Question 50: Skipped

What's the purpose of linked services in Azure Data Factory?

- To represent a processing step in a pipeline.
- To link data storage devices between on-prem and cloud environments.
- To represent a data store or a compute resource that can host execution of an activity.
  **(Correct)**
- To link data stores or computer resources together for the movement of data between resources.

**Explanation**

Linked services define the connection information needed for Data Factory to connect to external resources.

https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services

---

Question 51: Skipped

Within Azure Synapse Link for Azure Cosmos DB, which Column-oriented store is optimized for queries?

- Analytical store
  **(Correct)**
- Query store
- Transactional store
- Cosmos DB store

**Explanation**
An analytical store is a data store optimized for analytical queries.

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

---

Question 52: Skipped

**True or False:** By default, Azure Storage accounts automatically encrypt data-at-rest and data-in-transit. This will protect data-in-transit regardless if an authorized connection uses HTTP or HTTPS.

- False
  **(Correct)**
- True

**Explanation**
**Encryption in transit**

Keep your data secure by enabling *transport-level security* between Azure and the client. Always use `HTTPS` to secure communication over the public internet. When you call the REST APIs to access objects in storage accounts, you can enforce the use of `HTTPS` by requiring secure transfer for the storage account. After you enable secure transfer, connections that use `HTTP` will be refused. This flag will also enforce secure transfer over SMB by requiring SMB 3.0 for all file share mounts.

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

---

Question 53: Skipped

Where do you enable Azure Synapse Link for Azure Cosmos DB?

- In Azure Cosmos DB
  **(Correct)**
- Azure Portal
- In Azure Synapse Analytics
- In Azure Synapse Link

**Explanation**

When you enable Azure Synapse Link for Azure Cosmos DB it must be done in Azure Cosmos DB.

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

---

Question 54: Skipped

What are the two prerequisites for connecting Azure Databricks with Azure Synapse Analytics that apply to the Azure Synapse Analytics instance?

- Create a database master key and configure the firewall to enable Azure services to connect
**(Correct)**

- Generate a OTP to verify the account credentials, then set a master endpoint then configure the endpoint firewall to enable Azure services to connect.
- Use a correctly formatted `ConnectionString` and create a database master key
- Add the client IP address to the firewall's allowed IP addresses list and use the correctly formatted `ConnectionString`

**Explanation**
Azure Databricks is an Apache Spark–based analytics platform that supports SQL analytics and can be integrated with Azure Synapse to run high-performance analytics. It allows faster interactive processing of batch and streaming data and has built-in functions for machine learning and big data processing.

**The two prerequisites for connecting Azure Databricks with Azure Synapse Analytics that apply to the Azure Synapse Analytics instance are to create a database master key and configure the firewall to enable Azure services to connect.**

https://docs.databricks.com/data/data-sources/azure/synapse-analytics.html

---

Question 55: Skipped

**Scenario**: You are working on a new project and you are in a meeting to discuss which Azure data platform technology is best for your company.

**Requirement**: A globally distributed, multimodel database that can perform queries in less than a second.

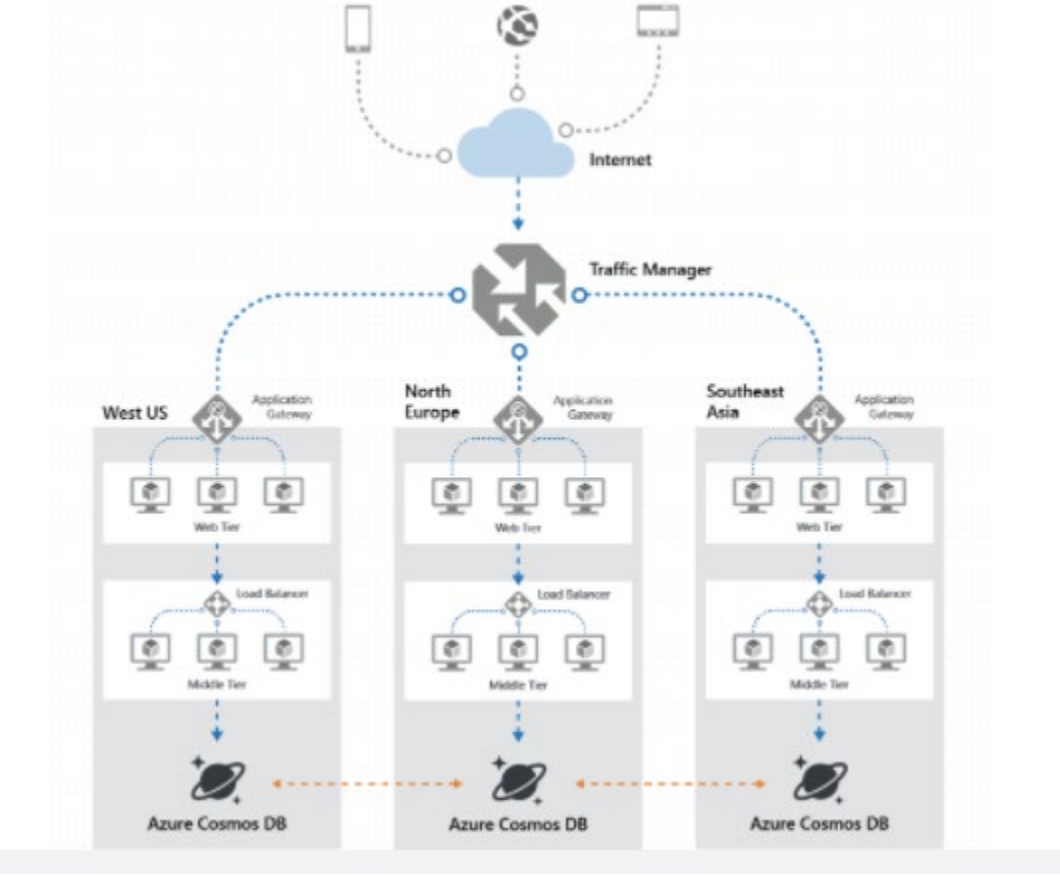
Which of the following should you choose?

- Azure SQL Database
- Azure Data Factory
- Azure SQL on VM
- Azure Cosmos DB
  **(Correct)**
- Azure SQL Data Warehouse
- Azure Databricks

**Explanation**
Azure Cosmos DB is a globally distributed, multimodel database that can offer subsecond query performance. Azure Cosmos DB transparently replicates the data to all the regions associated with your Cosmos account. Azure Cosmos DB is a globally distributed database service that's designed to provide low latency, elastic scalability of throughput, well-defined semantics for data consistency, and high availability. In short, if your application needs fast response time anywhere in the world, if it's required to be always online, and needs unlimited and elastic scalability of throughput and storage, you should build your application on Azure Cosmos DB.

You can configure your databases to be globally distributed and available in any of the Azure regions. To lower the latency, place the data close to where your users are. Choosing the required regions depends on the global reach of your application and where your users are located. Cosmos DB transparently replicates the data to all the regions associated with your Cosmos account. It provides a single system image of your globally distributed Azure Cosmos database and containers that your application can read and write to locally.

With Azure Cosmos DB, you can add or remove the regions associated with your account at any time. Your application doesn't need to be paused or redeployed to add or remove a region.




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

Question 56: Skipped

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

As a Data Engineer, you can transfer and move data in several ways. The most common tool is [?], which provides robust resources and nearly 100 enterprise connectors. [?] also allows you to transform data by using a wide variety of languages.

- Azure Data Factory
  **(Correct)**
- Azure Stream Analytics
- Azure Data Lake Storage
- Azure Data Catalogue
- Azure Databricks

**Explanation**
As a Data Engineer, you can transfer and move data in several ways. One way is to start an *Extract, Transform, and Load (ETL)* process.

Extraction sources can include databases, files, and streams. Each source has unique data formats that can be structured, semistructured, or unstructured. In Azure, data sources include Azure Cosmos DB, Azure Data Lake, files, and Azure Blob storage.

**ETL tools**

As a data engineer, you'll use several tools for ETL. The most common tool is Azure Data Factory, which provides robust resources and nearly 100 enterprise connectors. Data Factory also allows you to transform data by using a wide variety of languages.

You might find that you also need a repository to maintain information about your organization's data sources and dictionaries. Azure Data Catalog can store this information centrally.

**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

**Evolution from ETL**

Azure has opened the way for technologies that can handle unstructured data at an unlimited scale. This change has shifted the paradigm for loading and transforming data from ETL to extract, load, and transform (ELT).

The benefit of ELT is that you can store data in its original format, be it JSON, XML, PDF, or images. In ELT, you define the data's structure during the transformation phase, so you can use the source data in multiple downstream systems.

In an ELT process, data is extracted and loaded in its native format. This change reduces the time required to load the data into a destination system. The change also limits resource contention on the data sources.

The steps for the ELT process are the same as for the ETL process. They just follow a different order.

Another process like ELT is called extract, load, transform, and load (ELTL). The difference with ELTL is that it has a final load into a destination system.

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

---

Question 57: Skipped

Which notebook format is used in Databricks?

- `.notebook`
- DBC
  **(Correct)**
- `.spark`
- `.dbrk`

**Explanation**
The supported Databricks notebook format is the DBC file type.

**Notebook external formats**

Azure Databricks supports several notebook external formats:

• Source file: A file containing only source code statements with the extension `.scala`, `.py`, `.sq`l, or `.r`.

• HTML: An Azure Databricks notebook with the extension `.html`.

• DBC archive: A Databricks archive.

• IPython notebook: A Jupyter notebook with the extension `.ipynb`.

• RMarkdown: An R Markdown document with the extension `.Rmd`.

https://docs.microsoft.com/en-us/azure/databricks/notebooks/notebooks-manage

---

Question 58: Skipped

**Scenario:** O'Shaughnessy's is a fast food restaurant. The chain has stores nationwide and is rivalled by Big Belly Burgers. You have been hired by the company to advise on working with Microsoft Azure Synapse Analytics.

At the moment, you are leading a meeting where the topic at hand is designing an enterprise data warehouse.

The IT team at O'Shaughnessy's is working on a project to design and create an enterprise data warehouse in Azure Synapse Analytics which will contain a table named Customers. Customers will contain credit card information.

Because security is critical to O'Shaughnessy's, they have asked you to recommend a solution to provide salespeople with the ability to view all the entries in Customers but prevent all the salespeople from viewing or inferring the credit card information.

Which of the following techniques should you propose in your recommendation?

- Data masking
  **(Correct)**
- Column-level security
- Row-level security
- Always Encrypted

**Explanation**
*You should propose the use of Data masking in your recommendation because SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users.*

The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card.

Example: XXXX-XXXX-XXXX-1234 -

**Monitor and optimize data storage and data processing**

Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics support dynamic data masking. Dynamic data masking limits sensitive data exposure by masking it to non-privileged users.

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It’s a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.

For example, a service representative at a call centre might identify a caller by confirming several characters of their email address, but the complete email address shouldn't be revealed to the service representative. A masking rule can be defined that masks all the email address in the result set of any query. As another example, an appropriate data mask can be defined to protect personal data, so that a developer can query production environments for troubleshooting purposes without violating compliance regulations.

**Dynamic data masking basics**

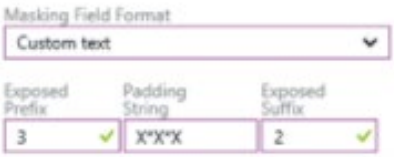
You set up a dynamic data masking policy in the Azure portal by selecting the **Dynamic Data Masking** blade under **Security** in your SQL Database configuration pane. This feature cannot be set using portal for SQL Managed Instance. For more information, see Dynamic Data Masking.

Dynamic data masking policy

**SQL users excluded from masking** - A set of SQL users or Azure AD identities that 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** - A set of rules that define the designated fields to be masked and the masking function that is used. The designated fields can be defined using a database schema name, table name, and column name.

**Masking functions** - A set of methods that control the exposure of data for different scenarios.

| Masking function | Masking logic |
| --- | --- |
| **Default** | **Full masking according to the data types of the designated fields**<br><br>• Use XXXX or fewer Xs if the size of the field is less than 4 characters for string data types (nchar, ntext, nvarchar).<br>• Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).<br>• Use 01-01-1900 for date/time data types (date, datetime2, datetime, datetimeoffset, smalldatetime, time).<br>• For SQL variant, the default value of the current type is used.<br>• For XML the document <masked/> is used.<br>• Use an empty value for special data types (timestamp table, hierarchyid, GUID, binary, image, varbinary spatial types). |
| **Credit card** | **Masking method, which exposes the last four digits of the designated fields** and adds a constant string as a prefix in the form of a credit card.<br><br>XXXX-XXXX-XXXX-1234 |
| **Email** | **Masking method, which exposes the first letter and replaces the domain with XXX.com** using a constant string prefix in the form of an email address.<br><br>aXX@XXXX.com |
| **Random number** | **Masking method, which generates a random number** according to the selected boundaries and actual data types. If the designated boundaries are equal, then the masking function is a constant number.<br>  |
| **Custom text** | **Masking method, which exposes the first and last characters** and adds a custom padding string in the middle. If the original string is shorter than the exposed prefix and suffix, only the padding string is used.<br>prefix[padding]suffix<br>  |

**Recommended fields to mask**

The DDM recommendations engine, flags certain fields from your database as potentially sensitive fields, which may be good candidates for masking. In the Dynamic Data Masking blade in the portal, you will see the recommended columns for your database. All you need to do is click **Add Mask** for one or more columns and then **Save** to apply a mask for these fields.

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

---

Question 59: Skipped

**Scenario:** You are working at OZcorp which is a multi-million dollar company run by Mayor Norman Osborn. Profits from the company are used to fund Norman's operatives, such as a police task force.

At the moment, you have been hired by OZcorp as a Microsoft Azure Synapse Analytics SME.

**Given:**

OZcorp has an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey.

• **Table - Sales:** The table is 600 GB in size. DateKey is used extensively in the `WHERE` clause queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Seventy-five percent of the records relate to one of forty regions.

• **Table - Invoice:** The table is 6 GB in size. DateKey and ProductKey are used extensively in the `WHERE` clause queries. RegionKey is used for grouping.

• There are 120 unique product keys and 65 unique region keys.

• Queries that use the data warehouse take a long time to complete.

**Required:**

The team plans to migrate the solution to use Azure Synapse Analytics and they need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.

Azure Synapse Analytics SME, the team looks to you for the best way forward. Which of the following should you recommend for the Sales table?

- Distribution type: Round-robin, Distribution column: ProductKey
- Distribution type: Hash-distributed, Distribution column: ProductKey **(Correct)**
- Distribution type: Round-robin, Distribution column: RegionKey
- Distribution type: Hash-distributed, Distribution column: RegionKey

**Explanation**
*You should recommend Hash-distributed for the Distribution type and ProductKey for the Distribution column.*

This is because ProductKey is used extensively in joins and Hash-distributed tables improve query performance on large fact tables.

**What is a distributed table?**

A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.

**Hash-distributed tables** improve query performance on large fact tables, and are the focus of this article. **Round-robin tables** are useful for improving loading speed. These design choices have a significant impact on improving query and loading performance.

Another table storage option is to replicate a small table across all the Compute nodes. For more information, see Design guidance for replicated tables. To quickly choose among the three options, see Distributed tables in the tables overview.

As part of table design, understand as much as possible about your data and how the data is queried. For example, consider these questions:

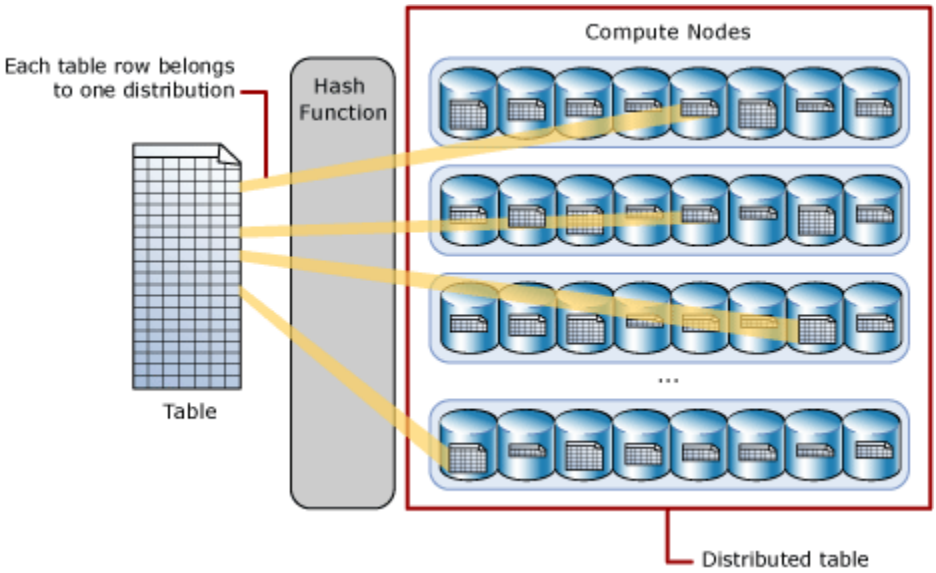
• How large is the table?

• How often is the table refreshed?

• Do I have fact and dimension tables in a dedicated SQL pool?

**Hash distributed**

A hash-distributed table distributes table rows across the Compute nodes by using a deterministic hash function to assign each row to one distribution.




Since identical values always hash to the same distribution, SQL Analytics has built-in knowledge of the row locations. In dedicated SQL pool this knowledge is used to minimize data movement during queries, which improves query performance.

Hash-distributed tables work well for large fact tables in a star schema. They can have very large numbers of rows and still achieve high performance. There are, of course, some design considerations that help you to get the performance the distributed system is designed to provide. Choosing a good distribution column is one such consideration that is described in this article.

Consider using a hash-distributed table when:

• The table size on disk is more than 2 GB.

• The table has frequent insert, update, and delete operations.

**Round-robin distributed**

A round-robin distributed table distributes table rows evenly across all distributions. The assignment of rows to distributions is random. Unlike hash-distributed tables, rows with equal values are not guaranteed to be assigned to the same distribution.

As a result, the system sometimes needs to invoke a data movement operation to better organize your data before it can resolve a query. This extra step can slow down your queries. For example, joining a round-robin table usually requires reshuffling the rows, which is a performance hit.

Consider using the round-robin distribution for your table in the following scenarios:

• When getting started as a simple starting point since it is the default

• If there is no obvious joining key

• If there is no good candidate column for hash distributing the table

• If the table does not share a common join key with other tables

• If the join is less significant than other joins in the query

• When the table is a temporary staging table

**Choosing a distribution column**

A hash-distributed table has a distribution column that is the hash key. For example, the following code creates a hash-distributed table with ProductKey as the distribution column.

```
SQL
CREATE TABLE [dbo].[FactInternetSales]
(   [ProductKey]            int          NOT NULL
,   [OrderDateKey]          int          NOT NULL
,   [CustomerKey]           int          NOT NULL
,   [PromotionKey]          int          NOT NULL
,   [SalesOrderNumber]      nvarchar(20) NOT NULL
,   [OrderQuantity]         smallint     NOT NULL
,   [UnitPrice]             money        NOT NULL
,   [SalesAmount]           money        NOT NULL
)
WITH
(   CLUSTERED COLUMNSTORE INDEX
,  DISTRIBUTION = HASH([ProductKey])
)
;
```

Data stored in the distribution column can be updated. Updates to data in the distribution column could result in data shuffle operation.

Choosing a distribution column is an important design decision since the values in this column determine how the rows are distributed. The best choice depends on several factors, and usually involves tradeoffs. Once a distribution column is chosen, you cannot change it.

If you didn't choose the best column the first time, you can use CREATE TABLE AS SELECT (CTAS) to re-create the table with a different distribution column.

**Choose a distribution column with data that distributes evenly**

For best performance, all of the distributions should have approximately the same number of rows. When one or more distributions have a disproportionate number of rows, some distributions finish their portion of a parallel query before others. Since the query can't complete until all distributions have finished processing, each query is only as fast as the slowest distribution.

Data skew means the data is not distributed evenly across the distributions

Processing skew means that some distributions take longer than others when running parallel queries. This can happen when the data is skewed.

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

**Has many unique values.** The column can have some duplicate values. However, all rows with the same value are assigned to the same distribution. Since there are 60 distributions, the column should have at least 60 unique values. Usually the number of unique values is much greater.

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

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

Choose a distribution column that minimizes data movement

To get the correct query result queries might move data from one Compute node to another. Data movement commonly happens when queries have joins and aggregations on distributed tables. Choosing a distribution column that helps minimize data movement is one of the most important strategies for optimizing performance of your dedicated SQL pool.

To minimize data movement, select a distribution column that:

Is used in `JOIN`, `GROUP BY`, `DISTINCT`, `OVER`, and `HAVING` clauses. When two large fact tables have frequent joins, query performance improves when you distribute both tables on one of the join columns. When a table is not used in joins, consider distributing the table on a column that is frequently in the `GROUP BY` clause.

Is *not* used in `WHERE` clauses. This could narrow the query to not run on all the distributions.

Is *not* a date column. `WHERE` clauses often filter by date. When this happens, all the processing could run on only a few distributions.

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

---

Question 60: Skipped

How many access keys are provided for accessing your Azure storage account?

- 1 per authorized user
- 1
- 2
  **(Correct)**
- 3
- 4

**Explanation**
Each storage account has two access keys. This lets you follow the best-practice guideline of periodically replacing the key used by your applications without incurring downtime.

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

Question 61: Skipped

In which version of SQL Server was SSIS Projects introduced?

- SQL Server 2014
- SQL Server 2008
- SQL Server 2016
- SQL Server 2012 **(Correct)**

**Explanation**
SSIS Projects was introduced in SQL Server 2012 and is the unit of deployment for SSIS solutions.

SQL Server 2012 was a major release for SSIS. It introduced the concept of the **project deployment model**, where entire projects with their packages are deployed to a server, instead of individual packages. The SSIS of SQL Server 2005 and 2008 is now referred to as the **(legacy) package deployment model**. SSIS 2012 made it easier to configure packages and it came with a centralized storage and management utility: the catalogue. We'll dive deeper into those topics later on in the tutorial.

SQL Server 2014 didn't have any changes for SSIS, but on the side new sources or transformations were added to the product. This was done by separate downloads trough CodePlex (an open-source code website) or through the SQL Server Feature Pack. Examples are the Azure feature pack (to connect to cloud sources and objects) and the balanced data distributor (to divide your data stream into multiple pipelines).

In SQL Server 2016 there were some updates to the SSIS product. Instead of deploying entire projects, you can new deploy packages individually again. There are additional sources – especially cloud and big data sources – and some important changes were made to the catalogue. You can find an overview of all new features here and here.

During all these years, SSIS has built itself a reputation for being a stable, robust and fast ETL tool with support for many sources. However, it's still mainly an on-premises solution, there is – at the time of writing – no real cloud alternative.

https://www.mssqltips.com/sqlservertutorial/9054/sql-server-integration-services-ssis-versions-and-tools/

---

Question 62: Skipped

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

A Control Activity in Data Factory is defined in JSON format as follows:

```
JSON
{
"name": "Control Activity Name",
"description": "description",
"type": "<ActivityType>",
"typeProperties":
{
},
"dependsOn":
{
}
}
```

Which of the JSON properties are required? (Select all that apply)

- dependsOn
- description
  **(Correct)**
- type
  **(Correct)**
- typeProperties
- name
  **(Correct)**

**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

**Activities and pipelines**

**Defining control activities**

A Control Activity in Data Factory is defined in JSON format as follows:

```
JSON
{
"name": "Control Activity Name",
"description": "description",
"type": "<ActivityType>",
"typeProperties":
{
},
"dependsOn":
{
}
}
```

The following describes properties in the above JSON:

**Property: name**

Name of the activity.

Required: Yes

**Property: description**

Text describing what the activity or is used for.

Required: Yes

**Property: type**

Defines the type of the activity.

Required: Yes

**Property: typeProperties**

Properties in the typeProperties section depend on each type of activity.

Required: No

**Property: dependsOn**

This property is used to define activity dependencies, and how subsequent activities depend on previous activities.

Required: No

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities

---

Question 63: Skipped

Which of the following are facets of Azure Databricks security? (Select four)

- Load Balancing
- IAM/Auth **(Correct)**
- Data Protection **(Correct)**
- VNet Peering
- Vault
- Compliance **(Correct)**
- Network **(Correct)**
- Encryption

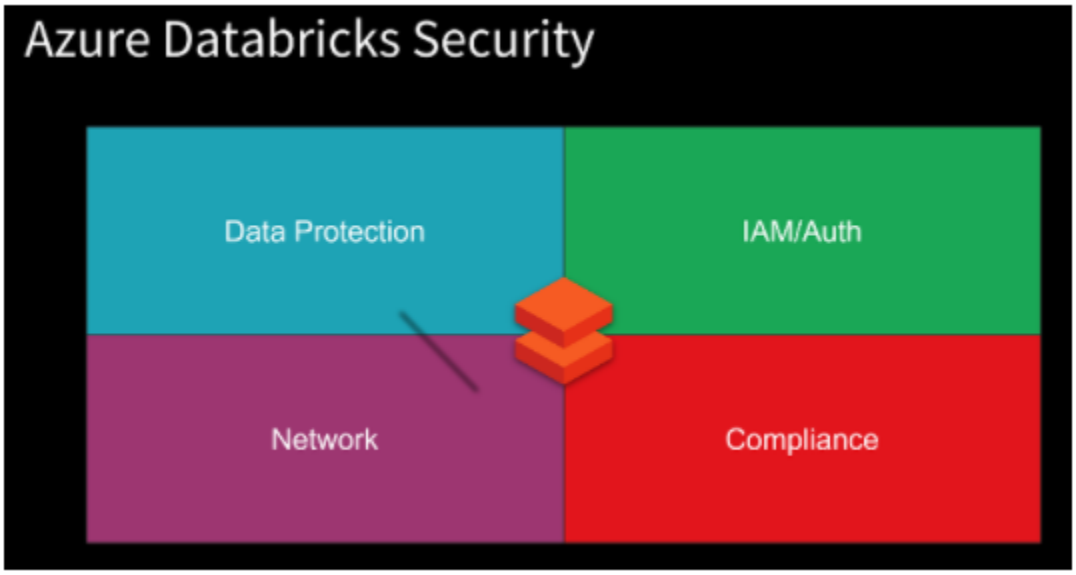
**Explanation**
**The following are the facets of Azure Databricks security:**

**• Data Protection**

**• IAM/Auth**

**• Network**

**• Compliance**




**Data Protection** is comprised of the following:

• Encryption at-rest – Service Managed Keys, User Managed Keys

• Encryption in-transit (Transport Layer Security - TLS)

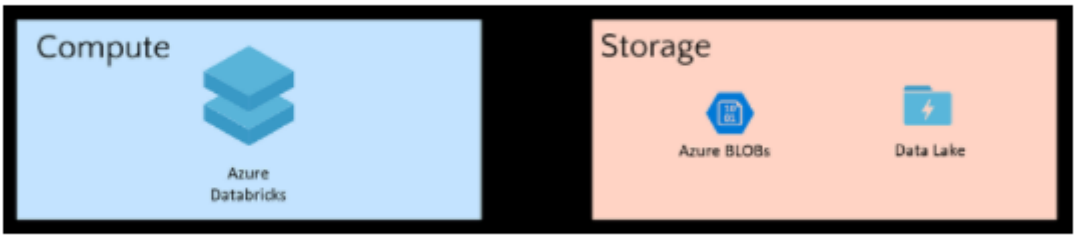
• File/Folder Level access control lists (ACLs) for Azure Active Directory (AAD) Users, Groups, Service Principals

• ACLs for Clusters, Folders, Notebooks, Tables, Jobs

• Secrets with Azure Key Vault

**Encryption at-rest**

Azure Databricks has separation of compute and storage.




Azure Databricks is a compute platform. It does not store data, except for notebooks. Clusters are transient in nature. They process the data then are terminated. All data is stored in the customer's subscription. Because the Azure storage services use server-side encryption, communication between these services and the Databricks clusters is seamless.

Storage Services such as Azure Storage Blobs and Azure Data Lake Storage (Gen1/2) provide:

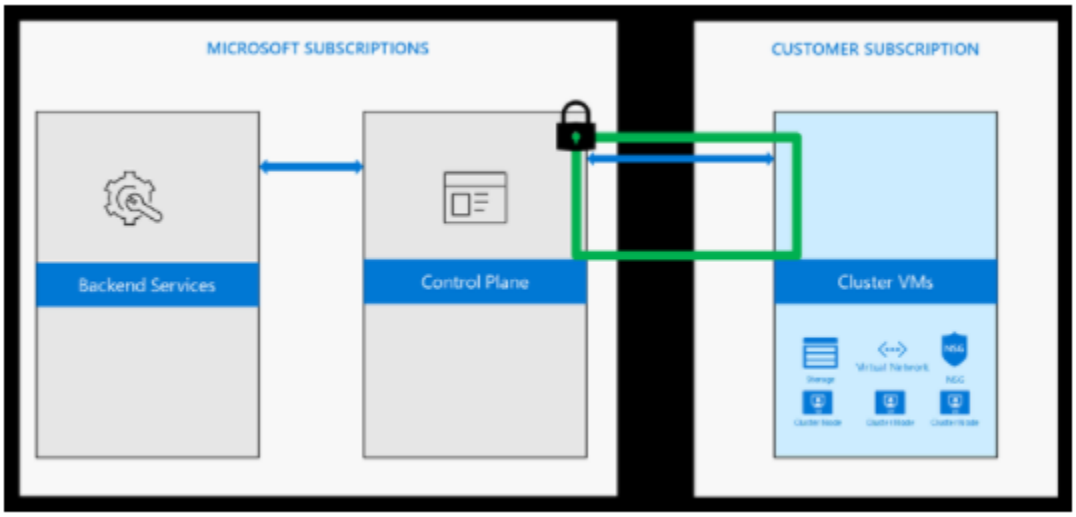
• Encryption of Data - Automatic server-side encryption in addition to encryption on storage attached to the VMs

• Customer Managed Keys - Bring your own keys with Key Vault integration

• File/Folder Level ACLs (Azure Data Lake Storage (Gen1/2))

**Encryption in-transit**

All the traffic from the Control Plane to the clusters in the customer subscription (Data Plane) is always encrypted with TLS.




When clusters access data from various Azure services, TLS is always used to ensure encryption in-transit.

When customers access notebooks via their web browsers, the connection is also secured with TLS.

**Access control - ADLS Passthrough**

When enabled, authentication automatically takes place in Azure Data Lake Storage (ADLS) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that one uses to log into Azure Databricks. Commands running on a configured cluster will be able to read and write data in ADLS without needing to configure service principal credentials. Any ACLs applied at the folder or file level in ADLS are enforced based on the user's identity.

ADLS Passthrough is configured when you create a cluster in the Azure Databricks workspace. ADLS Gen1 requires Databricks Runtime 5.1+. ADLS Gen2 requires 5.3+.

On a *standard cluster*, when you enable this setting you must set single user access to one of the Azure Active Directory (AAD) users in the Azure Databricks workspace. Only one user is allowed to run commands on this cluster when Credential Passthrough is enabled.


*High-concurrency clusters* can be shared by multiple users. When you enable ADLS Passthrough on this type of cluster, it does not require you to select a single user.


**Access control - Folders**

Access control is available only in the Premium SKU. By default, all users can create and modify workspace objects unless an administrator enables workspace access control. With workspace access control, individual permissions determine a user's abilities. This section describes the individual permissions and how to enable and configure workspace access control.

You can assign five permission levels to notebooks and folders: No Permissions, Read, Run, Edit, and Manage. The following tables lists the abilities for each permission.

| Ability | No Permissions | Read | Run | Edit | Manage |
|---|---|---|---|---|---|
| View items | | X | X | X | X |
| Create, clone, import, export items | | X | X | X | X |
| Run commands on notebooks | | | X | X | X |
| Attach/detach notebooks | | | X | X | X |
| Delete items | | | | X | X |
| Move/rename items | | | | X | X |
| Change permissions | | | | | X |

## Access control - Notebooks

| Ability | No Permissions | Read | Run | Edit | Manage |
|---|---|---|---|---|---|
| View cells | | X | X | X | X |
| Comment | | X | X | X | X |
| Run commands | | | X | X | X |
| Attach/detach notebooks | | | X | X | X |
| Edit cells | | | | X | X |
| Change permissions | | | | | X |

All notebooks in a folder inherit all permissions settings of that folder. For example, a user that has Run permission on a folder has Run permission on all notebooks in that folder.

To enable workspace access control:

• Go to the Admin Console.

• Select the Access Control tab.

• Click the Enable button next to Workspace Access Control.

• Click Confirm to confirm the change.

**Access control - Clusters**

All users can view libraries. To control who can attach libraries to clusters, manage access control on clusters.

By default, all users can create and modify clusters unless an administrator enables cluster access control. With cluster access control, permissions determine a user's abilities. There are four permission levels for a cluster: No Permissions, Can Attach To, Can Restart, and Can Manage:

| | | | | |
|---|---|---|---|---|
| View Spark UI | | x | x | x |
| View cluster metrics | | x | x | x |
| Terminate cluster | | | x | x |
| Start cluster | | | x | x |
| Restart cluster | | | x | x |
| Edit cluster | | | | x |
| Attach library to cluster | | | | x |
| Resize cluster | | | | x |
| Modify permissions | | | | x |

Note: You have Can Manage permission for any cluster that you create.

**Access control - Jobs**

To control who can run jobs and see the results of job runs, manage access control on jobs.

There are five permission levels for jobs: No Permissions, Can View, Can Manage Run, Is Owner, and Can Manage. The Can Manage permission is reserved for administrators.

| Ability | No Permissions | Can View | Can Manage Run | Is Owner | Can Manage (admin) |
|---|---|---|---|---|---|
| View job details and settings | x | x | x | x | x |
| View results, Spark UI, logs of a job run | | x | x | x | x |
| Run now | | | x | x | x |
| Cancel run | | | x | x | x |
| Edit job settings | | | | x | x |
| Modify permissions | | | | x | x |

**Access control - Tables**

Table access control (table ACLs) lets you programmatically grant and revoke access to your data from SQL, Python, and PySpark.

By default, all users have access to all data stored in a cluster's managed tables unless an administrator enables table access control for that cluster. Once table access control is enabled for a cluster, users can set permissions for data objects on that cluster.

Before you can grant or revoke privileges on data objects, an administrator must enable table access control for the cluster.

**View-based access control model**

The Azure Databricks view-based access control model defines the following privileges:

- `SELECT` – gives read access to an object.
- `CREATE` – gives ability to create an object (for example, a table in a database)
- `MODIFY` – gives ability to add/delete/modify data to/from an object.
- `READ_METADATA` – gives ability to view an object and its metadata.
- `CREATE_NAMED_FUNCTION` – gives ability to create a named UDF in an existing catalogue or database.
- `ALL PRIVILEGES` – gives all privileges (gets translated into all the above privileges)

The privileges above can apply to the following classes of objects:

- `CATALOG` - controls access to the entire data catalog.
- `DATABASE` - controls access to a database.
- `TABLE` - controls access to a managed or external table.
- `VIEW` - controls access to SQL views.
- `FUNCTION` - controls access to a named function.
- `ANONYMOUS FUNCTION` - controls access to anonymous or temporary functions.
- `ANY FILE` - controls access to the underlying filesystem.

**Secrets**

Using the Secrets APIs, Secrets can be securely stored including in an Azure Key Vault or Databricks backend. Authorized users can consume the secrets to access services.

Azure Databricks has two types of secret scopes: Key Vault-backed and Databricks-backed. These secret scopes allow you to store secrets, such as database connection strings, securely. If someone tries to output a secret to a notebook, it is replaced by `[REDACTED]`. This helps prevent someone from viewing the secret or accidentally leaking it when displaying or sharing the notebook.

As a best practice, instead of directly entering your credentials into a notebook, use Azure Databricks secrets to store your credentials and reference them in notebooks and jobs.

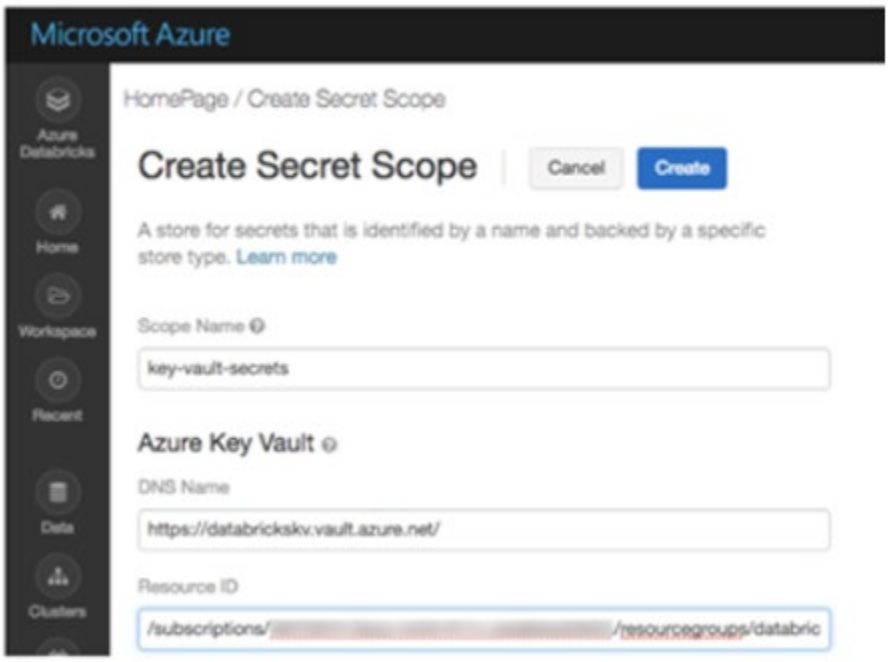
To set up secrets you:

• Create a secret scope. Secret scope names are case insensitive.

• Add secrets to the scope. Secret names are case insensitive.

• If you have the Azure Databricks Premium Plan, assign access control to the secret scope.

Screenshot of creating an Azure Key Vault-backed secret scope:




https://docs.microsoft.com/en-us/azure/databricks/scenarios/security-baseline

---

Question 64: Skipped

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

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

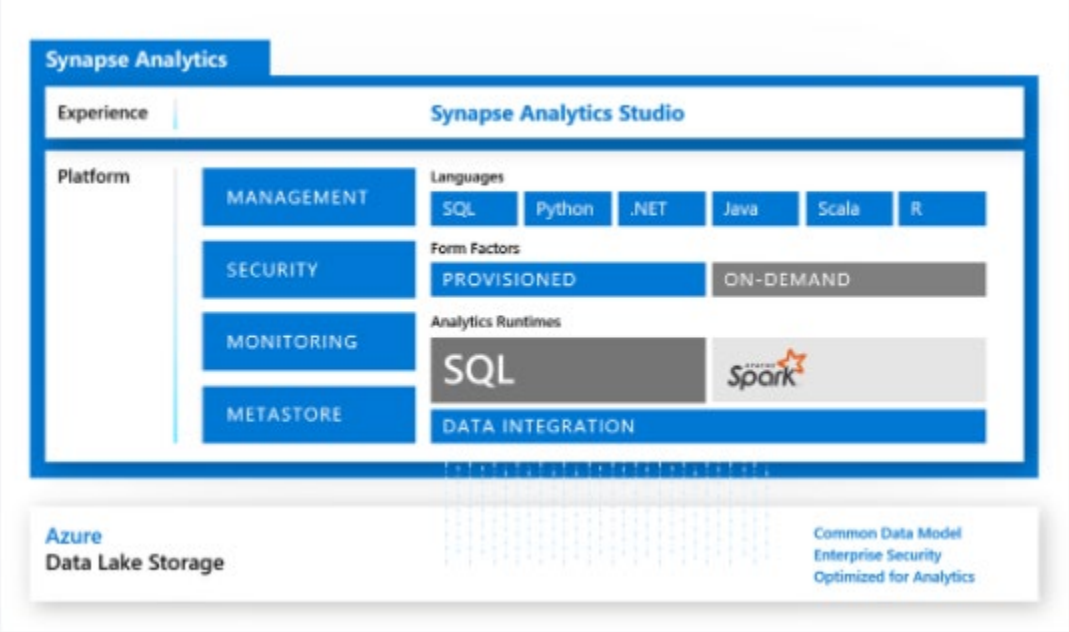
[?] leverages the capabilities of Azure Data Factory and is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale. Using this, you can create and schedule data-driven workflows that can ingest data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

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

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

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

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




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

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

**Data integration to integrate your data with Azure Synapse Pipelines**

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

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

Azure Synapse Analytics enables you to reach out to operational data using Azure Synapse Link, and is achieved without impacting the performance of the transactional data store. For this to happen, you have to enable the feature within both Azure Synapse

Analytics, and within the data store to which Azure Synapse Analytics will connect, such as Azure Cosmos DB. In the case of Azure Cosmos DB, this will create an analytical data store. As data changes in the transactional system, the changed data is fed to the analytical store in a Column store format from which Azure Synapse Link can query with no disruption to the source system.

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

---

Question 65: Skipped

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.

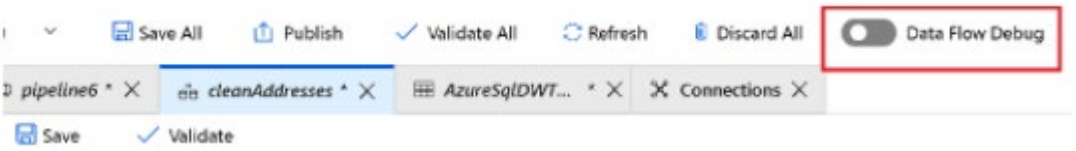
Clicking Debug will provision the Spark clusters required to interact with the Mapping Data Flow transformations. If you select `AutoResolveIntegrationRuntime`, what will be the result? (Select all that apply)

- A cluster with eight cores that will be available with a time to live value of 60 minutes.
**(Correct)**
- Data engineers can develop data transformation logic with or without writing code.
- It typically takes 5-7 minutes for the cluster to spin up.
**(Correct)**
- None of the listed options.
- The number of rows that are returned within the data previewer are fixed by the AutoResolve Agent.
- All the listed options.

**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.

During the building of Mapping Data Flows, you can interactively watch how the data transformations are executing so that you can debug them. To use this functionality, it is first necessary to turn on the “Data Flow Debug” feature.




Clicking Debug will provision the Spark clusters required to interact with the Mapping Data Flow transformations. On turning Debug on, you will be prompted to select the Integration Runtime that you require to use in the environment. **If you select AutoResolveIntegrationRuntime, a cluster with eight cores that will be available with a time to live value of 60 minutes.**

**Note: It typically takes 5-7 minutes for the cluster to spin up.** With this mode on and the Spark clusters running, you are able to build your data flow step by step and view the data as it runs through each transformation phase.

A Data Preview tab is available in Debug mode that will allow you to view the data at each stage of the pipeline. You can view the data after each transformation. The data previewer also provides the ability to actions on the data such as looking at descriptive statistics of the data, or the ability to modify the data.

Finally, you can use the debug settings to control the number of rows that are returned within the data previewer.

Note: It is recommended to limit the number of rows that returns enough to enable you to confirm that the data is correct. The bigger the data set, the longer it takes to return the results back. You can also use the Debug settings to specify any parameter values that should be used during the execution of the pipeline.

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

Question 66: Skipped

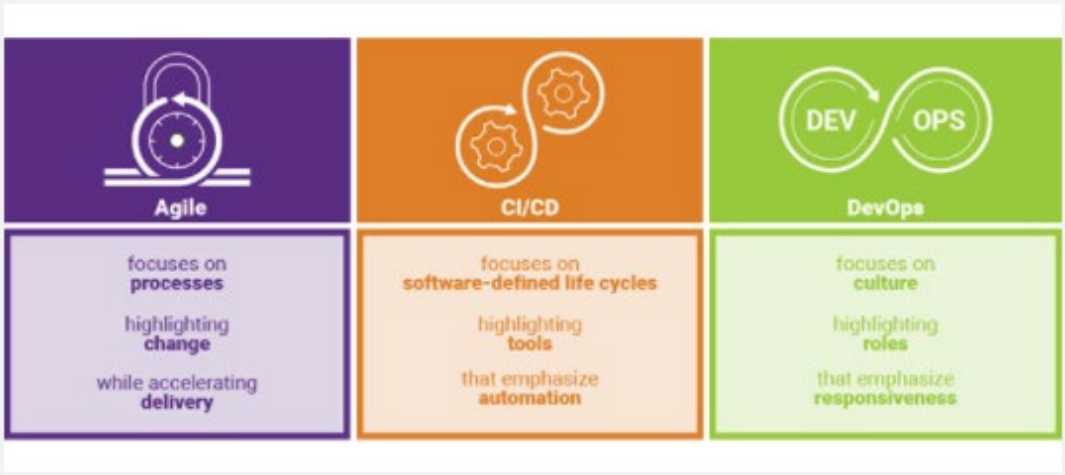
While Agile, CI/CD, and DevOps are different, they support one another

What does DevOps focus on?

- Practices
- Development process
- Culture
  **(Correct)**
- Strategy

**Explanation**

While Agile, CI/CD, and DevOps are different, they support one another. Agile focuses on the development process, CI/CD on practices, and DevOps on culture.




• **Agile** focuses on processes highlighting change while accelerating delivery.

• **CI/CD** focuses on software-defined life cycles highlighting tools that emphasize automation.

• **DevOps** focuses on culture highlighting roles that emphasize responsiveness.

https://www.synopsys.com/blogs/software-security/agile-cicd-devops-difference/

Azure DevOps is a collection of services that provide an end-to-end solution for the five core practices of DevOps: planning and tracking, development, build and test, delivery, and monitoring and operations.

It is possible to put an Azure Databricks Notebook under Version Control in an Azure Devops repo. Using Azure DevOps, you can then build Deployment pipelines to manage your release process.

**CI/CD with Azure DevOps**

Here are some of the features that make it well-suited to CI/CD with Azure Databricks.

• Integrated Git repositories

• Integration with other Azure services

• Automatic virtual machine management for testing builds

• Secure deployment

• Friendly GUI that generates (and accepts) various scripted files

**But what is CI/CD?**

**Continuous Integration**

Throughout the development cycle, developers commit code changes locally as they work on new features, bug fixes, etc. If the developers practice continuous integration, they merge their changes back to the main branch as often as possible. Each merge into the master branch triggers a build and automated tests that validate the code changes to ensure successful integration with other incoming changes. This process avoids integration headaches that frequently happen when people wait until the release day before they merge all their changes into the release branch.

**Continuous Delivery**

Continuous delivery builds on top of continuous integration to ensure you can successfully release new changes in a fast and consistent way. This is because, in addition to the automated builds and testing provided by continuous integration, the release process is automated to the point where you can deploy your application with the click of a button.

**Continuous Deployment**

Continuous deployment takes continuous delivery a step further by automatically deploying your application without human intervention. This means that merged changes pass through all stages of your production pipeline and, unless any of the tests fail, automatically release to production in a fully automated manner.

**Who benefits?**

*Everyone*. Once properly configured, automated testing and deployment can free up your engineering team and enable your data team to push their changes into production. For example:

• Data engineers can easily deploy changes to generate new tables for BI analysts.

• Data scientists can update models being used in production.

• Data analysts can modify scripts being used to generate dashboards.

In short, changes made to a Databricks notebook can be pushed to production with a simple mouse click (and then any amount of oversight that your DevOps team feels is appropriate).

https://docs.microsoft.com/en-us/azure/devops/user-guide/alm-devops-features?view=azure-devops

---

Question 67: Skipped

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

Azure Synapse Analytics is an integrated analytics platform, which combines data warehousing, big data analytics, data integration, and visualization into a single environment. Azure Synapse Analytics empowers users of all abilities to gain access and quick insights across all of their data, enabling a whole new level of performance and scale.

Descriptive analytics answers the question [?].

- Why is it happening?
- When will the modification made meet my goals?
- What is likely to happen in the future based on previous trends and patterns?"
- What is happening in my business?
  **(Correct)**

**Explanation**
Azure Synapse Analytics is an integrated analytics platform, which combines data warehousing, big data analytics, data integration, and visualization into a single environment. Azure Synapse Analytics empowers users of all abilities to gain access and quick insights across all of their data, enabling a whole new level of performance and scale.

**The range of analytical types that Azure Synapse Analytics can support include:**

**Descriptive analytics**

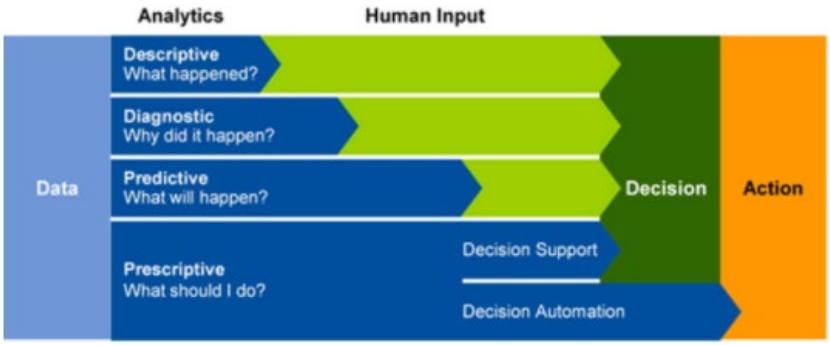
Descriptive analytics answers the question "What is happening in my business?" The data to answer this question is typically answered through the creation of a data warehouse. Azure Synapse Analytics leverages the dedicated SQL Pool capability that enables you to create a persisted data warehouse to perform this type of analysis. You can also make use of SQL Serverless to prepare data from files to create a data warehouse interactively to answer the question too.

**Diagnostic analytics**

Diagnostic analytics deals with answering the question "Why is it happening?" this may involve exploring information that already exists in a data warehouse, but typically

involves a wider search of your data estate to find more data to support this type of analysis. You can use the same SQL serverless capability within Azure Synapse Analytics that enables you to interactively explore data within a data lake. This can quickly enable a user to search for additional data that may help them to understand "Why is it happening?"

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




**Predictive analytics**

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

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

**Prescriptive analytics**

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

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

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

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

Question 68: Skipped

The pipelines in Azure Data Factory typically perform the which of the following steps? (Select four)

- Machine learning
- Monitor
**(Correct)**
- Connect and collect
**(Correct)**
- Publish
**(Correct)**
- Transform and enrich
**(Correct)**
- Migration
- DataCopy
- Treemap

**Explanation**
**Data-driven workflows**

The pipelines (data-driven workflows) in Azure Data Factory typically perform the following four steps:

**Connect and collect**

The first step in building an orchestration system is to define and connect all the required sources of data together, such as databases, file shares, and FTP web services. The next step is to ingest the data as needed to a centralized location for subsequent processing.

**Transform and enrich**

Compute services such as Databricks and Machine Learning can be used to prepare or produce transformed data on a maintainable and controlled schedule to feed production environments with cleansed and transformed data. In some instances, you

may even augment the source data with additional data to aid analysis, or consolidate it through a normalization process to be used in a Machine Learning experiment as an example.

**Publish**

After the raw data has been refined into a business-ready consumable form from the transform and enrich phase, you can load the data into Azure Data Warehouse, Azure SQL Database, Azure Cosmos DB, or whichever analytics engine your business users can point to from their business intelligence tools

**Monitor**

Azure Data Factory has built-in support for pipeline monitoring via Azure Monitor, API, PowerShell, Azure Monitor logs, and health panels on the Azure portal, to monitor the scheduled activities and pipelines for success and failure rates.

https://mindmajix.com/azure-data-factory

---

Question 69: Skipped

When a table is created, by default the data structure has no indexes and is called a heap. A well-designed indexing strategy can reduce disk I/O operations and consume less system resources therefore improving query performance, especially when using filtering, scans, and joins in a query.

Dedicated SQL Pools have which of the following indexing options available?

- Clustered columnstore indexes
  **(Correct)**
- Key indexes
- Hash indexes
- Clustered Rowstore Indexes
  **(Correct)**
- B-tree indexes
- Non-clustered index
  **(Correct)**

**Explanation**
When a table is created, by default the data structure has no indexes and is called a heap. A well-designed indexing strategy can reduce disk I/O operations and consume less system resources therefore improving query performance, especially when using filtering, scans, and joins in a query.

Dedicated SQL Pools have the following indexing options available:

**Clustered columnstore index**

Dedicated SQL Pools create a clustered columnstore index when no index options are specified on a table. Clustered columnstore indexes offer both the highest level of data compression as well as the best overall query performance. Clustered columnstore indexes will generally outperform clustered rowstore indexes or heap tables and are usually the best choice for large tables.

Additional compression on the data can be gained also with the index option `COLUMNSTORE_ARCHIVE`. These reduced sizes allow less memory to be used when

accessing and using the data as well as reducing the IOPs required to retrieve data from storage.

Columnstore works on segments of 1,024,000 rows that get compressed and optimized by column. This segmentation further helps to filter out and reduce the data accessed through leveraging metadata stored which summarizes the range and values within each segment during query optimization.

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

**Clustered index**

Clustered Rowstore Indexes define how the table itself is stored, ordered by the columns used for the Index. There can be only one clustered index on a table.

Clustered indexes are best for queries and joins that require ranges of data to be scanned, preferably in the same order that the index is defined.

**Non-clustered index**

A non-clustered index can be defined on a table or view with a clustered index or on a heap. Each index row in the non-clustered index contains the non-clustered key value and a row locator. This is a data structure separate/additional to the table or heap. You can create multiple non-clustered indexes on a table.

Non clustered indexes are best used when used for the columns in a join, group by statement or where clauses that return an exact match or few rows.

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?view=aps-pdw-2016-au7

---

Question 70: Skipped

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

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

**Business Requirements:**

BB wants to create a new analytics environment in Azure to meet the following requirements:

• See inventory levels across the stores. Data must be updated as close to real time as possible.

• Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

• Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

**Technical Requirements:**

BB identifies the following technical requirements:

• Minimize the number of different Azure services needed to achieve the business goals.

• Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by BB.

• Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

• Use Azure Active Directory (Azure AD) authentication whenever possible.

• Use the principle of least privilege when designing security.

• Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. BB wants to remove transient data from Data

• Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

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

• Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

**Planned Environment:**

BB plans to implement the following environment:

• The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

• Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

• Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

• Daily inventory data comes from a Microsoft SQL server located on a private network.

• BB currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

• BB will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

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

**The Ask:**

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

- Activity type: Event-based activity
- Activity type: Copy activity
  **(Correct)**
- Activity type: Lookup activity
- Activity type: Stored procedure activity

**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

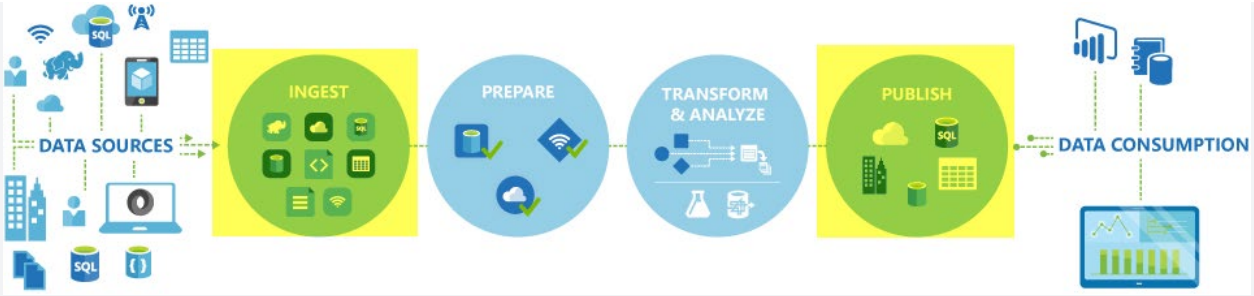
**Rational:**

• 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.

**Copy activity in Azure Data Factory**

In Azure Data Factory, you can use the Copy activity to copy data among data stores located on-premises and in the cloud. After you copy the data, you can use other activities to further transform and analyze it. You can also use the Copy activity to publish transformation and analysis results for business intelligence (BI) and application consumption.




The Copy activity is executed on an integration runtime. You can use different types of integration runtimes for different data copy scenarios:

When you're copying data between two data stores that are publicly accessible through the internet from any IP, you can use the Azure integration runtime for the copy activity. This integration runtime is secure, reliable, scalable, and globally available.

When you're copying data to and from data stores that are located on-premises or in a network with access control (for example, an Azure virtual network), you need to set up a self-hosted integration runtime.

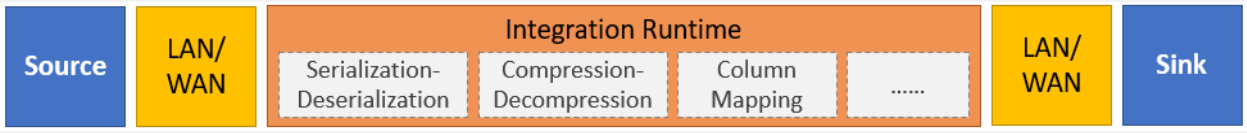
An integration runtime needs to be associated with each source and sink data store. For information about how the Copy activity determines which integration runtime to use, see Determining which IR to use.

To copy data from a source to a sink, the service that runs the Copy activity performs these steps:

Reads data from a source data store.

Performs serialization/deserialization, compression/decompression, column mapping, and so on. It performs these operations based on the configuration of the input dataset, output dataset, and Copy activity.

Writes data to the sink/destination data store.




https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview

---

Question 71: Skipped

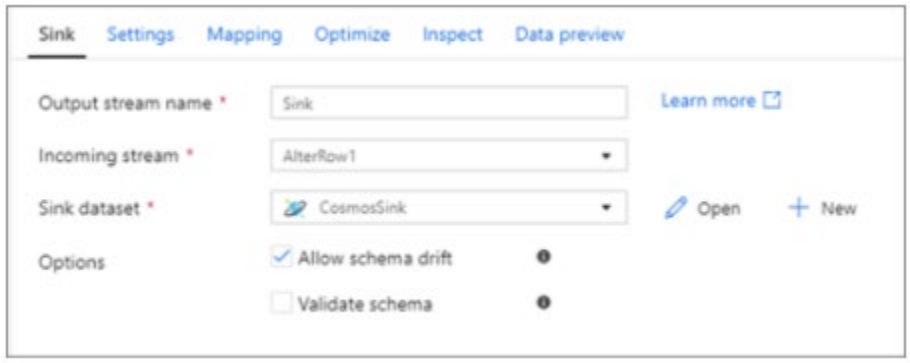
**Scenario**: While working on a project using Azure Data Factory, you are planning to load data into a data store or compute resource.

Which transformation in Mapping Data Flow is used to do this?

- Sink
**(Correct)**
- Field mapping
- Window
- Cache
- Source

**Explanation**
After you finish transforming your data, write it into a destination store by using the sink transformation. Every data flow requires at least one sink transformation, but you can write to as many sinks as necessary to complete your transformation flow. To write to additional sinks, create new streams via new branches and conditional splits.

Each sink transformation is associated with exactly one Azure Data Factory dataset object or linked service. The sink transformation determines the shape and location of the data you want to write to.




A Sink transformation allows you to choose a dataset definition for the destination output data. You can have as many sink transformations as your data flow requires.

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

Question 72: Skipped

In the context of analytics, data streams are event data generated by sensors or other sources that can be analyzed by another technology. Analyzing a data stream is typically done to measure the state change of a component or to capture information on an area of interest.

Which are approaches to processing data streams? (Select two)

- ☐ All the listed options
- ☐ Near real time
- ☐ Live
  **(Correct)**
- ☐ Multiprocessing
- ☐ On-demand
  **(Correct)**

**Explanation**
In the context of analytics, data streams are event data generated by sensors or other sources that can be analyzed by another technology. Analyzing a data stream is typically done to measure the state change of a component or to capture information on an area of interest. The intent being to:

• Continuously analyze data to detect issues and understand or respond to them.

• Understand component or system behaviour under various conditions to fuel further enhancements of said component or system.

•Trigger specific actions when certain thresholds are identified.

In today's world, data streams are ubiquitous. Companies can harness the latent knowledge in data streams to improve efficiencies and further innovation. Examples of use cases that analyze data streams include:

• Stock market trends.

• Monitoring data of water pipelines and electrical transmission and distribution systems by utility companies.

• Mechanical component health monitoring data in automotive and automobile industries.

• Monitoring data from industrial and manufacturing equipment.

• Sensor data in transportation, such as traffic management and highway toll lanes.

• Patient health monitoring data in the healthcare industry.

• Satellite data in the space industry.

• Fraud detection in the banking and finance industries.

• Sentiment analysis of social media posts.

**Approaches to data stream processing**

**There are two approaches to processing data streams: on-demand and live.**

Streaming data can be collected over time and persisted in storage as static data. The data can then be processed when convenient or during times when compute costs are lower. The downside to this approach is the cost of storing the data.

In contrast, live data streams have relatively low storage requirements. They also require more processing power to run computations in sliding windows over continuously incoming data to generate the insights.

https://www.simplilearn.com/what-is-data-processing-article

Question 73: Skipped

Continuous integration is the practice of testing each change made to your codebase automatically and as early as possible. Continuous delivery follows the testing that happens during continuous integration and pushes changes to a staging or production system. Below is a sample overview of the CI/CD lifecycle in an Azure data factory that's configured with Azure Repos Git.

The order of the activities has been shuffled.

a. A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes.

b. A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.

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

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

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

f. After a developer is satisfied with their changes, they create a pull request from their feature branch to the master or collaboration branch to get their changes reviewed by peers.

Select the correct sequence of events in the CI/CD lifecycle.

- a → b → c → f → d → e
- b → a → f → d → e → c
- a → f → d → b → c → e
- b → a → f → c → e → d **(Correct)**

**Explanation**
Continuous integration is the practice of testing each change made to your codebase automatically and as early as possible. Continuous delivery follows the testing that happens during continuous integration and pushes changes to a staging or production system. In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another. Azure Data Factory utilizes Azure Resource Manager templates to store the configuration of your various Azure Data Factory entities (pipelines, datasets, data flows, and so on). There are two suggested methods to promote a data factory to another environment:

• Automated deployment using Data Factory's integration with Azure Pipelines.

• Manually upload a Resource Manager template using Data Factory UX integration with Azure Resource Manager.

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

1. A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.

2. A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes.

3. After a developer is satisfied with their changes, they create a pull request from their feature branch to the master or collaboration branch to get their changes reviewed by peers.

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

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

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

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

Question 74: Skipped

Azure provides many ways to store your data. There are multiple database options like Azure SQL Database, Azure Cosmos DB, and Azure Table Storage. Azure offers multiple ways to store and send messages, such as Azure Queues and Event Hubs. You can even store loose files using services like Azure Files and Azure Blobs.

A storage account defines a policy that applies to all the storage services in the account.

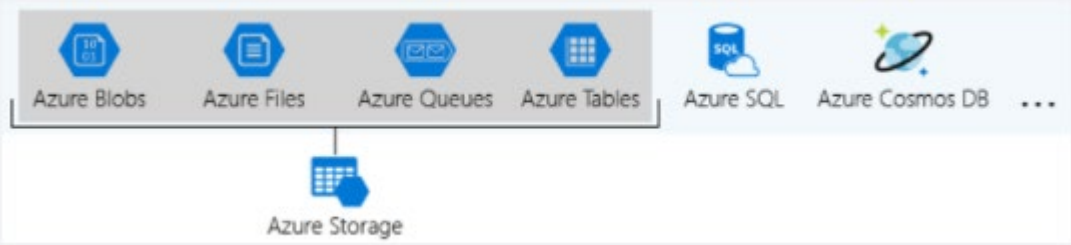
Which are the settings that are controlled by a storage account. (Select all that apply)

- ☐ Replication
  **(Correct)**
- ☐ Virtual networks
  **(Correct)**
- ☐ Secure transfer required
  **(Correct)**
- ☐ Location
  **(Correct)**
- ☐ Subscription
  **(Correct)**
- ☐ Access tier
  **(Correct)**
- ☐ Performance
  **(Correct)**

**Explanation**
**What is Azure Storage?**

Azure provides many ways to store your data. There are multiple database options like Azure SQL Database, Azure Cosmos DB, and Azure Table Storage. Azure offers multiple ways to store and send messages, such as Azure Queues and Event Hubs. You can even store loose files using services like Azure Files and Azure Blobs.

Azure selected four of these data services and placed them together under the name *Azure Storage*. The four services are Azure Blobs, Azure Files, Azure Queues, and Azure Tables. The following illustration shows the elements of Azure Storage.




These four were given special treatment because they are all primitive, cloud-based storage services and are often used together in the same application.

**Storage account settings**

A storage account defines a policy that applies to all the storage services in the account. For example, you could specify that all the contained services will be stored in the West US datacentre, accessible only over https, and billed to the sales department's subscription.

The settings that are controlled by a storage account are:

• **Subscription**: The Azure subscription that will be billed for the services in the account.

• **Location**: The datacentre that will store the services in the account.

• **Performance**: Determines the data services you can have in your storage account and the type of hardware disks used to store the data. **Standard** allows you to have any data service (Blob, File, Queue, Table) and uses magnetic disk drives. **Premium** introduces additional services for storing data. For example, storing unstructured object data as block blobs or append blobs, and specialized file storage used to store and create premium file shares. These storage accounts use solid-state drives (SSD) for storage.

• **Replication**: Determines the strategy used to make copies of your data to protect against hardware failure or natural disaster. At a minimum, Azure will automatically maintain three copies of your data within the data centre associated with the storage account. This is called locally-redundant storage (LRS), and guards against hardware failure but does not protect you from an event that incapacitates the entire datacentre.

You can upgrade to one of the other options such as geo-redundant storage (GRS) to get replication at different datacentres across the world.

• **Access tier**: Controls how quickly you will be able to access the blobs in this storage account. Hot gives quicker access than Cool, but at increased cost. This applies only to blobs, and serves as the default value for new blobs.

• **Secure transfer required**: A security feature that determines the supported protocols for access. Enabled requires HTTPs, while disabled allows HTTP.

• **Virtual networks**: A security feature that allows inbound access requests only from the virtual network(s) you specify.

https://www.c-sharpcorner.com/article/what-is-microsoft-azure-storage/

---

Question 75: Skipped

**Scenario**: You are configuring a new Azure Storage Account.

By default, what is the network rule set to?

- To allow all connection from a private IP address range.
- To allow all connections from all networks.
  **(Correct)**
- To deny all connections from all networks.
- None of the listed options.

**Explanation**
The default network rule is to allow all connections from all networks.

To secure your storage account, you should first configure a rule to deny access to traffic from all networks (including internet traffic) on the public endpoint, by default. Then, you should configure rules that grant access to traffic from specific VNets. You can also configure rules to grant access to traffic from selected public internet IP address ranges, enabling connections from specific internet or on-premises clients. This configuration enables you to build a secure network boundary for your applications.

You can combine firewall rules that allow access from specific virtual networks and from public IP address ranges on the same storage account. Storage firewall rules can be applied to existing storage accounts, or when creating new storage accounts.

Storage firewall rules apply to the public endpoint of a storage account. You don't need any firewall access rules to allow traffic for private endpoints of a storage account. The process of approving the creation of a private endpoint grants implicit access to traffic from the subnet that hosts the private endpoint.

Network rules are enforced on all network protocols for Azure storage, including REST and SMB. To access data using tools such as the Azure portal, Storage Explorer, and AZCopy, explicit network rules must be configured.

Once network rules are applied, they're enforced for all requests. SAS tokens that grant access to a specific IP address serve to limit the access of the token holder, but don't grant new access beyond configured network rules.

Virtual machine disk traffic (including mount and unmount operations, and disk IO) is not affected by network rules. REST access to page blobs is protected by network rules.

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

---

Question 76: Skipped

**True or False:** Unique credential creation is always required when utilizing the Azure Synapse Apache Spark Pool to Synapse SQL connector to enforce access control.

- True
- False
**(Correct)**

**Explanation**
The Authentication between the Azure Synapse Apache Spark Pool to Synapse SQL systems is made seamless due to the Azure Synapse Apache Spark Pool to Synapse SQL connector used in Azure Synapse Analytics. **The Token Service connects with Azure Active Directory to obtain security tokens for use when accessing the storage account or the data warehouse server.**

**For this reason, there's no need to create credentials or specify them in the connector API as long as Azure AD-Auth is configured at the storage account and the data warehouse server. If not, SQL Auth can be specified.** The only constraint that needs to be taken into account is that this connector is only working in scala.

There are some Prerequisites in order to authenticate namely:

• It needs to be a member of db_exporter role in the database or SQL pool from which you to transfer data to or from.

• It needs to be a member of the Storage Blob Data Contributor role on the default storage account.

If you want to create users, you need to connect to the SQL Pool database from which you want transfer data to/from.

Import statements are not needed since they are pre-loaded in case you use the notebook experience.

Once the authentication is in place, you are enabled to transfer data to or from a dedicated SQL pool attached within the workspace.

https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-secure-credentials-with-tokenlibrary?pivots=programming-language-csharp

Question 77: Skipped

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 the analytical processing, ingesting, synthesizing, and managing large sets of historical data.

- OLAP
**(Correct)**
- OLTP
- ADPS
- ETL
- ELT

**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.

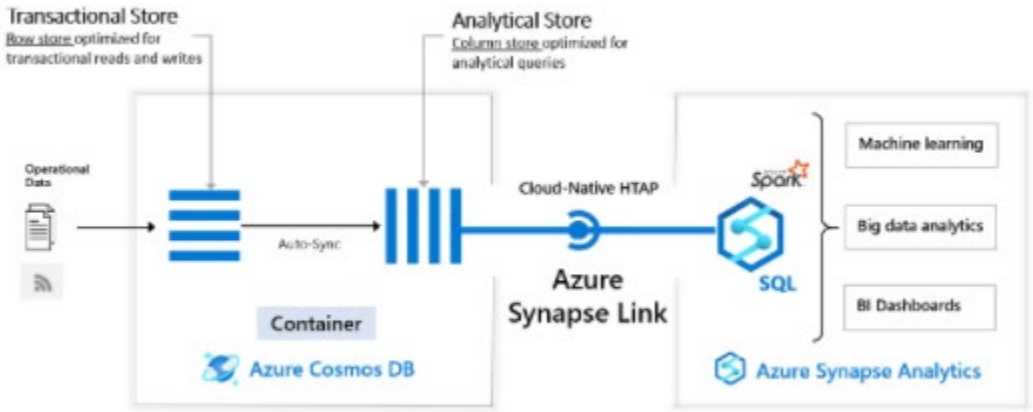
OLAP systems are optimized for the analytical processing, ingesting, synthesizing, and managing large sets of historical data. The data processed by OLAP systems largely originates from OLTP systems and needs to be loaded into the OLTP systems by means of batch processes commonly referred to as ETL (Extract, 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

---

Question 78: Skipped

Managing default network access rules for Azure Storage accounts can be done with which of the following?

- Azure Portal
**(Correct)**
- ARM templates
- Azure CLI
**(Correct)**
- Azure Designer
- PowerShell
**(Correct)**
- Azure CloudShell

**Explanation**
**Manage Default Network Access Rules**

Manage default network access rules for storage accounts through the Azure portal, PowerShell, or the Azure CLI.

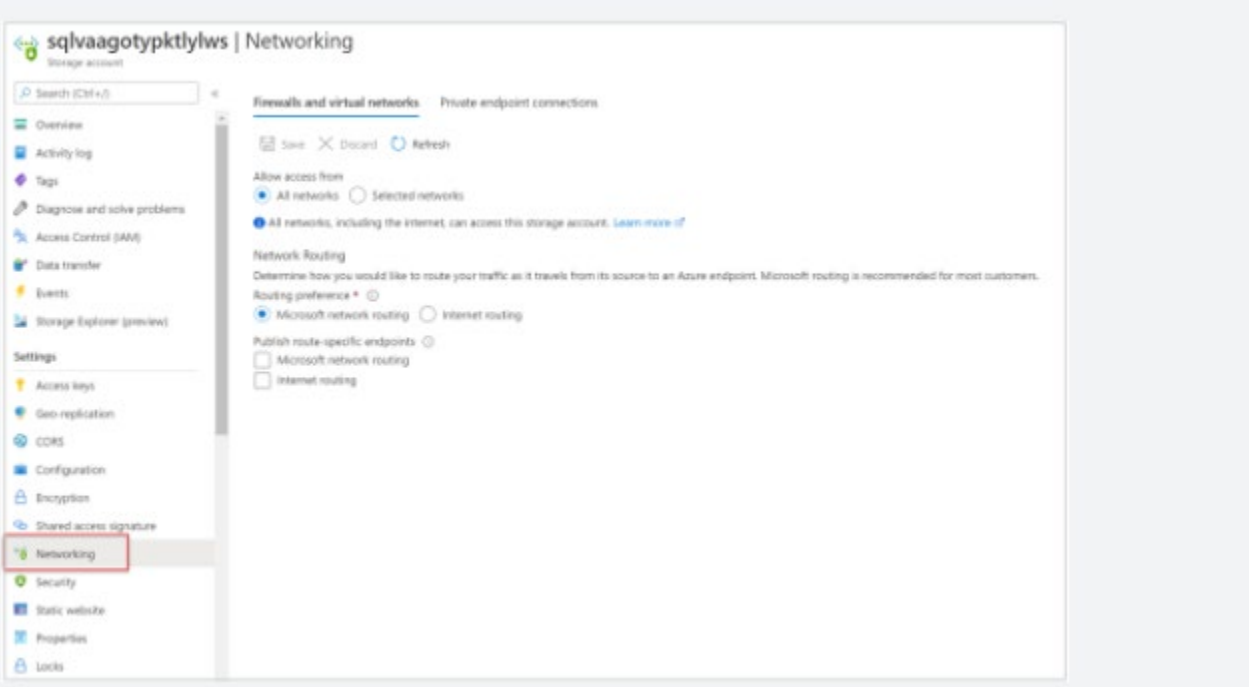
Follow these steps to change default network access in the Azure portal.

1. Go to the storage account you want to secure.

2. Select **Networking**.

3. To restrict traffic from selected networks, select **Selected networks**. To allow traffic from all networks, select **All networks**.

4. To apply your changes, select **Save**.




*Note: The keyphrase here is "Managing Default", not modifying specific instances.*

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

---

Question 79: Skipped

If you want to store data *without performing analysis on the data*, how should you set the Hierarchical Namespace option within the storage account of an Azure Blob storage account?

- Disabled **(Correct)**
- ON
- OFF
- Auto-scale
- Enabled

**Explanation**

In Azure Blob storage, you can store large amounts of unstructured ("object") data, in a single hierarchy, also known as a flat namespace. You can access this data by using `HTTP` or `HTTPs`. Azure Data Lake Storage Gen2 builds on blob storage and optimizes I/O of high-volume data by using hierarchical namespaces that you turned on in the previous exercise.

Hierarchical namespaces organize blob data into *directories* and stores metadata about each directory and the files within it. This structure allows operations, such as directory renames and deletes, to be performed in a single atomic operation. Flat namespaces, by contrast, require several operations proportionate to the number of objects in the structure.

**Azure Blob storage vs. Azure Data Lake Storage**

If you want to store data *without performing analysis on the data*, set the **Hierarchical Namespace** option to **Disabled** to set up the storage account as an Azure Blob storage account.

If you are performing analytics on the data, set up the storage account as an Azure Data Lake Storage Gen2 account by setting the **Hierarchical Namespace** option to **Enabled**. Because Azure Data Lake Storage Gen2 is integrated into the Azure Storage platform, applications can use either the Blob APIs or the Azure Data Lake Storage Gen2 file system APIs to access data.

https://blog.pragmaticworks.com/azure-data-lake-vs-azure-blob-storage-in-data-warehousing

Question 80: Skipped

Spark is a distributed computing environment. Therefore, work is parallelized across executors. At which two levels does this parallelization occur?

- The Driver and the Executor
- The Slot and the Task
- The Executor and the Task
- The Executor and the Slot
  **(Correct)**

**Explanation**
We parallelize at two levels:

• **The first level of parallelization is the Executor - a Java virtual machine running on a node, typically, one instance per node. Each Executor has a number of Slots to which parallelized Tasks can be assigned to it by the Driver.**

• The second level of parallelization is the Slot - the number of which is determined by the number of cores and CPUs of each node/executor.

**Executor**

The **executors** are responsible for actually executing the work that the **driver** assigns them. This means, each executor is responsible for only two things:

1.Executing code assigned to it by the driver

2.Reporting the state of the computation, on that executor, back to the driver node

**Cores/Slots/Threads**

• Each **Executor** has a number of **Slots** to which parallelized **Tasks** can be assigned to it by the **Driver**.

• So for example:

• If we have **3** identical home desktops (*nodes*) hooked up together in a LAN (like through your home router), each with i7 processors (**8** cores), then that's a **3** node Cluster:

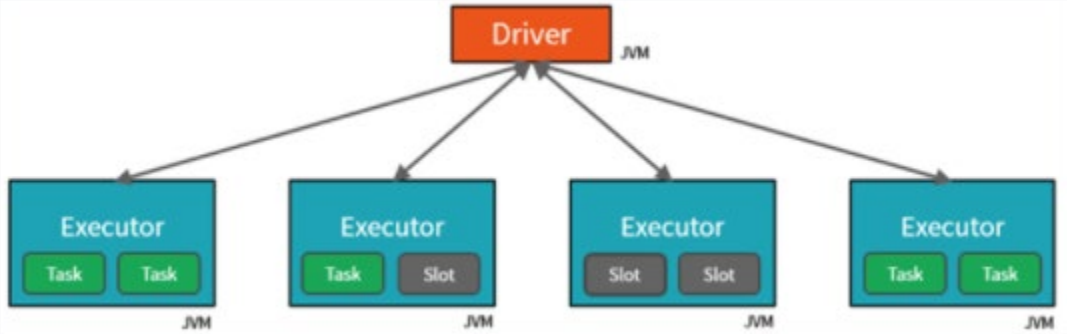
• **1** Driver node

• **2** Executor nodes

• The **8 cores per Executor node** means **8 Slots**, meaning the driver can assign each executor up to **8 Tasks**

• The idea is, an i7 CPU Core is manufactured by Intel such that it is capable of executing it's own Task independent of the other Cores, so **8 Cores = 8 Slots = 8 Tasks in parallel**

*For example: the diagram below is showing 2 Core Executor nodes:*




https://www.rakirahman.me/spark-certification-study-guide-part-1/

---

Question 81: Skipped

**Scenario**: While working on a project, the need arises to develop T-SQL scripts and notebooks in Azure Synapse Analytics.

Which of the following may be used to accomplish this?

- Azure Synapse Studio
  **(Correct)**
- DevTest Labs
- Databricks
- Azure Portal
- Data Lake

**Explanation**
Azure Synapse Studio is where you can develop T-SQL scripts and notebooks.

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

---

Question 82: Skipped

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

In Data Lake Storage Gen1, data engineers query data by using the [?] language.

- U-SQL
**(Correct)**
- ABS API
- M-SQL
- T-SQL
- ADLS API

**Explanation**
**Data Lake Storage Queries**

In Data Lake Storage Gen1, data engineers query data by using the U-SQL language. U-SQL is a language that combines declarative SQL with imperative C# to let you process data at any scale. Through the scalable, distributed-query capability of U-SQL, you can efficiently analyze data across relational stores such as Azure SQL Database. With U-SQL, you can process unstructured data by applying schema on read and inserting custom logic and UDFs. Additionally, U-SQL includes extensibility that gives you fine-grained control over how to execute at scale.

In Gen 2, use the Azure Blob Storage API or the Azure Data Lake System (ADLS) API.

https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-u-sql-get-started

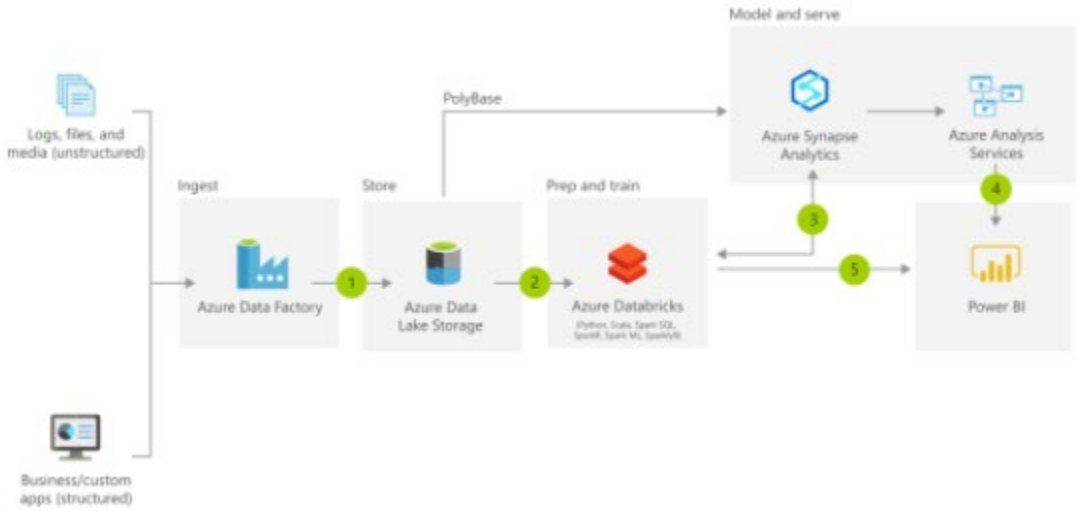
Question 83: Skipped

The pace of change in both the capabilities of technologies, and the elastic nature of cloud services has meant that new opportunities have been presented to evolve the data warehouse to handle modern workloads.

Which of the following are examples of these opportunities? (Select all that apply)

- ☐ Advanced analytics for all users
  **(Correct)**
- ☐ Static data velocities
- ☐ Increased flexibility for data volumes
  **(Correct)**
- ☐ Insights through analytical dashboards
  **(Correct)**
- ☐ New varieties of data
  **(Correct)**

**Explanation**
A modern data warehouse lets you bring together all your data at any scale easily, and means you can get insights through analytical dashboards, operational reports, or advanced analytics for all your users.




The pace of change in both the capabilities of technologies, and the elastic nature of cloud services has meant that new opportunities have been presented to evolve the data warehouse to handle modern workloads including:

**Increased volumes of data**

Microsoft Azure services have the capability to scale its capacity to meet the demands that an organization faces as its data grows. In traditional on-premises data, scaling on-premises servers is a non-trivial task that involves costs, procurement of additional hardware, as well as potential disruption to the business to meet the demand. With Azure, services such as Azure Synapse Analytics can be scaled at the click of a button, and can even be auto-scaled.

Staging data is also simplified using Azure Data Lake Store Gen2, which can store a wide variety of data in its raw format, making the process of ingesting data into a data warehouse much easier.

**New varieties of data**

Traditional data warehouse in the past have had difficulty in handling certain types of data. For example, extrapolating data from sources such as PDF files through to sound files where either too complex or cost prohibitive. The improvements in AI technologies such as Form Recognizer and Speech to Text Cognitive Services means that these types of data sources can now be passed through a cognitive service and outputted in a text-based format that can be stored in the Azure Data Lake Store Gen2, along with the source files themselves.

**Data velocities**

Traditional on-premises data warehouses in the main have dealt with the batch movement of data based on a schedule. Some organization may build real-time data warehouse if the business need is compelling and the organization can absorb the cost of the implementation. Azure has made it easier and much more cost effective to provision streaming services that can interact with a wide variety of services so that a modern data warehouse can deliver solutions in a batch or a real-time manner without the obstruction of cost.

https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/modern-data-warehouse

---

Question 84: Skipped

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

Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in [?].

- Historic batches
- Near real-time
  **(Correct)**
- Real-time
- Prediction mode

**Explanation**
Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real time.

With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data. The API continuously increments and updates the final data.

**Event Hubs and Spark Structured Streaming**

Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.

Azure Event Hubs can be integrated with Spark Structured Streaming to perform processing of messages in near real time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.

**Streaming concepts**

Stream processing is where you continuously incorporate new data into Data Lake storage and compute results. The streaming data comes in faster than it can be consumed when using traditional batch-related processing techniques. A stream of data is treated as a table to which data is continuously appended. Examples of such data include bank card transactions, Internet of Things (IoT) device data, and video game play events.

A streaming system consists of:

• Input sources such as Kafka, Azure Event Hubs, IoT Hub, files on a distributed system, or TCP-IP sockets

• Stream processing using Structured Streaming, forEach sinks, memory sinks, etc.

https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html

---

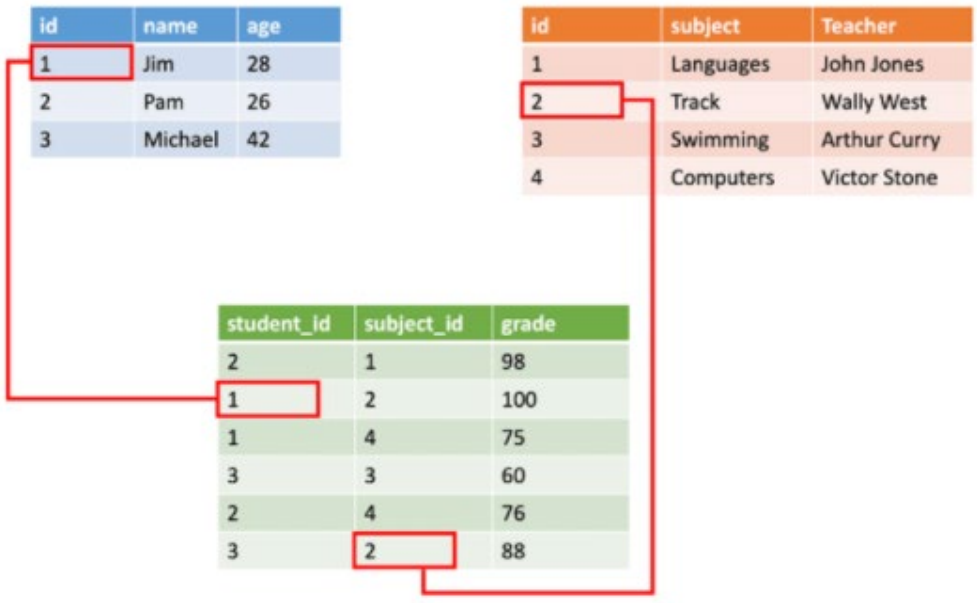
Question 85: Skipped

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

[A] data is typically tabular data that is represented by rows and columns in a database. Databases that hold tables in this form are called [B] databases.

- [A] Relational, [B] Structured
- [A] JSON, [B] Semi-Structured
- [A] Unstructured, [B] Binary
- [A] Structured, [B] Relational
  **(Correct)**

**Explanation**
Structured data is typically tabular data that is represented by rows and columns in a database. Databases that hold tables in this form are called *relational databases* (the mathematical term *relation* refers to an organized set of data held as a table).




https://f5a395285c.nxcli.net/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data

Question 86: Skipped

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

Bruce has come to you because his IT team needs advice on which API to use for the database model and type based on the following information.

**Specifications:**

• The application uses a NoSQL database to store data

• The database uses the key-value and wide-column NoSQL database type.

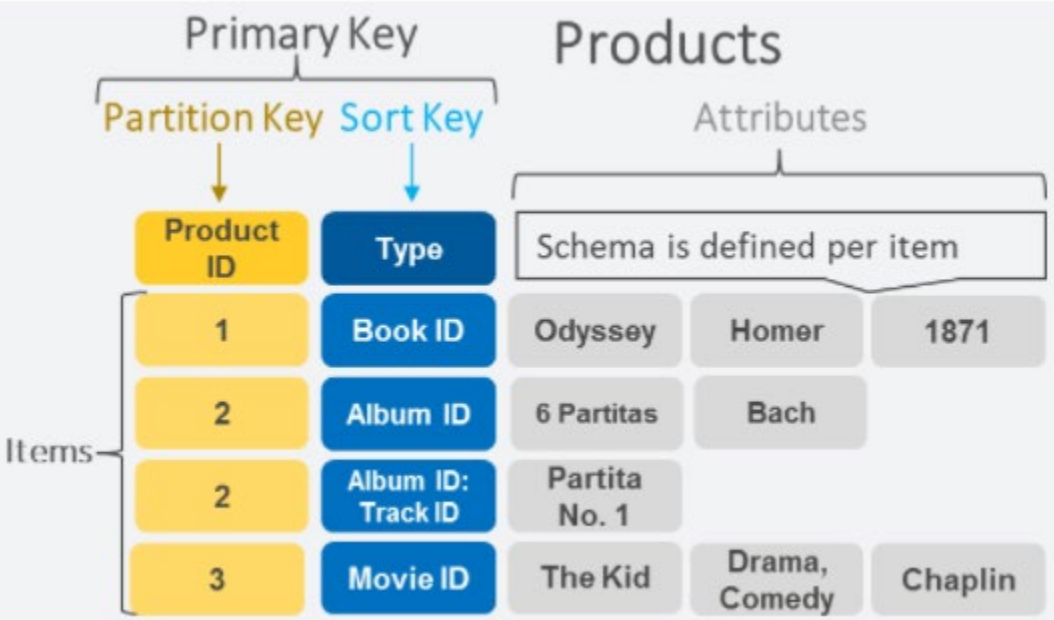
**Required:** Developers need to access data in the database using an API.

Which of the following APIs should you recommend to Bruce and his team?

- ☐ Table API
  **(Correct)**
- ☐ Cassandra API
  **(Correct)**
- ☐ Gremlin API
- ☐ MongoDB API
- ☐ SQL API

**Explanation**

A **key-value database** is a type of nonrelational database that uses a simple key-value method to store data. A key-value database stores data as a collection of key-value pairs in which a key serves as a unique identifier. Both keys and values can be anything, ranging from simple objects to complex compound objects. Key-value databases are highly partitionable and allow horizontal scaling at scales that other types of databases cannot achieve.




https://aws.amazon.com/nosql/key-value/

Azure Cosmos DB provides the Table API for applications that are written for Azure Table storage and that need premium capabilities like:

• Turnkey global distribution.

• Dedicated throughput worldwide (when using provisioned throughput).

• Single-digit millisecond latencies at the 99th percentile.

• Guaranteed high availability.

• Automatic secondary indexing.

Applications written for Azure Table storage can migrate to Azure Cosmos DB by using the Table API with no code changes and take advantage of premium capabilities. The Table API has client SDKs available for .NET, Java, Python, and Node.js.

∴ The database uses the key-value → Table API

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

**Wide Column Databases**, or Column Family Databases, refers to a category of NoSQL databases that works well for storing enormous amounts of data that can be collected. Its architecture uses persistent, sparse matrix, multi-dimensional mapping (row-value, column-value, and timestamp) in a tabular format meant for massive scalability (over and above the petabyte scale). Column Family stores do not follow the relational model, and they aren't optimized for joins.

Good Wide Column Database use cases include:

• Sensor Logs [Internet of Things (IOT)]

• User preferences

• Geographic information

• Reporting systems

• Time Series Data

• Logging and other write heavy applications

Wide Column Databases are not the preferred choice for applications with ad-hoc query patterns, high level aggregations and changing database requirements. This type of data store does not keep good data lineage.

https://www.dataversity.net/wide-column-database/

Azure Cosmos DB Cassandra API can be used as the data store for apps written for Apache Cassandra. This means that by using existing Apache drivers compliant with CQLv4, your existing Cassandra application can now communicate with the Azure Cosmos DB Cassandra API. In many cases, you can switch from using Apache Cassandra to using Azure Cosmos DB's Cassandra API, by just changing a connection string.

The Cassandra API enables you to interact with data stored in Azure Cosmos DB using the Cassandra Query Language (CQL) , Cassandra-based tools (like cqlsh) and Cassandra client drivers that you're already familiar with.

Wide-column stores store data together as columns instead of rows and are optimized for queries over large datasets. The most popular are Cassandra and HBase.

∴ The database uses wide-column NoSQL database type → Cassandra API

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

**Summary**

Key-value databases → Table API

Columnar databases → Cassandra API

Graph databases → Gremlin API

Document databases → SQL API MongoDB API

---

Question 87: Skipped

Within the context of Azure Databricks, sharing data from one worker to another can be a costly operation.

Spark has optimized this operation by using a format called [?] which prevents the need for expensive serialization and de-serialization of objects in order to get data from one JVM to another.

- Stage boundary
- Pipelining
- Tungsten **(Correct)**
- Stages
- Lineage
- Shuffles

**Explanation**
As opposed to narrow transformations, wide transformations cause data to shuffle between executors. This is because a wide transformation requires sharing data across workers. **Pipelining** helps us optimize our operations based on the differences between the two types of transformations.

**Pipelining**

• Pipelining is the idea of executing as many operations as possible on a single partition of data.

• Once a single partition of data is read into RAM, Spark will combine as many narrow operations as it can into a single **Task**

• Wide operations force a shuffle, conclude a stage, and end a pipeline.

**Shuffles**

A shuffle operation is triggered when data needs to move between executors.

To carry out the shuffle operation Spark needs to:

• Convert the data to the `UnsafeRow`, commonly referred to as **Tungsten Binary Format**.

• Write that data to disk on the local node - at this point the slot is free for the next task.

• Send that data across the wire to another executor

• Technically the Driver decides which executor gets which piece of data.

• Then the executor pulls the data it needs from the other executor's shuffle files.

• Copy the data back into RAM on the new executor

• The concept, if not the action, is just like the initial read "every“ DataFrame starts with.

• The main difference being it's the 2nd+ stage.

As we will see in a moment, this amounts to a free cache from what is effectively temp files.

Some actions induce in a shuffle. Good examples would include the operations count() and reduce(..).

**UnsafeRow (also known as Tungsten Binary Format)**

Sharing data from one worker to another can be a costly operation.

**Spark has optimized this operation by using a format called Tungsten.**

Tungsten prevents the need for expensive serialization and de-serialization of objects in order to get data from one JVM to another.

The data that is "shuffled" is in a format known as `UnsafeRow`, or more commonly, the Tungsten Binary Format.

`UnsafeRow` is the in-memory storage format for Spark SQL, DataFrames & Datasets.

Advantages include:

• Compactness:

• Column values are encoded using custom encoders, not as JVM objects (as with RDDs).

• The benefit of using Spark 2.x's custom encoders is that you get almost the same compactness as Java serialization, but significantly faster encoding/decoding speeds.

• Also, for custom data types, it is possible to write custom encoders from scratch.

• Efficiency: Spark can operate *directly out of Tungsten*, without first deserializing Tungsten data into JVM objects.

**How UnsafeRow works**

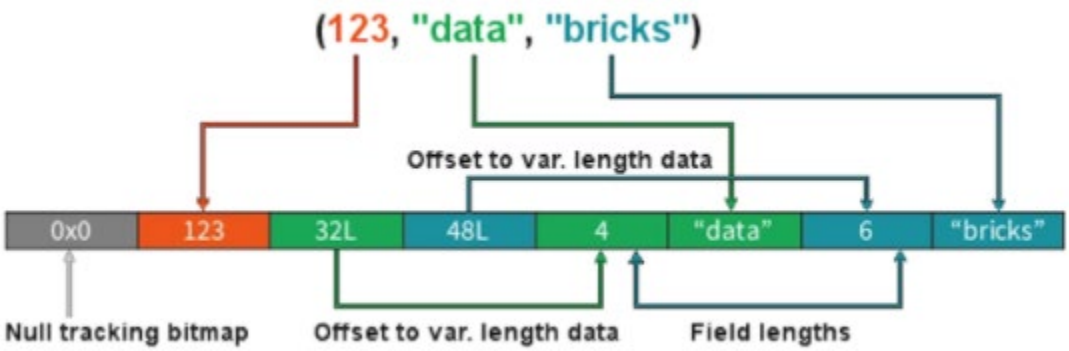
• The first field, "123", is stored in place as its primitive.

• The next 2 fields, "data" and "bricks", are strings and are of variable length.

• An offset for these two strings is stored in place (32L and 48L respectively shown in the picture below).

• The data stored in these two offset's are of format "length + data".

• At offset 32L, we store 4 + "data" and likewise at offset 48L we store 6 + "bricks".




**Stages**

• When we shuffle data, it creates what is known as a stage boundary.

• Stage boundaries represent a process bottleneck.

Take for example the following transformations:

**Step Transformation**

1 Read

2 Select

3 Filter

4 GroupBy

5 Select

6 Filter

7 Write

Spark will break this one job into two stages (steps 1-4b and steps 4c-7):

**Stage #1**

**Step Transformation**

1 Read

2 Select

3 Filter

4a GroupBy 1/2

4b shuffle write

**Stage #1**

Step Transformation

4c shuffle read

4d GroupBy 2/2

5 Select

6 Filter

7 Write

In **Stage #1**, Spark will create a pipeline of transformations in which the data is read into RAM (Step #1), and then perform steps #2, #3, #4a & #4b

All partitions must complete **Stage #1** before continuing to **Stage #2**

• It's not possible to group all records across all partitions until every task is completed.

• This is the point at which all the tasks must synchronize.

• This creates our bottleneck.

• Besides the bottleneck, this is also a significant performance hit: disk IO, network IO and more disk IO.

Once the data is shuffled, we can resume execution...

For **Stage #2**, Spark will again create a pipeline of transformations in which the shuffle data is read into RAM (Step #4c) and then perform transformations #4d, #5, #6 and finally the write action, step #7.

**Lineage**

From the developer's perspective, we start with a read and conclude (in this case) with a write:

**Step Transformation**

1 Read

2 Select

3 Filter

4 GroupBy

5 Select

6 Filter

7 Write

However, Spark starts with the action (write(..) in this case).

Next, it asks the question, what do I need to do first?

It then proceeds to determine which transformation precedes this step until it identifies the first transformation.

**Step Transformation**

7 Write Depends on #6

6 Filter Depends on #5

5 Select Depends on #4

4 GroupBy Depends on #3

3 Filter Depends on #2

2 Select Depends on #1

1 Read First

**Why Work Backwards?**

**Question:** So what is the benefit of working backward through your action's lineage?

**Answer:** It allows Spark to determine if it is necessary to execute every transformation.

Take another look at our example:

• Say we've executed this once already

• On the first execution, step #4 resulted in a shuffle

• Those shuffle files are on the various executors (src & dst)

• Because the transformations are immutable, no aspect of our lineage can change.

• That means the results of our last shuffle (if still available) can be reused.

**Step Transformation**

7 Write Depends on #6

6 Filter Depends on #5

5 Select Depends on #4

4 GroupBy <<< shuffle

3 Filter don't care

2 Select don't care

1 Read don't care

In this case, what we end up executing is only the operations from **Stage #2**.

This saves us the initial network read and all the transformations in **Stage #1**

**Step Transformation**

1 Read skipped

2 Select skipped

3 Filter skipped

4a GroupBy 1/2 skipped

4b shuffle write skipped

4c shuffle read -

4d GroupBy 2/2 -

5 Select -

6 Filter -

7 Write

**And Caching...**

The reuse of shuffle files (also known as our temp files) is just one example of Spark optimizing queries anywhere it can.

We cannot assume this will be available to us.

Shuffle files are by definition temporary files and will eventually be removed.

However, we cache data to explicitly accomplish the same thing that happens inadvertently with shuffle files.

In this case, the lineage plays the same role. Take for example:

**Step Transformation**

7 Write Depends on #6

6 Filter Depends on #5

5 Select <<< cache

4 GroupBy <<< shuffle files

3 Filter ?

2 Select ?

1 Read ?

In this case we cached the result of the select(..).

We never even get to the part of the lineage that involves the shuffle, let alone Stage #1.

Instead, we pick up with the cache and resume execution from there:

**Step Transformation**

1 Read skipped

2 Select skipped

3 Filter skipped

4a GroupBy 1/2 skipped

4b shuffle write skipped

4c shuffle read skipped

4d GroupBy 2/2 skipped

5a cache read -

5b Select -

6 Filter -

7 Write

https://databricks.com/blog/2015/04/28/project-tungsten-bringing-spark-closer-to-bare-metal.html

---

Question 88: Skipped

Where is the best place to monitor spark pools?

- Azure Monitor from the Azure Portal linked to your Azure Synapse Workspace
- Any of the listed options are equally proficient to monitor spark pools
- Monitor tab in Azure Synapse Studio within your Azure Synapse Workspace **(Correct)**
- Monitor tab in Azure Advisor linked to your Azure Synapse Workspace

**Explanation**

If you want to monitor your spark pools the best place to go to is to navigate to the Monitor tab in Azure Synapse Studio within your Azure Synapse Workspace.

The Monitor hub enables you to view pipeline and trigger runs, view the status of the various integration runtimes that is running, view Apache Spark jobs, SQL requests, and data flow debug activities.

We will focus on the Apache Spark Pools jobs within the Monitor Tab in Azure Synapse Studio that you can access through your workspace. The reason why is, if you want to see the status of a job or activity, it's exactly where you want to go.

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.

If you have created a pipeline, and you ran that pipeline, you can see all the pipeline run activities here. It is also possible to view run details where you can see input and outputs for the activities in the pipeline as well as error messages that might have occurred.

If you automated a pipeline run by setting up automated triggers, you can find the runs here as well. If you would like to create new triggers, schedules, or tumbling windows and event-based triggers to execute a pipeline, it is where you need to go.

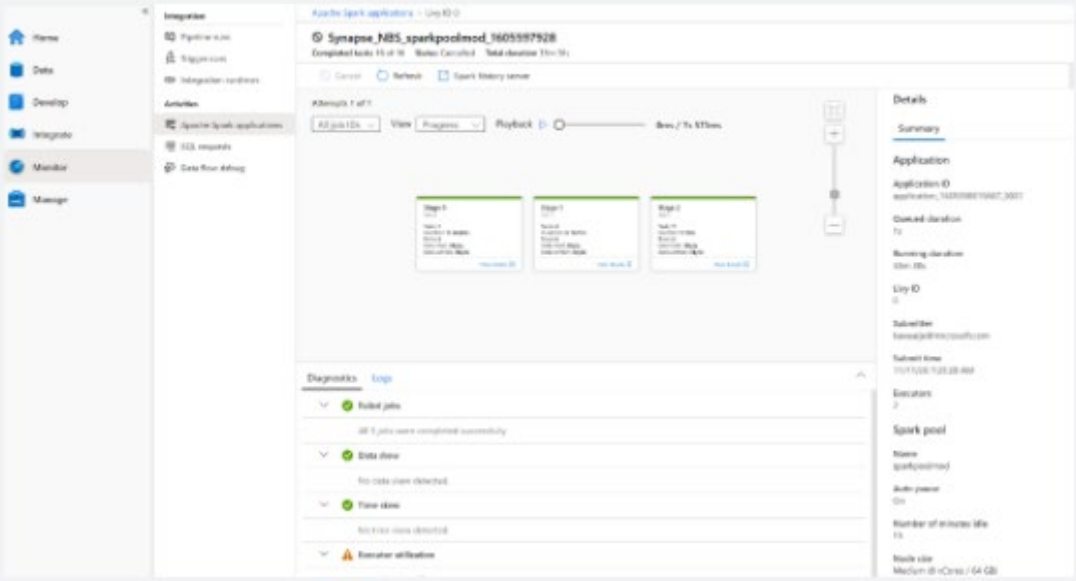
In relation to Apache Spark applications, you are able to see all the Spark applications that are running or have run in your workspace.

Let us deep dive into the monitor tab of the Synapse Studio environment within your Synapse Analytics Workspace.

Let's say you ran some Apache Spark activities, what do you do for monitoring?

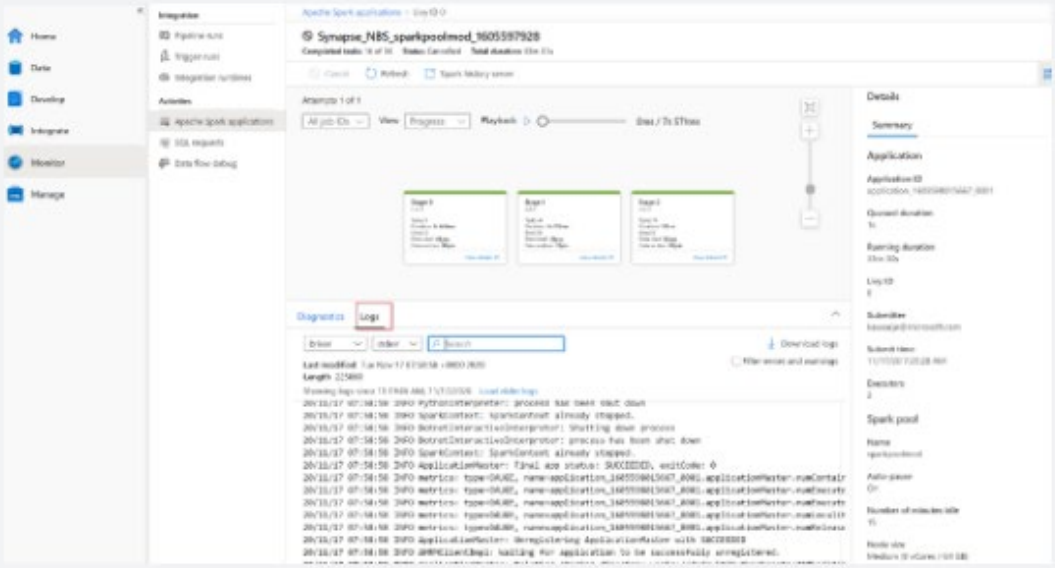
First, you should navigate to Monitor > Activities > Apache Spark applications. It's here where you can see all the Spark applications that are running or have run in your workspace. If you want to find out more about information about a Spark Application that is no longer running, you should click on the application name in the Monitor -> Apache Spark Application tab, name. Here you will find all the details of the spark application.

To give you a visual interpretation of how that looks like. see below:



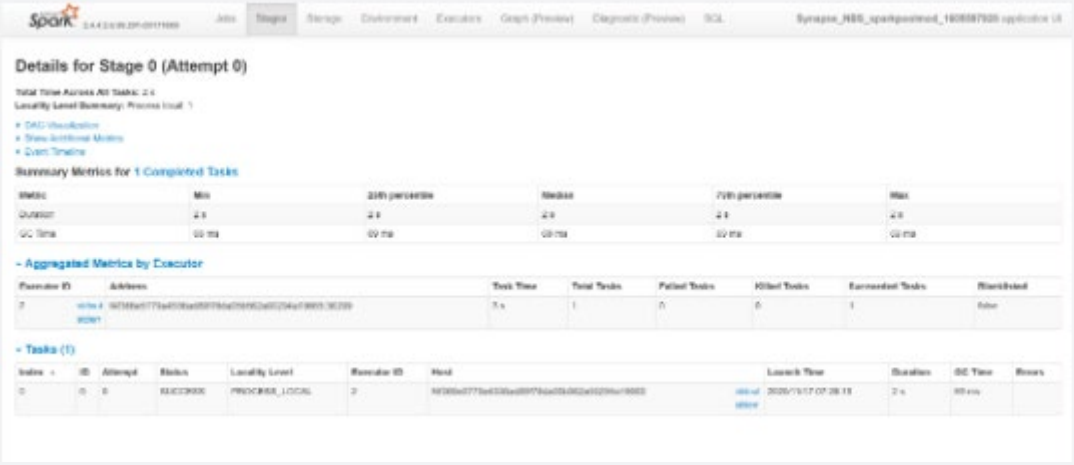

If you are familiar with Apache Spark, you can find the standard Apache Spark history server UI by clicking on Spark history server.

Not only can you check the diagnostics of the Spark application when you run, for example, a notebook attached to a spark pool, you can also check the logs if you navigate to the logs tab:



On the right-hand side, you'll find the details of the spark application as well as the running duration, number of executors, the spark pool details and many more.

In addition to that, if you want to view details for each stage, you can go to the View details tab of one of the stages that looks like the follows:




It will redirect you to the Apache Spark UI where you can find more details in relation to the stages of the Spark Pool.

https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/how-to-monitor-spark-pools

Question 89: Skipped

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

Bruce has come to you because his IT team plans to use Microsoft Azure Synapse Analytics and they are in need of expert guidance.

Wayne Enterprises has an Azure Active Directory (Azure AD) tenant which contains a security group named Group1. They also have an Azure Synapse Analytics dedicated SQL pool named bw1 that contains a schema named schema1.

**Required:**

• Grant Group1 read-only permissions to all the tables and views in schema1.

• The solution must use the principle of least privilege.

Bruce and the team have put together some options they are considering to fulfill the requirements:

a. Create a database role named Role1 and grant Role1 `SELECT` permissions to Schema1

b. Create user from external provider for Group1

c. Assign the Azure role-based access control (RBAC) Reader role for bw1 to Group1

d. Add user to the Role1

e. Create Role1 with `SELECT` on Schema1

f. Create a database user in bw1 that represents Group1 and uses the `FROM EXTERNAL PROVIDER` clause

g. Assign Role1 to the Group1 database use

Which sequence of actions should you recommend to the team to use?

b → e → d → c

- a → d → g → c
- b → e → d
  **(Correct)**
- b → d → c → d

**Explanation**

*The correct sequence of actions should you recommend to the team to use is: b → e → d*

*1. Create user from external provider for Group1*

*2. Create Role1 with select on schema1*

*3. Add user to the Role1*

**Authenticate to dedicated SQL pool (formerly SQL DW) in Azure Synapse Analytics**

To connect to a dedicated SQL pool (formerly SQL DW), you must pass in security credentials for authentication purposes. Upon establishing a connection, certain connection settings are configured as part of establishing your query session.

**SQL authentication**

To connect to dedicated SQL pool (formerly SQL DW), you must provide the following information:

• Fully qualified servername

• Specify SQL authentication

• Username

• Password

• Default database (optional)

By default, your connection connects to the *master* database and not your user database. To connect to your user database, you can choose to do one of two things:

• Specify the default database when registering your server with the SQL Server Object Explorer in SSDT, SSMS, or in your application connection string. For example, include the InitialCatalog parameter for an ODBC connection.

• Highlight the user database before creating a session in SSDT.

**Azure Active Directory authentication**

Azure Active Directory authentication is a mechanism of connecting to SQL pool by using identities in Azure Active Directory (Azure AD). With Azure Active Directory authentication, you can centrally manage the identities of database users and other Microsoft services in one central location. Central ID management provides a single place to manage dedicated SQL pool (formerly SQL DW) users and simplifies permission management.

**Benefits**

Azure Active Directory benefits include:

• Provides an alternative to SQL Server authentication.

• Helps stop the proliferation of user identities across servers.

• Allows password rotation in a single place

• Manage database permissions using external (Azure AD) groups.

• Eliminates storing passwords by enabling integrated Windows authentication and other forms of authentication supported by Azure Active Directory.

• Uses contained database users to authenticate identities at the database level.

• Supports token-based authentication for applications connecting to SQL pool.

• Supports Multi-Factor authentication through Active Directory Universal Authentication for various tools including SQL Server Management Studio and SQL Server Data Tools.

**Configuration steps**

Follow these steps to configure Azure Active Directory authentication.

• Create and populate an Azure Active Directory

• Optional: Associate or change the active directory that is currently associated with your Azure Subscription

• Create an Azure Active Directory administrator for Azure Synapse

• Configure your client computers

• Create contained database users in your database mapped to Azure AD identities

• Connect to your SQL pool by using Azure AD identities

**Find the details**

• The steps to configure and use Azure Active Directory authentication are nearly identical for Azure SQL Database and Synapse SQL in Azure Synapse. Follow the detailed steps in the topic Connecting to SQL Database or SQL Pool By Using Azure Active Directory Authentication.

• Create custom database roles and add users to the roles. Then grant granular permissions to the roles. For more information, see Getting Started with Database Engine Permissions.

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

---

Question 90: Skipped

What is an example of a branching activity used in control flows in Azure Data Factory?

- If-condition
  **(Correct)**
- Lookup- condition
- Where-condition
- Having-condition
- Until-condition

**Explanation**
**If Condition activity in Azure Data Factory**

The If Condition activity provides the same functionality that an if statement provides in programming languages. It executes a set of activities when the condition evaluates to true and another set of activities when the condition evaluates to false.

An example of a branching activity is the If-condition activity which is similar to an if-statement provided in programming languages.

https://docs.microsoft.com/en-us/azure/data-factory/control-flow-if-condition-activity