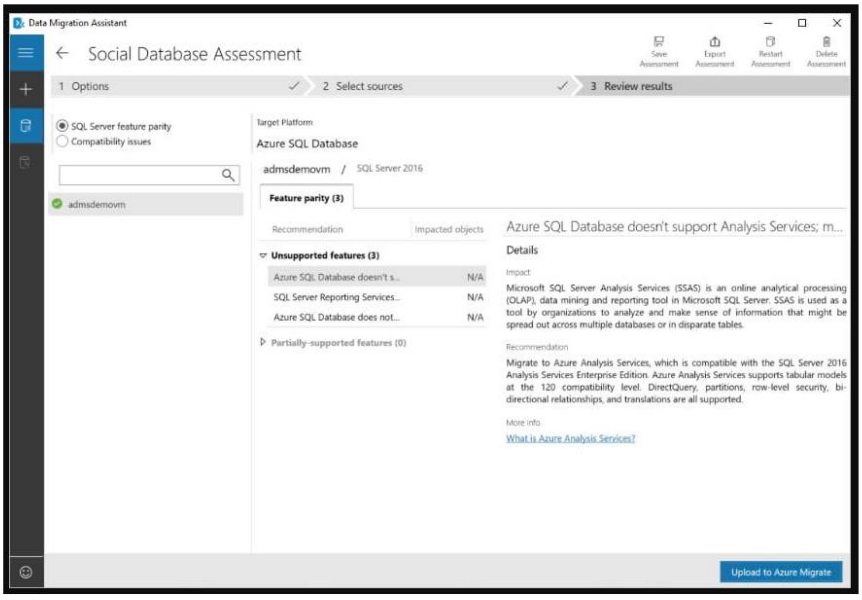
Azure Database Migration Service is a fully-managed Azure service that provides automated, seamless data migrations from multiple sources into the Azure data platforms.

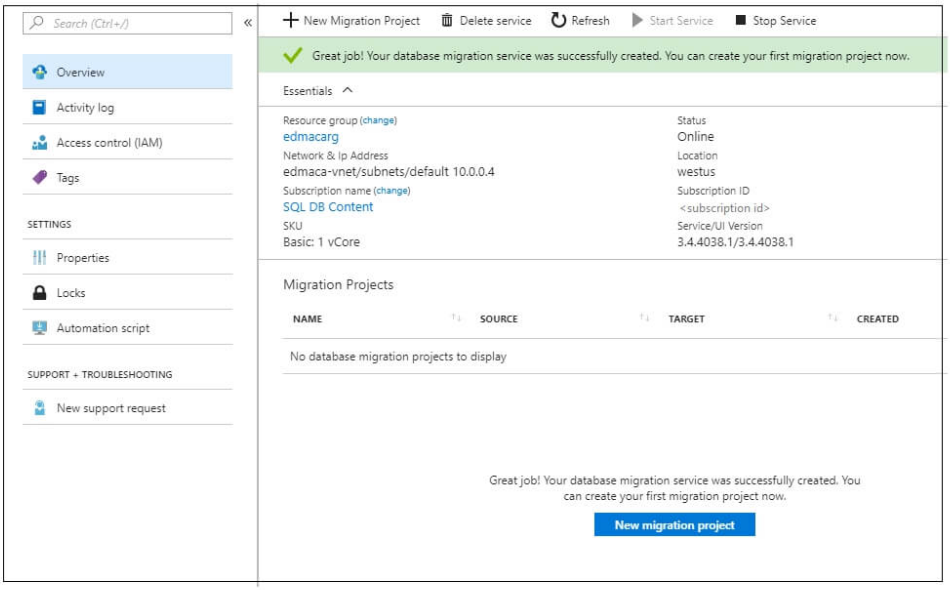
Database Migration Service runs on the Azure platform, as opposed to being a client application like Data Migration Assistant. It's capable of moving large amounts of data quickly and is not dependent upon installation of a client application. Database Migration Service can operate in two modes, offline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

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

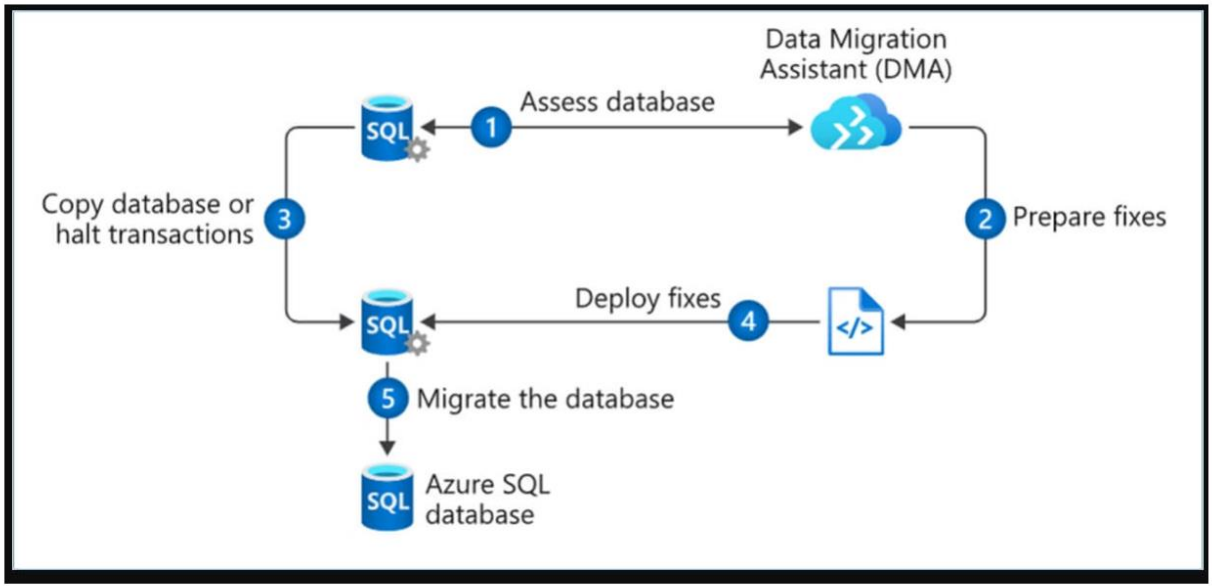
In online mode, your source database can remain in use while the bulk of the data is migrated. At the end of the migration, you'll take the source system offline momentarily while any final changes to the source are synced to the new Azure SQL Database. At this point, your applications can cut over to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

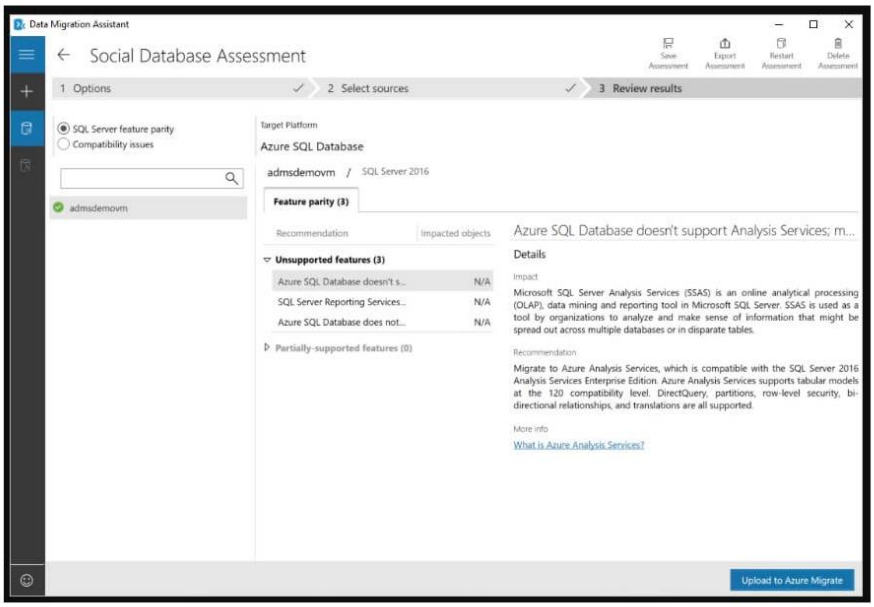

Data Migration Assistant (DMA)
Assess database
1
Prepare fixes
2
Copy database or halt transactions
3
Deploy fixes
4
Migrate the database
5
Azure SQL database

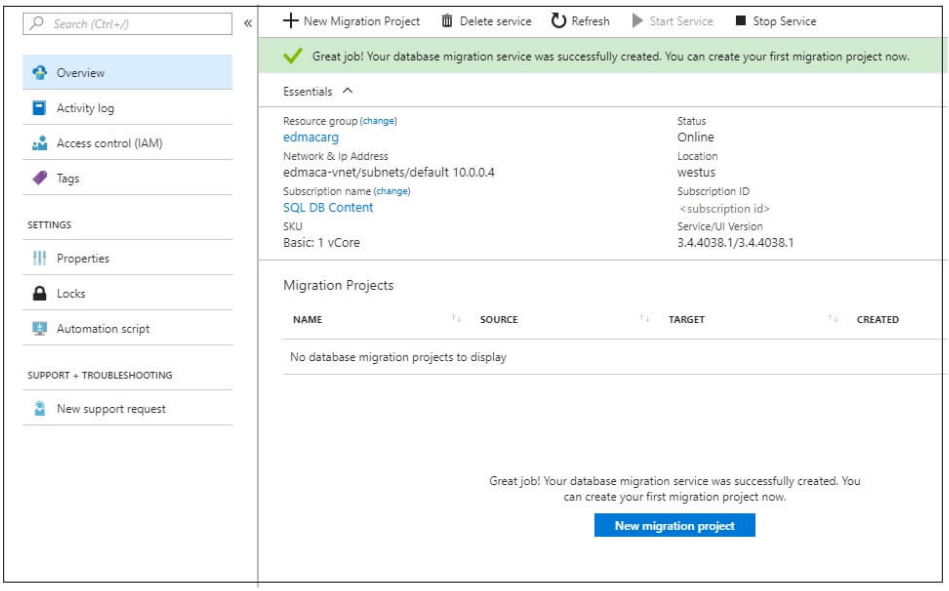

Data Migration Assistant
Social Database Assessment
Save Assessment
Export Assessment
Restart Assessment
Delete Assessment
1 Options
2 Select sources
3 Review results
SQL Server feature parity
Compatibility issues
admsdemovm
Target Platform
Azure SQL Database
admsdemovm / SQL Server 2016
Feature parity (3)
Recommendation
Impacted objects
Unsupported features (3)
Azure SQL Database doesn't s...
N/A
SQL Server Reporting Services...
N/A
Azure SQL Database does not...
N/A
Partially-supported features (0)
Azure SQL Database doesn't support Analysis Services; m...
Details
Impact
Microsoft SQL Server Analysis Services (SSAS) is an online analytical processing (OLAP), data mining and reporting tool in Microsoft SQL Server. SSAS is used as a tool by organizations to analyze and make sense of information that might be spread out across multiple databases or in disparate tables.
Recommendation
Migrate to Azure Analysis Services, which is compatible with the SQL Server 2016 Analysis Services Enterprise Edition. Azure Analysis Services supports tabular models at the 120 compatibility level. DirectQuery, partitions, row-level security, bi-directional relationships, and translations are all supported.
More info
What is Azure Analysis Services?
Upload to Azure Migrate



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







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

Q. 32 Activities within Azure Data Factory define the actions that will be performed on the data. Which are valid activity categories? (Select three)

- [ ] Control activities
- [x] Data movement activities
- [ ] Analytic activities
- [x] Data transformation activities
- [ ] Test Lab activities
- [x] Data storage activities

Report Error

Q. 32 Activities within Azure Data Factory define the actions that will be performed on the data. Which are valid activity categories? (Select three)

- [x] Data transformation activities

Explanation:- Activities within Azure Data Factory define the actions that will be performed on the data and there are three categories including
• Data movement activities
• Data transformation activities
• Control activities
Data movement activities
Data movement activities simply move data from one data store to another. You can use the Copy Activity to perform data movement activities, or by using JSON. There are a wide range of data stores that are supported as a source and as a sink. This list is ever increasing, and you can find the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-movement-activities
Data transformation activities
Data transformation activities can be performed natively within the authoring tool of Azure Data Factory using the Mapping Data Flow. Alternatively, you can call a compute resource to change or enhance data through transformation, or perform analysis of the data. These include compute technologies such as Azure Databricks, Azure Batch, SQL Database and Azure Synapse Analytics, Machine Learning Services, Azure Virtual machines and HDInsight. You can make use of any existing SQL Server Integration Services (SSIS) Packages stored in a catalogue to execute in Azure
As this list is always evolving, you can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-transformation-activiti
Control activities
When graphically authoring ADF solutions, you can use the control flow within the designed to orchestrate pipeline activities that include chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger. The current capabilities include:
• Execute Pipeline Activity
Execute Pipeline activity allows a Data Factory pipeline to invoke another pipeline.
• ForEachActivity
ForEach Activity defines a repeating control flow in your pipeline. This activity is used to iterate over a collection and executes specified activities in a loop. The loop implementation of this activity is similar to Foreach looping structure in programming languages.

• WebActivity
Web Activity can be used to call a custom REST endpoint from a Data Factory pipeline. You can pass datasets and linked services to be consumed and accessed by the activity.
• Lookup Activity
Lookup Activity can be used to read or look up a record/ table name/ value from any external source. This output can further be referenced by succeeding activities.
• Get Metadata Activity
GetMetadata activity can be used to retrieve metadata of any data in Azure Data Factory.
• Until Activity
Implements Do-Until loop that is similar to Do-Until looping structure in programming languages. It executes a set of activities in a loop until the condition associated with the activity evaluates to true. You can specify a timeout value for the until activity in Data Factory.
• If Condition Activity
The If Condition can be used to branch based on condition that evaluates to true or false. The If Condition activity provides the same functionality that an if statement provides in programming languages. It evaluates a set of activities when the condition evaluates to true and another set of activities when the condition evaluates to false.
• Wait Activity
When you use a Wait activity in a pipeline, the pipeline waits for the specified period of time before continuing with execution of subsequent activities.
You can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#control-activities

- [x] Data movement activities

Explanation:- Activities within Azure Data Factory define the actions that will be performed on the data and there are three categories including
• Data movement activities
• Data transformation activities
• Control activities
Data movement activities
Data movement activities simply move data from one data store to another. You can use the Copy Activity to perform data movement activities, or by using JSON. There are a wide range of data stores that are supported as a source and as a sink. This list is ever increasing, and you can find the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-movement-activities
Data transformation activities

Data transformation activities can be performed natively within the authoring tool of Azure Data Factory using the Mapping Data Flow. Alternatively, you can call a compute resource to change or enhance data through transformation, or perform analysis of the data. These include compute technologies such as Azure Databricks, Azure Batch, SQL Database and Azure Synapse Analytics, Machine Learning Services, Azure Virtual machines and HDInsight. You can make use of any existing SQL Server Integration Services (SSIS) Packages stored in a catalogue to execute in Azure
As this list is always evolving, you can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-transformation-activiti
Control activities
When graphically authoring ADF solutions, you can use the control flow within the designed to orchestrate pipeline activities that include chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger. The current capabilities include:
• Execute Pipeline Activity
Execute Pipeline activity allows a Data Factory pipeline to invoke another pipeline.
• ForEachActivity
ForEach Activity defines a repeating control flow in your pipeline. This activity is used to iterate over a collection and executes specified activities in a loop. The loop implementation of this activity is similar to Foreach looping structure in programming languages.
• WebActivity
Web Activity can be used to call a custom REST endpoint from a Data Factory pipeline. You can pass datasets and linked services to be consumed and accessed by the activity.
• Lookup Activity
Lookup Activity can be used to read or look up a record/ table name/ value from any external source. This output can further be referenced by succeeding activities.
• Get Metadata Activity
GetMetadata activity can be used to retrieve metadata of any data in Azure Data Factory.
• Until Activity
Implements Do-Until loop that is similar to Do-Until looping structure in programming languages. It executes a set of activities in a loop until the condition associated with the activity evaluates to true. You can specify a timeout value for the until activity in Data Factory.
• If Condition Activity
The If Condition can be used to branch based on condition that evaluates to true or false. The If Condition activity provides the same functionality that an if statement provides in programming languages. It evaluates a set of activities when the condition evaluates to true and another set of activities when the condition evaluates to false.

• Wait Activity
When you use a Wait activity in a pipeline, the pipeline waits for the specified period of time before continuing with execution of subsequent activities.
You can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#control-activities

- [x] Control activities

Explanation:- Activities within Azure Data Factory define the actions that will be performed on the data and there are three categories including
• Data movement activities
• Data transformation activities
• Control activities
Data movement activities
Data movement activities simply move data from one data store to another. You can use the Copy Activity to perform data movement activities, or by using JSON. There are a wide range of data stores that are supported as a source and as a sink. This list is ever increasing, and you can find the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-movement-activities
Data transformation activities
Data transformation activities can be performed natively within the authoring tool of Azure Data Factory using the Mapping Data Flow. Alternatively, you can call a compute resource to change or enhance data through transformation, or perform analysis of the data. These include compute technologies such as Azure Databricks, Azure Batch, SQL Database and Azure Synapse Analytics, Machine Learning Services, Azure Virtual machines and HDInsight. You can make use of any existing SQL Server Integration Services (SSIS) Packages stored in a catalogue to execute in Azure
As this list is always evolving, you can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#data-transformation-activiti
Control activities
When graphically authoring ADF solutions, you can use the control flow within the designed to orchestrate pipeline activities that include chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger. The current capabilities include:
• Execute Pipeline Activity
Execute Pipeline activity allows a Data Factory pipeline to invoke another pipeline.
• ForEachActivity
ForEach Activity defines a repeating control flow in your pipeline. This activity is used to iterate over a collection and executes specified activities in a loop. The loop implementation of this activity is similar to Foreach looping structure in programming languages.

• WebActivity
Web Activity can be used to call a custom REST endpoint from a Data Factory pipeline. You can pass datasets and linked services to be consumed and accessed by the activity.
• Lookup Activity
Lookup Activity can be used to read or look up a record/ table name/ value from any external source. This output can further be referenced by succeeding activities.
• Get Metadata Activity
GetMetadata activity can be used to retrieve metadata of any data in Azure Data Factory.
• Until Activity
Implements Do-Until loop that is similar to Do-Until looping structure in programming languages. It executes a set of activities in a loop until the condition associated with the activity evaluates to true. You can specify a timeout value for the until activity in Data Factory.
• If Condition Activity
The If Condition can be used to branch based on condition that evaluates to true or false. The If Condition activity provides the same functionality that an if statement provides in programming languages. It evaluates a set of activities when the condition evaluates to true and another set of activities when the condition evaluates to false.
• Wait Activity
When you use a Wait activity in a pipeline, the pipeline waits for the specified period of time before continuing with execution of subsequent activities.
You can get the latest information here: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities#control-activities

- [ ] Analytic activities
- [ ] Data storage activities
- [ ] Test Lab activities

Report Error

Q. 33 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 Synapse Link
- ○ Azure Stream Analytics
- ◉ Azure Synapse Spark pools
- ○ None of the listed options
- ○ Azure Synapse Studio
- ○ Azure Data Warehouse

Report Error

◉ None of the listed options

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

Q. 34

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 SQL Datawarehouse
- ○ Azure Cosmos DB
- ◉ Azure Data Catalog
- ○ Azure Storage Explorer
- ○ Azure Data Factory
- ○ Azure Data Lake Storage

Report Error

Q. 34

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 SQL Datawarehouse
- ○ Azure Data Catalog
- ◉ Azure Data Factory

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

- ○ Azure Data Lake Storage
- ○ Azure Storage Explorer
- ○ Azure Cosmos DB

Report Error

Q. 35

Scenario: You have been contracted by Wayne Enterprises, a company owned by Bruce Wayne with market value of over twenty seven million dollars.

Bruce founded Wayne Enterprises shortly after he created the Wayne Foundation and he became the president and chairman of the company.

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

Required:
• A single storage account must be used to store all operations (includes all reads, writes and deletes)
• Retention policy dictates that an on-premises copy must exist for all historical operations

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

- [x] Use the storage client to download log data from $logs/table
- [ ] Configure the storage account to log read, write and delete operations for service type queue
- [x] Configure the storage account to log read, write and delete operations for service-type table
- [ ] Use the AzCopy tool to download log data from $logs/blob
- [x] Configure the storage account to log read, write and delete operations for service type Blob

Report Error

- [x] Use the AzCopy tool to download log data from $logs/blob

Explanation:- Storage Logging logs request data in a set of blobs in a blob container named $logs in your storage account. This container does not show up if you list all the blob containers in your account but you can see its contents if you access it directly.
Storage Analytics logs detailed information about successful and failed requests to a storage service. This information can be used to monitor individual requests and to diagnose issues with a storage service. Requests are logged on a best-effort basis. This means that most requests will result in a log record, but the completeness and timeliness of Storage Analytics logs are not guaranteed.
Storage Analytics logging is not enabled by default for your storage account. You can enable it in the Azure portal or by using PowerShell, or Azure CLI. For step-by-step guidance, see Enable and manage Azure Storage Analytics logs (classic).
You can also enable Storage Analytics logs programmatically via the REST API or the client library. Use the Get Blob Service Properties, Get Queue Service Properties, and Get Table Service Properties operations to enable Storage Analytics for each service. To see an example that enables Storage Analytics logs by using .NET, see Enable logs
Log entries are created only if there are requests made against the service endpoint. For example, if a storage account has activity in its Blob endpoint but not in its Table or Queue endpoints, only logs pertaining to the Blob service will be created.
https://docs.microsoft.com/en-us/rest/api/storageservices/enabling-storage-logging-and-accessing-log-data
To view and analyze your log data, you should download the blobs that contain the log data you are interested in to a local machine. Many storage-browsing tools enable you to download blobs from your storage account; you can also use the Azure Storage team provided command-line Azure Copy Tool (AzCopy) to download your log data.
AzCopy is a command-line utility that you can use to copy blobs or files to or from a storage account. This article helps you download AzCopy, connect to your storage account, and then transfer files.
https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10
https://www.youtube.com/watch?v=GJYAgi5eYYE

- [ ] Use the storage client to download log data from $logs/table
- [ ] Configure the storage account to log read, write and delete operations for service-type table
- [ ] Configure the storage account to log read, write and delete operations for service type queue

- [x] Configure the storage account to log read, write and delete operations for service type Blob

Explanation:- Storage Logging logs request data in a set of blobs in a blob container named $logs in your storage account. This container does not show up if you list all the blob containers in your account but you can see its contents if you access it directly.
Storage Analytics logs detailed information about successful and failed requests to a storage service. This information can be used to monitor individual requests and to diagnose issues with a storage service. Requests are logged on a best-effort basis. This means that most requests will result in a log record, but the completeness and timeliness of Storage Analytics logs are not guaranteed.
Storage Analytics logging is not enabled by default for your storage account. You can enable it in the Azure portal or by using PowerShell, or Azure CLI. For step-by-step guidance, see Enable and manage Azure Storage Analytics logs (classic).
You can also enable Storage Analytics logs programmatically via the REST API or the client library. Use the Get Blob Service Properties, Get Queue Service Properties, and Get Table Service Properties operations to enable Storage Analytics for each service. To see an example that enables Storage Analytics logs by using .NET, see Enable logs
Log entries are created only if there are requests made against the service endpoint. For example, if a storage account has activity in its Blob endpoint but not in its Table or Queue endpoints, only logs pertaining to the Blob service will be created.
https://docs.microsoft.com/en-us/rest/api/storageservices/enabling-storage-logging-and-accessing-log-data
To view and analyze your log data, you should download the blobs that contain the log data you are interested in to a local machine. Many storage-browsing tools enable you to download blobs from your storage account; you can also use the Azure Storage team provided command-line Azure Copy Tool (AzCopy) to download your log data.
AzCopy is a command-line utility that you can use to copy blobs or files to or from a storage account. This article helps you download AzCopy, connect to your storage account, and then transfer files.
https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10
https://www.youtube.com/watch?v=GJYAgi5eYYE

Report Error

Q. 36 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?

- (●) All of these
- ( ) SOC2, Type 2
- ( ) ISO 27001
- ( ) AICPA
- ( ) HITRUST
- ( ) PCI DSS

Report Error

◉ All of these

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.

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.

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:

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:

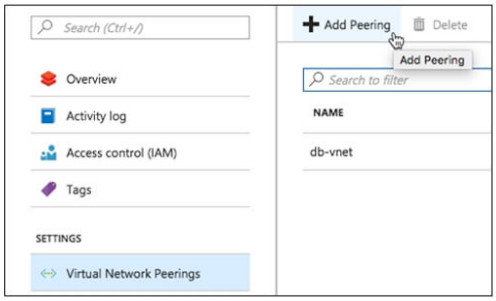
• HITRUST
• AICPA
• PCI DSS
• ISO 27001
• ISO 27018
• HIPAA (Covered by MSFT Business Associates Agreement (BAA))
• SOC2, Type 2

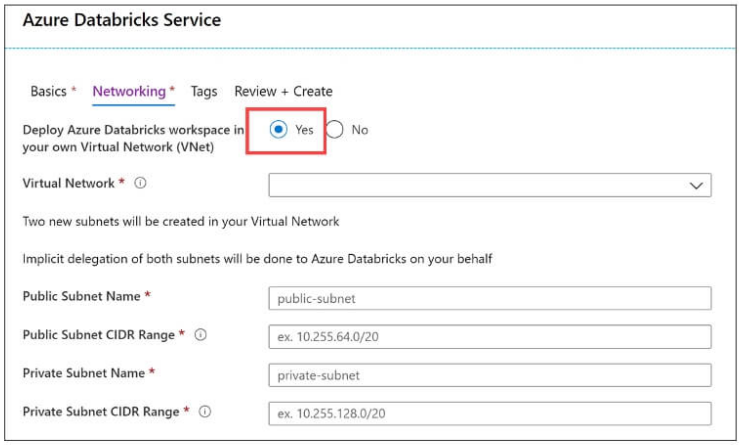
Audit logs

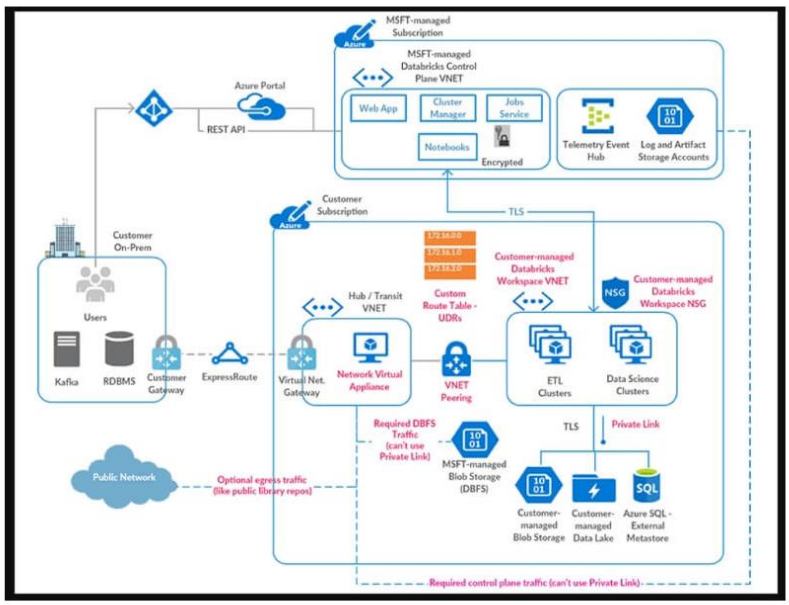
Databricks provides comprehensive end-to-end audit logs of activities performed by Databricks users, allowing your enterprise to monitor detailed Databricks usage patterns. Azure Monitor integration enables you to capture the audit logs and make then centrally available and fully searchable.

Services / Entities included are:

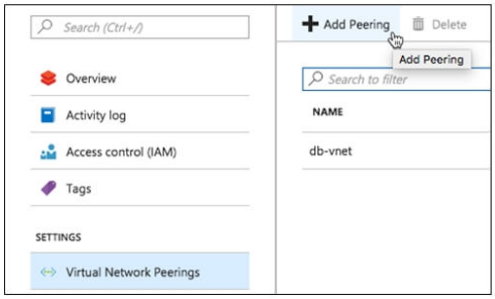
- Accounts
- Clusters
- DBFS
- Genie
- Jobs
- ACLs
- SSH
- Tables

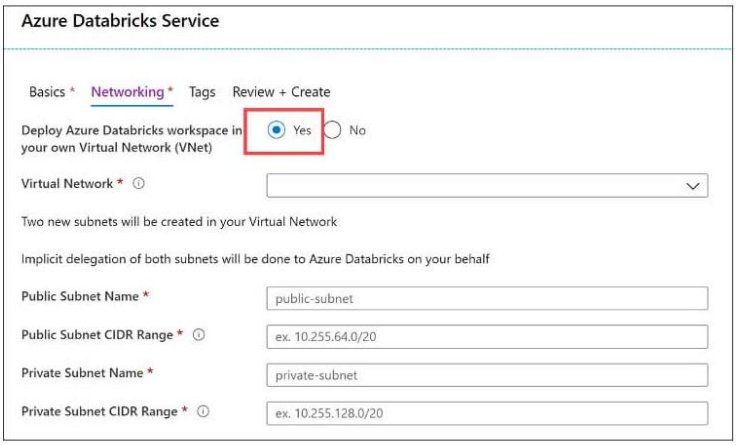


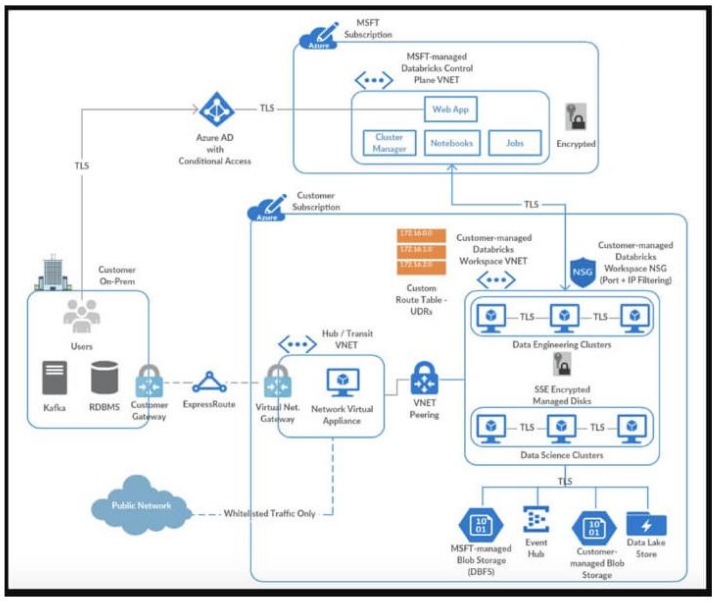


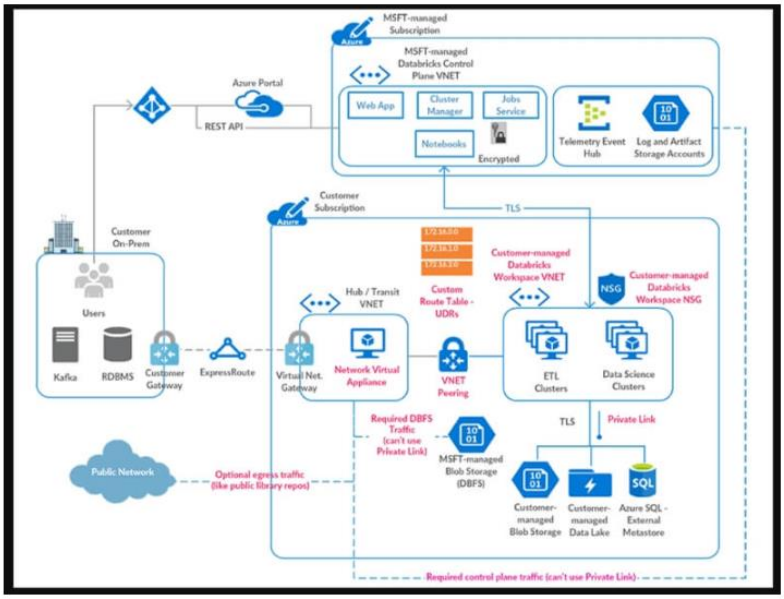


https://docs.microsoft.com/en-us/azure/security/fundamentals/network-overview









- ○ PCI DSS
- ○ HITRUST
- ○ SOC2, Type 2
- ○ ISO 27001

Report Error

Q. 37 In Spark Structured Streaming, what method should be used to read streaming data into a DataFrame?

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

Report Error

Q. 37 In Spark Structured Streaming, what method should be used to read streaming data into a DataFrame?

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

Report Error

Q. 38 What is an Azure Key Vault-backed secret scope?

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

Report Error

Q. 38 What is an Azure Key Vault-backed secret scope?

- ◉ A Databricks secret scope that is backed by Azure Key Vault instead of Databricks.

Explanation:- A secret scope is provided by Azure Databricks and can be backed by either Databricks or Azure Key Vault.
https://docs.microsoft.com/en-us/azure/databricks/security/secrets/secret-scopes

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

Report Error

Q. 39

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 EventStream
- ○ Azure Multistream Processing
- ◉ Azure Stream Analytics
- ○ Azure StreamSets

Report Error

- ○ Azure Multistream Processing
- ○ Azure EventStream
- ○ Azure StreamSets
- ◉ Azure Stream Analytics

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.

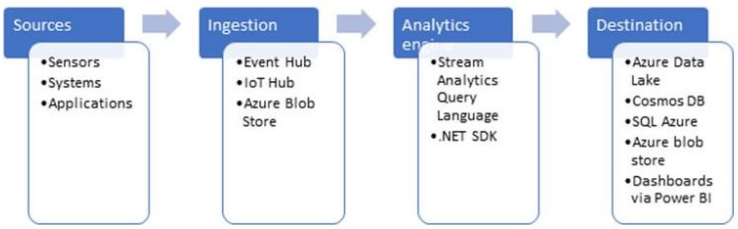
• Event consumer: A destination of the output from the stream analytics engine. The target can be storage, such as Azure Data Lake, Azure Cosmos DB, Azure SQL Database, or Azure Blob storage, or dashboards powered by Power BI.

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



Report Error

Q. 40 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.
- ( ) 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.

Report Error

Q. 40 What optimization does the following command perform: OPTIMIZE Students ZORDER BY Grade?

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

- ( ) 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 updates a graph that routes requests to the appropriate files.

Report Error

Q. 41

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

Which of the following are valid options for transforming data within Azure Data Factory? (Select three)

- [x] SSIS Packages
- [ ] Data Storage Activities
- [ ] Control Resources
- [ ] Data Movement Flows
- [x] Compute Resources
- [x] Mapping Data Flows

Report Error

☑ Mapping Data Flows

Explanation:- Just as Azure Data Factory provides a variety of methods for ingesting data, it also provides a range of methods to perform transformations. You can pick a method that matches the skillsets of your team or takes advantage of existing technologies that you already have in your data estate. There is also the opportunity to perform transformations without writing code at all using the Mapping Data Flow.
Transforming data using Mapping Data Flow
Mapping Data Flows provide an environment for building a wide range of data transformations visually without the need to use code. The resulting data flows that are created are subsequently executed on scaled-out Apache Spark clusters that are automatically provisioned when you execute the Mapping Data Flow. Mapping Data Flows also provides the capability to monitor the execution of the transformations so that you can view how the transformations are progressing, or to understand any errors that may occur
Transforming data using compute resources
Azure Data Factory can also call on compute resources to transform data by a data platform service that may be better suited to the job. A great example of this is that Azure Data Factory can create a pipeline to an analytical data platform such as Spark pools in an Azure Synapse Analytics instance to perform a complex calculation using python. Another example could be to send data to an Azure SQL Database instance to execute a stored procedure using Transact-SQL. There is a wide range of compute resource, and the associated activities that they can perform as shown in the following table:
Compute environment: On-demand HDInsight cluster or your own HDInsight cluster
Activities: Hive, Pig, Spark, MapReduce, Hadoop Streaming
Compute environment: Azure Batch
Activities: Custom activities
Compute environment: Azure Machine Learning Studio Machine
Activities: Learning activities: Batch Execution and Update Resource
Compute environment: Azure Machine Learning
Activities: Azure Machine Learning Execute Pipeline
Compute environment: Azure Data Lake Analytics
Activities: Data Lake Analytics U-SQL
Compute environment: Azure SQL, Azure SQL Data Warehouse, SQL Server
Activities: Stored Procedure
Compute environment: Azure Databricks
Activities: Notebook, Jar, Python
Compute environment: Azure Function
Activities: Azure Function activity

Transforming data using SQL Server Integration Services (SSIS) packages
Many organizations have decades of development investment in SSIS packages that contain both ingestion and transformation logic from on-premises and cloud data stores. Azure Data Factory provides the ability to lift and shift existing SSIS workload, by creating an Azure-SSIS Integration Runtime to natively execute SSIS packages. Using Azure-SSIS Integration Runtime will enable you to deploy and manage your existing SSIS packages with little to no change using familiar tools such as SQL Server Data Tools (SSDT) and SQL Server Management Studio (SSMS), just like using SSIS on premises.
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

☐ Data Storage Activities

☑ Compute Resources

Explanation:- Just as Azure Data Factory provides a variety of methods for ingesting data, it also provides a range of methods to perform transformations. You can pick a method that matches the skillsets of your team or takes advantage of existing technologies that you already have in your data estate. There is also the opportunity to perform transformations without writing code at all using the Mapping Data Flow.
Transforming data using Mapping Data Flow
Mapping Data Flows provide an environment for building a wide range of data transformations visually without the need to use code. The resulting data flows that are created are subsequently executed on scaled-out Apache Spark clusters that are automatically provisioned when you execute the Mapping Data Flow. Mapping Data Flows also provides the capability to monitor the execution of the transformations so that you can view how the transformations are progressing, or to understand any errors that may occur
Transforming data using compute resources
Azure Data Factory can also call on compute resources to transform data by a data platform service that may be better suited to the job. A great example of this is that Azure Data Factory can create a pipeline to an analytical data platform such as Spark pools in an Azure Synapse Analytics instance to perform a complex calculation using python. Another example could be to send data to an Azure SQL Database instance to execute a stored procedure using Transact-SQL. There is a wide range of compute resource, and the associated activities that they can perform as shown in the following table:
Compute environment: On-demand HDInsight cluster or your own HDInsight cluster
Activities: Hive, Pig, Spark, MapReduce, Hadoop Streaming
Compute environment: Azure Batch
Activities: Custom activities
Compute environment: Azure Machine Learning Studio Machine
Activities: Learning activities: Batch Execution and Update Resource
Compute environment: Azure Machine Learning

Activities: Azure Machine Learning Execute Pipeline
Compute environment: Azure Data Lake Analytics
Activities: Data Lake Analytics U-SQL
Compute environment: Azure SQL, Azure SQL Data Warehouse, SQL Server
Activities: Stored Procedure
Compute environment: Azure Databricks
Activities: Notebook, Jar, Python
Compute environment: Azure Function
Activities: Azure Function activity
Transforming data using SQL Server Integration Services (SSIS) packages
Many organizations have decades of development investment in SSIS packages that contain both ingestion and transformation logic from on-premises and cloud data stores. Azure Data Factory provides the ability to lift and shift existing SSIS workload, by creating an Azure-SSIS Integration Runtime to natively execute SSIS packages. Using Azure-SSIS Integration Runtime will enable you to deploy and manage your existing SSIS packages with little to no change using familiar tools such as SQL Server Data Tools (SSDT) and SQL Server Management Studio (SSMS), just like using SSIS on premises.
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

☐ Data Movement Flows

☑ SSIS Packages

Explanation:- Just as Azure Data Factory provides a variety of methods for ingesting data, it also provides a range of methods to perform transformations. You can pick a method that matches the skillsets of your team or takes advantage of existing technologies that you already have in your data estate. There is also the opportunity to perform transformations without writing code at all using the Mapping Data Flow.
Transforming data using Mapping Data Flow
Mapping Data Flows provide an environment for building a wide range of data transformations visually without the need to use code. The resulting data flows that are created are subsequently executed on scaled-out Apache Spark clusters that are automatically provisioned when you execute the Mapping Data Flow. Mapping Data Flows also provides the capability to monitor the execution of the transformations so that you can view how the transformations are progressing, or to understand any errors that may occur
Transforming data using compute resources

Azure Data Factory can also call on compute resources to transform data by a data platform service that may be better suited to the job. A great example of this is that Azure Data Factory can create a pipeline to an analytical data platform such as Spark pools in an Azure Synapse Analytics instance to perform a complex calculation using python. Another example could be to send data to an Azure SQL Database instance to execute a stored procedure using Transact-SQL. There is a wide range of compute resource, and the associated activities that they can perform as shown in the following table:
Compute environment: On-demand HDInsight cluster or your own HDInsight cluster
Activities: Hive, Pig, Spark, MapReduce, Hadoop Streaming
Compute environment: Azure Batch
Activities: Custom activities
Compute environment: Azure Machine Learning Studio Machine
Activities: Learning activities: Batch Execution and Update Resource
Compute environment: Azure Machine Learning
Activities: Azure Machine Learning Execute Pipeline
Compute environment: Azure Data Lake Analytics
Activities: Data Lake Analytics U-SQL
Compute environment: Azure SQL, Azure SQL Data Warehouse, SQL Server
Activities: Stored Procedure
Compute environment: Azure Databricks
Activities: Notebook, Jar, Python
Compute environment: Azure Function
Activities: Azure Function activity
Transforming data using SQL Server Integration Services (SSIS) packages
Many organizations have decades of development investment in SSIS packages that contain both ingestion and transformation logic from on-premises and cloud data stores. Azure Data Factory provides the ability to lift and shift existing SSIS workload, by creating an Azure-SSIS Integration Runtime to natively execute SSIS packages. Using Azure-SSIS Integration Runtime will enable you to deploy and manage your existing SSIS packages with little to no change using familiar tools such as SQL Server Data Tools (SSDT) and SQL Server Management Studio (SSMS), just like using SSIS on premises.
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

☐ Control Resources

Q. 42 What is a step in flattening a nested schema?

- ○ COPY data
- ◉ Explode Arrays
- ○ CREATE parquet file
- ○ LOAD CSV file

Report Error

◉ Explode Arrays

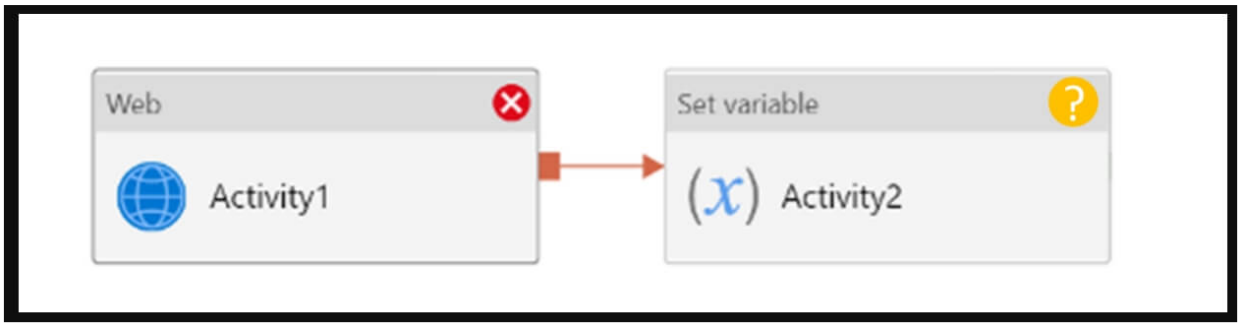
Explanation:- Explode Arrays is a third step in flattening nested schema's. It is necessary to transform the array in the data frame into a new dataframe where the column that you want to select is defined.
A DataFrame creates a data structure and it's one of the core data structures in Spark. In Spark, it is seen as a distributed collection of data that is organized into columns that have names.
Some use cases for transforming complex data types are as follows:
• Complex data types are increasingly common and represent a challenge for data engineers as analyzing nested schema and arrays tend to include time-consuming and complex SQL queries.
• It can be difficult to rename or cast the nested columns data type.
• Performance issues arise when working with deeply nested objects.
• Data Engineers need to understand how to efficiently process complex data types and make them easily accessible to everyone.
Synapse Spark can be used to read and transform objects into a flat structure through data frames. Synapse SQL serverless can be used to query such objects directly and return those results as a regular table. With Synapse Spark, it's easy to transform nested structures into columns and array elements into multiple rows.
In the overview below, the steps show the techniques involved to deal with complex data types
• Step 1: Define a function for flattening We define a function to flatten the nested schema.
• Step 2: Flatten nested schema Use the function to flatten the nested schema of the data frame (df) into a new data frame.
• Step 3: Explode Arrays Transform the array in the data frame into a new dataframe where you also define the column that you want to select.
• Step 4: Flatten child nested Schema Use the function you create to flatten the nested schema of the data frame into a new data frame.
https://medium.com/@saikrishna_55717/flattening-nested-data-json-xml-using-apache-spark-75fa4c8ea2a7

Report Error

Q. 43

Scenario: We are working on a project which has a pipeline with two activities where Activity2 has a failure dependency on Activity1.

What will the result be of the pipeline?



- ○ This pipeline reports skipped.
- ○ This pipeline reports completed.
- ◉ This pipeline reports failure.
- ○ This pipeline reports success.

Report Error

◉ This pipeline reports success.

Explanation:- If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.

Azure Data Factory

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

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

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

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

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

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

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

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

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

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

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

The four dependency conditions are:

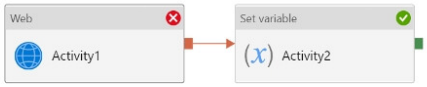
• Succeeded

• Failed

• Skipped

• Completed

For example, if a pipeline has an Activity A, followed by an Activity B and Activity B has as a dependency condition on Activity A 'Succeeded', then Activity B will only run if Activity A has the status of succeeded.
If you have multiple activities in a pipeline and subsequent activities are not dependent on previous activities, the activities may run in parallel.
https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/



Report Error

Q. 44

See the following code:

What is this code template used to setup?

```
PowerShell
$SubscriptionId = "add your subscription here"

Add-AzureRmAccount
Set-AzureRmContext -SubscriptionId $SubscriptionId

Register-AzureRmResourceProvider -ProviderNamespace Microsoft.DataFactory

$resourceGroupName = "cto_ignite"
$rglocation = "West US 2"

New-AzureRmDataFactoryV2 -ResourceGroupName $resourceGroupName -Name "ctoigniteADF" -Location $rglocation
```

- ○ Azure SQL Datawarehouse
- ◉ Azure Data Factory
- ○ Azure Network Security Groups
- ○ Azure Synapse Spark
- ○ Azure Linked Service
- ○ Azure Private Endpoint

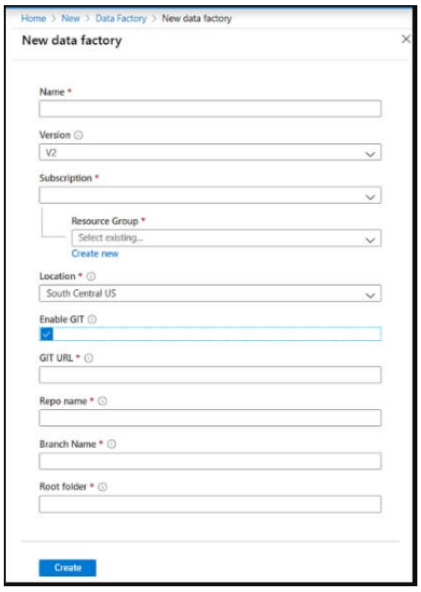
Report Error

- [x] Azure Data Factory

Explanation:- It is easy to set up Azure Data Factory from within the Azure portal, you only require the following information:
• Name: The name of the Azure Data Factory instance
• Subscription: The subscription in which the ADF instance is created
• Resource group: The resource group where the ADF instance will reside
• Version: select V2 for the latest features
• Location: The datacentre location in which the instance is stored
Enable Git provides the capability to integrate the code that you create with a Git repository enabling you to source control the code that you would create. Define the GIT url, repository name, branch name, and the root folder.
Alternatively, there are a number of different ways that you can provision the service programmatically. In this example you can see PowerShell at work to set up the environment.
PowerShell

```
################################################################################
## PART I: Creating an Azure Data Factory ##
################################################################################
# Sign in to Azure and set the WINDOWS AZURE subscription to work with
$SubscriptionId = "add your subscription in the quotes"
Add-AzureRmAccount
Set-AzureRmContext -SubscriptionId $SubscriptionId
# register the Microsoft Azure Data Factory resource provider
Register-AzureRmResourceProvider -ProviderNamespace Microsoft.DataFactory
# DEFINE RESOURCE GROUP NAME AND LOCATION PARAMETERS
$resourceGroupName = "cto_ignite"
$rglocation = "West US 2"
# CREATE AZURE DATA FACTORY
New-AzureRmDataFactoryV2 -ResourceGroupName $resourceGroupName -Name "ctoigniteADF" -Location $rglocation
```

https://docs.microsoft.com/en-us/azure/data-factory/quickstart-create-data-factory-portal



- [ ] Azure Linked Service
- [ ] Azure Synapse Spark
- [ ] Azure Private Endpoint
- [ ] Azure Network Security Groups

Report Error

Q. 45

Scenario: You are working at a bank setting up a database which will be used by all employee-levels of the bank. At the moment, you are setting up permissions for service representatives in a call centre.

Often, due to compliance, the caller has to identify themselves by giving them the last four digits of their credit card number that they may have an issue with. These data items cannot be fully exposed to the service representative in that call centre.

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

- ○ Dynamic Data Masking
- ○ Row-level security
- ○ Table-level security
- ◉ Column-level security

Report Error

- ○ Table-level security
- ○ Column-level security
- ◉ Dynamic Data Masking

Explanation:- If you would define a masking rule, that masks all but the last four digits for example of that credit card number, you would get a query that only gives as a result the last four digits of the credit card number.

Dynamic Data Masking

Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics support Dynamic Data Masking. It's all in the name, Dynamic Data Masking is masking and ensures limited data exposure to non-privileged users, such that they can't see it. It also helps you in preventing unauthorized access to sensitive data. The way Dynamic Data Masking does it, is helping customers to designate how much of the sensitive data to reveal such that it has minimal impact on the application layer. Dynamic Data Masking is a policy-based security feature. It will hide the sensitive data in a result set of a query that runs over designated database fields. However, the data in the database will not be changed.

Let's give you an example how it works. Let's say you work at a bank as a service representative in a call centre. Sometime, due to compliance, the caller has to identify themselves by giving them several digits of their credit card number that they might have an issue with. However, these data items, should not be fully exposed to the service representative in that call centre, answering the call. If you would define a masking rule, that masks all but the last four digits for example of that credit card number, you would get a query that only gives as a result the last four digits of the credit card number.

If the caller, for example, also had to provide the representative with personal information, that should not be seen by the developer that can query the production environments in order to troubleshoot, you should appropriately mask data in order to protect the given personal data such that compliance is not violated.

For Azure Synapse Analytics, the way to set up a Dynamic Data Masking policy is using PowerShell or the REST API. Bear in mind that it won't be possible for Azure Synapse Analytics to set the Dynamic Data Masking policy in the Azure portal through selecting the Dynamic Data Masking page under Security in the SQL DB configuration pane. You need to set it up using PowerShell or REST API as mentioned before. However, the configuration of the Dynamic Data Masking policy can be done by the Azure SQL Database admin, server admin, or SQL Security Manager roles.

In Azure Synapse Analytics, you can find Dynamic Data Masking here.

Looking into Dynamic Data Masking Policies:

• SQL users are excluded from masking

A couple of SQL users or Azure AD identities can get unmasked data in the SQL query results. Users with administrator privileges are always excluded from masking, and see the original data without any mask.
• Masking rules - Masking rules are a set of rules that define the designated fields to be masked including the masking function that is used. The designated fields can be defined using a database schema name, table name, and column name.
• Masking functions - Masking functions are a set of methods that control the exposure of data for different scenarios.
Dynamic Data Masking for your database in Azure Synapse Analytics using PowerShell cmdlets
• Data masking policies
• Get-AzSqlDatabaseDataMaskingPolicy
The Get-AzSqlDatabaseDataMaskingPolicy gets the data masking policy for a database.
The syntax for the Get-AzSqlDatabaseDataMaskingPolicy in PowerShell is as follows:
PowerShell
Get-AzSqlDatabaseDataMaskingPolicy [-ServerName] [-DatabaseName]
[-ResourceGroupName] [-DefaultProfile] [-WhatIf] [-Confirm]
[]
What the Get-AzSqlDatabaseDataMaskingPolicy cmdlet does, is getting the data masking policy of an Azure SQL database.
To use this cmdlet in PowerShell, you'd have to specify the following parameters to identify the database:
• ResourceGroupName: name of the resource group you deployed the database in
• ServerName: sql server name
• DatabaseName : name of the database
This cmdlet is also supported by the SQL Server Stretch Database service on Azure.
• Set-AzSqlDatabaseDataMaskingPolicy
The Set-AzSqlDatabaseDataMaskingPolicy sets data masking for a database.
The syntax for the Set-AzSqlDatabaseDataMaskingPolicy in PowerShell is as follows:
PowerShell
Set-AzSqlDatabaseDataMaskingPolicy [-PassThru] [-PrivilegedUsers] [-DataMaskingState]
[-ServerName] [-DatabaseName] [-ResourceGroupName]
[-DefaultProfile] [-WhatIf] [-Confirm] []
What the Set-AzSqlDatabaseDataMaskingPolicy cmdlet does is setting the data masking policy for an Azure SQL database.
To use this cmdlet in PowerShell, you'd have to specify the following parameters to identify the database:
•ResourceGroupName: name of the resource group that you deployed the database in

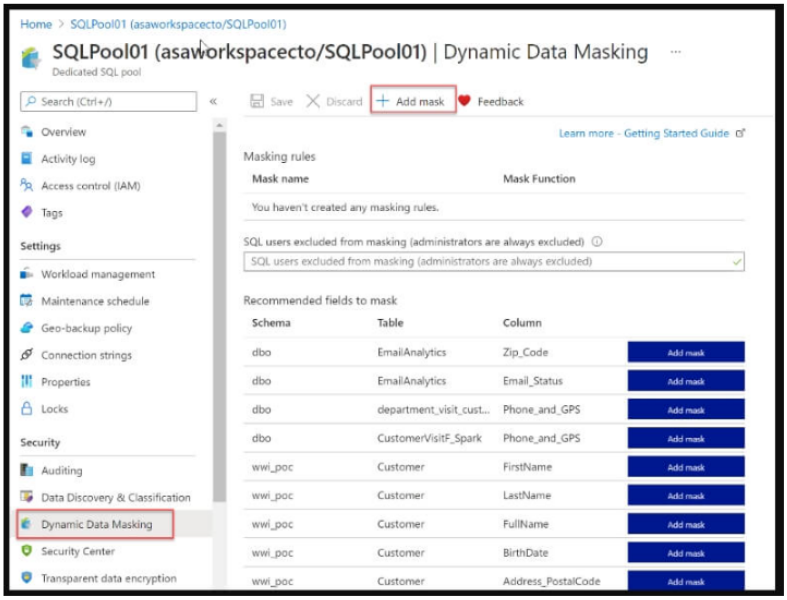
• ServerName : sql server name
• DatabaseName : name of the database
You'd also have to specify the RuleId parameter to specify which rule this cmdlet returns.
If you do not provide RuleId, all the data masking rules for that Azure SQL database are returned.
This cmdlet is also supported by the SQL Server Stretch Database service on Azure.
• New-AzSqlDatabaseDataMaskingRule
The New-AzSqlDatabaseDataMaskingRule creates a data masking rule for a database.
The syntax for the New-AzSqlDatabaseDataMaskingRule in PowerShell is as follows:
PowerShell
New-AzSqlDatabaseDataMaskingRule -MaskingFunction [-PrefixSize] [-ReplacementString]
[-SuffixSize] [-NumberFrom] [-NumberTo] [-PassThru] -SchemaName
-TableName -ColumnName [-ServerName] [-DatabaseName]
[-ResourceGroupName] [-DefaultProfile] [-WhatIf] [-Confirm]
[]
What the New-AzSqlDatabaseDataMaskingRule cmdlet does is creating a data masking rule for an Azure SQL database.
To use this cmdlet in PowerShell, you'd have to specify the following parameters to identify the rule:
• ResourceGroupName: name of the resource group that you deployed the database in
• ServerName : sql server name
• DatabaseName : name of the database
Providing the TableName and ColumnName is necessary in order to specify the target of the rule.
The MaskingFunction parameter is necessary to define how the data is masked.
If MaskingFunction has a value of Number or Text, you can specify the NumberFrom and NumberTo parameters, for number masking, or the PrefixSize, ReplacementString, and SuffixSize for text masking.
If the command succeeds and the PassThru parameter is used, the cmdlet returns an object describing the data masking rule properties in addition to the rule identifiers.
Rule identifiers can be, for example, ResourceGroupName, ServerName, DatabaseName, and RuleID.
This cmdlet is also supported by the SQL Server Stretch Database service on Azure.
• Remove-AzSqlDatabaseDataMaskingRule
The Remove-AzSqlDatabaseDataMaskingRule removes a data masking rule from a database.
The syntax for the Remove-AzSqlDatabaseDataMaskingRule in PowerShell is as follows:

PowerShell
Remove-AzSqlDatabaseDataMaskingRule [-PassThru] [-Force] -SchemaName -TableName
-ColumnName [-ServerName] [-DatabaseName] [-ResourceGroupName]
[-DefaultProfile] [-WhatIf] [-Confirm] []
What the Remove-AzSqlDatabaseDataMaskingRule cmdlet does, is it removes a specific data masking rule from an Azure SQL database.
To use this cmdlet in PowerShell, you'd have to specify the following parameters to identify the rule that needs to be removed:
• ResourceGroupName: name of the resource group that you deployed the database in
• ServerName : sql server name
• DatabaseName : name of the database
• RuleId : identifier of the rule
This cmdlet is also supported by the SQL Server Stretch Database service on Azure.
• Set-AzSqlDatabaseDataMaskingRule
The Set-AzSqlDatabaseDataMaskingRule Sets the properties of a data masking rule for a database.
The syntax for the Set-AzSqlDatabaseDataMaskingRule in PowerShell is as follows:
PowerShell
Set-AzSqlDatabaseDataMaskingRule [-MaskingFunction] [-PrefixSize]
[-ReplacementString] [-SuffixSize] [-NumberFrom] [-NumberTo] [-PassThru]
-SchemaName -TableName -ColumnName [-ServerName] [-DatabaseName]
[-ResourceGroupName] [-DefaultProfile] [-WhatIf] [-Confirm]
[]
What the Set-AzSqlDatabaseDataMaskingRule cmdlet does is setting a data masking rule for an Azure SQL database.
To use this cmdlet in PowerShell, you'd have to specify the following parameters to identify the rule:
• ResourceGroupName: name of the resource group that you deployed the database in
• ServerName : sql server name
• DatabaseName : name of the database
• RuleId : identifier of the rule
You can provide any of the parameters of SchemaName, TableName, and ColumnName to retarget the rule.
Specify the MaskingFunction parameter to modify how the data is masked.
If you specify a value of Number or Text for MaskingFunction, you can specify the NumberFrom and NumberTo parameters for number masking or the PrefixSize, ReplacementString, and SuffixSize parameters for text masking.

If the command succeeds, and if you specify the PassThru parameter, the cmdlet returns an object that describes the data masking rule properties and the rule identifiers.	-
Rule identifiers can be, ResourceGroupName, ServerName, DatabaseName, and RuleId.	
This cmdlet is also supported by the SQL Server Stretch Database service on Azure.	-
Set up Dynamic Data Masking for your database in Azure Synapse Analytics using the REST API	
For setting up Dynamic Data Masking in Azure Synapse Analytics, the other possibility is make use of the REST API.	
It will enable to programmatically manage data masking policy and rules.	
The REST API will support the following operations:	
• Data masking policies	
• Create Or Update	
The Create Or Update masking policy using the REST API will create or update a database data masking policy.	
In HTTP the following request can be made:	
HTTP	
PUT	
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/databases/{databaseName}/da	
api-version=2014-04-01	
The following parameters need to be passed through:	
• SubscriptionID: the ID of the subscription	
• ResourceGroupName: name of the resource group that you deployed the database in	
• ServerName : sql server name	
• DatabaseName : name of the database	
• dataMaskingPolicyName: the name of the data masking policy	
• api version: version of the api that is used.	
• Get	
The Get policy, Gets a database data masking policy.	
In HTTP the following request can be made:	
HTTP	
GET	
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/databases/{databaseName}/da	
api-version=2014-04-01	

The following parameters need to be passed through:
• SubscriptionID: the ID of the subscription
• ResourceGroupName: name of the resource group that you deployed the database in
• ServerName : sql server name
• DatabaseName : name of the database
• dataMaskingPolicyName: the name of the data masking policy
• api version: version of the api that is used.
• Data masking rules
• Create Or Update
The Create or Update masking rule creates or updates a database data masking rule.
In HTTP the following request can be made:
HTTP
PUT
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/databases/{databaseName}/d
api-version=2014-04-01
The following parameters need to be passed through:
• SubscriptionID: the ID of the subscription
• ResourceGroupName: name of the resource group that you deployed the database in
• ServerName : sql server name
• DatabaseName : name of the database
• dataMaskingPolicyName: the name of the data masking policy
• dataMaskingRuleName: the name of the rule for data masking
• api version: version of the api that is used.
• List By Database
The List By Database request gets a list of database data masking rules.
In HTTP the following request can be made:
HTTP
GET
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/databases/{databaseName}/d
api-version=2014-04-01

The following parameters need to be passed through:
• SubscriptionID: the ID of the subscription
• ResourceGroupName: name of the resource group that you deployed the database in
• ServerName : sql server name
• DatabaseName : name of the database
• dataMaskingPolicyName: the name of the data masking policy
• api version: version of the api that is used.
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview



○ Row-level security

Q. 46

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

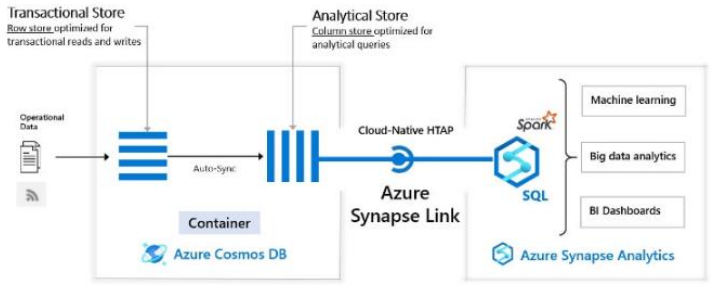
Many business application architectures separate transactional and analytical processing into separate systems with data stored and processed on separate infrastructures. [?] systems are optimized for dealing with discrete system or user requests immediately and responding as quickly as possible.

- (●) OLAP
- ( ) OLTP
- ( ) ADPS
- ( ) ELT
- ( ) ETL

Report Error

- ( ) ETL
- ( ) ELT
- ( ) OLAP
- (●) OLTP

Explanation:- Many business application architectures separate transactional and analytical processing into separate systems with data stored and processed on separate infrastructures. These infrastructures are commonly referred to as OLTP (online transaction processing) systems working with operational data, and OLAP (online analytical processing) systems working with historical data, with each system is optimized for their specific task.
OLTP systems are optimized for dealing with discrete system or user requests immediately and responding as quickly as possible.
OLAP systems are optimized for the analytical processing, ingesting, synthesizing, and managing large sets of historical data. The data processed by OLAP systems largely originates from OLTP systems and needs to be loaded into the OLTP systems by means of batch processes commonly referred to as ETL (Extract, Transform, and Load) jobs.
Due to their complexity and the need to physically copy large amounts of data, this creates a delay in data being available to provide insights by way of the OLAP systems.
As more and more businesses move to digital processes, they increasingly recognize the value of being able to respond to opportunities by making faster and well-informed decisions. HTAP (Hybrid Transactional/Analytical processing) enables business to run advanced analytics in near-real-time on data stored and processed by OLTP systems.
Azure Synapse Link for Azure Cosmos DB
Azure Synapse Link for Azure Cosmos DB is a cloud-native HTAP capability that enables you to run near-real-time analytics over operational data stored in Azure Cosmos DB. Azure Synapse Link creates a tight seamless integration between Azure Cosmos DB and Azure Synapse Analytics.
Azure Cosmos DB provides both a transactional store optimized for transactional workloads and an analytical store optimized for analytical workloads and a fully managed autosync process to keep the data within these stores in sync.
Azure Synapse Analytics provides both a SQL Serverless query engine for querying the analytical store using familiar T-SQL and an Apache Spark query engine for leveraging the analytical store using your choice of Scala, Java, Python or SQL and provides a user-friendly notebook experience.
Together Azure Cosmos DB and Synapse Analytics enable organizations to generate and consume insights from their operational data in near-real time, using the query and analytics tools of their choice. All of this is achieved without the need for complex ETL pipelines and without affecting the performance of their OLTP systems using Azure Cosmos DB.
https://docs.microsoft.com/en-us/azure/cosmos-db/synapse-link



ADPS

Report Error

Q. 47

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 import the daily inventory data from the SQL server to Azure Data Lake Storage.

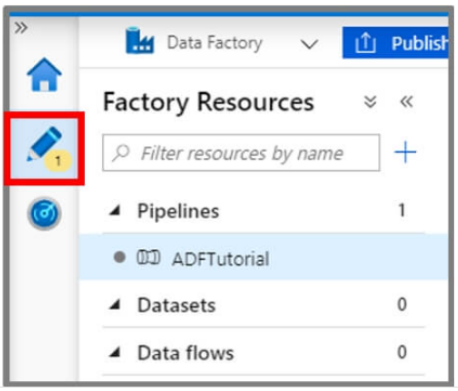
Which Azure Data Factory components should you recommend for the trigger type?

- ○ Scaling window trigger
- ○ Event-based trigger
- ○ Schedule trigger
- ◉ Tumbling window trigger

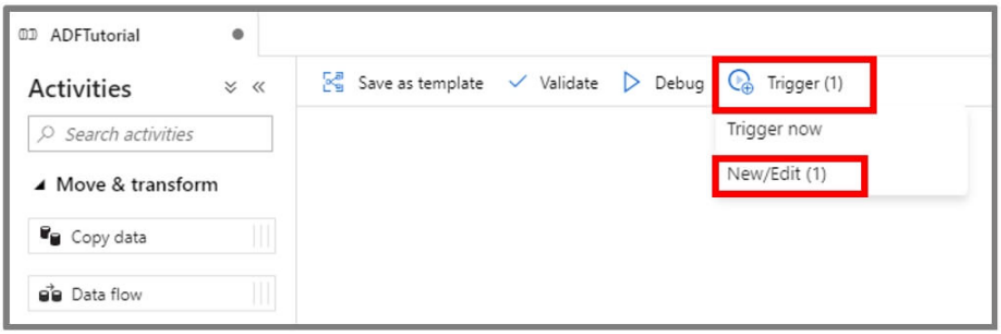
Report Error

◉ Schedule trigger

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.
Create a trigger that runs a pipeline on a schedule
When creating a schedule trigger, you specify a schedule (start date, recurrence, end date etc.) for the trigger, and associate with a pipeline. Pipelines and triggers have a many-to-many relationship. Multiple triggers can kick off a single pipeline. A single trigger can kick off multiple pipelines.
Note: For a complete walkthrough of creating a pipeline and a schedule trigger, which associates the trigger with the pipeline, and runs and monitors the pipeline, see Quickstart: create a data factory using Data Factory UI.
Data Factory UI
You can create a schedule trigger to schedule a pipeline to run periodically (hourly, daily, etc.).
1. Switch to the Edit tab, shown with a pencil symbol.

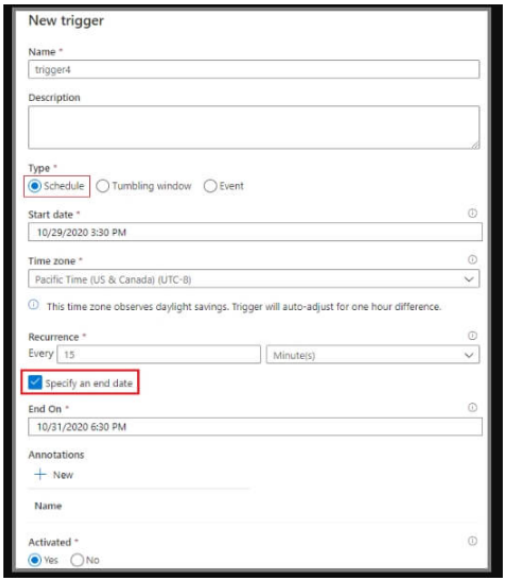


2. Select Trigger on the menu, then select New/Edit.

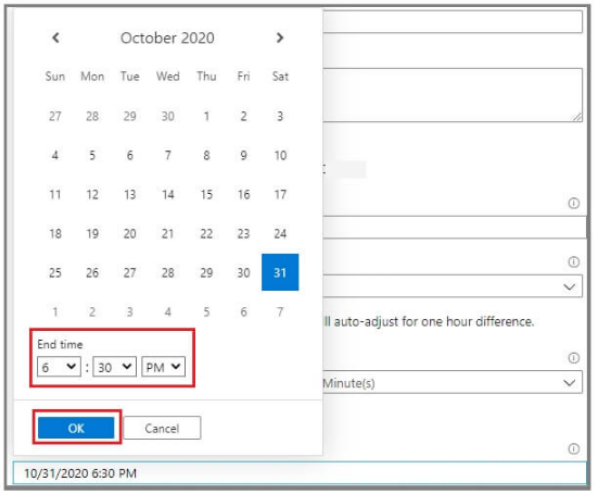


3. On the Add Triggers page, select Choose trigger..., then select +New.

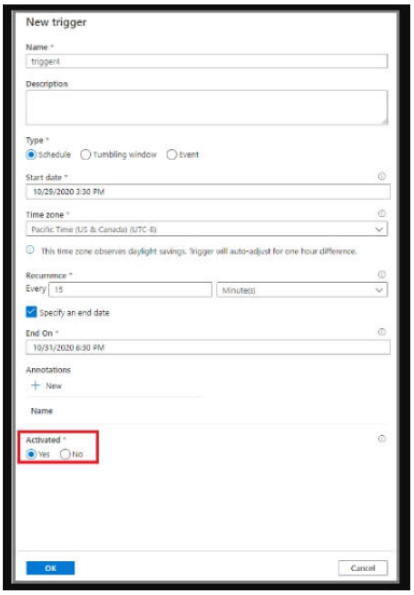
4. On the New Trigger page, do the following steps:



• Confirm that Schedule is selected for Type.
• Specify the start datetime of the trigger for Start Date. It's set to the current datetime in Coordinated Universal Time (UTC) by default.
• Specify the time zone that the trigger will be created in. The time zone setting will apply to Start Date, End Date, and Schedule Execution Times in Advanced recurrence options. Changing Time Zone setting will not automatically change your start date. Make sure the Start Date is correct in the specified time zone. Please note that Scheduled Execution time of Trigger will be considered post the Start Date (Ensure Start Date is atleast 1minute lesser than the Execution time else it will trigger pipeline in next recurrence).
Note: For time zones that observe daylight saving, trigger time will auto-adjust for the twice a year change. To opt out of the daylight saving change, please select a time zone that does not observe daylight saving, for instance UTC
• Specify Recurrence for the trigger. Select one of the values from the drop-down list (Every minute, Hourly, Daily, Weekly, and Monthly). Enter the multiplier in the text box. For example, if you want the trigger to run once for every 15 minutes, you select Every Minute, and enter 15 in the text box.
• To specify an end date time, select Specify an End Date, and specify Ends On, then select OK. There is a cost associated with each pipeline run. If you are testing, you may want to ensure that the pipeline is triggered only a couple of times. However, ensure that there is enough time for the pipeline to run between the publish time and the end time. The trigger comes into effect only after you publish the solution to Data Factory, not when you save the trigger in the UI.
5. In the New Trigger window, select Yes in the Activated option, then select OK. You can use this checkbox to deactivate the trigger later.

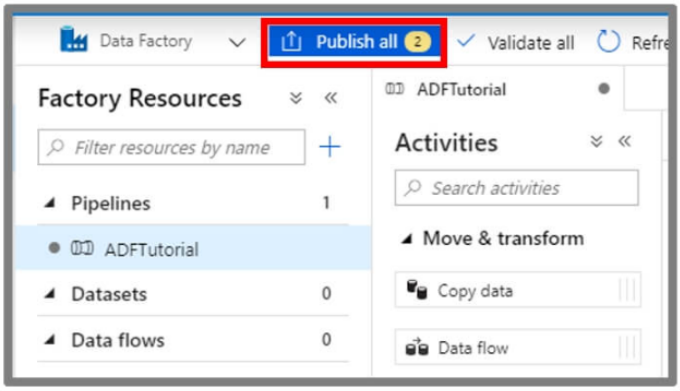


6. In the New Trigger window, review the warning message, then select OK.

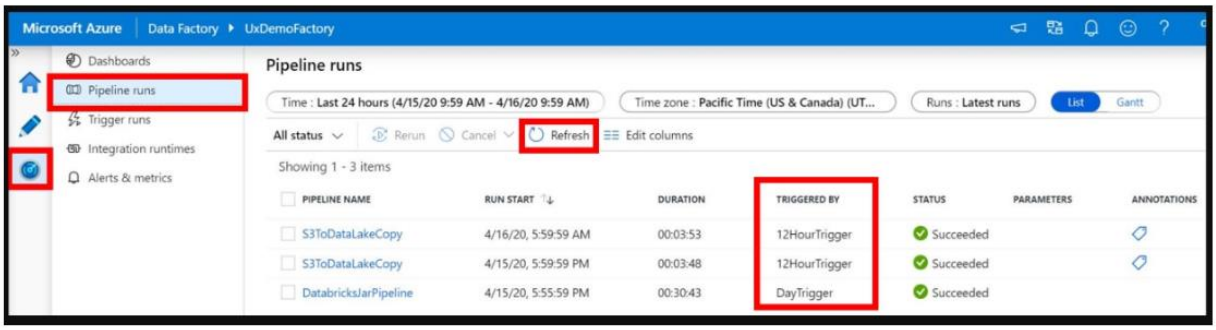


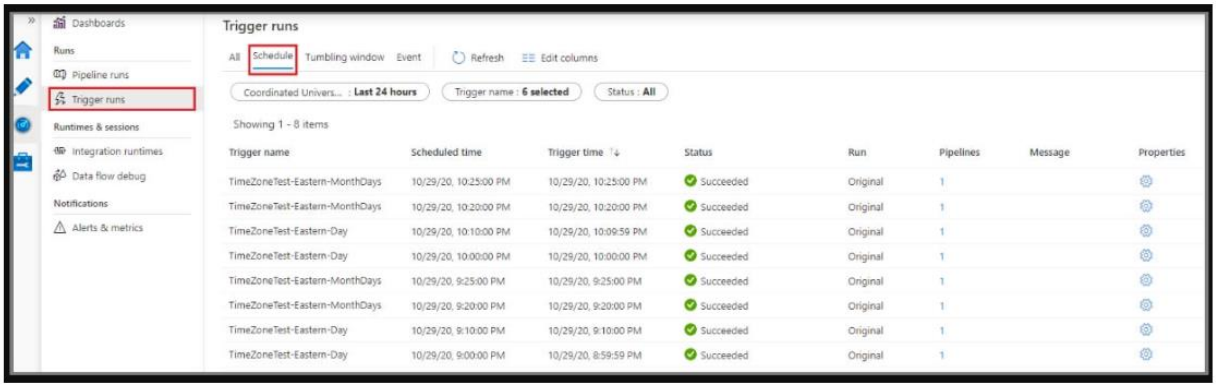
7. Select Publish all to publish the changes to Data Factory. Until you publish the changes to Data Factory, the trigger doesn't start triggering the pipeline runs.

8. Switch to the Pipeline runs tab on the left, then select Refresh to refresh the list. You will see the pipeline runs triggered by the scheduled trigger. Notice the values in the Triggered By column. If you use the Trigger Now option, you will see the manual trigger run in the list.

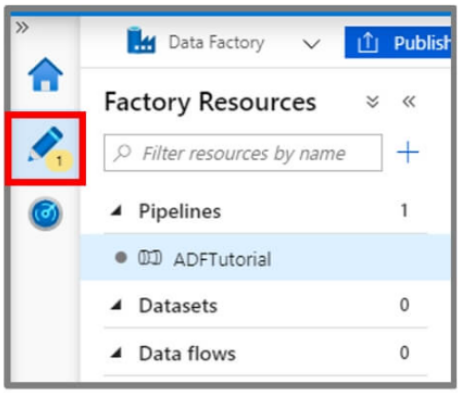


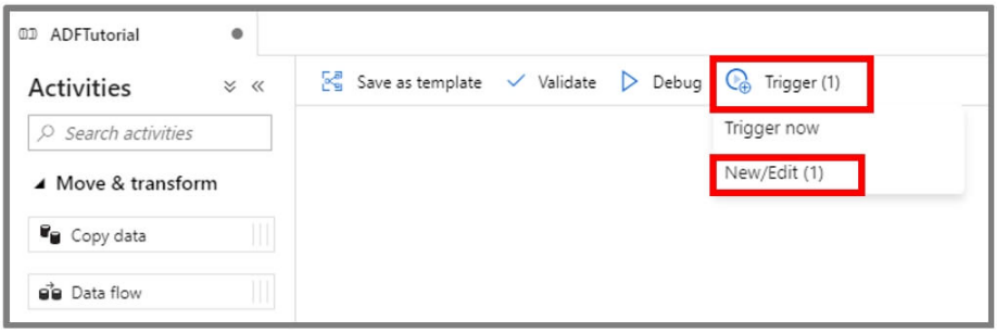
9. Switch to the Trigger Runs \ Schedule view.



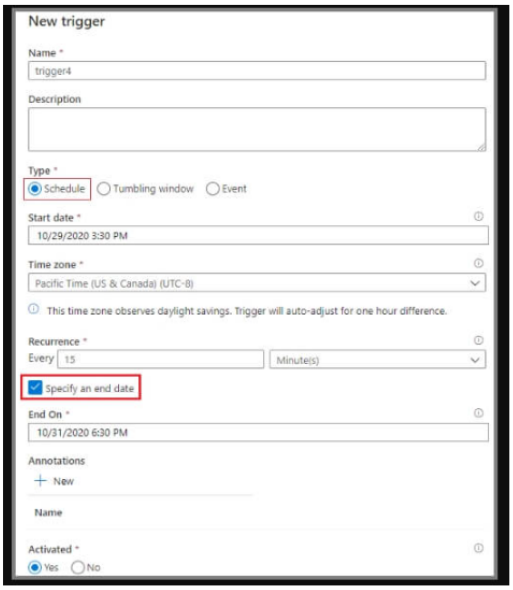


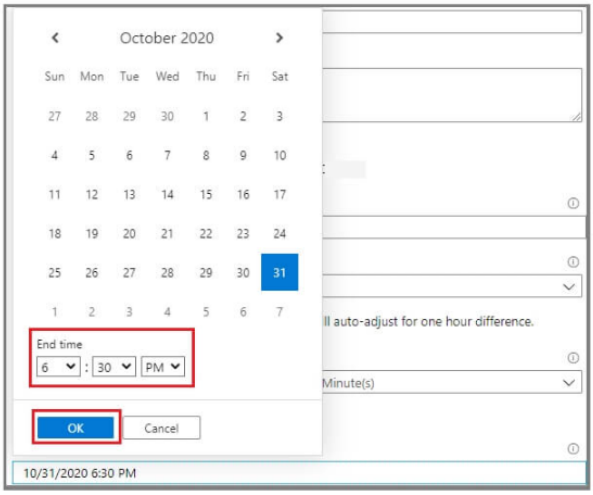
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger

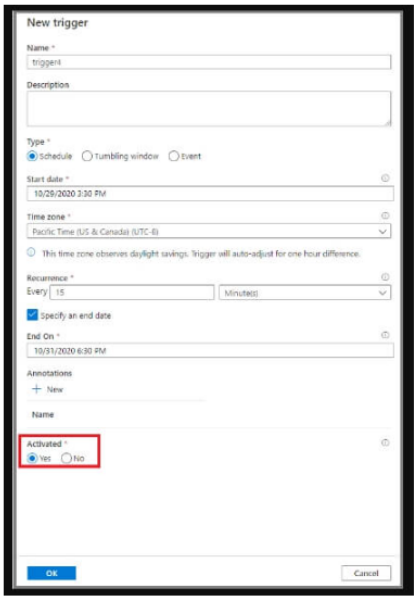

Data Factory
Publish
Factory Resources
Filter resources by name
Pipelines 1
ADFTutorial
Datasets 0
Data flows 0


ADFTutorial
Activities
Search activities
Move & transform
Copy data
Data flow
Save as template
Validate
Debug
Trigger (1)
Trigger now
New/Edit (1)

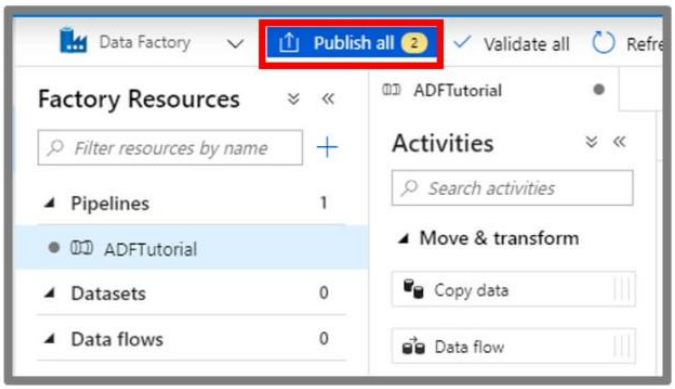

Add triggers
Choose trigger...
Search
New

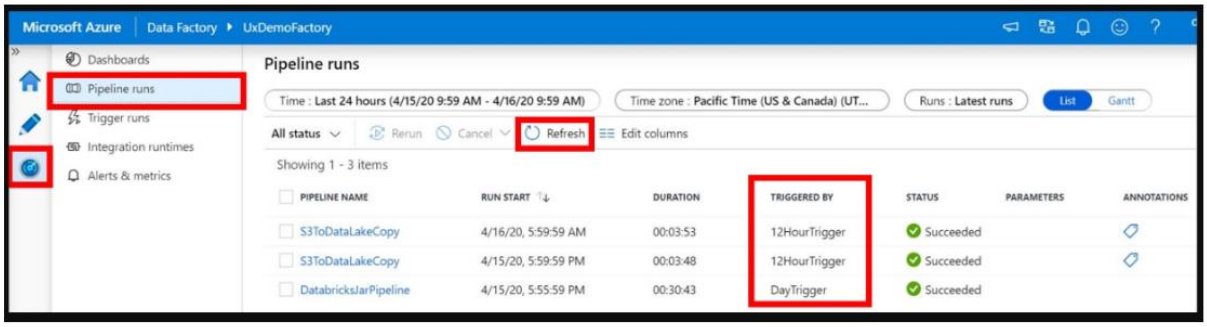

New trigger
Name *
trigger4
Description
Type *
Schedule
Tumbling window
Event
Start date *
10/29/2020 3:30 PM
Time zone *
Pacific Time (US & Canada) (UTC-8)
This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.
Recurrence *
Every 15
Minute(s)
Specify an end date
End On *
10/31/2020 6:30 PM
Annotations
New
Name
Activated *
Yes
No

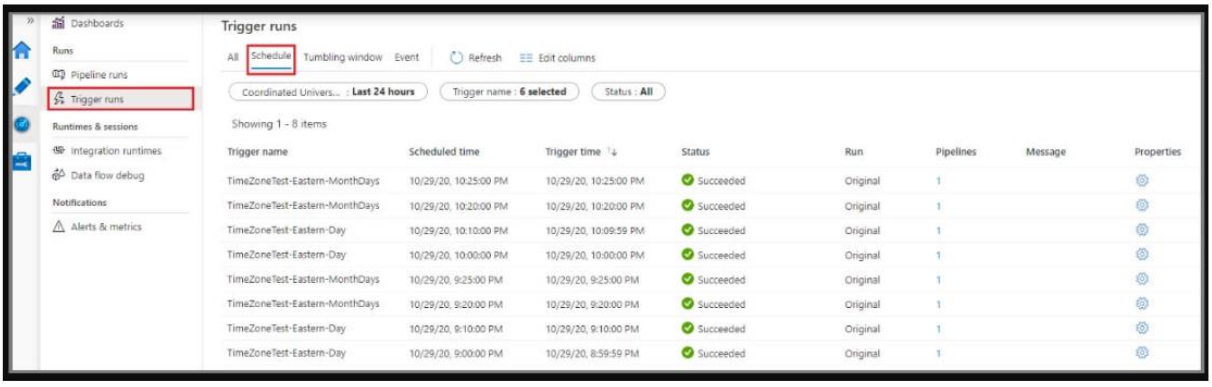

October 2020
Sun Mon Tue Wed Thu Fri Sat
27 28 29 30 1 2 3
4 5 6 7 8 9 10
11 12 13 14 15 16 17
18 19 20 21 22 23 24
25 26 27 28 29 30 31
1 2 3 4 5 6 7
End time
6 : 30 PM
OK
Cancel
ll auto-adjust for one hour difference.
Minute(s)
10/31/2020 6:30 PM


New trigger
Name *
trigger1
Description
Type *
Schedule
Tumbling window
Event
Start date *
10/29/2020 3:30 PM
Time zone *
Pacific Time (US & Canada) (UTC-8)
This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.
Recurrence *
Every
15
Minute(s)
Specify an end date
End On *
10/31/2020 6:30 PM
Annotations
New
Name
Activated *
Yes
No
OK
Cancel


New trigger
Trigger Run Parameters
NAME
TYPE
VALUE
This pipeline has no parameters
Make sure to "Publish" for trigger to be activated after clicking "OK"
OK
Cancel







- ○ Event-based trigger
- ○ Tumbling window trigger
- ○ Scaling window trigger

Report Error

Q. 48

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

Which transformations type is best described by:
"A Sort transformation that orders the data."

- (●) None of the listed options.
- ( ) Schema modifier transformations
- ( ) Multiple inputs/outputs transformations
- ( ) Row modifier transformations

Report Error

- ( ) Schema modifier transformations
- (●) Row modifier transformations

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:
Category Name: Schema modifier transformations
Description: These types of transformations will make a modification to a sink destination by creating new columns based on the action of the transformation. An example of this is the Derived Column transformation that will create a new column based on the operations performed on existing column.
Category Name: Row modifier transformations
Description: These types of transformations impact how the rows are presented in the destination. An example of this is a Sort transformation that orders the data.
Category Name: Multiple inputs/outputs transformations
Description: These types of transformations will generate new data pipelines or merge pipelines into one. An example of this is the Union transformation that combines multiple data streams.
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

- ( ) None of the listed options.
- ( ) Multiple inputs/outputs transformations

Report Error

Q. 49

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. A single Azure subscription can host up to [A] storage accounts, each of which can hold [B] TB of data.

- ○ [A] 500, [B] 500
- ○ [A] 200, [B] 500
- ○ [A] 500, [B] 1000
- ◉ [A] 250, [B] 500

Report Error

- ○ [A] 500, [B] 500
- ○ [A] 200, [B] 500
- ◉ [A] 250, [B] 500

Explanation:- Scale targets for standard storage accounts
The following table describes default limits for Azure general-purpose v1, v2, Blob storage, and block blob storage accounts. The ingress limit refers to all data that is sent to a storage account. The egress limit refers to all data that is received from a storage account.
https://docs.microsoft.com/en-us/azure/storage/common/scalability-targets-standard-account

Resource	Limit
Number of storage accounts per region per subscription, including standard, and premium storage accounts.	250
Maximum storage account capacity	5 PiB [1]
Maximum number of blob containers, blobs, file shares, tables, queues, entities, or messages per storage account	No limit
Maximum request rate[1] per storage account	20,000 requests per second
Maximum ingress[1] per storage account (US, Europe regions)	10 Gbps
Maximum ingress[1] per storage account (regions other than US and Europe)	5 Gbps if RA-GRS/GRS is enabled, 10 Gbps for LRS/ZRS[2]
Maximum egress for general-purpose v2 and Blob storage accounts (all regions)	50 Gbps
Maximum egress for general-purpose v1 storage accounts (US regions)	20 Gbps if RA-GRS/GRS is enabled, 30 Gbps for LRS/ZRS[2]
Maximum egress for general-purpose v1 storage accounts (non-US regions)	10 Gbps if RA-GRS/GRS is enabled, 15 Gbps for LRS/ZRS[2]
Maximum number of virtual network rules per storage account	200
Maximum number of IP address rules per storage account	200

Q. 50

Correct or Incorrect : Mapping data flows are visually displayed data transformations in Azure Data Factory.

Data flows allow data engineers to develop data transformation logic with or without writing code.

- ○ Incorrect
- ◉ Correct

Report Error

◉ Incorrect

Explanation:-

Transforming data with the Mapping Data Flow

You can natively perform data transformations with Azure Data Factory code free using the Mapping Data Flow task. Mapping Data Flows provide a fully visual experience with no coding required. Your data flows will run on your own execution cluster for scaled-out data processing. Data flow activities can be operationalized via existing Data Factory scheduling, control, flow, and monitoring capabilities.

When building data flows, you can enable debug mode, which turns on a small interactive Spark cluster. Turn on debug mode by toggling the slider at the top of the authoring module. Debug clusters take a few minutes to warm up, but can be used to interactively preview the output of your transformation logic.

With the Mapping Data Flow added, and the Spark cluster running, this will enable you to perform the transformation, and run and preview the data. No coding is required as Azure Data Factory handles all the code translation, path optimization, and execution of your data flow jobs.

Adding source data to the Mapping Data Flow

Open the Mapping Data Flow canvas. Click on the Add Source button in the Data Flow canvas. In the source dataset dropdown, select your data source, n this case the ADLS Gen2 dataset is used in this example

There are a couple of points to note:

• If your dataset is pointing at a folder with other files and you only want to use one file, you may need to create another dataset or utilize parameterization to make sure only a specific file is read

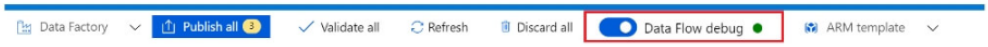
• If you have not imported your schema in your ADLS, but have already ingested your data, go to the dataset's 'Schema' tab and click 'Import schema' so that your data flow knows the schema projection.

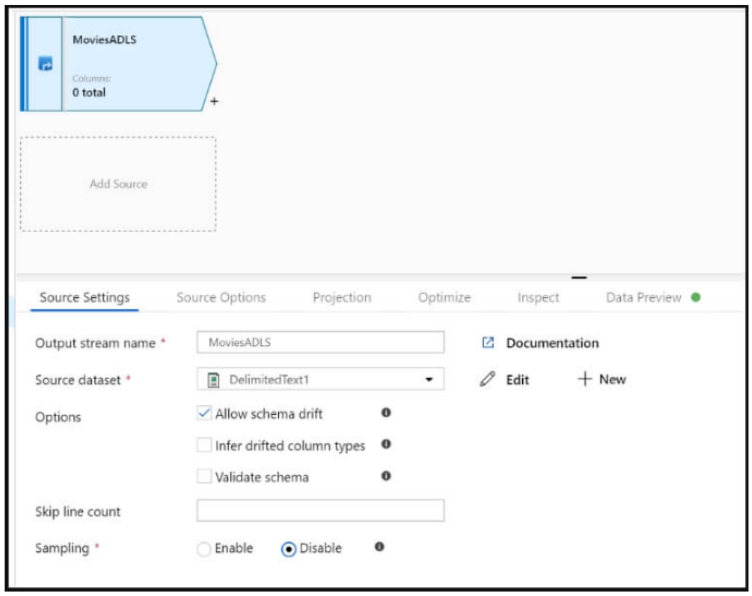
Mapping Data Flow follows an extract, load, transform (ELT) approach and works with staging datasets that are all in Azure. Currently the following datasets can be used in a source transformation:

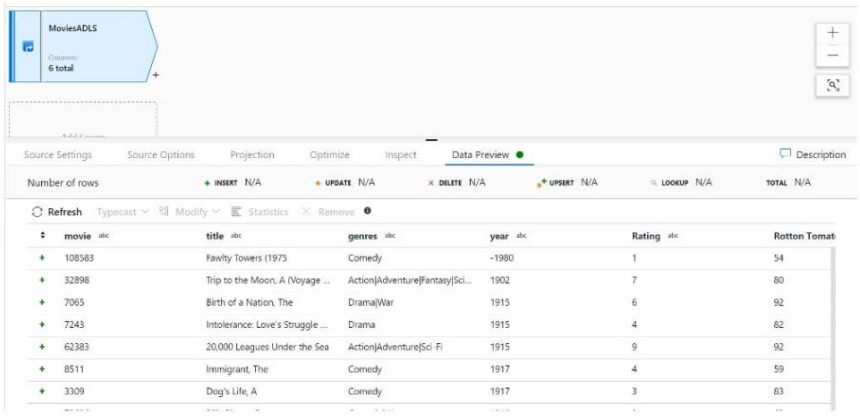
• Azure Blob Storage (JSON, Avro, Text, Parquet)
• Azure Data Lake Storage Gen1 (JSON, Avro, Text, Parquet)
• Azure Data Lake Storage Gen2 (JSON, Avro, Text, Parquet)
• Azure Synapse Analytics
• Azure SQL Database
• Azure CosmosDB

Azure Data Factory has access to over 80 native connectors. To include data from those other sources in your data flow, use the Copy Activity to load that data into one of the supported staging areas.

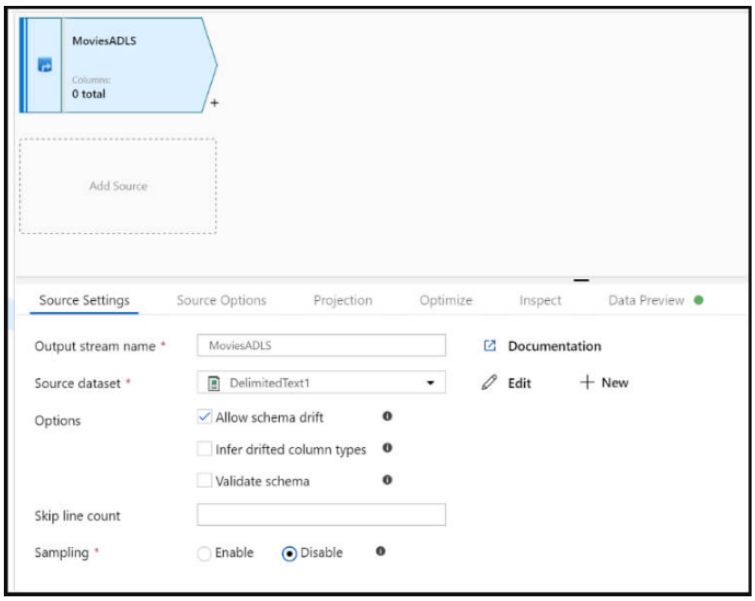
Once your debug cluster is warmed up, verify your data is loaded correctly via the Data Preview tab. Once you click the refresh button, Mapping Data Flow will show a snapshot of what your data looks like when it is at each transformation.

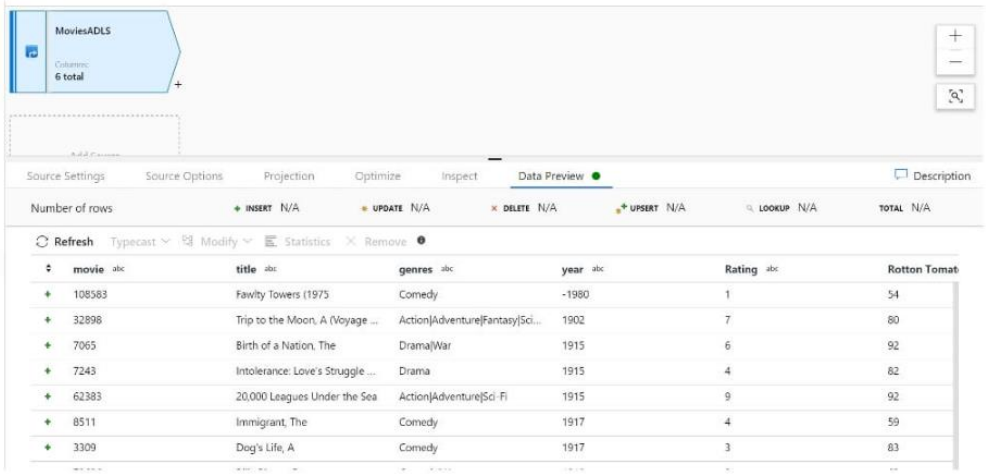






https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-overview





Correct

Report Error

Q. 51

Azure Data Lake Storage combines a file system with a storage platform to help you quickly identify insights into your data. Data Lake Storage Gen2 builds on Azure Blob storage capabilities to optimize it specifically for analytics workloads.

You can set permissions at a directory level or file level for the data stored within the data lake. This security is configurable through technologies such as Hive and Spark, or utilities such as Azure Storage Explorer. All data that is stored is encrypted at rest by using either Microsoft or customer-managed keys.

Data Lake Storage Gen2 supports which of the following to enhance security?

- [ ] AWS
- [x] POSIX
- [ ] GRS
- [x] HDFS
- [x] ACLs
- [ ] LRS

Report Error

- [ ] AWS
- [ ] LRS
- [ ] HDFS

- [x] POSIX

Explanation:- A data lake is a repository of data that is stored in its natural format, usually as blobs or files. Azure Data Lake Storage is a comprehensive, scalable, and cost-effective data lake solution for big data analytics built into Azure.

Azure Data Lake Storage combines a file system with a storage platform to help you quickly identify insights into your data. Data Lake Storage Gen2 builds on Azure Blob storage capabilities to optimize it specifically for analytics workloads. This integration enables analytics performance, the tiering and data lifecycle management capabilities of Blob storage, and the high-availability, security, and durability capabilities of Azure Storage.

The variety and volume of data that is generated and analyzed today is increasing. Companies have multiple sources of data, from websites to Point of Sale (POS) systems, and more recently from social media sites to Internet of Things (IoT) devices. Each source provides an essential aspect of data that needs to be collected, analyzed, and potentially acted upon.

Benefits

Data Lake Storage Gen2 is designed to deal with this variety and volume of data at exabyte scale while securely handling hundreds of gigabytes of throughput. With this, you can use Data Lake Storage Gen2 as the basis for both real-time and batch solutions. Here is a list of additional benefits that Data Lake Storage Gen2 brings:

Hadoop compatible access

A benefit of Data Lake Storage Gen2 is that you can treat the data as if it's stored in a Hadoop Distributed File System. With this feature, you can store the data in one place and access it through compute technologies including Azure Databricks, Azure HDInsight, and Azure Synapse Analytics without moving the data between environments.

Security

Data Lake Storage Gen2 supports access control lists (ACLs) and Portable Operating System Interface (POSIX) permissions. You can set permissions at a directory level or file level for the data stored within the data lake. This security is configurable through technologies such as Hive and Spark, or utilities such as Azure Storage Explorer. All data that is stored is encrypted at rest by using either Microsoft or customer-managed keys.

Performance

Azure Data Lake Storage organizes the stored data into a hierarchy of directories and subdirectories, much like a file system, for easier navigation. As a result, data processing requires less computational resources, reducing both the time and cost.

Data redundancy

Data Lake Storage Gen2 takes advantage of the Azure Blob replication models that provide data redundancy in a single data centre with locally redundant storage (LRS), or to a secondary region by using the Geo-redundant storage (GRS) option. This feature ensures that your data is always available and protected if catastrophe strikes.

https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-overview

- [x] ACLs

Explanation:- A data lake is a repository of data that is stored in its natural format, usually as blobs or files. Azure Data Lake Storage is a comprehensive, scalable, and cost-effective data lake solution for big data analytics built into Azure.
Azure Data Lake Storage combines a file system with a storage platform to help you quickly identify insights into your data. Data Lake Storage Gen2 builds on Azure Blob storage capabilities to optimize it specifically for analytics workloads. This integration enables analytics performance, the tiering and data lifecycle management capabilities of Blob storage, and the high-availability, security, and durability capabilities of Azure Storage.
The variety and volume of data that is generated and analyzed today is increasing. Companies have multiple sources of data, from websites to Point of Sale (POS) systems, and more recently from social media sites to Internet of Things (IoT) devices. Each source provides an essential aspect of data that needs to be collected, analyzed, and potentially acted upon.
Benefits
Data Lake Storage Gen2 is designed to deal with this variety and volume of data at exabyte scale while securely handling hundreds of gigabytes of throughput. With this, you can use Data Lake Storage Gen2 as the basis for both real-time and batch solutions. Here is a list of additional benefits that Data Lake Storage Gen2 brings:
Hadoop compatible access
A benefit of Data Lake Storage Gen2 is that you can treat the data as if it's stored in a Hadoop Distributed File System. With this feature, you can store the data in one place and access it through compute technologies including Azure Databricks, Azure HDInsight, and Azure Synapse Analytics without moving the data between environments.
Security
Data Lake Storage Gen2 supports access control lists (ACLs) and Portable Operating System Interface (POSIX) permissions. You can set permissions at a directory level or file level for the data stored within the data lake. This security is configurable through technologies such as Hive and Spark, or utilities such as Azure Storage Explorer. All data that is stored is encrypted at rest by using either Microsoft or customer-managed keys.
Performance
Azure Data Lake Storage organizes the stored data into a hierarchy of directories and subdirectories, much like a file system, for easier navigation. As a result, data processing requires less computational resources, reducing both the time and cost.
Data redundancy
Data Lake Storage Gen2 takes advantage of the Azure Blob replication models that provide data redundancy in a single data centre with locally redundant storage (LRS), or to a secondary region by using the Geo-redundant storage (GRS) option. This feature ensures that your data is always available and protected if catastrophe strikes.
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-overview

- [ ] GRS

Report Error

Q. 52

Scenario: Pennyworth's Haberdashery is a clothing retailer based in London. The company has 2,000 retail stores across the EU and an emerging online presence. The network contains an Active Directory forest named pennyworths.com. The forest it integrated with an Azure Active Directory (Azure AD) tenant named pennyworths.com. Pennyworth's has an Azure subscription associated to the pennyworths.com Azure AD tenant.

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

You have been hired as a consultant by Alfred Pennyworth to advise on very important projects within the company.

During your assessment of the IT environment, you estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

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

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

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

The e-commerce department at Pennyworth's develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements
Pennyworth's plans to implement the following changes:
• Load the sales transaction dataset to Azure Synapse Analytics.
• Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.
• Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Pennyworth's identifies the following requirements for the sales transaction dataset:
• Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
• Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
• Implement a surrogate key to account for changes to the retail store addresses.
• Ensure that data storage costs and performance are predictable.
• Minimize how long it takes to remove old records.

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

Data Integration Requirements
Pennyworth's identifies the following requirements for data integration:
• Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data.
• Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

The Ask:
Alfred places a great importance on this project and asks you to work closely with the team to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements.

Which of the following should you advise the team to create?

- ○ A table that has an IDENTITY property.
- ○ A user-defined SEQUENCE object.
- ◉ A system-versioned temporal table.
- ○ A table that has a FOREIGN KEY constraint.

Report Error

◉ A table that has an IDENTITY property.

Explanation:- The best way to implement a surrogate key to account for changes to the retail store addresses is to create a table that has an IDENTITY property.

A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Using IDENTITY to create surrogate keys using dedicated SQL pool in Azure Synapse Analytics

What is a surrogate key?

A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Note: In Azure Synapse Analytics, the IDENTITY value increases on its own in each distribution and does not overlap with IDENTITY values in other distributions. The IDENTITY value in Synapse is not guaranteed to be unique if the user explicitly inserts a duplicate value with "SET IDENTITY_INSERT ON" or reseeds IDENTITY. For details, see CREATE TABLE (Transact-SQL) IDENTITY (Property).

Creating a table with an IDENTITY column

The IDENTITY property is designed to scale out across all the distributions in the dedicated SQL pool without affecting load performance. Therefore, the implementation of IDENTITY is oriented toward achieving these goals.

You can define a table as having the IDENTITY property when you first create the table by using syntax that is similar to the following statement:

```
SQL
CREATE TABLE dbo.T1
( C1 INT IDENTITY(1,1) NOT NULL
, C2 INT NULL
)
WITH
( DISTRIBUTION = HASH(C2)
, CLUSTERED COLUMNSTORE INDEX
)
;
```

In the preceding example, two rows landed in distribution 1. The first row has the surrogate value of 1 in column C1, and the second row has the surrogate value of 61. Both of these values were generated by the IDENTITY property. However, the allocation of the values is not contiguous. This behavior is by design.

Skewed data

The range of values for the data type are spread evenly across the distributions. If a distributed table suffers from skewed data, then the range of values available to the datatype can be exhausted prematurely. For example, if all the data ends up in a single distribution, then effectively the table has access to only one-sixtieth of the values of the data type. For this reason, the IDENTITY property is limited to INT and BIGINT data types only.

SELECT..INTO

When an existing IDENTITY column is selected into a new table, the new column inherits the IDENTITY property, unless one of the following conditions is true:

- The SELECT statement contains a join.
- Multiple SELECT statements are joined by using UNION.
- The IDENTITY column is listed more than one time in the SELECT list.
- The IDENTITY column is part of an expression.

If any one of these conditions is true, the column is created NOT NULL instead of inheriting the IDENTITY property.

CREATE TABLE AS SELECT

CREATE TABLE AS SELECT (CTAS) follows the same SQL Server behavior that's documented for SELECT..INTO. However, you can't specify an IDENTITY property in the column definition of the CREATE TABLE part of the statement. You also can't use the IDENTITY function in the SELECT part of the CTAS. To populate a table, you need to use CREATE TABLE to define the table followed by INSERT..SELECT to populate it.

Explicitly inserting values into an IDENTITY column

Dedicated SQL pool supports SET IDENTITY_INSERT ON|OFF syntax. You can use this syntax to explicitly insert values into the IDENTITY column.

Many data modelers like to use predefined negative values for certain rows in their dimensions. An example is the -1 or "unknown member" row.

The next script shows how to explicitly add this row by using SET IDENTITY_INSERT:

```
SQL
SET IDENTITY_INSERT dbo.T1 ON;
INSERT INTO dbo.T1
( C1
, C2
)
VALUES (-1,'UNKNOWN')
;
SET IDENTITY_INSERT dbo.T1 OFF;
SELECT *
FROM dbo.T1
;
```

The presence of the IDENTITY property has some implications to your data-loading code. This section highlights some basic patterns for loading data into tables by using IDENTITY.

To load data into a table and generate a surrogate key by using IDENTITY, create the table and then use INSERT..SELECT or INSERT..VALUES to perform the load.

The following example highlights the basic pattern:

SQL

```
--CREATE TABLE with IDENTITY
CREATE TABLE dbo.T1
( C1 INT IDENTITY(1,1)
, C2 VARCHAR(30)
)
WITH
( DISTRIBUTION = HASH(C2)
, CLUSTERED COLUMNSTORE INDEX
)
;
--Use INSERT..SELECT to populate the table from an external table
INSERT INTO dbo.T1
(C2)
SELECT C2
FROM ext.T1
;
SELECT *
FROM dbo.T1
;
DBCC PDW_SHOWSPACEUSED('dbo.T1');
```

Note: It's not possible to use CREATE TABLE AS SELECT currently when loading data into a table with an IDENTITY column.

System views

You can use the sys.identity_columns catalog view to identify a column that has the IDENTITY property.

To help you better understand the database schema, this example shows how to integrate sys.identity_column` with other system catalog views:

SQL

```
SELECT sm.name
, tb.name
, co.name
, CASE WHEN ic.column_id IS NOT NULL
```

```
THEN 1
ELSE 0
END AS is_identity
FROM sys.schemas AS sm
JOIN sys.tables AS tb ON sm.schema_id = tb.schema_id
JOIN sys.columns AS co ON tb.object_id = co.object_id
LEFT JOIN sys.identity_columns AS ic ON co.object_id = ic.object_id
AND co.column_id = ic.column_id
WHERE sm.name = 'dbo'
AND tb.name = 'T1'
;
```

Limitations

The IDENTITY property can't be used:

- When the column data type is not INT or BIGINT
- When the column is also the distribution key
- When the table is an external table

The following related functions are not supported in dedicated SQL pool:

- IDENTITY()
- @@IDENTITY
- SCOPE_IDENTITY
- IDENT_CURRENT
- IDENT_INCR
- IDENT_SEED

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

- ○ A table that has a FOREIGN KEY constraint.
- ○ A system-versioned temporal table.
- ○ A user-defined SEQUENCE object.

Report Error

Q. 53

There are a range of network security steps that you should consider to secure Azure Synapse Analytics. One of the first aspects that you will consider is securing access to the service itself.

This can be achieved by creating the following network objects including:

- Firewall rules
- Virtual networks
- Private endpoints

Which of the following are benefits of using a managed workspace virtual network? (Select all that apply)

- ☑ It ensures that your workspace is a consolidated network with your other workspaces.
- ☐ You don't need to create a subnet for your Spark clusters based on peak load.
- ☑ With a Managed workspace Virtual Network, you can offload the burden of managing the Virtual Network to Azure Synapse.
- ☑ Managed workspace Virtual Network along with Managed private endpoints protects against data exfiltration.
- ☐ You don't have to configure inbound NSG rules on your own Virtual Networks to allow Azure Synapse management traffic to enter your Virtual Network.

Report Error

- [ ] It ensures that your workspace is a consolidated network with your other workspaces.
- [x] With a Managed workspace Virtual Network, you can offload the burden of managing the Virtual Network to Azure Synapse.

Explanation:-

There are a range of network security steps that you should consider to secure Azure Synapse Analytics. One of the first aspects that you will consider is securing access to the service itself. This can be achieved by creating the following network objects including:

• Firewall rules
• Virtual networks
• Private endpoints

Firewall rules

Firewall rules enable you to define the type of traffic that is allowed or denied access to an Azure Synapse workspace using the originating IP address of the client that is trying to access the Azure Synapse Workspace. IP firewall rules configured at the workspace level apply to all public endpoints of the workspace including dedicated SQL pools, serverless SQL pool, and the development endpoint.

You can choose to allow connections from all IP addresses as you are creating the Azure Synapse Workspaces, although this is not recommended as it does not allow for control access to the workspace. Instead, within the Azure portal, you can configure specific IP address ranges and associate them with a rule name so that you have greater control.

Make sure that the firewall on your network and local computer allows outgoing communication on TCP ports 80, 443 and 1443 for Synapse Studio.

Also, you need to allow outgoing communication on UDP port 53 for Synapse Studio. To connect using tools such as SSMS and Power BI, you must allow outgoing communication on TCP port 1433.

https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-ip-firewall

Virtual networks

Azure Virtual Network (VNet) enables private networks in Azure. VNet enables many types of Azure resources, such as Azure Synapse Analytics, to securely communicate with other virtual networks, the internet, and on-premises networks. When you create your Azure Synapse workspace, you can choose to associate it to a Microsoft Azure Virtual Network. The Virtual Network associated with your workspace is managed by Azure Synapse. This Virtual Network is called a Managed workspace Virtual Network.

Using a managed workspace virtual network provides the following benefits:

• With a Managed workspace Virtual Network, you can offload the burden of managing the Virtual Network to Azure Synapse.
• You don't have to configure inbound NSG rules on your own Virtual Networks to allow Azure Synapse management traffic to enter your Virtual Network. Misconfiguration of these NSG rules causes service disruption for customers.
• You don't need to create a subnet for your Spark clusters based on peak load.
• Managed workspace Virtual Network along with Managed private endpoints protects against data exfiltration. You can only create Managed private endpoints in a workspace that has a Managed workspace Virtual Network associated with it.
• It ensures that your workspace is network isolated from other workspaces.

If your workspace has a Managed workspace Virtual Network, Data integration and Spark resources are deployed in it. A Managed workspace Virtual Network also provides user-level isolation for Spark activities because each Spark cluster is in its own subnet.

Dedicated SQL pool and serverless SQL pool are multi-tenant capabilities and therefore reside outside of the Managed workspace Virtual Network. Intra-workspace communication to dedicated SQL pool and serverless SQL pool use Azure private links. These private links are automatically created for you when you create a workspace with a Managed workspace Virtual Network associated to it.

You can only choose to enable managed virtual networks as you are creating the Azure Synapse Workspaces.

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

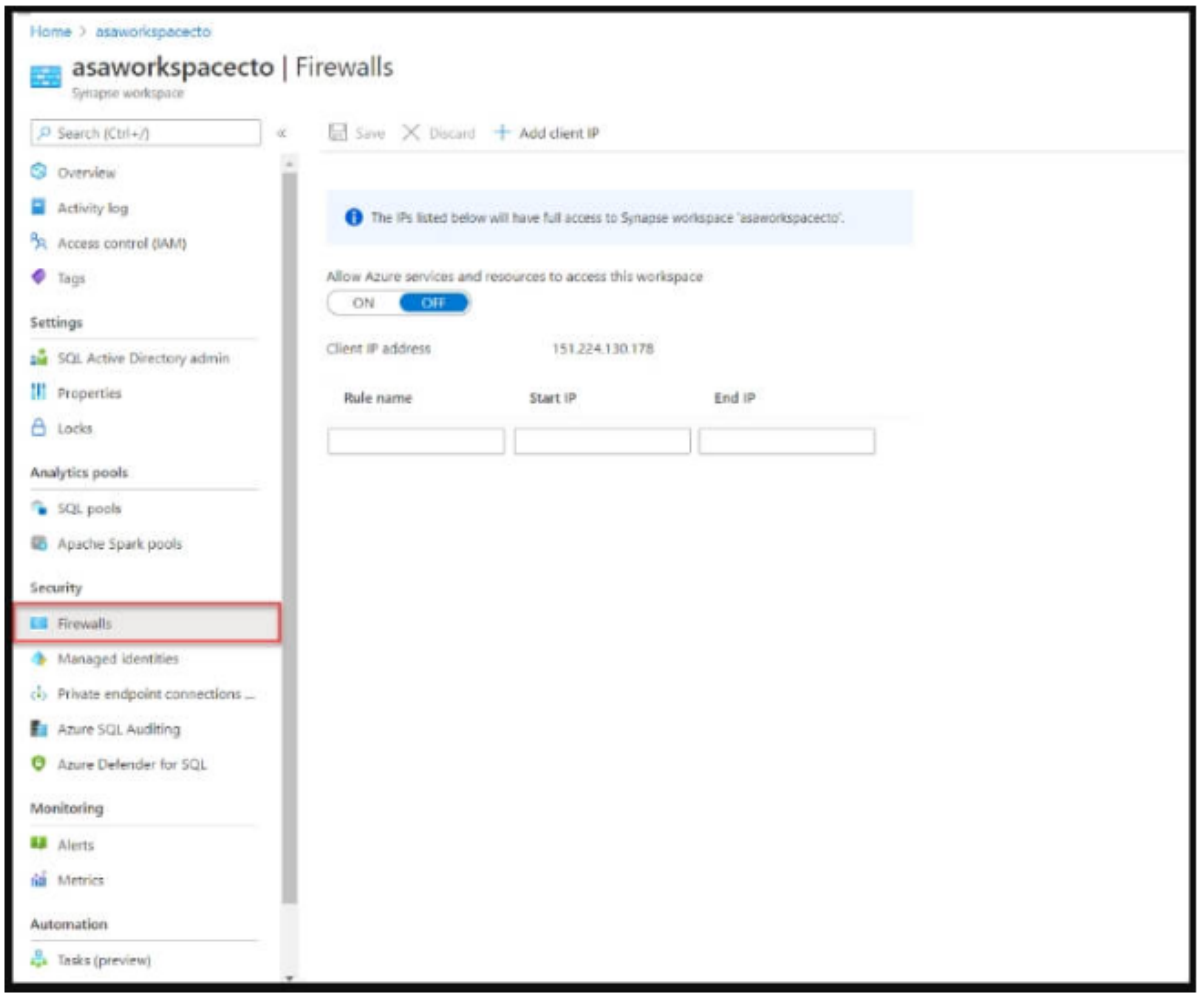
Private endpoints

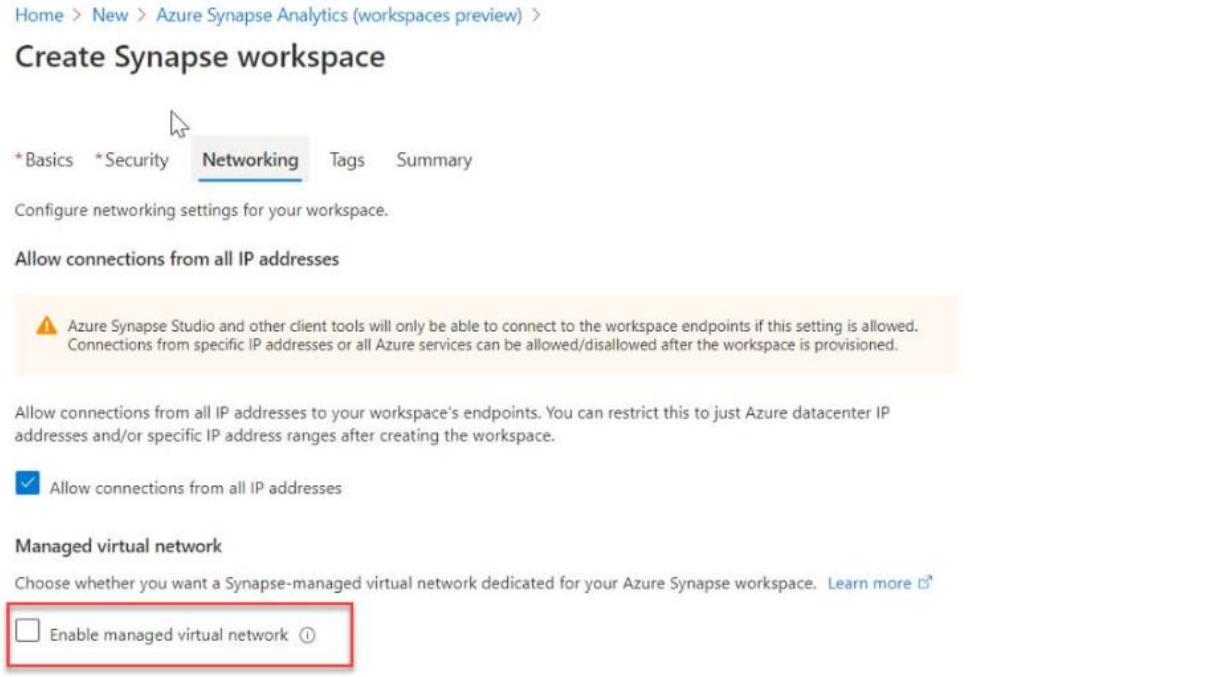
Azure Synapse Analytics enables you to connect upto its various components through endpoints. You can set up managed private endpoints to access these components in a secure manner known as private links. This can only be achieved in an Azure Synapse workspace with a Managed workspace Virtual Network. Private link enables you to access Azure services (such as Azure Storage and Azure Cosmos DB) and Azure hosted customer/partner services from your Azure Virtual Network securely.

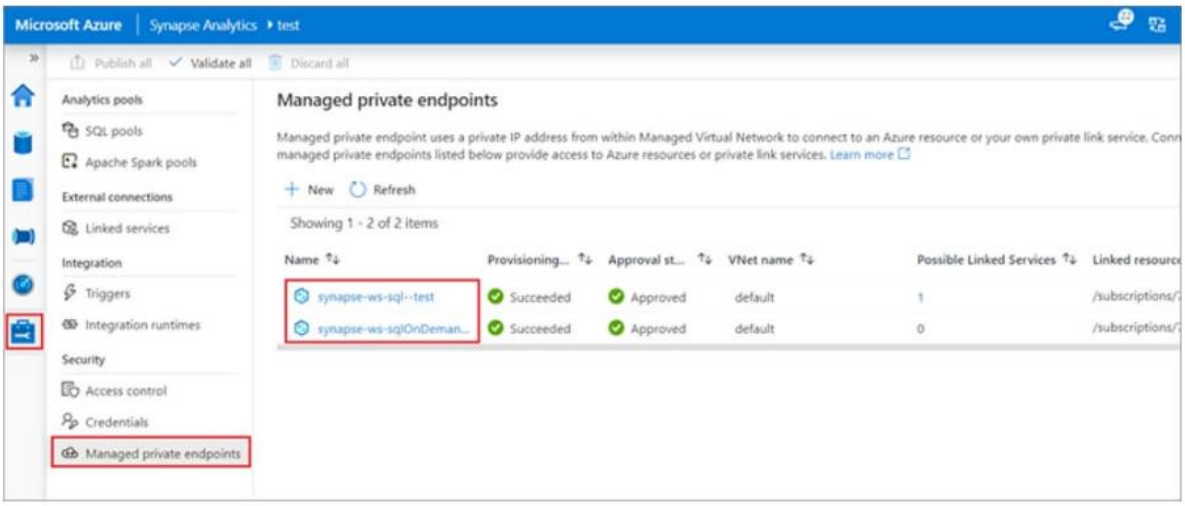
When you use a private link, traffic between your Virtual Network and workspace traverses entirely over the Microsoft backbone network. Private Link protects against data exfiltration risks. You establish a private link to a resource by creating a private endpoint.

Private endpoint uses a private IP address from your Virtual Network to effectively bring the service into your Virtual Network. Private endpoints are mapped to a specific resource in Azure and not the entire service. Customers can limit connectivity to a specific resource approved by their organization.

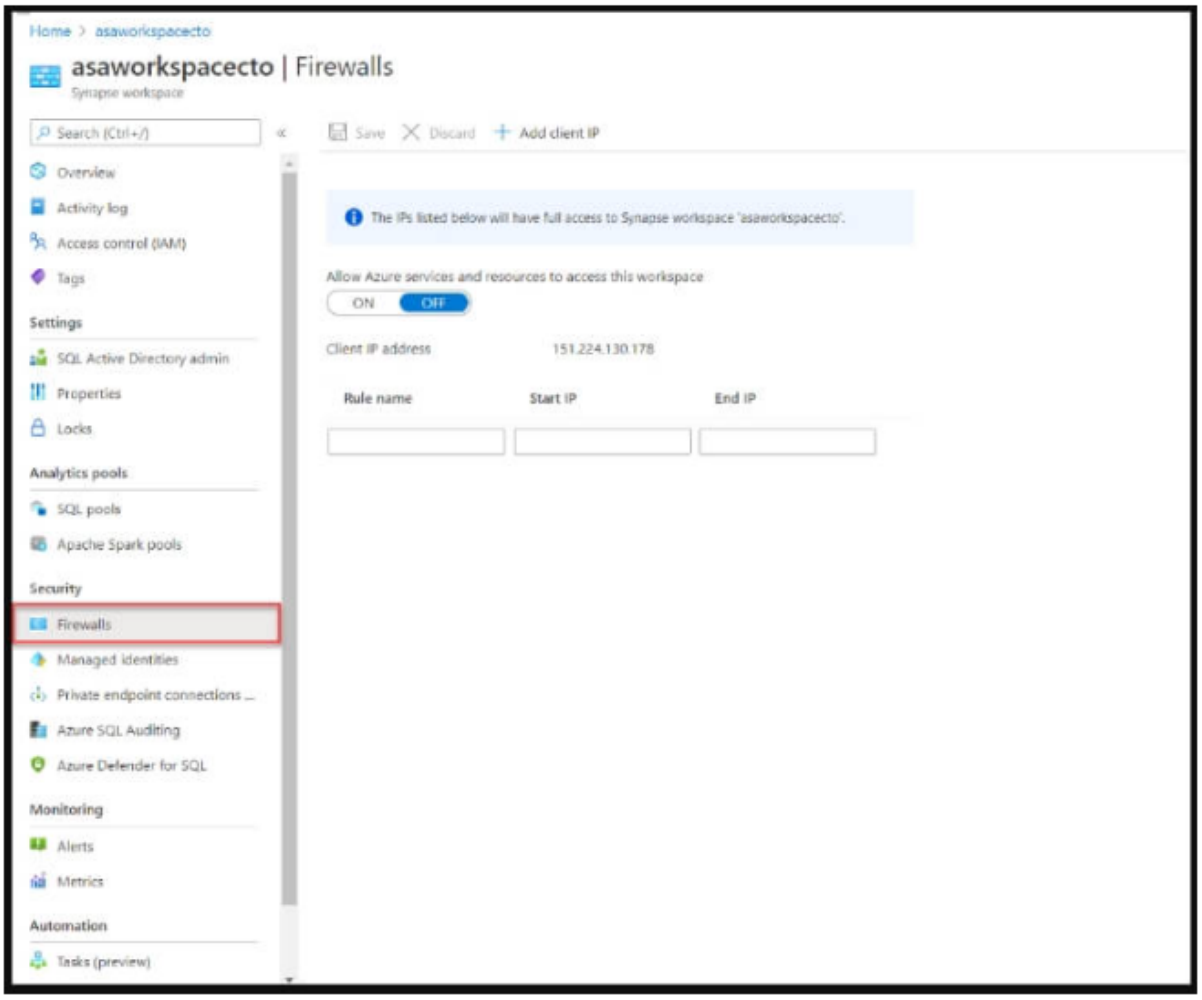
You can manage the private endpoints in the Azure Synapse Studio manage hub.


Home > asaworkspacecto
asaworkspacecto | Firewalls
Synapse workspace
Search (Ctrl+/)
Save
Discard
Add client IP
Overview
Activity log
Access control (IAM)
Tags
Settings
SQL Active Directory admin
Properties
Locks
Analytics pools
SQL pools
Apache Spark pools
Security
Firewalls
Managed identities
Private endpoint connections ...
Azure SQL Auditing
Azure Defender for SQL
Monitoring
Alerts
Metrics
Automation
Tasks (preview)
The IPs listed below will have full access to Synapse workspace 'asaworkspacecto'.
Allow Azure services and resources to access this workspace
ON
OFF
Client IP address
151.224.130.178
Rule name
Start IP
End IP





https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-private-endpoints


Home > asaworkspacecto
asaworkspacecto | Firewalls
Synapse workspace
Search (Ctrl+/)
Save
Discard
Add client IP
Overview
Activity log
Access control (IAM)
Tags
Settings
SQL Active Directory admin
Properties
Locks
Analytics pools
SQL pools
Apache Spark pools
Security
Firewalls
Managed identities
Private endpoint connections ...
Azure SQL Auditing
Azure Defender for SQL
Monitoring
Alerts
Metrics
Automation
Tasks (preview)
The IPs listed below will have full access to Synapse workspace 'asaworkspacecto'.
Allow Azure services and resources to access this workspace
ON
OFF
Client IP address
151.224.130.178
Rule name
Start IP
End IP

Home > New > Azure Synapse Analytics (workspaces preview) >

Create Synapse workspace

*Basics *Security Networking Tags Summary

Configure networking settings for your workspace.

Allow connections from all IP addresses

⚠ Azure Synapse Studio and other client tools will only be able to connect to the workspace endpoints if this setting is allowed. Connections from specific IP addresses or all Azure services can be allowed/disallowed after the workspace is provisioned.

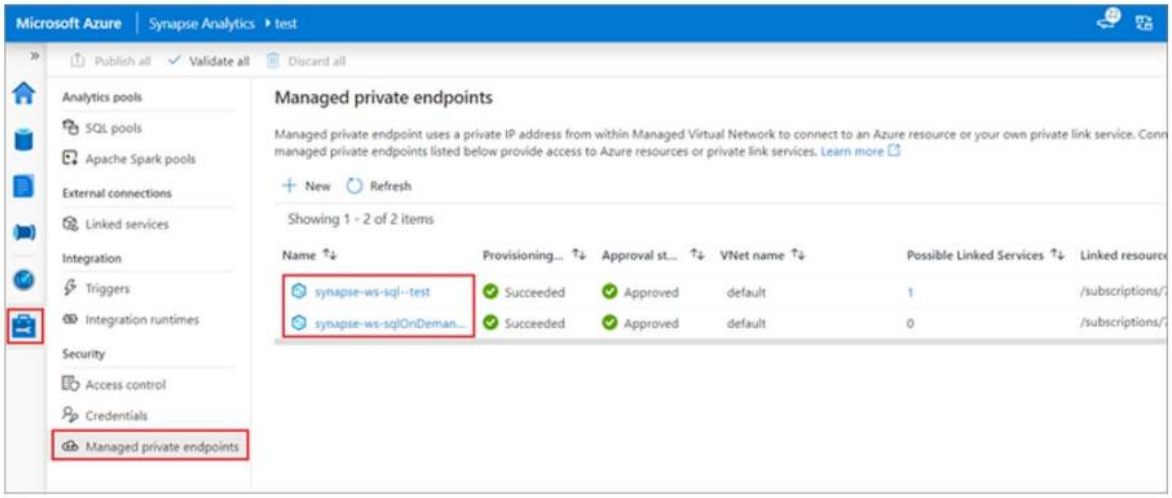
Allow connections from all IP addresses to your workspace's endpoints. You can restrict this to just Azure datacenter IP addresses and/or specific IP address ranges after creating the workspace.

☑ Allow connections from all IP addresses

Managed virtual network

Choose whether you want a Synapse-managed virtual network dedicated for your Azure Synapse workspace. Learn more

☐ Enable managed virtual network ⓘ



☑ You don't have to configure inbound NSG rules on your own Virtual Networks to allow Azure Synapse management traffic to enter your Virtual Network.

Explanation:- There are a range of network security steps that you should consider to secure Azure Synapse Analytics. One of the first aspects that you will consider is securing access to the service itself. This can be achieved by creating the following network objects including:

• Firewall rules
• Virtual networks
• Private endpoints

Firewall rules

Firewall rules enable you to define the type of traffic that is allowed or denied access to an Azure Synapse workspace using the originating IP address of the client that is trying to access the Azure Synapse Workspace. IP firewall rules configured at the workspace level apply to all public endpoints of the workspace including dedicated SQL pools, serverless SQL pool, and the development endpoint.

You can choose to allow connections from all IP addresses as you are creating the Azure Synapse Workspaces, although this is not recommended as it does not allow for control access to the workspace. Instead, within the Azure portal, you can configure specific IP address ranges and associate them with a rule name so that you have greater control.

Make sure that the firewall on your network and local computer allows outgoing communication on TCP ports 80, 443 and 1443 for Synapse Studio.

Also, you need to allow outgoing communication on UDP port 53 for Synapse Studio. To connect using tools such as SSMS and Power BI, you must allow outgoing communication on TCP port 1433.

https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-ip-firewall

Virtual networks

Azure Virtual Network (VNet) enables private networks in Azure. VNet enables many types of Azure resources, such as Azure Synapse Analytics, to securely communicate with other virtual networks, the internet, and on-premises networks. When you create your Azure Synapse workspace, you can choose to associate it to a Microsoft Azure Virtual Network. The Virtual Network associated with your workspace is managed by Azure Synapse. This Virtual Network is called a Managed workspace Virtual Network.

Using a managed workspace virtual network provides the following benefits:

• With a Managed workspace Virtual Network, you can offload the burden of managing the Virtual Network to Azure Synapse.
• You don't have to configure inbound NSG rules on your own Virtual Networks to allow Azure Synapse management traffic to enter your Virtual Network. Misconfiguration of these NSG rules causes service disruption for customers.
• You don't need to create a subnet for your Spark clusters based on peak load.
• Managed workspace Virtual Network along with Managed private endpoints protects against data exfiltration. You can only create Managed private endpoints in a workspace that has a Managed workspace Virtual Network associated with it.
• It ensures that your workspace is network isolated from other workspaces.

If your workspace has a Managed workspace Virtual Network, Data integration and Spark resources are deployed in it. A Managed workspace Virtual Network also provides user-level isolation for Spark activities because each Spark cluster is in its own subnet.

Dedicated SQL pool and serverless SQL pool are multi-tenant capabilities and therefore reside outside of the Managed workspace Virtual Network. Intra-workspace communication to dedicated SQL pool and serverless SQL pool use Azure private links. These private links are automatically created for you when you create a workspace with a Managed workspace Virtual Network associated to it.
You can only choose to enable managed virtual networks as you are creating the Azure Synapse Workspaces.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-vne
Private endpoints
Azure Synapse Analytics enables you to connect upto its various components through endpoints. You can set up managed private endpoints to access these components in a secure manner known as private links. This can only be achieved in an Azure Synapse workspace with a Managed workspace Virtual Network. Private link enables you to access Azure services (such as Azure Storage and Azure Cosmos DB) and Azure hosted customer/partner services from your Azure Virtual Network securely.
When you use a private link, traffic between your Virtual Network and workspace traverses entirely over the Microsoft backbone network. Private Link protects against data exfiltration risks. You establish a private link to a resource by creating a private endpoint.
Private endpoint uses a private IP address from your Virtual Network to effectively bring the service into your Virtual Network. Private endpoints are mapped to a specific resource in Azure and not the entire service. Customers can limit connectivity to a specific resource approved by their organization.
You can manage the private endpoints in the Azure Synapse Studio manage hub.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-private-endpoints

- [x] You don't need to create a subnet for your Spark clusters based on peak load.

Explanation:- There are a range of network security steps that you should consider to secure Azure Synapse Analytics. One of the first aspects that you will consider is securing access to the service itself. This can be achieved by creating the following network objects including:
• Firewall rules
• Virtual networks
• Private endpoints
Firewall rules
Firewall rules enable you to define the type of traffic that is allowed or denied access to an Azure Synapse workspace using the originating IP address of the client that is trying to access the Azure Synapse Workspace. IP firewall rules configured at the workspace level apply to all public endpoints of the workspace including dedicated SQL pools, serverless SQL pool, and the development endpoint.
You can choose to allow connections from all IP addresses as you are creating the Azure Synapse Workspaces, although this is not recommended as it does not allow for control access to the workspace. Instead, within the Azure portal, you can configure specific IP address ranges and associate them with a rule name so that you have greater control.
Make sure that the firewall on your network and local computer allows outgoing communication on TCP ports 80, 443 and 1443 for Synapse Studio.
Also, you need to allow outgoing communication on UDP port 53 for Synapse Studio. To connect using tools such as SSMS and Power BI, you must allow outgoing communication on TCP port 1433.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-ip-firewall
Virtual networks

Azure Virtual Network (VNet) enables private networks in Azure. VNet enables many types of Azure resources, such as Azure Synapse Analytics, to securely communicate with other virtual networks, the internet, and on-premises networks. When you create your Azure Synapse workspace, you can choose to associate it to a Microsoft Azure Virtual Network. The Virtual Network associated with your workspace is managed by Azure Synapse. This Virtual Network is called a Managed workspace Virtual Network.
Using a managed workspace virtual network provides the following benefits:
• With a Managed workspace Virtual Network, you can offload the burden of managing the Virtual Network to Azure Synapse.
• You don't have to configure inbound NSG rules on your own Virtual Networks to allow Azure Synapse management traffic to enter your Virtual Network. Misconfiguration of these NSG rules causes service disruption for customers.
• You don't need to create a subnet for your Spark clusters based on peak load.
• Managed workspace Virtual Network along with Managed private endpoints protects against data exfiltration. You can only create Managed private endpoints in a workspace that has a Managed workspace Virtual Network associated with it.
• It ensures that your workspace is network isolated from other workspaces.
If your workspace has a Managed workspace Virtual Network, Data integration and Spark resources are deployed in it. A Managed workspace Virtual Network also provides user-level isolation for Spark activities because each Spark cluster is in its own subnet.
Dedicated SQL pool and serverless SQL pool are multi-tenant capabilities and therefore reside outside of the Managed workspace Virtual Network. Intra-workspace communication to dedicated SQL pool and serverless SQL pool use Azure private links. These private links are automatically created for you when you create a workspace with a Managed workspace Virtual Network associated to it.
You can only choose to enable managed virtual networks as you are creating the Azure Synapse Workspaces.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-vne
Private endpoints
Azure Synapse Analytics enables you to connect upto its various components through endpoints. You can set up managed private endpoints to access these components in a secure manner known as private links. This can only be achieved in an Azure Synapse workspace with a Managed workspace Virtual Network. Private link enables you to access Azure services (such as Azure Storage and Azure Cosmos DB) and Azure hosted customer/partner services from your Azure Virtual Network securely.
When you use a private link, traffic between your Virtual Network and workspace traverses entirely over the Microsoft backbone network. Private Link protects against data exfiltration risks. You establish a private link to a resource by creating a private endpoint.
Private endpoint uses a private IP address from your Virtual Network to effectively bring the service into your Virtual Network. Private endpoints are mapped to a specific resource in Azure and not the entire service. Customers can limit connectivity to a specific resource approved by their organization.
You can manage the private endpoints in the Azure Synapse Studio manage hub.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-private-endpoints

☑ Managed workspace Virtual Network along with Managed private endpoints protects against data exfiltration.

Explanation:- There are a range of network security steps that you should consider to secure Azure Synapse Analytics. One of the first aspects that you will consider is securing access to the service itself. This can be achieved by creating the following network objects including:
• Firewall rules
• Virtual networks
• Private endpoints
Firewall rules
Firewall rules enable you to define the type of traffic that is allowed or denied access to an Azure Synapse workspace using the originating IP address of the client that is trying to access the Azure Synapse Workspace. IP firewall rules configured at the workspace level apply to all public endpoints of the workspace including dedicated SQL pools, serverless SQL pool, and the development endpoint.
You can choose to allow connections from all IP addresses as you are creating the Azure Synapse Workspaces, although this is not recommended as it does not allow for control access to the workspace. Instead, within the Azure portal, you can configure specific IP address ranges and associate them with a rule name so that you have greater control.
Make sure that the firewall on your network and local computer allows outgoing communication on TCP ports 80, 443 and 1443 for Synapse Studio.
Also, you need to allow outgoing communication on UDP port 53 for Synapse Studio. To connect using tools such as SSMS and Power BI, you must allow outgoing communication on TCP port 1433.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-ip-firewall
Virtual networks
Azure Virtual Network (VNet) enables private networks in Azure. VNet enables many types of Azure resources, such as Azure Synapse Analytics, to securely communicate with other virtual networks, the internet, and on-premises networks. When you create your Azure Synapse workspace, you can choose to associate it to a Microsoft Azure Virtual Network. The Virtual Network associated with your workspace is managed by Azure Synapse. This Virtual Network is called a Managed workspace Virtual Network.
Using a managed workspace virtual network provides the following benefits:
• With a Managed workspace Virtual Network, you can offload the burden of managing the Virtual Network to Azure Synapse.
• You don't have to configure inbound NSG rules on your own Virtual Networks to allow Azure Synapse management traffic to enter your Virtual Network. Misconfiguration of these NSG rules causes service disruption for customers.
• You don't need to create a subnet for your Spark clusters based on peak load.
• Managed workspace Virtual Network along with Managed private endpoints protects against data exfiltration. You can only create Managed private endpoints in a workspace that has a Managed workspace Virtual Network associated with it.
• It ensures that your workspace is network isolated from other workspaces.
If your workspace has a Managed workspace Virtual Network, Data integration and Spark resources are deployed in it. A Managed workspace Virtual Network also provides user-level isolation for Spark activities because each Spark cluster is in its own subnet.

Dedicated SQL pool and serverless SQL pool are multi-tenant capabilities and therefore reside outside of the Managed workspace Virtual Network. Intra-workspace communication to dedicated SQL pool and serverless SQL pool use Azure private links. These private links are automatically created for you when you create a workspace with a Managed workspace Virtual Network associated to it.
You can only choose to enable managed virtual networks as you are creating the Azure Synapse Workspaces.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-vne
Private endpoints
Azure Synapse Analytics enables you to connect upto its various components through endpoints. You can set up managed private endpoints to access these components in a secure manner known as private links. This can only be achieved in an Azure Synapse workspace with a Managed workspace Virtual Network. Private link enables you to access Azure services (such as Azure Storage and Azure Cosmos DB) and Azure hosted customer/partner services from your Azure Virtual Network securely.
When you use a private link, traffic between your Virtual Network and workspace traverses entirely over the Microsoft backbone network. Private Link protects against data exfiltration risks. You establish a private link to a resource by creating a private endpoint.
Private endpoint uses a private IP address from your Virtual Network to effectively bring the service into your Virtual Network. Private endpoints are mapped to a specific resource in Azure and not the entire service. Customers can limit connectivity to a specific resource approved by their organization.
You can manage the private endpoints in the Azure Synapse Studio manage hub.
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-private-endpoints

Report Error

Q. 54

As great as data lakes are at inexpensively storing our raw data, they also bring with them performance challenges:

• Too many small or very big files - more time opening & closing files rather than reading contents (worse with streaming).
• Partitioning also known as "poor man's indexing"- breaks down if you picked the wrong fields or when data has many dimensions, high cardinality columns.
• No caching - cloud storage throughput is low (cloud object storage is 20-50MB/s/core vs 300MB/s/core for local SSDs).

As a solution to the challenges with Data Lakes noted above, [?] is a file format that can help you build a data lake comprised of one or many tables in [?] format. [?] integrates tightly with Apache Spark, and uses an open format that is based on Parquet. Because it is an open-source format, [?] is also supported by other data platforms, including Azure Synapse Analytics.

- ○ Delta Lake
- ◉ Data Organizer
- ○ Data Sea
- ○ Augmenter

Report Error

◉ Delta Lake

Explanation:-

Delta Lake is a transactional storage layer designed specifically to work with Apache Spark and Databricks File System (DBFS). At the core of Delta Lake is an optimized Spark table. It stores your data as Apache Parquet files in DBFS and maintains a transaction log that efficiently tracks changes to the table.

Data lakes

A data lake is a storage repository that inexpensively stores a vast amount of raw data, both current and historical, in native formats such as XML, JSON, CSV, and Parquet. It may contain operational relational databases with live transactional data.

Enterprises have been spending millions of dollars getting data into data lakes with Apache Spark. The aspiration is to do data science and ML on all that data using Apache Spark.

But the data is not ready for data science & ML. The majority of these projects are failing due to unreliable data!

The challenge with data lakes

Why are these projects struggling with reliability and performance?

To extract meaningful information from a data lake, you must solve problems such as:

• Schema enforcement when new tables are introduced.
• Table repairs when any new data is inserted into the data lake.
• Frequent refreshes of metadata.
• Bottlenecks of small file sizes for distributed computations.
• Difficulty sorting data by an index if data is spread across many files and partitioned.

There are also data reliability challenges with data lakes:

• Failed production jobs leave data in corrupt state requiring tedious recovery.
• Lack of schema enforcement creates inconsistent and low quality data.
• Lack of consistency makes it almost impossible to mix appends and reads, batch and streaming.

As great as data lakes are at inexpensively storing our raw data, they also bring with them performance challenges:

• Too many small or very big files - more time opening & closing files rather than reading contents (worse with streaming).
• Partitioning also known as "poor man's indexing"- breaks down if you picked the wrong fields or when data has many dimensions, high cardinality columns.
• No caching - cloud storage throughput is low (cloud object storage is 20-50MB/s/core vs 300MB/s/core for local SSDs).

The solution: Delta Lake

Delta Lake is a file format that can help you build a data lake comprised of one or many tables in Delta Lake format. Delta Lake integrates tightly with Apache Spark, and uses an open format that is based on Parquet. Because it is an open-source format, Delta Lake is also supported by other data platforms, including Azure Synapse Analytics.

Delta Lake makes data ready for analytics.

Delta Lake is an open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads.

You can read and write data that's stored in Delta Lake by using Apache Spark SQL batch and streaming APIs. These are the same familiar APIs that you use to work with Hive tables or DBFS directories. Delta Lake provides the following functionality:

ACID Transactions: Data lakes typically have multiple data pipelines reading and writing data concurrently, and data engineers have to go through a tedious process to ensure data integrity, due to the lack of transactions. Delta Lake brings ACID transactions to your data lakes. It provides serializability, the strongest level of isolation level.
Scalable Metadata Handling: In big data, even the metadata itself can be "big data". Delta Lake treats metadata just like data, leveraging Spark's distributed processing power to handle all its metadata. As a result, Delta Lake can handle petabyte-scale tables with billions of partitions and files at ease.
Time Travel (data versioning): Delta Lake provides snapshots of data enabling developers to access and revert to earlier versions of data for audits, rollbacks or to reproduce experiments.
Open Format: All data in Delta Lake is stored in Apache Parquet format enabling Delta Lake to leverage the efficient compression and encoding schemes that are native to Parquet.
Unified Batch and Streaming Source and Sink: A table in Delta Lake is both a batch table, as well as a streaming source and sink. Streaming data ingest, batch historic backfill, and interactive queries all just work out of the box.
Schema Enforcement: Delta Lake provides the ability to specify your schema and enforce it. This helps ensure that the data types are correct and required columns are present, preventing bad data from causing data corruption.
Schema Evolution: Big data is continuously changing. Delta Lake enables you to make changes to a table schema that can be applied automatically, without the need for cumbersome DDL.
100% Compatible with Apache Spark API: Developers can use Delta Lake with their existing data pipelines with minimal change as it is fully compatible with Spark, the commonly used big data processing engine.
Get started with Delta using Spark APIs
Delta Lake is included with Azure Databricks. You can start using it today. To quickly get started with Delta Lake, do the following:
Instead of parquet...
Python
```
CREATE TABLE ...
USING parquet
...
dataframe
.write
.format("parquet")
.save("/data")
```
... simply say delta
Python
```
CREATE TABLE ...
USING delta
...
dataframe
.write
.format("delta")
.save("/data")
```
Using Delta with your existing Parquet tables

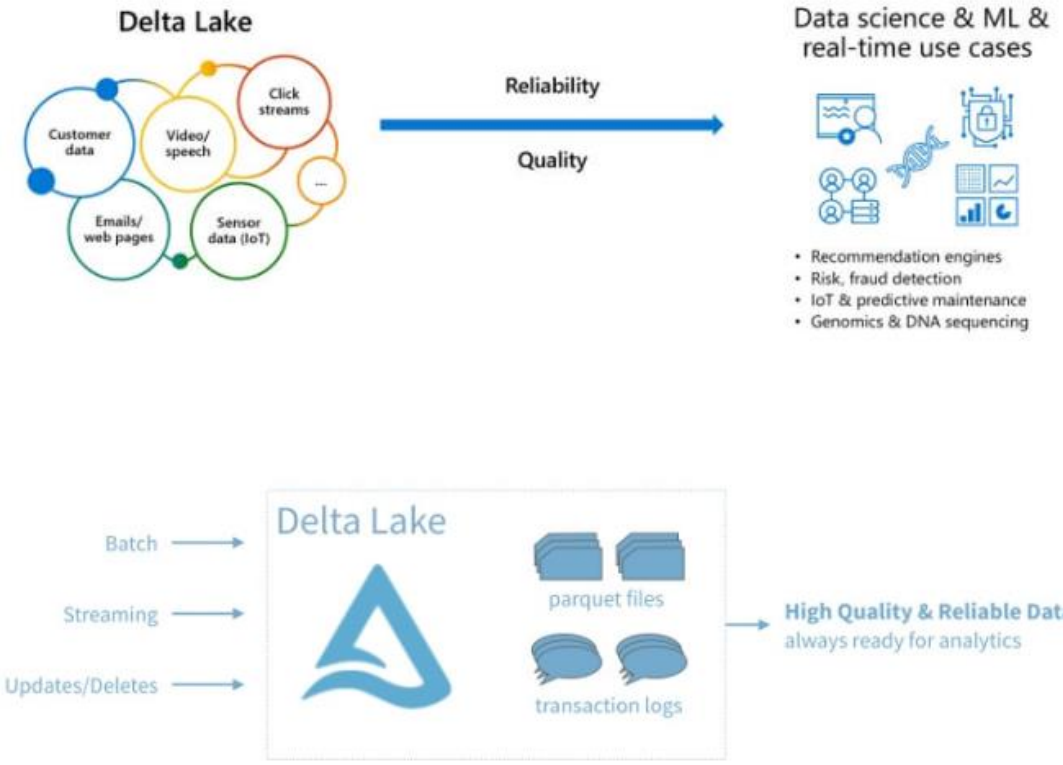
Step 1: Convert Parquet to Delta tables:
Python
CONVERT TO DELTA parquet.`path/to/table` [NO STATISTICS]
[PARTITIONED BY (col_name1 col_type1, col_name2 col_type2, ...)]

Step 2: Optimize layout for fast queries:

Python
OPTIMIZE events
WHERE date >= current_timestamp() - INTERVAL 1 day
ZORDER BY (eventType)
Basic syntax
Two of the core features of Delta Lake are performing upserts (insert/updates) and Time Travel operations.
To UPSERT means to "UPdate" and "inSERT". In other words, UPSERT is literally TWO operations. It is not supported in traditional data lakes, as running an UPDATE could invalidate data that is accessed by the subsequent INSERT operation.
Using Delta Lake, however, we can do UPSERTS. Delta Lake combines these operations to guarantee atomicity to
• INSERT a row
• if the row already exists, UPDATE the row.
Upsert syntax
Upserting, or merging, in Delta Lake provides fine-grained updates of your data. The following syntax shows how to perform an Upsert:
SQL
MERGE INTO customers -- Delta table
USING updates
ON customers.customerId = source.customerId
WHEN MATCHED THEN
UPDATE SET address = updates.address
WHEN NOT MATCHED
THEN INSERT (customerId, address) VALUES (updates.customerId, updates.address)
Time Travel syntax
Because Delta Lake is version controlled, you have the option to query past versions of the data. Using a single file storage system, you now have access to several versions your historical data, ensuring that your data analysts will be able to replicate their reports (and compare aggregate changes over time) and your data scientists will be able to replicate their experiments.
Other time travel use cases are:
• Re-creating analyses, reports, or outputs (for example, the output of a machine learning model). This could be useful for debugging or auditing, especially in regulated industries.
• Writing complex temporal queries.
• Fixing mistakes in your data.
• Providing snapshot isolation for a set of queries for fast changing tables.

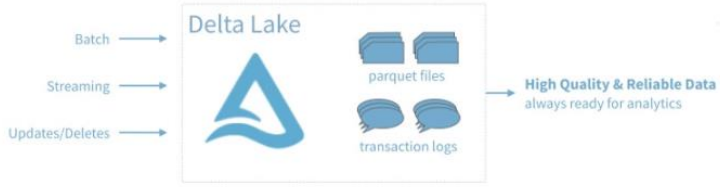
Example of using time travel to reproduce experiments and reports:
SQL
SELECT count(*) FROM events
TIMESTAMP AS OF timestamp
SELECT count(*) FROM events
VERSION AS OF version
Python
spark.read.format("delta").option("timestampAsOf", timestamp_string).load("/events/")
If you need to rollback accidental or bad writes:
SQL
INSERT INTO my_table
SELECT * FROM my_table TIMESTAMP AS OF
date_sub(current_date(), 1)

Delta Lake: Makes data ready for analytics



https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-what-is-delta-lake

Delta Lake: Makes data ready for analytics



- Augmenter
- Data Organizer
- Data Sea

Report Error

Q. 55

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.

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.

Correct or Incorrect : You can write to the default DBFS file storage as needed, but you cannot change the Blob Storage account settings.

- Incorrect
- Correct

Report Error

○ Incorrect

◉ 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.

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.

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.

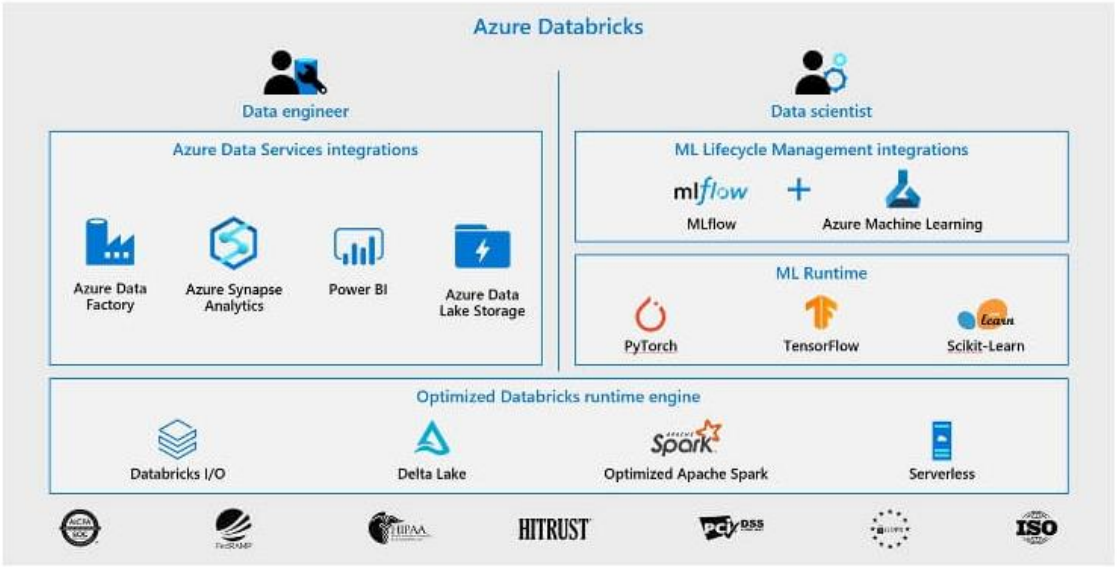
The diagram above shows a standard deployment that contains the boundaries between the Control Plane and the Data Plane with the Azure components deployed to each. At the top of the diagram is the Control Plane that exists within the Microsoft subscription. The customer subscription is at the bottom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

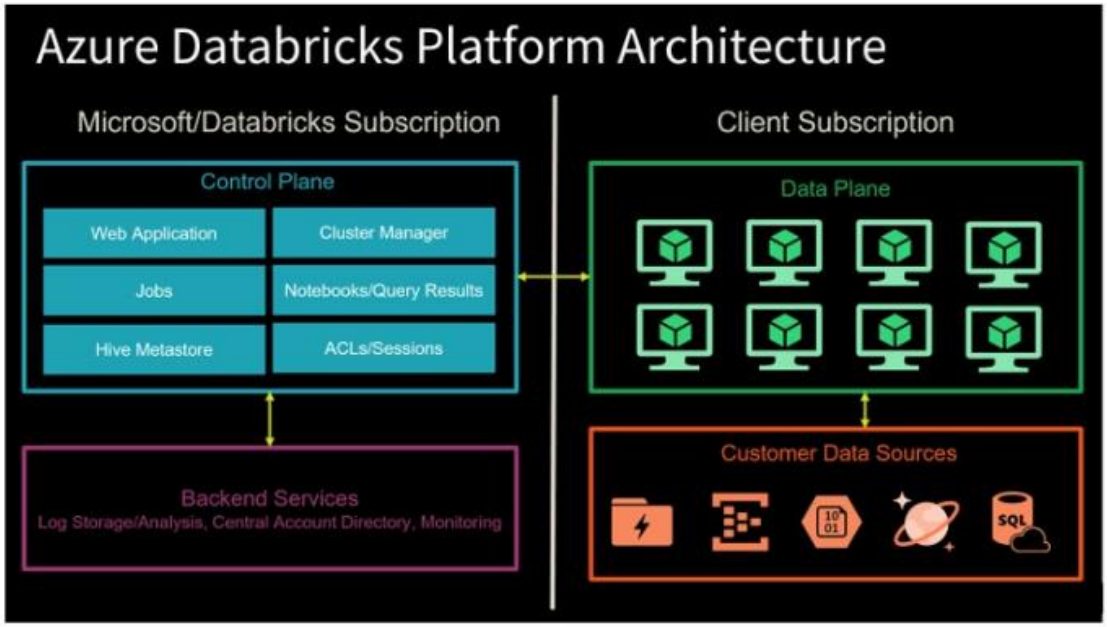
The Blob Storage account provides default file storage within the workspace (databricks file system (DBFS)). This resource and all other Microsoft-managed resources are completely locked from changes made by the customer. All other resources within the customer subscription are customer-managed and can 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.

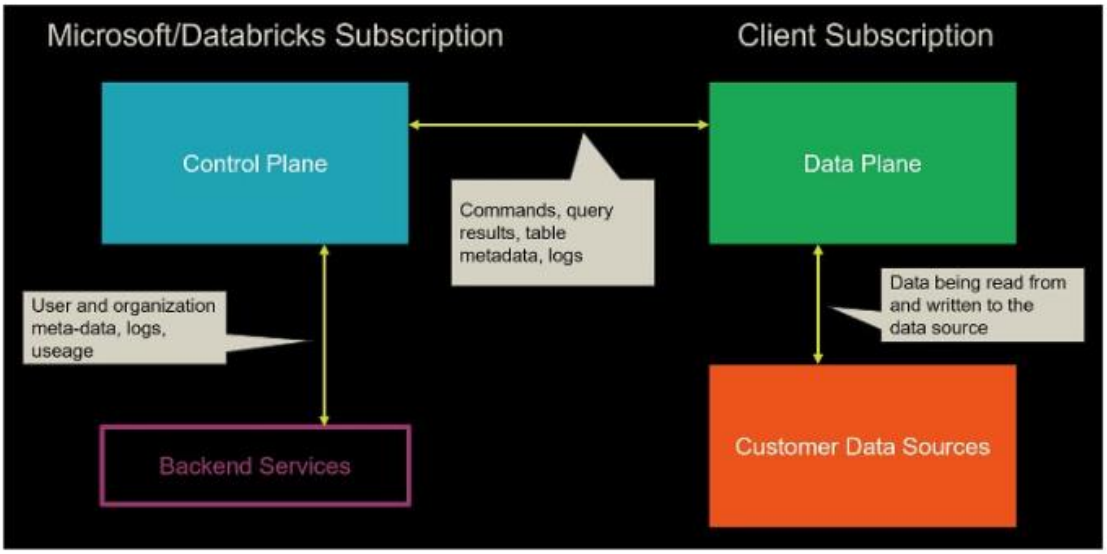
To clarify, you can write to the default DBFS file storage as needed, but you cannot change the Blob Storage account settings since the account is managed by the Microsoft-managed Control Plane. As a best practice, only use the default storage for temporary files and mount additional storage accounts (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.

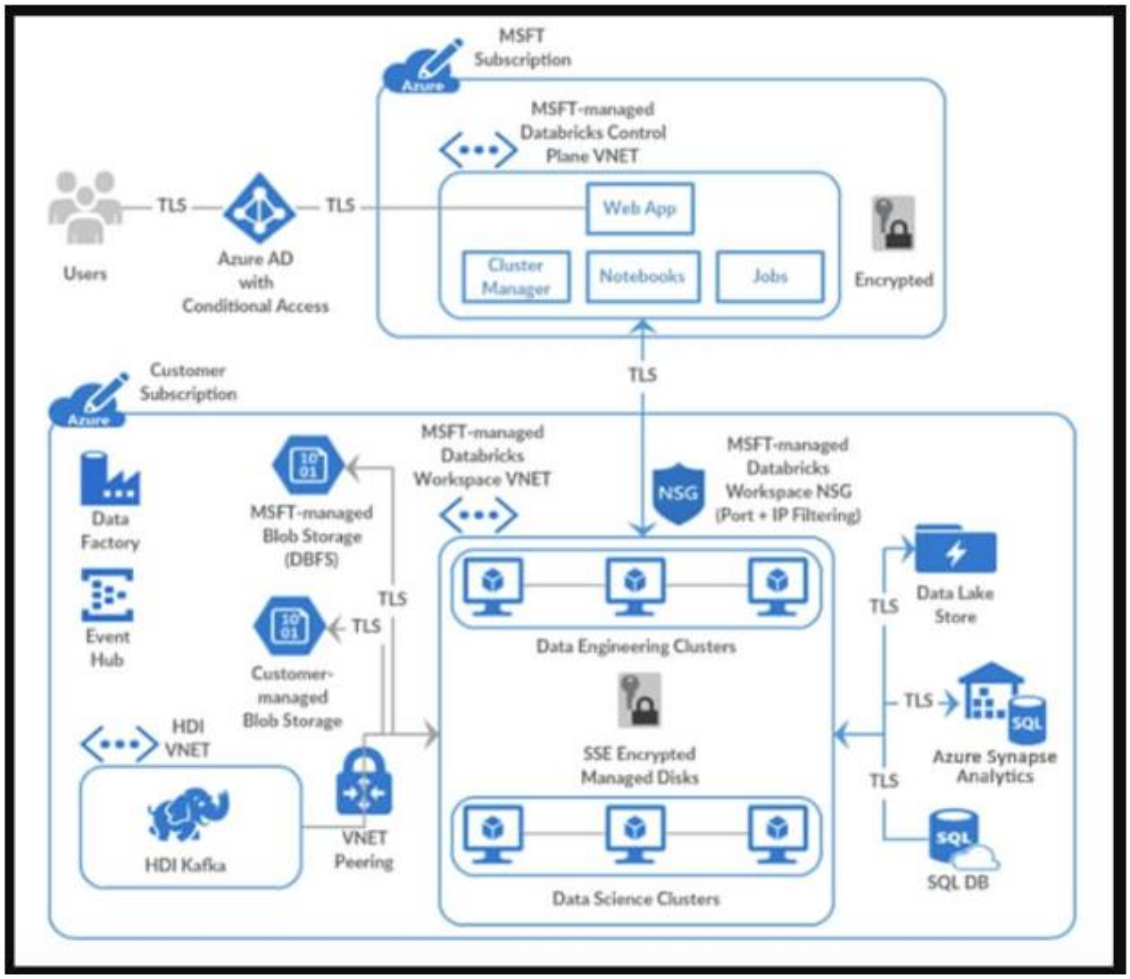
If you need advanced network connectivity, such as custom VNet peering and VNet injection, you could deploy Azure Databricks Data Plane resources within your own VNet.



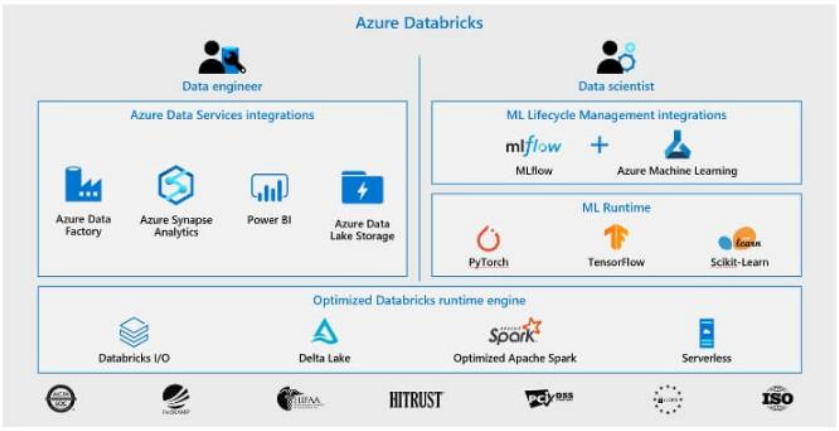
NAME	TYPE	LOCATION	
03a67d3205c04e2aa8604531d8946956	Virtual machine	East US 2	•••
03a67d3205c04e2aa8604531d8946956_OsDisk_1_d0553bd1c27948fa90f088b6c3d09251	Disk	East US 2	•••
03a67d3205c04e2aa8604531d8946956-containerRootVolume	Disk	East US 2	•••
03a67d3205c04e2aa8604531d8946956-privateNIC	Network interface	East US 2	•••
03a67d3205c04e2aa8604531d8946956-publicIP	Public IP address	East US 2	•••
03a67d3205c04e2aa8604531d8946956-publicNIC	Network interface	East US 2	•••
2300d7f9bf814f6ea728c4e54032bc2a-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	•••
dbstoragezkbo4ipeo56z2	Storage account	East US 2	•••
workers-sg	Network security group	East US 2	•••
workers-vnet	Virtual network	East US 2	•••


Azure Databricks Platform Architecture
Microsoft/Databricks Subscription
Client Subscription
Control Plane
Web Application
Cluster Manager
Jobs
Notebooks/Query Results
Hive Metastore
ACLs/Sessions
Data Plane
Customer Data Sources
Backend Services
Log Storage/Analysis, Central Account Directory, Monitoring

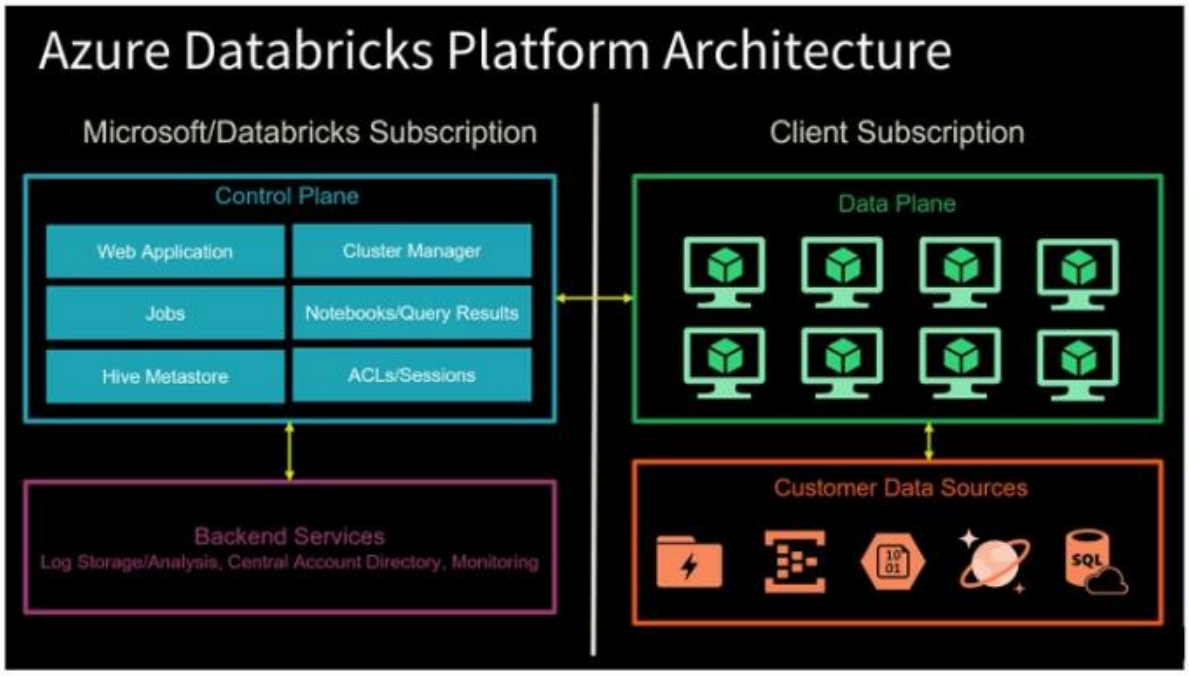

Microsoft/Databricks Subscription
Client Subscription
Control Plane
Data Plane
Commands, query results, table metadata, logs
User and organization meta-data, logs, useage
Data being read from and written to the data source
Backend Services
Customer Data Sources

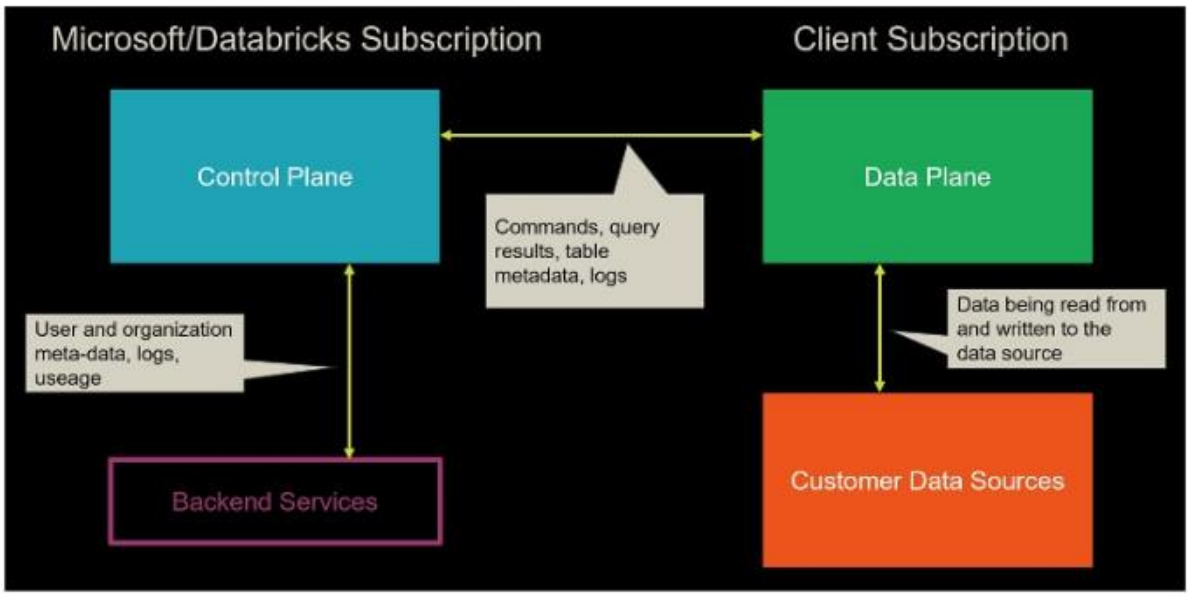


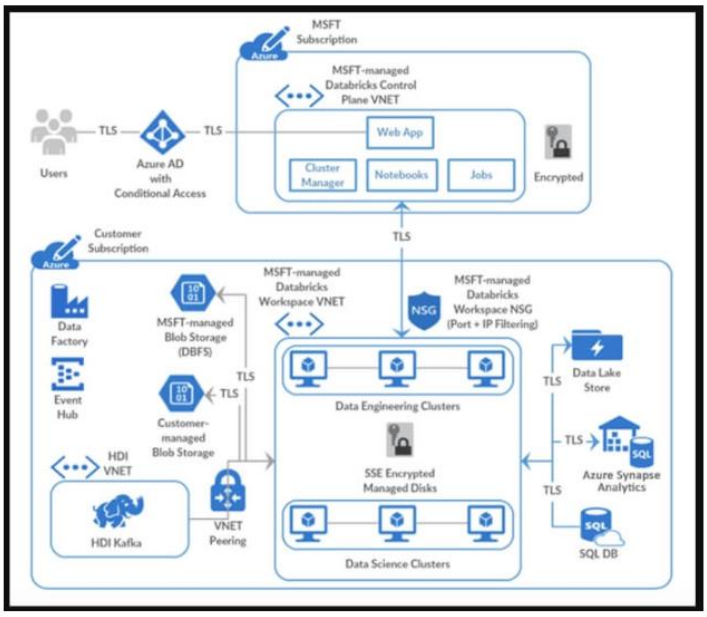
https://docs.databricks.com/getting-started/overview.html



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
2300d7f9bf814f6ea728c4e54032bc2a-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
dbstoragezkbo4ipeo56z2	Storage account	East US 2
workers-sg	Network security group	East US 2
workers-vnet	Virtual network	East US 2


Azure Databricks Platform Architecture
Microsoft/Databricks Subscription
Client Subscription
Control Plane
Web Application
Cluster Manager
Jobs
Notebooks/Query Results
Hive Metastore
ACLs/Sessions
Data Plane
Customer Data Sources
Backend Services
Log Storage/Analysis, Central Account Directory, Monitoring


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



Report Error

Q. 56 What size does OPTIMIZE compact small files to?

- (●) Around 500 MB
- ( ) Around 100 MB
- ( ) Around 2 GB
- ( ) Around 1 GB

Report Error

Q. 56 What size does OPTIMIZE compact small files to?

- ( ) Around 2 GB
- ( ) Around 100 MB
- ( ) Around 500 MB
- (●) Around 1 GB

Explanation:- The OPTIMIZE command compacts small files to around 1GB. The Spark optimization team determined this value to be a good compromise between speed and performance.
https://docs.databricks.com/spark/latest/spark-sql/language-manual/delta-optimize.html

Report Error

Q. 57

Scenario: You are working as a consultant at Avengers Security and at the moment, you are working with the data engineering team which manages Azure HDInsight clusters at the company. The group spends an enormous amount of time creating and destroying clusters each day due to the fact that the majority of the data pipeline process runs in minutes.

Required: Utilize a solution which will deploy multiple HDInsight clusters with minimal effort.

Which of the following should recommend to the IT team to implement?

- ○ Azure Traffic Manager
- ○ Azure Databricks
- ◉ Azure Resource Manager templates
- ○ Azure PowerShell

Report Error

Q. 57

Scenario: You are working as a consultant at Avengers Security and at the moment, you are working with the data engineering team which manages Azure HDInsight clusters at the company. The group spends an enormous amount of time creating and destroying clusters each day due to the fact that the majority of the data pipeline process runs in minutes.

Required: Utilize a solution which will deploy multiple HDInsight clusters with minimal effort.

Which of the following should recommend to the IT team to implement?

- ○ Azure PowerShell
- ○ Azure Databricks
- ○ Azure Traffic Manager
- ◉ Azure Resource Manager templates

Explanation:- A Resource Manager template makes it easy to create the following resources for your application in a single, coordinated operation:
• HDInsight clusters and their dependent resources (such as the default storage account).
• Other resources (such as Azure SQL Database to use Apache Sqoop).
In the template, you define the resources that are needed for the application. You also specify deployment parameters to input values for different environments.
The template consists of JSON and expressions that you use to construct values for your deployment.
https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-create-linux-clusters-arm-templates

Report Error

Q. 58

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

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.

Correct or Incorrect : Configuring a git repository allows you to save changes, letting you only publish when you have tested your changes to your satisfaction.

- ○ Incorrect
- ◉ Correct

Report Error

Incorrect

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.

Advantages of Git integration

Below is a list of some of the advantages git integration provides to the authoring experience:

• Source control: As your data factory workloads become crucial, you would want to integrate your factory with Git to leverage several source control benefits like the following:

• Ability to track/audit changes.

• Ability to revert changes that introduced bugs.

• Partial saves: When authoring against the data factory service, you can't save changes as a draft and all publishes must pass data factory validation. Whether your pipelines are not finished or you simply don't want to lose changes if your computer crashes, git integration allows for incremental changes of data factory resources regardless of what state they are in. Configuring a git repository allows you to save changes, letting you only publish when you have tested your changes to your satisfaction.

• Collaboration and control: If you have multiple team members contributing to the same factory, you may want to let your teammates collaborate with each other via a code review process. You can also set up your factory such that not every contributor has equal permissions. Some team members may only be allowed to make changes via Git and only certain people in the team are allowed to publish the changes to the factory.

• Better CI/CD: If you are deploying to multiple environments with a continuous delivery process, git integration makes certain actions easier. Some of these actions include:

• Configure your release pipeline to trigger automatically as soon as there are any changes made to your 'dev' factory.

• Customize the properties in your factory that are available as parameters in the Resource Manager template. It can be useful to keep only the required set of properties as parameters, and have everything else hard coded.

• Better Performance: An average factory with git integration loads 10 times faster than one authoring against the data factory service. This performance improvement is because resources are downloaded via Git.

Connect to a Git repository

There are different ways to connect a Git repository to your data factory for both Azure Repos and GitHub. After you connect to a Git repository, you can view and manage your configuration in the management hub under Git configuration in the Source control section.

Configuration method 1: Home page

In the Azure Data Factory home page, select Set up Code Repository.

Configuration method 2: Authoring canvas

In the Azure Data Factory UX authoring canvas, select the Data Factory drop-down menu, and then select Set up Code Repository.

Configuration method 3: Management hub

Go to the management hub in the Azure Data Factory UX. Select Git configuration in the Source control section. If you have no repository connected, click Set up code repository.

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

Report Error

Q. 59 What is the name of the application architecture that enables near real-time querying to provide insights?

- ○ OLAP
- ○ ELT
- ◉ ADPS
- ○ ETL
- ○ HTAP
- ○ OLTP

Report Error

Q. 59 What is the name of the application architecture that enables near real-time querying to provide insights?

- ◉ HTAP

Explanation:- HTAP stands for Hybrid Transactional and Analytical Processing that enable you to gain insights from operational systems without impacting the performance of the operational system.
https://www.zdnet.com/article/what-is-hybrid-transactionanalytical-processing-htap/

- ○ ADPS
- ○ OLAP
- ○ ELT
- ○ OLTP
- ○ ETL

Report Error

Q. 60 How do you list files in DBFS within a notebook?

- ○ ls /my-file-path
- ◉ %fs ls /my-file-path
- ○ %fs dir /my-file-path
- ○ %dfs ls /my-file-path

Report Error

Q. 60 How do you list files in DBFS within a notebook?

- [x] %fs ls /my-file-path

Explanation:- DBFS and local driver node paths
You can work with files on DBFS or on the local driver node of the cluster. You can access the file system using magic commands such as %fs or %sh. You add the file system magic to the cell before executing the ls command.
https://docs.microsoft.com/en-us/azure/databricks/data/databricks-file-system

- [ ] %fs dir /my-file-path
- [ ] %dfs ls /my-file-path
- [ ] ls /my-file-path

Report Error