Q. 1

Init Scripts provide a way to configure cluster's nodes. It is recommended to favour Cluster Scoped Init Scripts over Global and Named scripts.
Which of the following is best described by:

"By placing the init script in /databricks/init folder, you force the script's execution every time any cluster is created or restarted by users of the workspace."

- ○ Cluster Scoped
- ○ Cluster Named
- ○ Interactive
- ◉ Global

Explanation:- Favour cluster scoped init scripts over global and named scripts
Init Scripts provide a way to configure cluster's nodes and to perform custom installs. Init scripts can be used in the following modes:
• Global: by placing the Init script in /databricks/init folder, you force the script's execution every time any cluster is created or restarted by users of the workspace.
• Cluster Named (deprecated): you can limit the init script to run only on for a specific cluster's creation and restarts by placing it in /databricks/init/ folder.
• Cluster Scoped: in this mode, the Init script is not tied to any cluster by its name and its automatic execution is not a virtue of its dbfs location. Rather, you specify the script in cluster's configuration by either writing it directly in the cluster configuration UI or storing it on Databricks File System (DBFS) and specifying the path in Cluster Create API. Any location under DBFS /databricks folder except /databricks/init can be used for this purpose, such as: /databricks//set-env-var.sh
You should treat Init scripts with extreme caution because they can easily lead to intractable cluster launch failures. If you really need them, please use the Cluster Scoped execution mode as much as possible because:
• ADB executes the script's body in each cluster node. Thus, a successful cluster launch and subsequent operation are predicated on all nodal Init scripts executing in a timely manner without any errors and reporting a zero exit code. This process is highly error prone, especially for scripts downloading artifacts from an external service over unreliable and/or misconfigured networks.
• Because Global and Cluster Named Init scripts execute automatically due to their placement in a special DBFS location, it is easy to overlook that they could be causing a cluster to not launch. By specifying the Init script in the Configuration, there's a higher chance that you'll consider them while debugging launch failures.
Use cluster log delivery feature to manage logs
By default, Cluster logs are sent to default DBFS but you should consider sending the logs to a blob store location under your control using the Cluster Log Delivery feature. The Cluster Logs contain logs emitted by user code, as well as Spark framework's Driver and Executor logs. Sending them to a blob store controlled by yourself is recommended over default DBFS location because:
• ADB's automatic 30-day default DBFS log purging policy might be too short for certain compliance scenarios. A blob store location in your subscription will be free from such policies.
• You can ship logs to other tools only if they are present in your storage account and a resource group governed by you. The root DBFS, although present in your subscription, is launched inside a Microsoft Azure managed resource group and is protected by a read lock. Because of this lock, the logs are only accessible by privileged Azure Databricks framework code. However, constructing a pipeline to ship the logs to downstream log analytics tools requires logs to be in a lock-free location first.
https://github.com/Azure/AzureDatabricksBestPractices/blob/master/toc.md

Q. 2

Dynamic Management Views provide a programmatic experience for monitoring the Azure Synapse Analytics SQL pool activity by using the Transact-SQL language.

What type of information or assistance do the views provide? (Select all that apply)

- ○ SQL execution requests and queries
- ○ Data movement service activity
- ○ Troubleshoot workload performance bottlenecks
- ○ Resource blocking and locking activity
- ○ Connection information and activity
- ◉ All of these

Explanation:- Dynamic Management Views provide a programmatic experience for monitoring the Azure Synapse Analytics SQL pool activity by using the Transact-SQL language. The views that are provided, not only enable you to troubleshoot and identify performance bottlenecks with the workloads working on your system, but they are also used by other services such as Azure Advisor to provide recommendations about Azure Synapse Analytics.
There are over 90 Dynamic Management Views that can queried against dedicated SQL pools to retrieve information about the following areas of the service:
• Connection information and activity
• SQL execution requests and queries
• Index and statistics information
• Resource blocking and locking activity
• Data movement service activity
• Errors
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor

Report Error

Q. 3 What can cause a slower performance on join or shuffle jobs?

- ○ Use the cache option
- ○ Bucketing
- ○ Enablement of autoscaling
- ◉ Data skew

Explanation:- The data skew is one of the most common reasons why your Apache Spark job is underperforming. Data skew can cause a slower performance on join or shuffle jobs due to asymmetry in your job data.
Spark is a distributed system, and as such, it divides the data into multiple pieces, called partitions, moves them into the different cluster nodes, and processes them in parallel. If one of these partitions happens to be much larger than others, the node processing it may experience the resource issues and slow down entire execution. This kind of data imbalance is called a data skew.
The size of the partitions depends on the factors, like partitioning configuration of the source files, the number of CPU cores and the nature of your query. The most common scenarios, involving the data skew problems, include the aggregation and join queries, where the grouping or joining field has unequally distributed keys (i.e. few keys have much more rows, than the remaining keys). In this scenario, Spark will send the rows with the same key to the same partition and cause data skew issues.
A traditional Apache Spark UI has some dashboards to determine data skew issues. In addition to that, Azure Synapse Analytics introduced nice data skew diagnosis tools.
https://www.mssqltips.com/sqlservertip/6747/azure-synapse-analytics-analyze-data-skew-issues/

Report Error

Q. 4

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

When a table is created, by default the data structure has no indexes and is called a(n) [?].

- (●) Heap

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

- ( ) NoMap object
- ( ) N-tree
- ( ) Open table

Report Error

Q. 5

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

Azure provides many ways to store your data. A storage account is a(n) [?] that groups a set of Azure Storage services together.

- ○ Structured dataset
- ◉ Container

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.

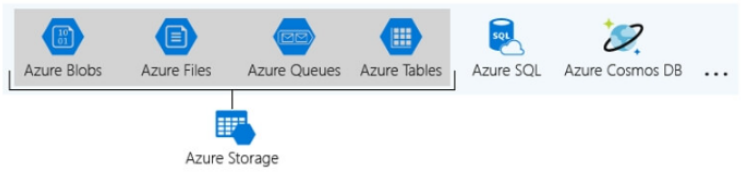
What is a storage account?

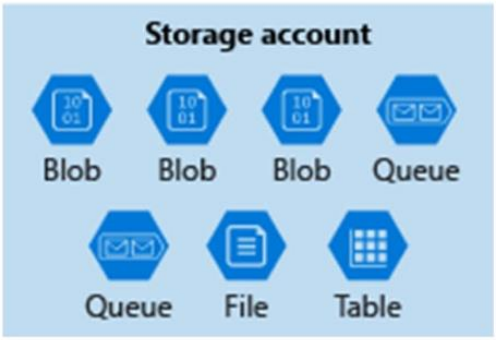
A storage account is a container that groups a set of Azure Storage services together. Only data services from Azure Storage can be included in a storage account (Azure Blobs, Azure Files, Azure Queues, and Azure Tables). The following illustration shows a storage account containing several data services.

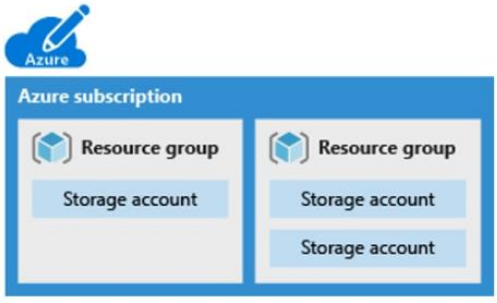
Combining data services into a storage account lets you manage them as a group. The settings you specify when you create the account, or any that you change after creation, are applied to everything in the account. Deleting the storage account deletes all of the data stored inside it.

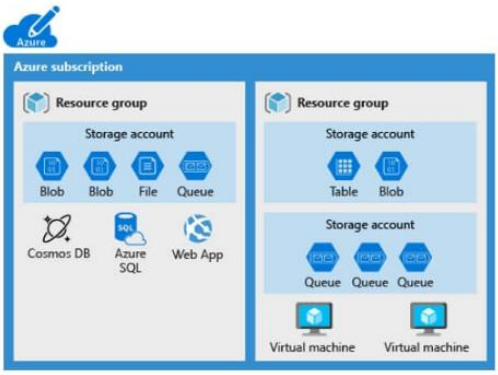
A storage account is an Azure resource and is included in a resource group. The following illustration shows an Azure subscription containing multiple resource groups, where each group contains one or more storage accounts.

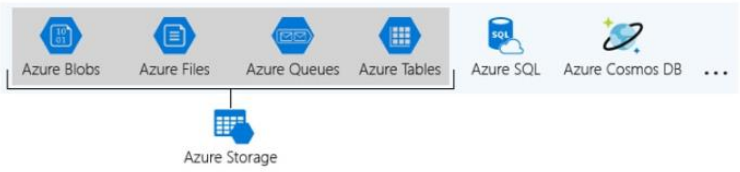
Other Azure data services like Azure SQL and Azure Cosmos DB are managed as independent Azure resources and cannot be included in a storage account. The following illustration shows a typical arrangement: Blobs, Files, Queues, and Tables are inside storage accounts, while other services are not.

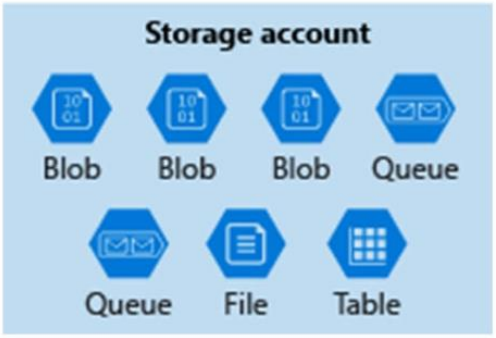


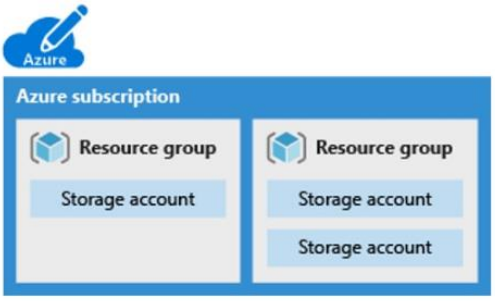


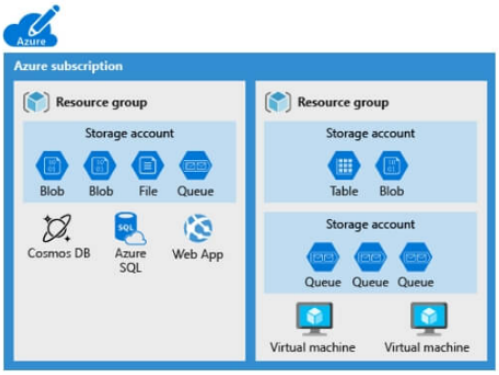



Azure
Azure subscription
Resource group
Storage account
Blob
Blob
File
Queue
Cosmos DB
Azure SQL
Web App
Resource group
Storage account
Table
Blob
Storage account
Queue
Queue
Queue
Virtual machine
Virtual machine


Azure Blobs
Azure Files
Azure Queues
Azure Tables
Azure SQL
Azure Cosmos DB
...
Azure Storage


Storage account
Blob
Blob
Blob
Queue
Queue
File
Table


Azure
Azure subscription
Resource group
Storage account
Resource group
Storage account
Storage account



- ○ VM
- ○ Unstructured dataset
- ○ Blob

Report Error

Q. 6

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

Transactional databases are often called [?] systems. These systems commonly support lots of users, have quick response times, and handle large volumes of data.

- ○ Automated Data Processing Structured (ADPS)
- ○ Extract, transform, and load (ETL)
- ○ OLAP (Online Analytical Processing)
- ○ Atomicity, Consistency, Isolation, and Durability (ACID)
- ◉ OLTP (Online Transaction Processing)

Explanation:- A transaction is a logical group of database operations that execute together.
Here's the question to ask yourself regarding whether you need to use transactions in your application: Will a change to one piece of data in your dataset impact another? If the answer is yes, then you'll need support for transactions in your database service.
Transactions are often defined by a set of four requirements, referred to as ACID guarantees. ACID stands for Atomicity, Consistency, Isolation, and Durability:
• Atomicity means a transaction must execute exactly once and must be atomic; either all of the work is done, or none of it is. Operations within a transaction usually share a common intent and are interdependent.
• Consistency ensures that the data is consistent both before and after the transaction.
• Isolation ensures that one transaction is not impacted by another transaction.
• Durability means that the changes made due to the transaction are permanently saved in the system. Committed data is saved by the system so that even in the event of a failure and system restart, the data is available in its correct state.
When a database offers ACID guarantees, these principles are applied to any transactions in a consistent manner.
OLTP vs OLAP
Transactional databases are often called OLTP (Online Transaction Processing) systems. OLTP systems commonly support lots of users, have quick response times, and handle large volumes of data. They are also highly available (meaning they have very minimal downtime), and typically handle small or relatively simple transactions.
On the contrary, OLAP (Online Analytical Processing) systems commonly support fewer users, have longer response times, can be less available, and typically handle large and complex transactions.
The terms OLTP and OLAP aren't used as frequently as they used to be, but understanding them makes it easier to categorize the needs of your application.
Now that you're familiar with transactions, OLTP, and OLAP, let's walk through each of the data sets in the online retail scenario, and determine the need for transactions.
https://www.guru99.com/oltp-vs-olap.html

- ○ Extract, load, and transform (ELT)

Report Error

Q. 7

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

[?] is typically used to automate the process of extracting, transforming, and loading the data through a batch process against structured and unstructured data sources.

- ○ Azure Conductor
- ◉ Azure Data Factory

Explanation:- Modern Data Warehouse workloads:
A Modern Data Warehouse is a centralized data store that provides descriptive analytics and decision support services across the whole enterprise using structured, unstructured, or streaming data sources. Data flows into the warehouse from multiple transactional systems, relational databases, and other data sources on a periodic basis. The stored data is used for historical and trend analysis reporting. The data warehouse acts as a central repository for many subject areas and contains the "single source of truth."
Azure Data factory is typically used to automate the process of extracting, transforming, and loading the data through a batch process against structured and unstructured data sources.
Advanced Analytical Workloads
You can perform advanced analytics in the form of predictive or preemptive analytics using a range of Azure data platform services. Azure Data Factory provides the integration from source systems into a Data Lake store, and can initiate compute resources such as Azure Databricks, or HDInsight to use the data to perform the advanced analytical work
https://cloudblogs.microsoft.com/industry-blog/en-gb/technetuk/2020/08/25/data-orchestration-with-azure-data-factory/

- ○ Azure Stored Procedure
- ○ Azure Orchestrator
- ○ Azure Designer
- ○ Azure PowerShell

Report Error

Q. 8

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

What does CI/CD focus on?

◯ Practices

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

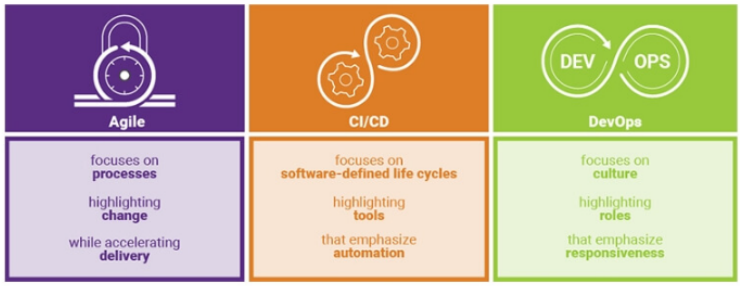
Continuous delivery builds on top of continuous integration to ensure you can successfully release new changes in a fast and consistent way. This is because, in addition to the automated builds and testing provided by continuous integration, the release process is automated to the point where you can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



- ◉ Development process
- ○ Culture
- ○ Strategy

Report Error

Q. 9

Microsoft Azure Storage is a managed service that provides durable, secure, and scalable storage in the cloud. You can create an Azure storage account using the Azure Portal, Azure PowerShell, or Azure CLI. Azure Storage provides three distinct account options with different pricing and features supported.

Which of the Azure Storage account options is best described by:
"Support all of the latest features for blobs, files, queues, and tables. Pricing has been designed to deliver the lowest per gigabyte prices."

- ○ Block
- ○ Page
- ◉ GPv2

Explanation:- Create a storage account
You can create an Azure storage account using the Azure portal, Azure PowerShell, or Azure CLI. Azure Storage provides three distinct account options with different pricing and features supported.
General-purpose v1 (GPv1)
General-purpose v1 (GPv1) accounts provide access to all Azure Storage services but may not have the latest features or the lowest per gigabyte pricing. For example, cool storage and archive storage are not supported in GPv1. Pricing is lower for GPv1 transactions, so workloads with high churn or high read rates may benefit from this account type.
General-purpose v2 (GPv2)
General-purpose v2 (GPv2) accounts are storage accounts that support all of the latest features for blobs, files, queues, and tables. Pricing for GPv2 accounts has been designed to deliver the lowest per gigabyte prices.
Blob storage accounts
A legacy account type, blob storage accounts support all the same block blob features as GPv2, but they are limited to supporting only block and append blobs. Pricing is broadly similar to pricing for general-purpose v2 accounts.
https://docs.microsoft.com/en-us/azure/storage/common/storage-account-overview

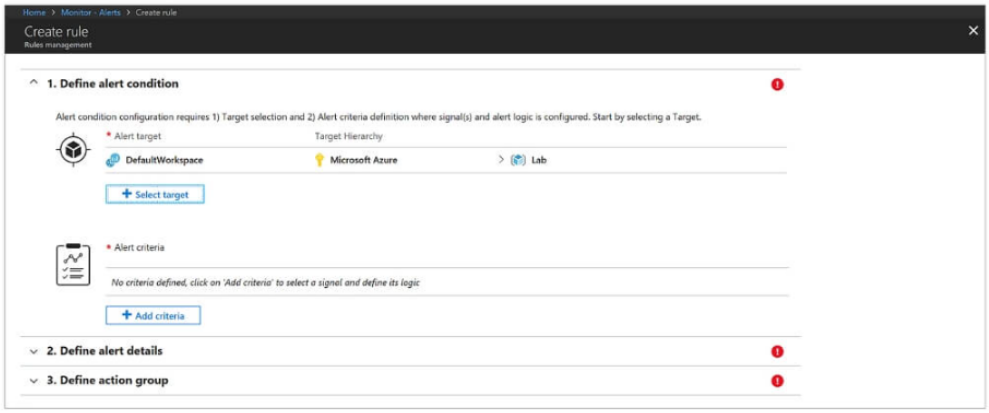
- Append
- Queue
- Blob storage accounts

Report Error

Q. 10 Which feature in alerts can be used to determine how an alert is fired?

- Add severity
- Add rule
- Add criteria

Explanation:- Azure Data Factory Alerts provide an automated response that can be beneficial to monitor and audit Azure Data Factory activity. These alerts are very proactive and more efficient than manual monitoring operations. Alerts can be fired on both success and failure of a pipeline based on the rule configuration.
Alert Rule
Azure Data Factory Alerts use an alert rule which states the criteria upon which the alerts should trigger. We can enable or disable the alert rules.
• The add criteria feature enables you to determine how an alert is fired.
https://docs.microsoft.com/en-us/azure/azure-monitor/alerts/tutorial-response



- Add specifications

Report Error

Q. 11

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

A(n) [?] schema must be defined before query time.

- ○ Hybrid data type
- ◉ Structured data type

Explanation:- Structured data
In relational database systems like Microsoft SQL Server, Azure SQL Database, and Azure SQL Data Warehouse, data structure is defined at design time. Data structure is designed in the form of tables. This means it's designed before any information is loaded into the system. The data structure includes the relational model, table structure, column width, and data types.
Relational systems react slowly to changes in data requirements because the structural database needs to change every time a data requirement changes. When new columns are added, you might need to bulk-update all existing records to populate the new column throughout the table.
Relational systems typically use a querying language such as Transact-SQL (T-SQL).
https://k21academy.com/microsoft-azure/dp-900/relational-and-non-relational-datastores/

- ○ Azure Cosmos DB data type
- ○ Unstructured data type

Report Error

Q. 12

Spark pools in Azure Synapse Analytics is one of Microsoft's implementation of Apache Spark.

Which of the following are true about Spark pools in Azure Synapse Analytics? (Select all that apply)

- ○ Once connected, Sparkle gets the executors on nodes in the pool. Those processes run computations and store data on your local machine.
- ○ The SparkContext is able to connect to the cluster manager, which allocates resources across applications. The cluster manager is Adobe Hadoop WOOL.
- ◉ Spark applications act as independent sets of processes on a pool. It is coordinated by the SParkContext object in a main (driver) program

Explanation:-
Apache Spark in Azure Synapse Analytics
Spark pools in Azure Synapse Analytics is one of Microsoft's implementation of Apache Spark, version Spark 2.4 for the Azure cloud.
Azure Synapse Analytics enables you to have a one-stop shop for your Analytics environment. With the addition of Spark Pools in Azure Synapse Analytics, it is now also possible to benefit from the features of Apache Spark in the same environment where you can set up your data warehousing solution. The spark pools within Azure Synapse Analytics are compatible with different Azure Storage solutions such as ADLS Gen2 and Blob Storage. It is imperative to know that currently providing Spark pools in an Azure Synapse Analytics workspace preview environment, is provided without a service level agreement and therefore not (yet) recommended for production workloads. In addition, some of the official Apache Spark documentation relies on using the spark console. At this moment, the spark console is not available on Azure Synapse Spark, so therefore it is highly recommended to use the notebook or IntelliJ experiences instead.
Spark Pools in Azure Synapse Analytics, a fully managed and integrated Spark service
Benefits of Spark Pools in Azure Synapse Analytics are listed below:
• Speed and Efficiency: Quick start-up time for nodes, automatic shut-down when instances are not used within 5 min after last job, unless there is a live notebook connection.
• Ease of creation: Creating a spark pool can be done through the Azure portal, PowerShell, or .NET SDK for Azure Synapse Analytics.
• Ease of use: Within the Azure Synapse Analytics workspace, you can connect directly to the Spark pool and interact with the integrated notebook experience, or use custom notebooks derived from Nteract. Notebook integration helps you in developing interactive data processing and visualization pipelines.
• REST APIs: In order to monitor and submit jobs remotely, you can use Apache Livy as Rest API Spark job server.
• Integration with third-party IDEs: Azure Synapse Analytics provides an IDE for IntelliJ to create and submit applications to the spark pool
• Pre-loaded Anaconda libraries: Over 200 Anaconda libraries pre-installed on the spark pool.
• Scalability: Possibility for autoscale, such that pools can be up/down scaled as required by adding or removing nodes.
Spark pools in Azure Synapse include the following components that are available on the pools by default.

• Spark Core. Includes Spark Core, Spark SQL, GraphX, and MLlib.
• Anaconda
• Apache Livy
• Nteract notebook

The supported languages and runtime versions for Apache spark and dependent components in Azure Synapse analytics can be found here:

• Apache Spark components in Azure Synapse Analytics

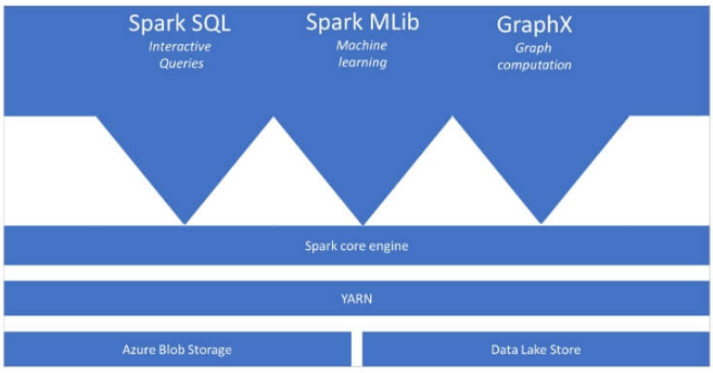
Spark pool architecture

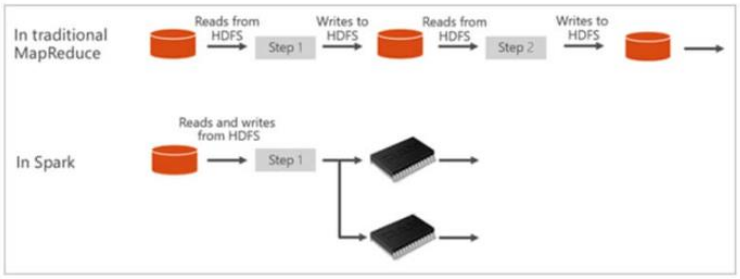
It is imperative to understand the components of Spark by understanding how Spark runs on Synapse Analytics. The different spark applications act as independent sets of processes on a pool. It is coordinated by the SParkContext object in a main (driver) program.

The SparkContext is able to connect to the cluster manager, which allocates resources across applications. The cluster manager is Apache Hadoop YARN. Once connected, Spark gets the executors on nodes in the pool. Those processes run computations and store data for your application. What follows is that your application code (defined by JAR or Python files passed to SparkContext) will be sent to the executors. Finally, SparkContext is able to send tasks to the executors to run.

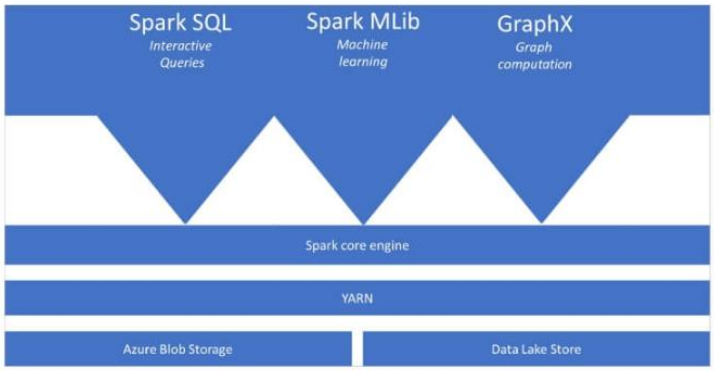
The SparkContext runs the user's so your main function. What is then will do is execute the various parallel operations on the nodes. Then, the SparkContext will collect all the results of the operations that were sent to the nodes. The nodes are able to read and write data from and to the file system. Like mentioned in the introduction, the nodes caches the transformed data in-memory as Resilient Distributed Datasets (RDDs).

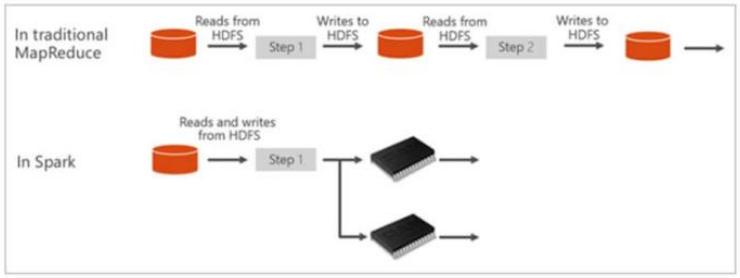
The SparkContext connects to the Spark pool in Synapse Analytics. It is responsible for converting an application to a directed acyclic graph (DAG). The graph consists of individual tasks that get executed within an executor process on the nodes. Each application gets its own executor processes, which stay up for the duration of the whole application and run tasks in multiple threads.

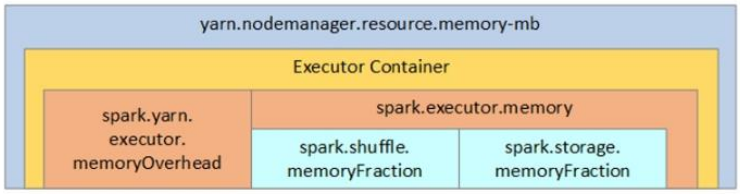




https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-overview







○ The SparkContext connects to the Sparkle pool in Synapse Analytics. It is responsible for converting an application to an Excel file.

Report Error

Q. 13

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

Correct or Incorrect : It is possible to use multiple languages in one notebook.

◉ Correct

Explanation:- Within creating a notebook, you need to specify the pool that needs to be attached to the notebook that is, a SQL or Spark pool. When it comes to the languages, a notebook has to be set with a primary language.
The primary languages available within the notebook environment are:
• PySpark (Python)
• Spark (Scala)
• .NET Spark (C#)
• Spark SQL
However, it is possible to use multiple languages in one notebook by specifying the language using a magic command at the beginning of a cell. The following table lists the magic commands to switch cell languages:
It is not possible to reference data or variables directly across different languages in a Synapse Studio notebook. In Spark, it is possible to reference a temporary table across languages.
https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-development-using-notebooks?tabs=classical

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

○ Incorrect

Q. 14

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

Within Azure Synapse SQL, [?] stores a copy of the result set on the control node so that queries do not need to pull data from the storage subsystem or compute nodes.

- ○ VM caching
- ○ Site caching
- ○ Server caching
- ◉ Result-set caching

Explanation:- Enable result-set caching when you expect results from queries to return the same values.
This option stores a copy of the result set on the control node so that queries do not need to pull data from the storage subsystem or compute nodes. The capacity for the result set cache is 1 TB and the data within the result-set cache is expired and purged after 48 hours of not being accessed.
Azure Synapse SQL automatically caches query results in the user database for repetitive use. Result-set caching allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage.
To enable result set caching, run this command when connecting to the MASTER database.
SQL
ALTER DATABASE [database_name]
SET RESULT_SET_CACHING ON;
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching

- ○ Browser caching

Report Error

Q. 15 Correct or Incorrect : Access keys are the easiest approach to authenticating access to a storage account which provide full access to anything in the storage account, similar to a root password on a computer.

- ◉ Correct

Explanation:- Shared Access Signatures (SAS)
Access keys are the easiest approach to authenticating access to a storage account. However they provide full access to anything in the storage account, similar to a root password on a computer.
Storage accounts offer a separate authentication mechanism called shared access signatures that support expiration and limited permissions for scenarios where you need to grant limited access. You should use this approach when you are allowing other users to read and write data to your storage account. There are links to our documentation on this advanced topic at the end of the module.
https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

- ○ Incorrect

Report Error

Q. 16

Scenario: Queen Consolidated was overtaken by Raymond Carson Palmer and rebranded as Palmer Technologies. Now that Ray is overseeing the operations at Palmer, Ray has decided to move away from on-prem datacentres to Azure. Ray and the IT team are developing a new data engineering solutions for a company.

The current project is dealing with social media and has the following requirements.

Required:
• Real-time Twitter feed analysis of posts which contain specific keywords and must be stored as well as processed on MS Azure then displayed using MS Power BI.

Ray and the IT team have put together a list of actions they think need to be performed to meet the needs of the project, but they are not sure on the order to execute. Below is a list of the actions they are considering.

Proposed Actions:
a. Create an HDInisght cluster with the Spark cluster type.
b. Create a Jupyter Notebook.
c. Run a job that uses the Spark Streaming API to ingest data from Twitter.
d. Create a Runbook.
e. Create an HDInsight cluster with the Spark cluster type.
f. Create a HVAC table.
g. Load the HVAC table into Power BI Desktop

As you are the Azure SME, Ray and the team look to you for direction on selecting the required items and putting them in the proper order.

Which of the below contains the correct items in the correct sequence to meet the requirements?

- (●) a,b,f,c,g

Explanation:- Step 1: Create an HDInisght cluster with the Spark cluster type.
Step 2: Create a Jyputer Notebook.
Step 3: Create HVAC table.
The Jupyter Notebook that you created in the previous step includes code to create an HVAC table.
Step 4: Run a job that uses the Spark Streaming API to ingest data from Twitter.
Step 5: Load the HVAC table into Power BI Desktop.
You use Power BI to create visualizations, reports, and dashboards from the Spark cluster data.
https://www.youtube.com/watch?v=_RJ0VjZ2-og
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-use-with-data-lake-store

- ( ) e,a,c,g,d
- ( ) f,b,d,a,g,c
- ( ) b,a,e,f,c,g

Report Error

Q. 17 What does the APPROX_COUNT_DISTINCT Transact-SQL function do?

- ○ None of the listed options.
- ○ Approximate execution using Hyperlog accuracy
- ◉ Calculates the approximate number of distinct records in a relational database.

Explanation:- APPROX_COUNT_DISTINCT (Transact-SQL)
This function returns the approximate number of unique non-null values in a group.
Syntax
syntaxsql
APPROX_COUNT_DISTINCT (expression)
Arguments
expression
An expression of any type, except image, sql_variant, ntext, or text.
Return types
bigint
Remarks
APPROX_COUNT_DISTINCT(expression) evaluates an expression for each row in a group, and returns the approximate number of unique non-null values in a group. This function is designed to provide aggregations across large data sets where responsiveness is more critical than absolute precision.
APPROX_COUNT_DISTINCT is designed for use in big data scenarios and is optimized for the following conditions:
Access of data sets that are millions of rows or higher and
Aggregation of a column or columns that have many distinct values
The function implementation guarantees up to a 2% error rate within a 97% probability.
APPROX_COUNT_DISTINCT requires less memory than an exhaustive COUNT DISTINCT operation. Given the smaller memory footprint, APPROX_COUNT_DISTINCT is less likely to spill memory to disk compared to a precise COUNT DISTINCT operation. To learn more about the algorithm used to achieve this, see HyperLogLog.
https://docs.microsoft.com/en-us/sql/t-sql/functions/approx-count-distinct-transact-sql?view=sql-server-ver15

- ○ Approximate count on distinct executions within a specified time period on a specific endpoint.
- ○ Calculates the approximate number of distinct records in a non-relational database.

Q. 18

Scenario: You are working as a consultant at Avengers Security and the IT team has developed a data ingestion process to import data to a Microsoft Azure SQL Data Warehouse. They are using an Azure Data Lake Gen 2 storage account to store the data to be ingested. The data to be ingested resides in parquet files.

Required: Load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.
The Avengers IT team has proposed the following solution:
1. Create an external data source pointing to the Azure storage account
2. Create an external file format and external table using the external data source
3. Load the data using the INSERT ... SELECT statement

Will the solution proposed by the Avengers IT team meet the requirement?

- ○ Correct
- ◉ Incorrect

Explanation:- The proposed solution will not meet the requirement. They need to create an external file format and external table using the external data source. To load the data, use the CREATE TABLE ... AS SELECT statement.
Use polybase by defining external tables
Using Transact-SQL, you can use PolyBase to access files that are located directly on Azure Storage as if they were structured tables within your SQL Pool. You define an external data source pointing to the location of the file or the folder the files reside in, the external file format, which can be GZip compressed delimited text, ORC, Parquet or JSON, and then the external table with the column attributes that map to the structure from the external files.
Create an import database
The first step in using PolyBase is to create a database-scoped credential that secures the credentials to the blob storage. Create a master key first, and then use this key to encrypt the database-scoped credential named AzureStorageCredential.
1. Paste the following code into the query window. Replace the SECRET value with the access key you retrieved in the previous exercise.

SQL

```
CREATE MASTER KEY;
CREATE DATABASE SCOPED CREDENTIAL AzureStorageCredential
WITH
IDENTITY = 'demodwStorage',
SECRET = 'THE-VALUE-OF-THE-ACCESS-KEY' -- put key1's value here
;
```

2. Select Run to run the query. It should report Query succeeded: Affected rows: 0.

Create an external data source connection

Use the database-scoped credential to create an external data source named AzureStorage. Note the location URL point to the container named data-files that you created in the blob storage. The type Hadoop is used for both Hadoop-based and Azure Blob storage-based access.

1. Paste the following code into the query window. Replace the LOCATION value with your correct value from the previous exercise.

SQL

```
CREATE EXTERNAL DATA SOURCE AzureStorage
WITH (
TYPE = HADOOP,
LOCATION = 'wasbs://data-files@demodwstorage.blob.core.windows.net',
CREDENTIAL = AzureStorageCredential
);
```

2. Select Run to run the query. It reports Query succeeded: Affected rows: 0..

Define the import file format

Define the external file format named TextFile. This name indicates to PolyBase that the format of the text file is DelimitedText and the field terminator is a comma.

1. Paste the following code into the query window.

SQL

```
CREATE EXTERNAL FILE FORMAT TextFile
WITH (
FORMAT_TYPE = DelimitedText,
FORMAT_OPTIONS (FIELD_TERMINATOR = ',')
);
```

2. Select Run to run the query. It reports Query succeeded: Affected rows: 0..

Create a temporary table

Create an external table named dbo.temp with the column definition for your table. At the bottom of the query, use a WITH clause to call the data source definition named AzureStorage, as previously defined, and the file format named TextFile, as previously defined. The location denotes that the files for the load are in the root folder of the data source.

Note: External tables are in-memory tables that don't persist onto the physical disk. External tables can be queried like any other table.

The table definition must match the fields defined in the input file. There are 12 defined columns, with data types that match the input file data.

1. Add the following code into the Visual Studio window underneath the previous code.

SQL

```
-- Create a temp table to hold the imported data
CREATE EXTERNAL TABLE dbo.Temp (
[Date] datetime2(3) NULL,
[DateKey] decimal(38, 0) NULL,
[MonthKey] decimal(38, 0) NULL,
[Month] nvarchar(100) NULL,
[Quarter] nvarchar(100) NULL,
[Year] decimal(38, 0) NULL,
[Year-Quarter] nvarchar(100) NULL,
[Year-Month] nvarchar(100) NULL,
[Year-MonthKey] nvarchar(100) NULL,
[WeekDayKey] decimal(38, 0) NULL,
[WeekDay] nvarchar(100) NULL,
[Day Of Month] decimal(38, 0) NULL
)
WITH (
LOCATION='../',
DATA_SOURCE=AzureStorage,
FILE_FORMAT=TextFile
);
```

2. Select Run to run the query. It takes a few seconds to complete and reports Query succeeded: Affected rows: 0..

Create a destination table

Create a physical table in the Azure Synapse Analytics database. In the following example, you create a table named dbo.StageDate. The table has a clustered column store index defined on all the columns. It uses a table geometry of round_robin by design because round_robin is the best table geometry to use for loading data.

1. Paste the following code into the query window.

SQL

```
-- Load the data from Azure Blob storage to Azure Synapse Analytics
CREATE TABLE [dbo].[StageDate]
WITH (
CLUSTERED COLUMNSTORE INDEX,
DISTRIBUTION = ROUND_ROBIN
)
AS
SELECT * FROM [dbo].[Temp];
```

2. Select Run to run the query. It takes a few seconds to complete and reports Query succeeded: Affected rows: 0..

Add statistics onto columns to improve query performance

As an optional step, create statistics on columns that feature in queries to improve the query performance against the table.

1. Paste the following code into the query window.

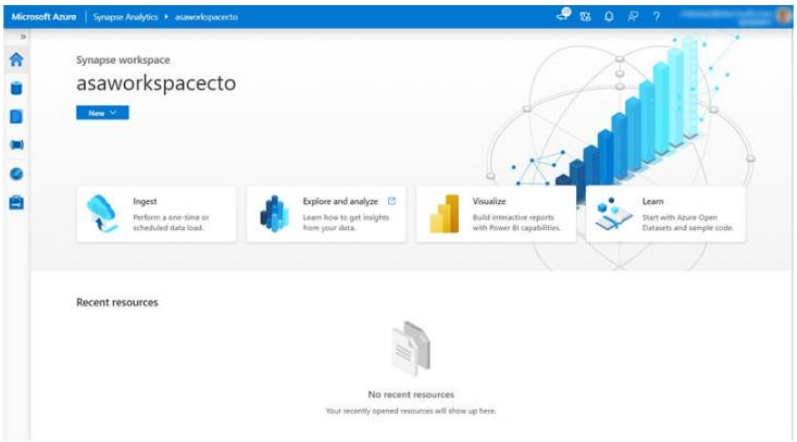
SQL

```
-- Create statistics on the new data
CREATE STATISTICS [DateKey] on [StageDate] ([DateKey]);
CREATE STATISTICS [Quarter] on [StageDate] ([Quarter]);
CREATE STATISTICS [Month] on [StageDate] ([Month]);
```

2. Select Run to run the query. It reports Query succeeded: Affected rows: 0..

You've loaded your first staging table in Azure Synapse Analytics. From here, you can write further Transact-SQL queries to perform transformations into dimension and fact tables. Try it out by querying the StageDate table in the query explorer or in another query tool. Refresh the view on the left to see the new table or tables that you created. Reuse the previous steps in a persistent SQL script to load additional data, as necessary.

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store



Q. 19

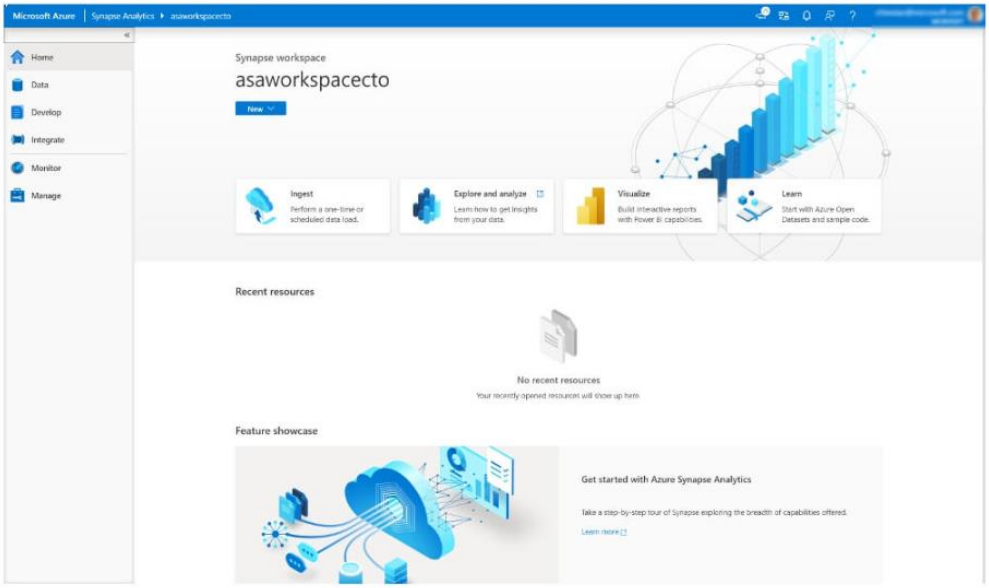
You can integrate your Azure Synapse Analytics workspace with a new Power BI workspace so that you can get you data from within Azure Synapse Analytics visualized in a Power BI report or dashboard.

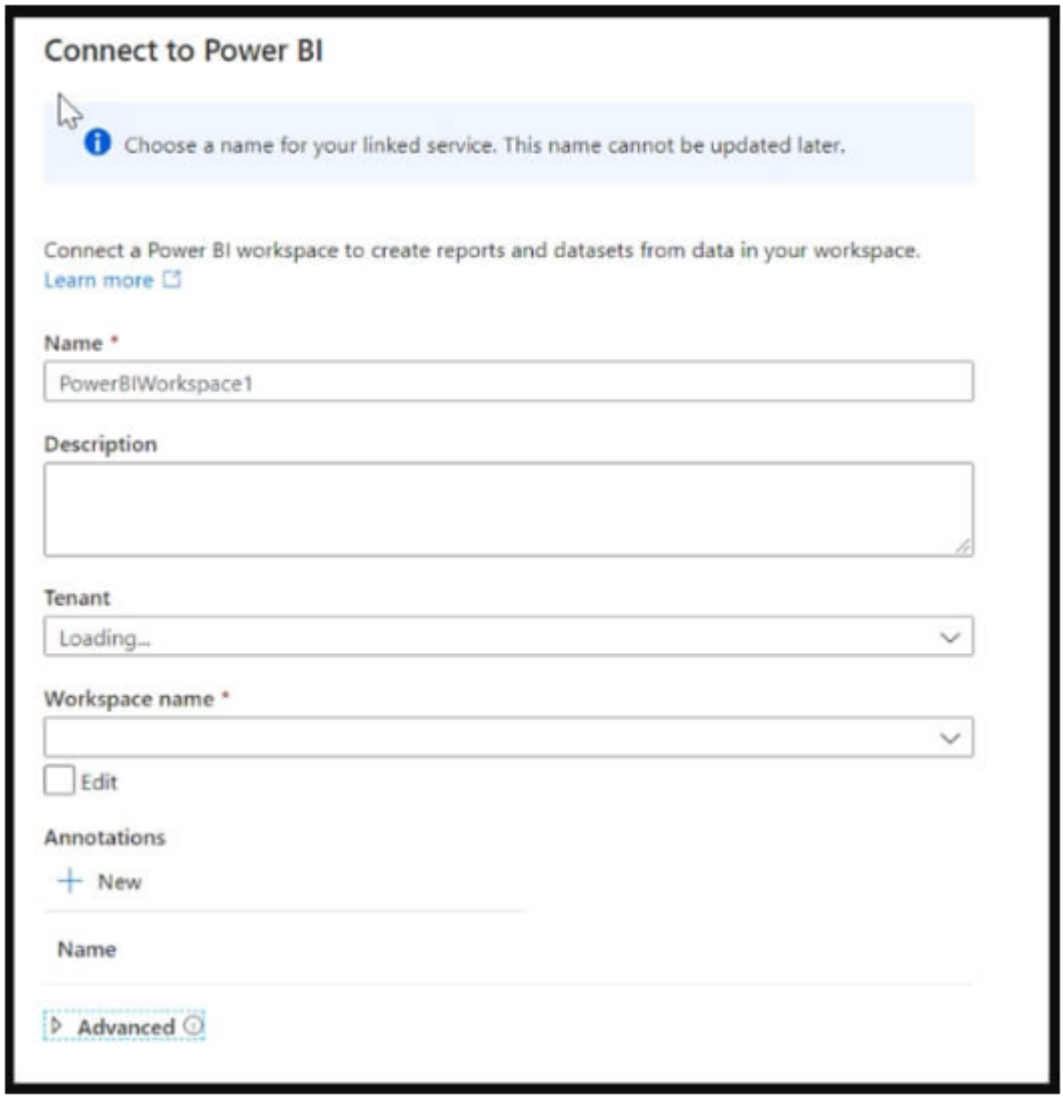
Which icon should you click on the home page of Azure Synapse Studio to begin the integration?

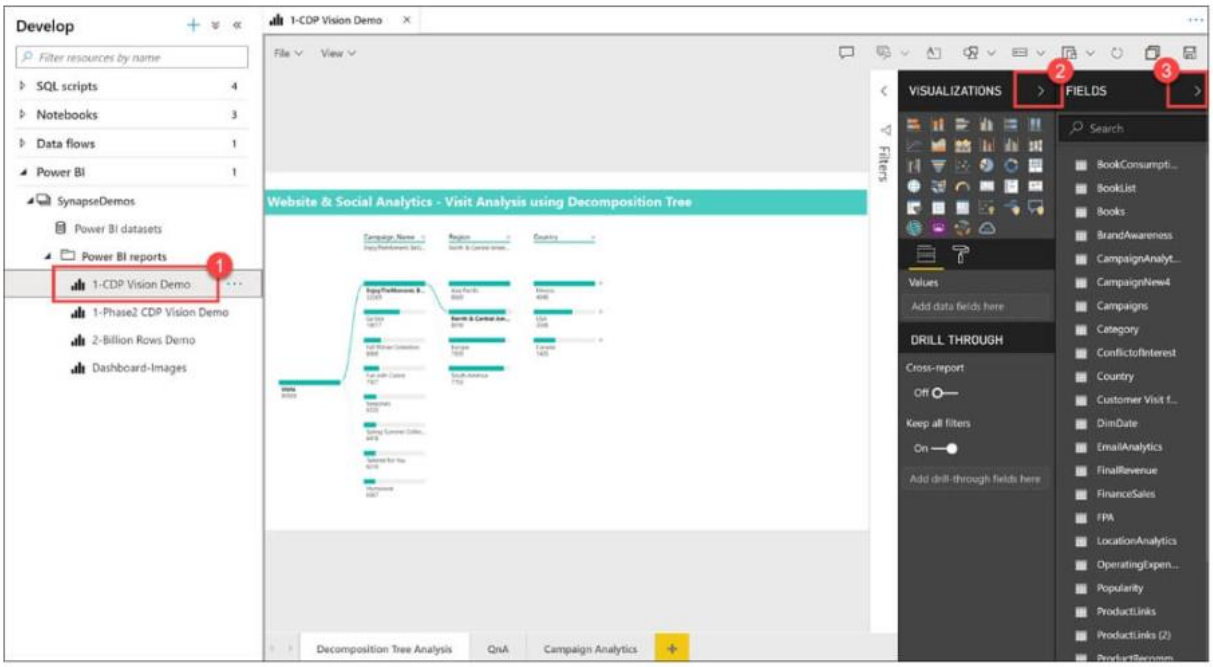
- ○ Explore and analyze
- ◉ None of the listed options

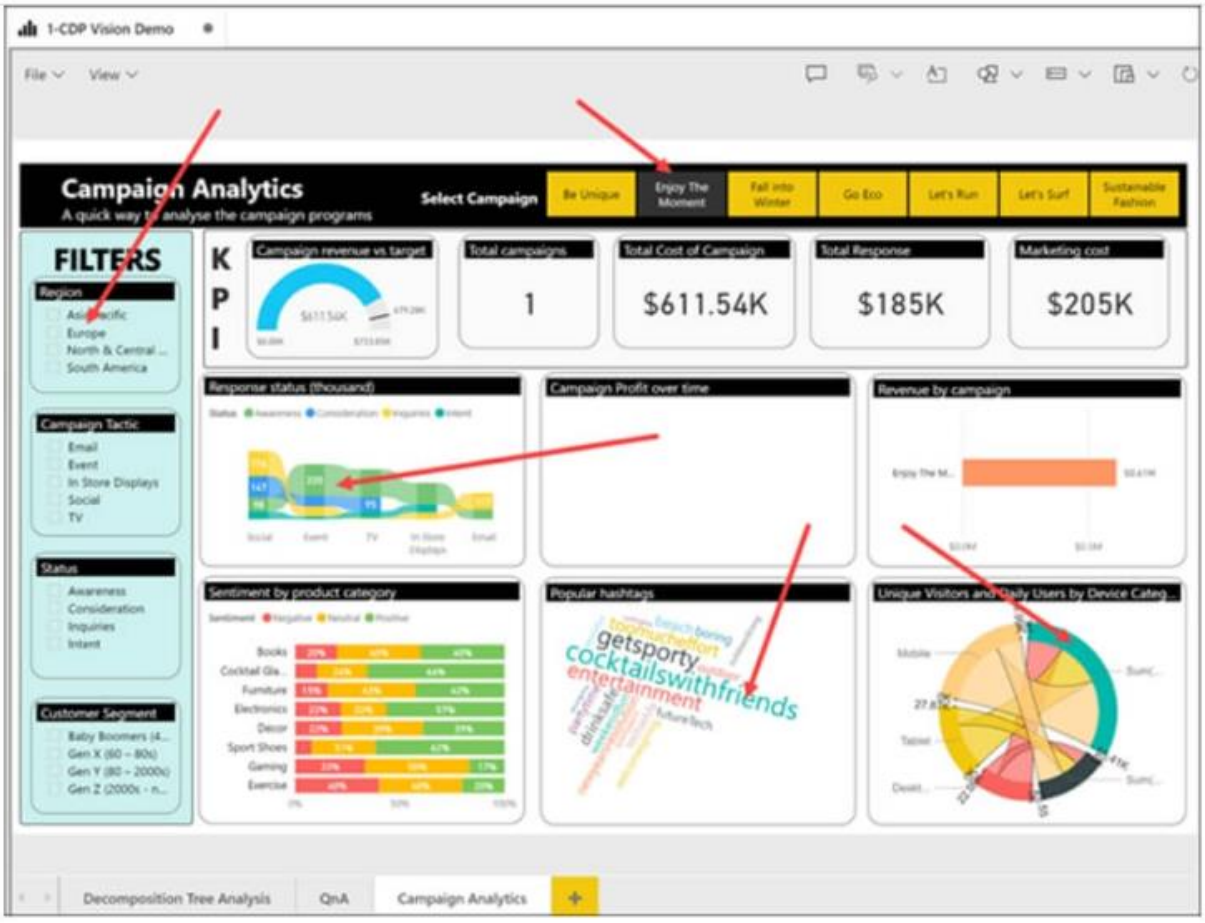
Explanation:-

You can integrate your Azure Synapse Analytics workspace with a new Power BI workspace so that you can get you data from within Azure Synapse Analytics visualized in a Power BI report or dashboard.
You can perform this step by clicking on the visualize icon on the home page of Azure Synapse Studio.
Which will bring up the Connect to Power BI screen.
From here you can define a name and description for the Power BI Workspace. Then you would select the Tenant and Workspace name. Once you have connected to your workspace, you will be able to access the existing reports in the Power BI workspace in the Develop hub in Azure Synapse Studio.
Expand Power BI, expand SynapseDemos, expand Power BI reports, then select 1-CDP Vision Demo (1). Select the arrows to collapse the **Visualizations pane (2) and the Fields pane (3) to increase the report size.
As you can see, we can create, edit, and view Power BI reports from within Synapse Studio! As a business analyst, data engineer, or developer, you no longer need to open another browser window, sign in to Power BI, and toggle back and forth between environments.
Select a Campaign Name and Region within the Decomposition Tree Analysis tab to explore the data. If you hover over an item, you will see a tool tip.
Select the Campaign Analytics tab at the bottom of the report.
The Campaign Analytics report combines data from the various data sources to create a compelling visualization of valuable data within an interactive interface.
You can select various filters, campaigns, and chart values to filter the results. Select an item to for the second time to deselect it.
Select Power BI datasets (1) in the left-hand menu, hover over the 2-Billion Rows Demo dataset and select the New Power BI report icon (2).
Here is how we can create a brand new Power BI report from a dataset that is part of the linked Power BI workspace, from within Synapse Studio.
Expand the Category table, then drag-and-drop the Category field on to the report canvas. This creates a new Table visualization that shows the categories.
Select a blank area on the report canvas to deselect the table, then select the Pie chart visualization.
Expand the ProdChamp table. Drag Campaign onto the Legend field, then drag ProductID onto the Values field. Resize the pie chart and hover over the pie slices to see the tool tips.
We have very quickly created a new Power BI report, using data stored within our Synapse Analytics workspace, without ever leaving the studio.

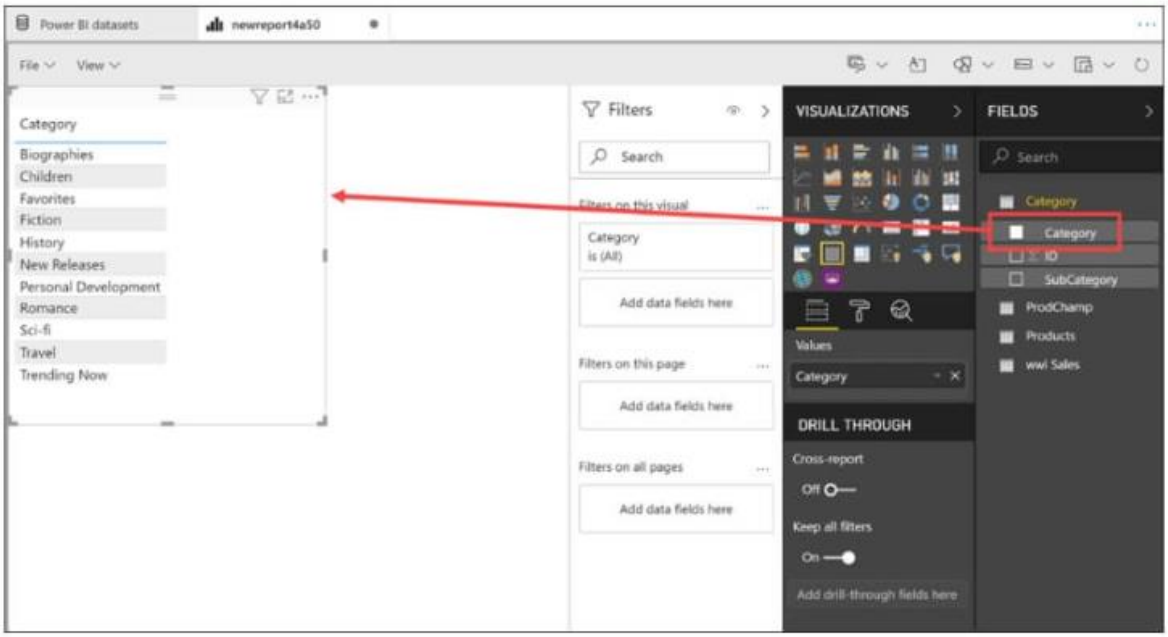


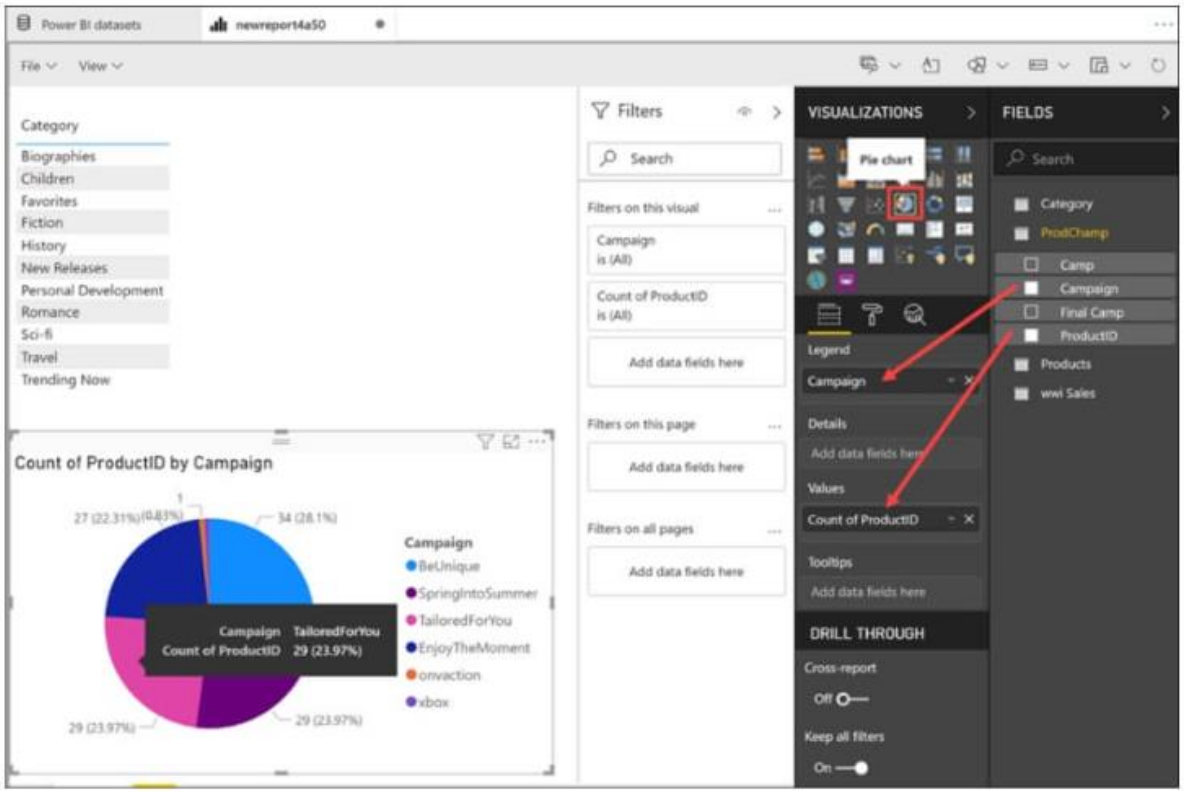

Connect to Power BI
Choose a name for your linked service. This name cannot be updated later.
Connect a Power BI workspace to create reports and datasets from data in your workspace.
Learn more
Name *
PowerBIWorkspace1
Description
Tenant
Loading...
Workspace name *
Edit
Annotations
New
Name
Advanced


Develop
Filter resources by name
SQL scripts 4
Notebooks 3
Data flows 1
Power BI 1
SynapseDemos
Power BI datasets
Power BI reports
1-CDP Vision Demo
1-Phase2 CDP Vision Demo
2-Billion Rows Demo
Dashboard-Images
1
1-CDP Vision Demo
File
View
Website & Social Analytics - Visit Analysis using Decomposition Tree
Decomposition Tree Analysis
QnA
Campaign Analytics
VISUALIZATIONS
2
FIELDS
3
Filters
Values
Add data fields here
DRILL THROUGH
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Search
BookConsumpti...
BookList
Books
BrandAwareness
CampaignAnalyt...
CampaignNew4
Campaigns
Category
ConflictofInterest
Country
Customer Visit f...
DimDate
EmailAnalytics
FinalRevenue
FinanceSales
FPA
LocationAnalytics
OperatingExpen...
Popularity
ProductLinks
ProductLinks (2)

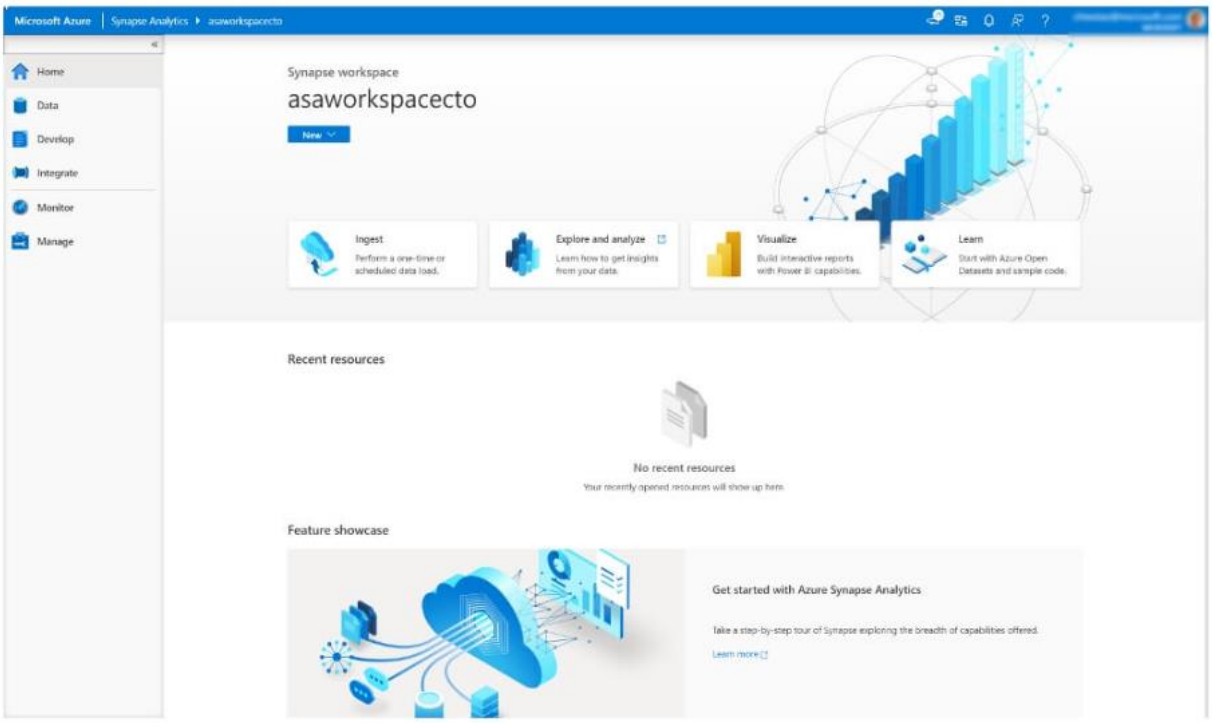

1-CDP Vision Demo
File
View
Campaign Analytics
A quick way to analyse the campaign programs
Select Campaign
Be Unique
Enjoy The Moment
Fall into Winter
Go Eco
Let's Run
Let's Surf
Sustainable Fashion
FILTERS
Region
Asia Pacific
Europe
North & Central ...
South America
Campaign Tactic
Email
Event
In Store Displays
Social
TV
Status
Awareness
Consideration
Inquiries
Intent
Customer Segment
Baby Boomers (4...
Gen X (60 - 80s)
Gen Y (80 - 2000s)
Gen Z (2000s - n...
KPI
Campaign revenue vs target
$611.54K
Total campaigns
1
Total Cost of Campaign
$611.54K
Total Response
$185K
Marketing cost
$205K
Response status (thousand)
Campaign Profit over time
Revenue by campaign
Sentiment by product category
Books
Furniture
Electronics
Decor
Sport Shoes
Gaming
Exercise
Popular hashtags
getsporty
cocktailswithfriends
entertainment
Unique Visitors and Daily Users by Device Categ...
Mobile
Tablet
Desktop
Decomposition Tree Analysis
QnA
Campaign Analytics

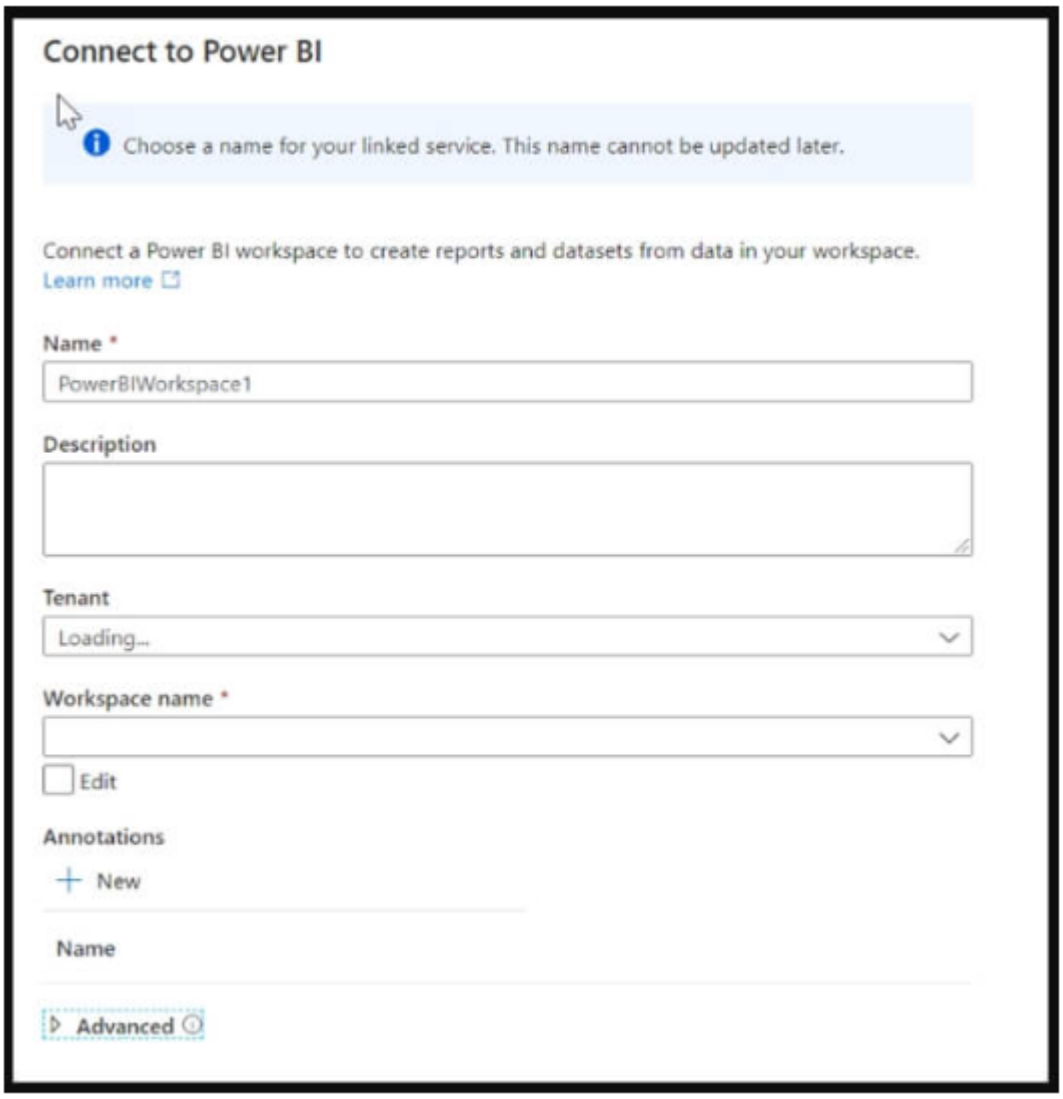

Power BI
SynapseDemos
Power BI datasets
Power BI reports
1-CDP Vision Demo
1-Phase2 CDP Vision Demo
2-Billion Rows Demo
Dashboard-Images
Name
1-CDP Vision Demo
1-Phase2 CDP Vision Demo
2-Billion Rows Demo
Dashboard-Images
New Power BI report
1
2

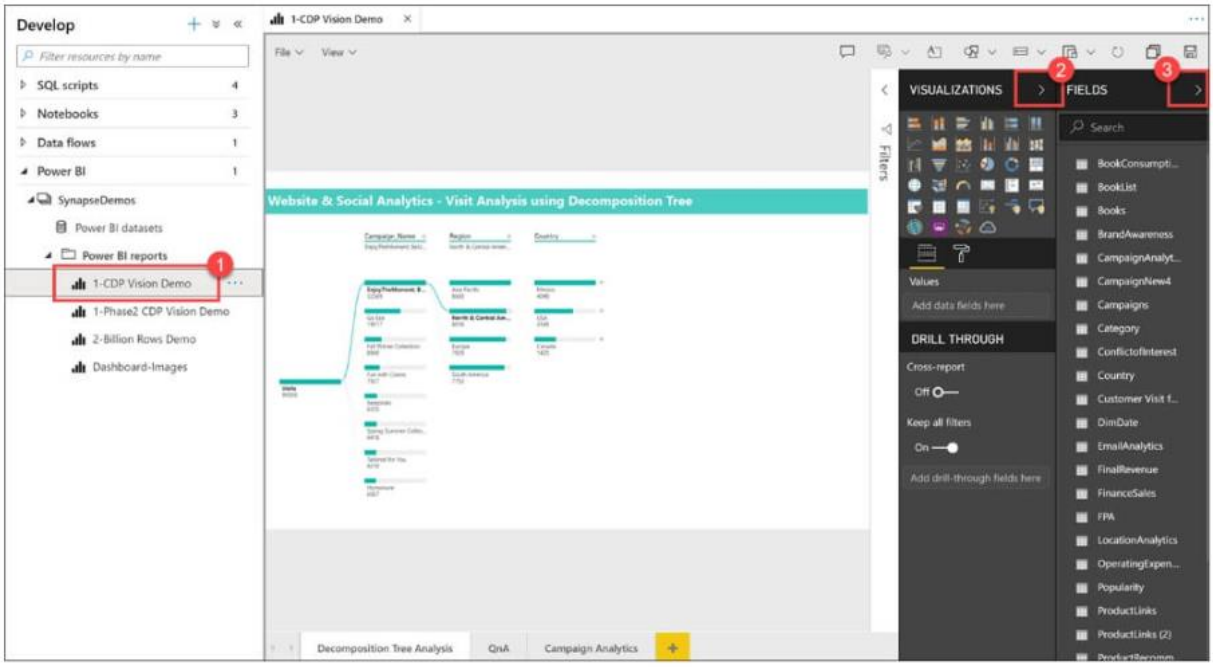

Power BI datasets
newreport4a50
File
View
Category
Biographies
Children
Favorites
Fiction
History
New Releases
Personal Development
Romance
Sci-fi
Travel
Trending Now
Filters
Search
Filters on this visual
Category
is (All)
Add data fields here
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
VISUALIZATIONS
Values
Category
DRILL THROUGH
Cross-report
Off
Keep all filters
On
Add drill-through fields here
FIELDS
Search
Category
Category
ID
SubCategory
ProdChamp
Products
wwi Sales

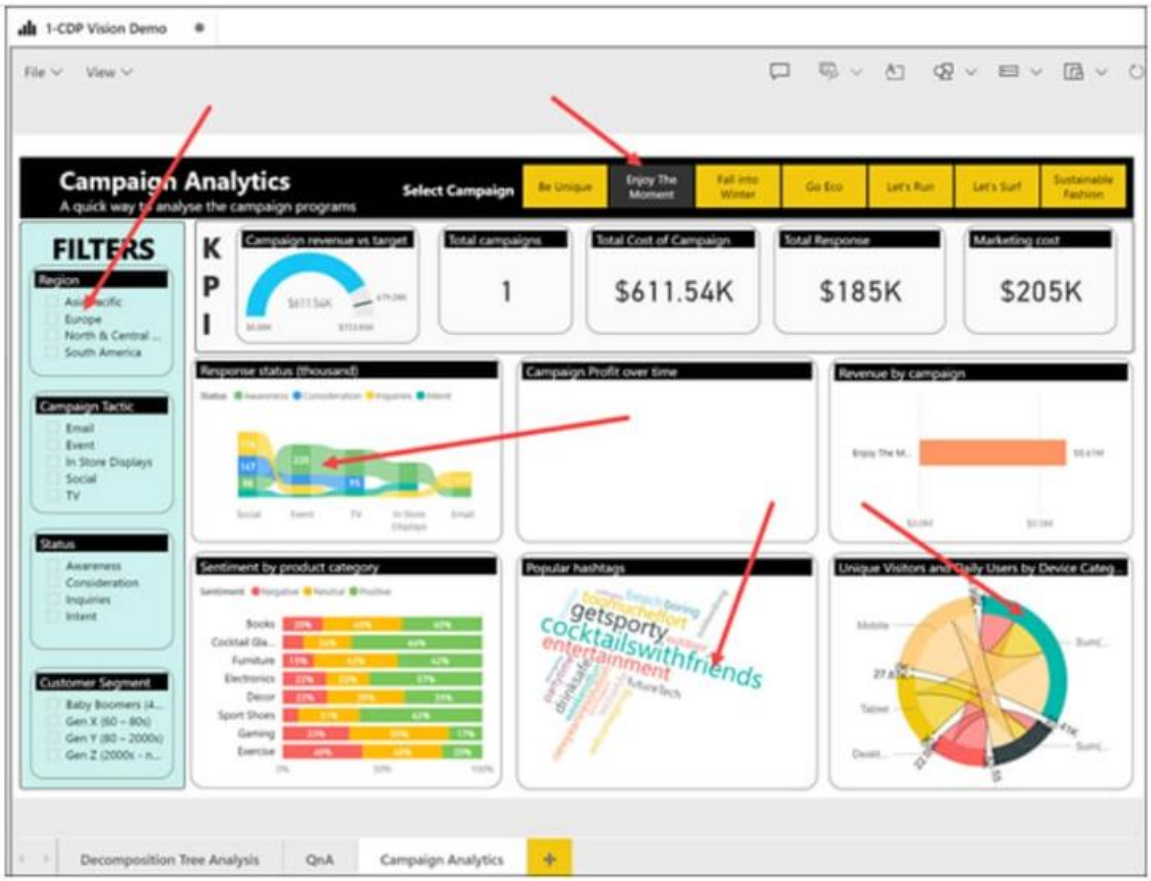


https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-power-bi

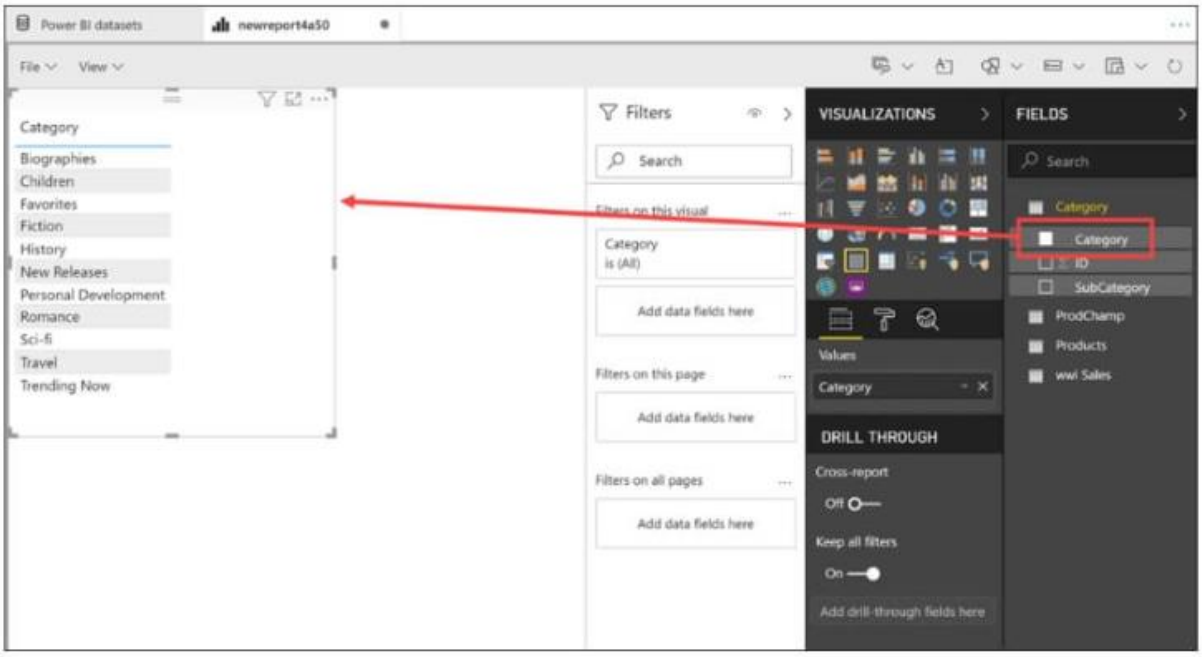

Microsoft Azure
Synapse Analytics
asaworkspacecto
Home
Data
Develop
Integrate
Monitor
Manage
Synapse workspace
asaworkspacecto
New
Ingest
Perform a one-time or scheduled data load.
Explore and analyze
Learn how to get insights from your data.
Visualize
Build interactive reports with Power BI capabilities.
Learn
Start with Azure Open Datasets and sample code.
Recent resources
No recent resources
Your recently opened resources will show up here.
Feature showcase
Get started with Azure Synapse Analytics
Take a step-by-step tour of Synapse exploring the breadth of capabilities offered.
Learn more

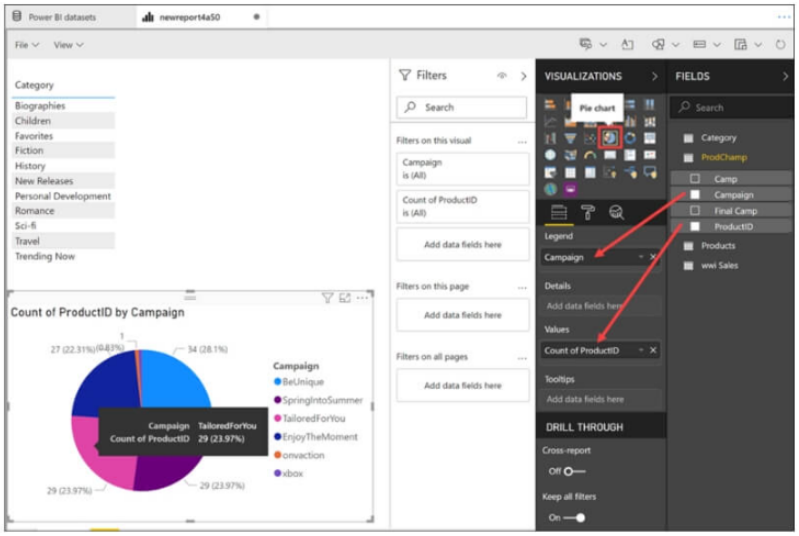

Connect to Power BI
Choose a name for your linked service. This name cannot be updated later.
Connect a Power BI workspace to create reports and datasets from data in your workspace.
Learn more
Name *
PowerBIWorkspace1
Description
Tenant
Loading...
Workspace name *
Edit
Annotations
New
Name
Advanced


Develop
Filter resources by name
SQL scripts 4
Notebooks 3
Data flows 1
Power BI 1
SynapseDemos
Power BI datasets
Power BI reports
1-CDP Vision Demo
1-Phase2 CDP Vision Demo
2-Billion Rows Demo
Dashboard-Images
1
1-CDP Vision Demo
File
View
Website & Social Analytics - Visit Analysis using Decomposition Tree
Filters
VISUALIZATIONS
2
FIELDS
3
Search
BookConsumpti...
BookList
Books
BrandAwareness
CampaignAnalyt...
CampaignNew4
Campaigns
Category
ConflictofInterest
Country
Customer Visit f...
DimDate
EmailAnalytics
FinalRevenue
FinanceSales
FPA
LocationAnalytics
OperatingExpen...
Popularity
ProductLinks
ProductLinks (2)
Values
Add data fields here
DRILL THROUGH
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Decomposition Tree Analysis
QnA
Campaign Analytics


1-CDP Vision Demo
File
View
Campaign Analytics
A quick way to analyse the campaign programs
Select Campaign
Be Unique
Enjoy The Moment
Fall into Winter
Go Eco
Let's Run
Let's Surf
Sustainable Fashion
FILTERS
Region
Asia Pacific
Europe
North & Central ...
South America
Campaign Tactic
Email
Event
In Store Displays
Social
TV
Status
Awareness
Consideration
Inquiries
Intent
Customer Segment
Baby Boomers (4...
Gen X (60 – 80s)
Gen Y (80 – 2000s)
Gen Z (2000s - n...
KPI
Campaign revenue vs target
Total campaigns
1
Total Cost of Campaign
$611.54K
Total Response
$185K
Marketing cost
$205K
Response status (thousand)
Campaign Profit over time
Revenue by campaign
Sentiment by product category
Popular hashtags
Unique Visitors and Daily Users by Device Categ...
getsporty
cocktailswithfriends
entertainment
Decomposition Tree Analysis
QnA
Campaign Analytics


Power BI
SynapseDemos
Power BI datasets
Power BI reports
1-CDP Vision Demo
1-Phase2 CDP Vision Demo
2-Billion Rows Demo
Dashboard-Images
Name
1-CDP Vision Demo
1-Phase2 CDP Vision Demo
2-Billion Rows Demo
Dashboard-Images
New Power BI report





- ○ Import
- ○ Connect BI
- ○ Ingest

Report Error

Q. 20

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.

Proposed Solution:
The team has chosen to use the following:
• Table - Sales: Distribution type: Hash-distributed, Distribution column: ProductKey
• Table - Invoice: Distribution type: Round-robin, Distribution column: RegionKey

Azure Synapse Analytics SME, the team looks to you for reassurance that they made the right choices.

Did they?

- ○ Correct
- ◉ incorrect

Explanation:- No, the team did not choose the correct option; both hashes are > 2GB. The Invoice table RegionKey cannot be used with Round-robin distribution as Round-robin does not take a distribution key. Hash-distributed for the Distribution type and ProductKey for the Distribution column is correct for the Sales table.
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.

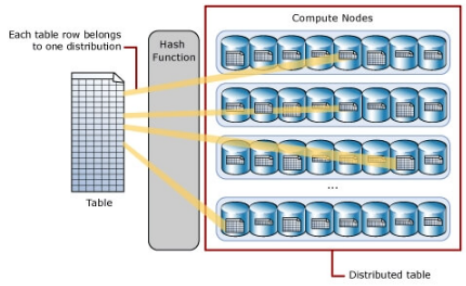
If you didn't choose the best column the first time, you can use CREATE TABLE AS SELECT (CTAS) to re-create the table with a different distribution column.

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



Report Error

Q. 21

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

Security administrators can control data access by using [?] within Data Lake Storage. Built-in security groups include ReadOnlyUsers, WriteAccessUsers, and FullAccessUsers.

- ○ Active Directory Security Groups

Explanation:- Data Lake Storage Data Security
Because Data Lake Storage supports Azure Active Directory ACLs, security administrators can control data access by using the familiar Active Directory Security Groups. Role-based access control (RBAC) is available both in Gen1 and Gen2. Built-in security groups include ReadOnlyUsers, WriteAccessUsers, and FullAccessUsers.
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices

- ○ Active Directory Application Groups
- ◉ AD OAuth
- ○ AD Desired State Configuration (ADDSC)

Report Error

Q. 22 Correct or Incorrect : Azure Blob Storage is the least expensive method to store data and one of it best features is that it allows for querying the data directly within the Blob environment.

- (●) Incorrect

Explanation:- Azure Blob Storage Queries
If you create a storage account as a Blob store, you can't query the data directly. To query it, either move the data to a store that supports queries or set up the Azure Storage account for a data lake storage account. Azure Blob storage has no API to query data within the blob - it's just dumb storage. You're essentially limited to reading, deserializing and grabbing your value(s).
https://stackoverflow.com/questions/38721458/query-blobs-in-blob-storage

- ( ) Correct

Report Error

Q. 23

All data written to Azure Storage is automatically encrypted by Storage Service Encryption (SSE) with a 256-bit Advanced Encryption Standard (AES) cipher, and is FIPS 140-2 compliant.

Correct or Incorrect : For virtual machines (VMs), Azure lets you encrypt virtual hard disks (VHDs) by using Azure Disk Encryption. If someone gets access to the VHD image and downloads it, they can't access the data on the VHD unless they have an Azure Storage account as well. If a bad actor restores the image within their own Azure environment, they will have access to the data on the image.

- ( ) Correct
- (●) Incorrect

Explanation:- Encryption at rest
All data written to Azure Storage is automatically encrypted by Storage Service Encryption (SSE) with a 256-bit Advanced Encryption Standard (AES) cipher, and is FIPS 140-2 compliant. SSE automatically encrypts data when writing it to Azure Storage. When you read data from Azure Storage, Azure Storage decrypts the data before returning it. This process incurs no additional charges and doesn't degrade performance. It can't be disabled.
For virtual machines (VMs), Azure lets you encrypt virtual hard disks (VHDs) by using Azure Disk Encryption. This encryption uses BitLocker for Windows images, and it uses dm-crypt for Linux.
Azure Key Vault stores the keys automatically to help you control and manage the disk-encryption keys and secrets. So even if someone gets access to the VHD image and downloads it, they can't access the data on the VHD.
https://docs.microsoft.com/en-us/azure/storage/common/storage-service-encryption

Report Error

Q. 24

Scenario: Dr. Karl Malus works for the Power Broker Corporation (PBC) founded by Curtiss Jackson, using technology to service various countries and their military efforts. You have been contracted by the company to assist Dr. Malus with their Microsoft Azure Synapse projects.

PBC has an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.

Dr. Malus is looking for a recommendation for a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

Required:
• Track the usage of encryption keys.
• Maintain the access of client apps to Pool1 in the event of an Azure datacentre outage that affects the availability of the encryption keys.

Which of the following should you include in the recommendation for the "Track the usage of encryption key" requirement?

- (●) TDE with customer-managed keys

Explanation:- You should include in "TDE with customer-managed keys" in the recommendation for the first requirement listed.
Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.
After you create one or more key vaults, you'll likely want to monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide. For step by step guidance on setting this up, see How to enable Key Vault logging.
What is logged:
All authenticated REST API requests, including failed requests as a result of access permissions, system errors, or bad requests.
Operations on the key vault itself, including creation, deletion, setting key vault access policies, and updating key vault attributes such as tags.
Operations on keys and secrets in the key vault, including: Creating, modifying, or deleting these keys or secrets. Signing, verifying, encrypting, decrypting, wrapping and unwrapping keys, getting secrets, and listing keys and secrets (and their versions).
Unauthenticated requests that result in a 401 response. Examples are requests that don't have a bearer token, that are malformed or expired, or that have an invalid token.
Azure Event Grid notification events for the following conditions: expired, near expiration, and changed vault access policy (the new version event isn't logged). Events are logged even if there's an event subscription created on the key vault.
You can access your logging information 10 minutes (at most) after the key vault operation. In most cases, it will be quicker than this. It's up to you to manage your logs in your storage account:

Use standard Azure access control methods in your storage account to secure your logs by restricting who can access them.
Delete logs that you no longer want to keep in your storage account.
https://docs.microsoft.com/en-us/azure/key-vault/general/logging?tabs=Vault

- ( ) Always Encrypted
- ( ) Any of the options listed will meet the requirement
- ( ) None of the options listed will meet the requirement
- ( ) TDE with platform-managed keys

Report Error

Q. 25

Scenario: Honest Eddie's Car Dealership is an establishment in South Carolina USA, which is dedicated to the purchase and sale of cars and light trucks. Currently the IT team is working looking into distribution options for a product dimension table.

Which of the following distribution options should Eddie's IT team use where a sales fact table will contain billions of records?

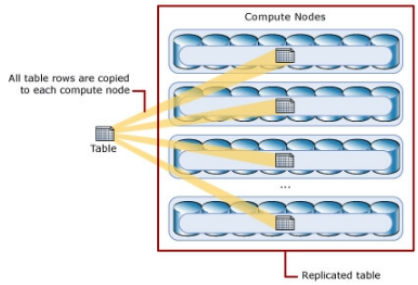
◯ DISTRIBUTION = REPLICATE

Explanation:- Replicate will result in a copy of the table on each compute node, which performs well with joins to the distributed fact table.
https://youtu.be/1VS_F37GI9U
What is a replicated table?
A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit. If the data is static and does not change, you can replicate larger tables.
The following diagram shows a replicated table that is accessible on each Compute node. In SQL pool, the replicated table is fully copied to a distribution database on each compute node.
Replicated tables work well for dimension tables in a star schema. Dimension tables are typically joined to fact tables which are distributed differently than the dimension table. Dimensions are usually of a size that makes it feasible to store and maintain multiple copies. Dimensions store descriptive data that changes slowly, such as customer name and address, and product details. The slowly changing nature of the data leads to less maintenance of the replicated table.
Consider using a replicated table when:
• The table size on disk is less than 2 GB, regardless of the number of rows. To find the size of a table, you can use the DBCC PDW_SHOWSPACEUSED command: DBCC PDW_SHOWSPACEUSED('ReplTableCandidate').
• The table is used in joins that would otherwise require data movement. When joining tables that are not distributed on the same column, such as a hash-distributed table to a round-robin table, data movement is required to complete the query. If one of the tables is small, consider a replicated table. We recommend using replicated tables instead of round-robin tables in most cases. To view data movement operations in query plans, use sys.dm_pdw_request_steps. The BroadcastMoveOperation is the typical data movement operation that can be eliminated by using a replicated table.
Replicated tables may not yield the best query performance when:
• The table has frequent insert, update, and delete operations. The data manipulation language (DML) operations require a rebuild of the replicated table. Rebuilding frequently can cause slower performance.
• The SQL pool is scaled frequently. Scaling a SQL pool changes the number of Compute nodes, which incurs rebuilding the replicated table.

• The table has a large number of columns, but data operations typically access only a small number of columns. In this scenario, instead of replicating the entire table, it might be more effective to distribute the table, and then create an index on the frequently accessed columns. When a query requires data movement, SQL pool only moves data for the requested columns.
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.
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables



- ○ DISTRIBUTION = ROUND_ROBIN([SalesOrderNumber])
- ○ DISTRIBUTION = HEAP
- ◉ DISTRIBUTION = HASH([SalesOrderNumber])

Report Error

Q. 26 Which of the following are supported connectors for built-in parameterization? (Select all that apply)

- ☑ Azure Key Vault

Explanation:- Azure Synapse Analytics is a supported connector for built-in parameterization for Linked Services in Azure Data Factory.
Supported linked service types
You can parameterize any type of linked service. When authoring linked service on UI, Data Factory provides built-in parameterization experience for the following types of linked services. In linked service creation/edit blade, you can find options to new parameters and add dynamic content.
• Amazon Redshift
• Amazon S3
• Azure Cosmos DB (SQL API)
• Azure Database for MySQL
• Azure Databricks
• Azure Key Vault
• Azure SQL Database
• Azure SQL Managed Instance
• Azure Synapse Analytics
• MySQL
• Oracle
• SQL Server
• Generic HTTP
• Generic REST
For other linked service types that are not in above list, you can parameterize the linked service by editing the JSON on UI:
• In linked service creation/edit blade ? expand "Advanced" at the bottom ? check "Specify dynamic contents in JSON format" checkbox ? specify the linked service JSON payload.
• Or, after you create a linked service without parameterization, in Management hub ? Linked services ? find the specific linked service ? click "Code" (button "{}") to edit the JSON.
Refer to the JSON sample to add parameters section to define parameters and reference the parameter using @{linkedService().paraName}.
https://docs.microsoft.com/en-us/azure/data-factory/parameterize-linked-services

- ☐ Azure Data Lake Storage Gen1
- ☐ Azure Data Lake Storage Gen2

☑ Azure Synapse Analytics

Explanation:- Azure Synapse Analytics is a supported connector for built-in parameterization for Linked Services in Azure Data Factory.
Supported linked service types
You can parameterize any type of linked service. When authoring linked service on UI, Data Factory provides built-in parameterization experience for the following types of linked services. In linked service creation/edit blade, you can find options to new parameters and add dynamic content.
• Amazon Redshift
• Amazon S3
• Azure Cosmos DB (SQL API)
• Azure Database for MySQL
• Azure Databricks
• Azure Key Vault
• Azure SQL Database
• Azure SQL Managed Instance
• Azure Synapse Analytics
• MySQL
• Oracle
• SQL Server
• Generic HTTP
• Generic REST
For other linked service types that are not in above list, you can parameterize the linked service by editing the JSON on UI:
• In linked service creation/edit blade ? expand "Advanced" at the bottom ? check "Specify dynamic contents in JSON format" checkbox ? specify the linked service JSON payload.
• Or, after you create a linked service without parameterization, in Management hub ? Linked services ? find the specific linked service ? click "Code" (button "{}") to edit the JSON.
Refer to the JSON sample to add parameters section to define parameters and reference the parameter using @{linkedService().paraName}.
https://docs.microsoft.com/en-us/azure/data-factory/parameterize-linked-services

☑ Amazon S3

Explanation:- Azure Synapse Analytics is a supported connector for built-in parameterization for Linked Services in Azure Data Factory.
Supported linked service types
You can parameterize any type of linked service. When authoring linked service on UI, Data Factory provides built-in parameterization experience for the following types of linked services. In linked service creation/edit blade, you can find options to new parameters and add dynamic content.
• Amazon Redshift
• Amazon S3
• Azure Cosmos DB (SQL API)
• Azure Database for MySQL
• Azure Databricks
• Azure Key Vault
• Azure SQL Database
• Azure SQL Managed Instance
• Azure Synapse Analytics
• MySQL
• Oracle
• SQL Server
• Generic HTTP
• Generic REST
For other linked service types that are not in above list, you can parameterize the linked service by editing the JSON on UI:
• In linked service creation/edit blade ? expand "Advanced" at the bottom ? check "Specify dynamic contents in JSON format" checkbox ? specify the linked service JSON payload.
• Or, after you create a linked service without parameterization, in Management hub ? Linked services ? find the specific linked service ? click "Code" (button "{}") to edit the JSON.
Refer to the JSON sample to add parameters section to define parameters and reference the parameter using @{linkedService().paraName}.
https://docs.microsoft.com/en-us/azure/data-factory/parameterize-linked-services

Q. 27 What function provides a rowset view over a JSON document?

- ○ WITH
- ○ VIEWRSET
- ○ OPENROWSET
- ◉ OPENJSON

Explanation:- OPENJSON (Transact-SQL) is a table-valued function that parses JSON text and returns objects and properties from the JSON input as rows and columns. In other words, OPENJSON provides a rowset view over a JSON document. You can explicitly specify the columns in the rowset and the JSON property paths used to populate the columns. Since OPENJSON returns a set of rows, you can use OPENJSON in the FROM clause of a Transact-SQL statement just as you can use any other table, view, or table-valued function.
Use OPENJSON to import JSON data into SQL Server, or to convert JSON data to relational format for an app or service that can't consume JSON directly.
The OPENJSON function provides a rowset view over a JSON document.
https://docs.microsoft.com/en-us/sql/t-sql/functions/openjson-transact-sql?view=sql-server-ver15

Report Error

Q. 28

Scenario: You are working in an Azure Databricks workspace and you want to filter based on the end of a column value using the Column Class. Specifically, you are looking at a column named verb and filtered by words ending with "ing".

Which command filters based on the end of a column value as required?

- ☑ df.filter(col("verb").endswith("ing"))

Explanation:- The Column Class supports both the endswith() method and the like() method (example - col("verb").like("%ing")).
https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html

- ☐ df.filter("verb like '%ing'")
- ☐ df.filter("verb like '_ing'")
- ☑ df.filter().col("verb").like("%ing")

Report Error

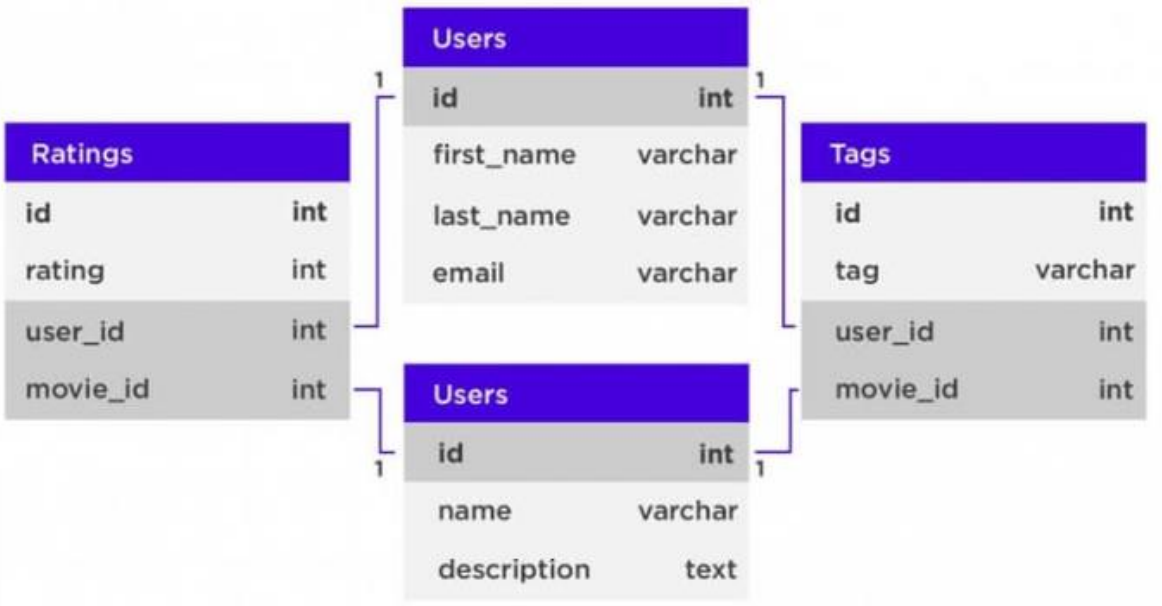
Q. 29

Which of the below have the following characteristics?

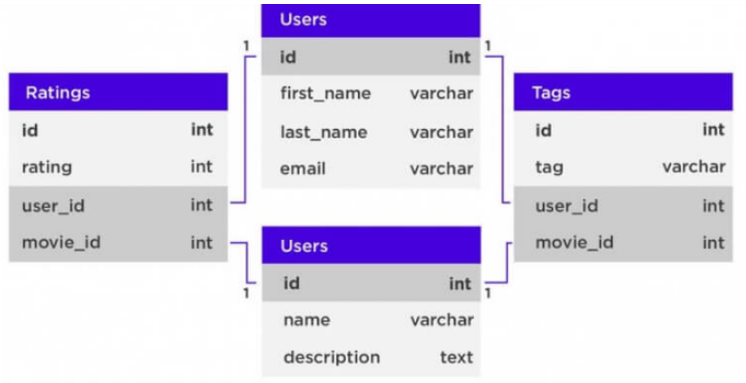
• Provide undoubtedly the most well-understood model for holding data.
• The simplest structure of columns and tables makes them very easy to use initially, but the inflexible structure can cause some problems.
• We can communicate with relational databases using SQL.

- ○ Key-Value
- ○ Non-Relational
- ◉ Relational

Explanation:-
Relational Data
• Relational databases provide undoubtedly the most well-understood model for holding data.
• The simplest structure of columns and tables makes them very easy to use initially, but the inflexible structure can cause some problems.
• We can communicate with relational databases using Structured Query Language (SQL).
• SQL allows the joining of tables using a few lines of code, with a structure most beginner employees can learn very fast.
• Examples of relational databases:
• MySQL
• PostgreSQL
• Db2
• SQL Server


Users
id int
first_name varchar
last_name varchar
email varchar
Ratings
id int
rating int
user_id int
movie_id int
Tags
id int
tag varchar
user_id int
movie_id int
Users
id int
name varchar
description text
1
1
1
1


Microsoft SQL Server
MySQL
IBM DB2
PostgreSQL
ORACLE
DB Management


Users
id int
first_name varchar
last_name varchar
email varchar
Ratings
id int
rating int
user_id int
movie_id int
Tags
id int
tag varchar
user_id int
movie_id int
Users
id int
name varchar
description text
1
1
1
1


Microsoft SQL Server
MySQL
IBM DB2
PostgreSQL
ORACLE
DB Management

- ○ JSON

Q. 30

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

[?] is a fully managed cloud service. Analysts, data scientists, developers, and others use [?] to discover, understand, and consume data sources. It features a crowdsourcing model of metadata and annotations.

In this central location, an organization's users contribute their knowledge to build a community of data sources that are owned by the organization.

- ○ Azure Data Factory
- ◉ Azure Data Catalog

Explanation:- Azure Data Catalog
Analysts, data scientists, developers, and others use Data Catalog to discover, understand, and consume data sources. Data Catalog features a crowdsourcing model of metadata and annotations. In this central location, an organization's users contribute their knowledge to build a community of data sources that are owned by the organization.
Data Catalog is a fully managed cloud service. Users discover and explore data sources, and they help the organization document information about their data sources.
https://docs.microsoft.com/en-us/azure/data-catalog/overview

- ○ Azure Storage Explorer
- ○ Azure Databricks
- ○ Azure SQL Datawarehouse
- ○ Azure Cosmos DB

Report Error

Q. 31

A data warehouse that is built on a Massively Parallel Processing (MPP) system is built for processing and analyzing large datasets. As such they perform well with larger batch type loads and updates that can be distributed across the compute nodes and storage.

Which of the following is the best approach if singleton or smaller transaction batch loads must be added to an MPP data warehouse?

- ○ All the approaches are equally valid.
- ◉ Develop two processes: one that writes the outputs of an INSERT statement to a file, and then another process to periodically load this file.

Explanation:- A data warehouse that is built on a Massively Parallel Processing (MPP) system is built for processing and analyzing large datasets. As such they perform well with larger batch type loads and updates that can be distributed across the compute nodes and storage.
Singleton or smaller transaction batch loads should be grouped into larger batches to optimize the Synapse SQL Pools processing capabilities. To be clear, A one-off load to a small table with an INSERT statement may be the best approach, if it is a one-off.
However, if you need to load thousands or millions of rows throughout the day, then singleton INSERTs aren't optimal against an MPP system. One way to solve this issue is to develop one process that writes the outputs of an INSERT statement to a file, and then another process to periodically load this file to take advantage of the parallelism that Azure Synapse Analytics.
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-best-practices

- ○ None of the listed options.
- ○ Manually create an append file with a trigger that once the contents of the manually created file reach a predetermined size, an automation process will be triggered to append the data to the data warehouse.
- ○ Develop a process that writes the outputs of an INSERT statement to a to the target file automatically, avoiding the need to do the INSERT manually.

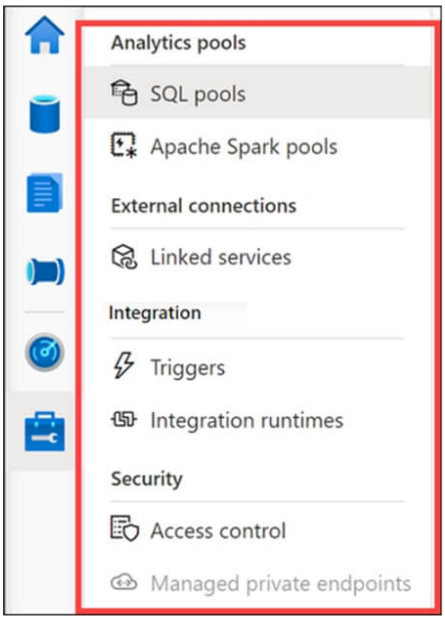
Report Error

Q. 32 Which hub is where you can grant access to Synapse workspace and resources?

◉ Manage hub

Explanation:- In Azure Synapse Studio, the Manage hub enables you to perform some of the same actions available in the Azure portal, such as managing SQL and Spark pools. However, there is a lot more you can do in this hub that you cannot do anywhere else, such as managing Linked Services and integration runtimes, and creating pipeline triggers.

• SQL pools. Lists the provisioned SQL pools and on-demand SQL serverless pools for the workspace. You can add new pools or hover over a SQL pool to pause or scale it. You should pause a SQL pool when it's not being used to save costs.
• Apache Spark pools. Lets you manage your Spark pools by configuring the auto-pause and auto-scale settings. You can provision a new Apache Spark pool from this blade.
• Linked services. Enables you to manage connections to external resources. Here you can see linked services for our data lake storage account, Azure Key Vault, Power BI, and Synapse Analytics. Task: Select + New to show how many types of linked services you can add.
• Triggers. Provides you a central location to create or remove pipeline triggers. Alternatively, you can add triggers from the pipeline.
• Integration runtimes. Lists the IR for the workspace, which serves as the compute infrastructure for data integration capabilities, like those provided by pipelines. Task: Hover over the integration runtimes to show the monitoring, code, and delete (if applicable) links. Click on a code link to show how you can modify the parameters in JSON format, including the TTL (time to live) setting for the IR.
• Access control. This is where you go to add and remove users to one of three security groups: workspace admin, SQL admin, and Apache Spark for Azure Synapse Analytics admin.
• Managed private endpoints. This is where you manage private endpoints, which use a private IP address from within a virtual network to connect to an Azure service or your own private link service. Connections using private endpoints listed here provide access to Synapse workspace endpoints (SQL, SqlOndemand and Dev).
https://techcommunity.microsoft.com/t5/azure-synapse-analytics/explore-the-manage-hub-in-synapse-studio-to-provision-and-secure/ba-p/1987788



○ Create hub

○ Data hub

○ None of the listed options

○ Monitor hub

Report Error

Q. 33

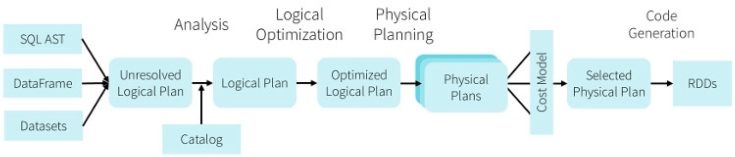
Because the Databricks API is declarative, a large number of optimizations are available to us. Among the most powerful components of Spark are Spark SQL. At its core lies the Catalyst optimizer.

When you execute code, Spark SQL uses Catalyst's general tree transformation framework in four phases, as shown below:
1. analyzing a logical plan to resolve references
2. logical plan optimization
3. physical planning
4. code generation to compile parts of the query to Java bytecode

In the physical planning phase, Catalyst may generate multiple plans and compare them based on [?].

- ○ Region
- ○ Rules
- ◉ Cost

Explanation:- Because the Databricks API is declarative, a large number of optimizations are available to us.
Some of the examples include:
• Optimizing data type for storage
• Rewriting queries for performance
• Predicate push downs
Among the most powerful components of Spark are Spark SQL. At its core lies the Catalyst optimizer. This extensible query optimizer supports both rule-based and cost-based optimization.
When you execute code, Spark SQL uses Catalyst's general tree transformation framework in four phases, as shown below:
1. analyzing a logical plan to resolve references
2. logical plan optimization
3. physical planning
4. code generation to compile parts of the query to Java bytecode
In the physical planning phase, Catalyst may generate multiple plans and compare them based on cost. All other phases are purely rule-based.
Catalyst is based on functional programming constructs in Scala and designed with these key two purposes:
• Easily add new optimization techniques and features to Spark SQL
• Enable external developers to extend the optimizer (e.g. adding data source specific rules, support for new data types, etc.)
https://data-flair.training/blogs/spark-sql-optimization/



- ○ Permissions

Report Error

Q. 34

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

When using JSON notation, the activities section can have one or more activity defined within it.

They have the following top-level structure:

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

- [ ] typeProperties
- [x] linkedServiceName

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 activities
When using JSON notation, the activities section can have one or more activities defined within it. There are two main types of activities: Execution and Control Activities. Execution (also known as Compute) activities include data movement and data transformation activities. They have the following top-level structure:

JSON

```
{
"name": "Execution Activity Name",
"description": "description",
"type": "",
"typeProperties":
{
},
"linkedServiceName": "MyLinkedService",
"policy":
{
},
"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: linkedServiceName

Name of the linked service used by the activity.

Required: Yes for HDInsight, Machine Learning Batch Scoring Activity and Stored Procedure Activity

Required: Yes for HDInsight, Machine Learning Batch Scoring Activity and Stored Procedure Activity
Property: typeProperties
Properties in the typeProperties section depend on each type of activity.
Required: No
Property: policy
Policies that affect the run-time behaviour of the activity. This property includes timeout and retry behaviour.
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

- [x] name

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 activities
When using JSON notation, the activities section can have one or more activities defined within it. There are two main types of activities: Execution and Control Activities. Execution (also known as Compute) activities include data movement and data transformation activities. They have the following top-level structure:

```
JSON
{
"name": "Execution Activity Name",
"description": "description",
"type": "",
"typeProperties":
{
},
"linkedServiceName": "MyLinkedService",
"policy":
{
},
"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: linkedServiceName

Name of the linked service used by the activity.

Required: Yes for HDInsight, Machine Learning Batch Scoring Activity and Stored Procedure Activity

Property: typeProperties

Properties in the typeProperties section depend on each type of activity.
Required: No
Property: policy
Policies that affect the run-time behaviour of the activity. This property includes timeout and retry behaviour.
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

☑ description

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 activities
When using JSON notation, the activities section can have one or more activities defined within it. There are two main types of activities: Execution and Control Activities. Execution (also known as Compute) activities include data movement and data transformation activities. They have the following top-level structure:

JSON

```
{
"name": "Execution Activity Name",
"description": "description",
"type": "",
"typeProperties":
{
},
"linkedServiceName": "MyLinkedService",
"policy":
{
},
"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: linkedServiceName

Name of the linked service used by the activity.

Required: Yes for HDInsight, Machine Learning Batch Scoring Activity and Stored Procedure Activity

Property: typeProperties

Properties in the typeProperties section depend on each type of activity.
Required: No
Property: policy
Policies that affect the run-time behaviour of the activity. This property includes timeout and retry behaviour.
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

- [ ] policy
- [x] type

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 activities
When using JSON notation, the activities section can have one or more activities defined within it. There are two main types of activities: Execution and Control Activities. Execution (also known as Compute) activities include data movement and data transformation activities. They have the following top-level structure:

```
JSON
{
"name": "Execution Activity Name",
"description": "description",
"type": "",
"typeProperties":
{
},
"linkedServiceName": "MyLinkedService",
"policy":
{
},
"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: linkedServiceName

Name of the linked service used by the activity.

Required: Yes for HDInsight, Machine Learning Batch Scoring Activity and Stored Procedure Activity

Property: typeProperties

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

Required: No

Property: policy

Policies that affect the run-time behaviour of the activity. This property includes timeout and retry behaviour.

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

Q. 35

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

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

Data Factory stores pipeline-run data for [?] days.

- 10
- 15
- 21
- 45 ✓

Explanation:- Monitor using Azure Monitor
Azure Monitor provides base-level infrastructure metrics and logs for most Azure services. Azure diagnostic logs are emitted by a resource and provide rich, frequent data about the operation of that resource. Azure Data Factory (ADF) can write diagnostic logs in Azure Monitor.
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets.
• Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.
• Event Hub: Stream the logs to Azure Event Hubs. The logs become input to a partner service/custom analytics solution like Power BI.
• Log Analytics: Analyze the logs with Log Analytics. The Data Factory integration with Azure Monitor is useful in the following scenarios:
• You want to write complex queries on a rich set of metrics that are published by Data Factory to Monitor. You can create custom alerts on these queries via Monitor.
• You want to monitor across data factories. You can route data from multiple data factories to a single Monitor workspace.
You can also use a storage account or event-hub namespace that isn't in the subscription of the resource that emits logs. The user who configures the setting must have appropriate Azure role-based access control (Azure RBAC) access to both subscriptions.
https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

- 30

Report Error

Q. 36

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

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

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

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

When we shuffle data, it creates what is known as a stage boundary which represents a process bottleneck which Spark will break this one job into two stages.

In Stage #1, Spark will create a pipeline of transformations in which the data is read into RAM.

For Stage #2, Spark will again create a pipeline of transformations in which the shuffle data is read into RAM.

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).

What is the main benefit of working backward through your action's lineage?

◉ It allows Spark to determine if it is necessary to execute every transformation.

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

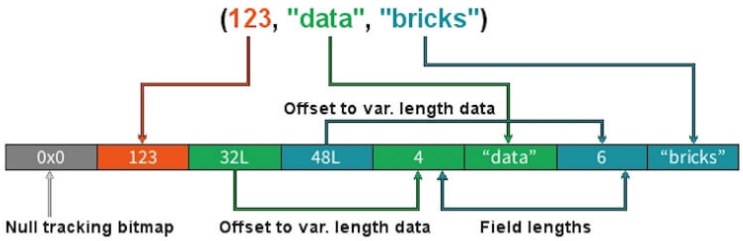
Spark will break this one job into two stages (steps 1-4b and steps 4c-7):

Spark will break this one job into two stages (steps 1-4b and steps 4c-7):
Stage #1
Step Transformation
1 Read
2 Select
3 Filter
4a GroupBy 1/2
4b shuffle write
Stage #1
Step Transformation
4c shuffle read
4d GroupBy 2/2
5 Select
6 Filter
7 Write
In Stage #1, Spark will create a pipeline of transformations in which the data is read into RAM (Step #1), and then perform steps #2, #3, #4a & #4b
All partitions must complete Stage #1 before continuing to Stage #2
• It's not possible to group all records across all partitions until every task is completed.
• This is the point at which all the tasks must synchronize.
• This creates our bottleneck.
• Besides the bottleneck, this is also a significant performance hit: disk IO, network IO and more disk IO.
Once the data is shuffled, we can resume execution...
For Stage #2, Spark will again create a pipeline of transformations in which the shuffle data is read into RAM (Step #4c) and then perform transformations #4d, #5, #6 and finally the write action, step #7.
Lineage
From the developer's perspective, we start with a read and conclude (in this case) with a write:

Step Transformation
1 Read
2 Select
3 Filter
4 GroupBy
5 Select
6 Filter
7 Write
However, Spark starts with the action (write(..) in this case).
Next, it asks the question, what do I need to do first?
It then proceeds to determine which transformation precedes this step until it identifies the first transformation.
Step Transformation
7 Write Depends on #6
6 Filter Depends on #5
5 Select Depends on #4
4 GroupBy Depends on #3
3 Filter Depends on #2
2 Select Depends on #1
1 Read First
Why Work Backwards?
Question: So what is the benefit of working backward through your action's lineage?
Answer: It allows Spark to determine if it is necessary to execute every transformation.
Take another look at our example:
• Say we've executed this once already
• On the first execution, step #4 resulted in a shuffle
• Those shuffle files are on the various executors (src & dst)
• Because the transformations are immutable, no aspect of our lineage can change.
• That means the results of our last shuffle (if still available) can be reused.
Why Work Backwards?

Step Transformation
7 Write Depends on #6
6 Filter Depends on #5
5 Select Depends on #4
4 GroupBy <<< shuffle
3 Filter don't care
2 Select don't care
1 Read don't care
In this case, what we end up executing is only the operations from Stage #2.
This saves us the initial network read and all the transformations in Stage #1
Step Transformation
1 Read skipped
2 Select skipped
3 Filter skipped
4a GroupBy 1/2 skipped
4b shuffle write skipped
4c shuffle read -
4d GroupBy 2/2 -
5 Select -
6 Filter -
7 Write
And Caching...
The reuse of shuffle files (also known as our temp files) is just one example of Spark optimizing queries anywhere it can.
We cannot assume this will be available to us.
Shuffle files are by definition temporary files and will eventually be removed.
However, we cache data to explicitly accomplish the same thing that happens inadvertently with shuffle files.
In this case, the lineage plays the same role. Take for example:
Step Transformation
7 Write Depends on #6
6 Filter Depends on #5

5 Select <<< cache
4 GroupBy <<< shuffle files
3 Filter ?
2 Select ?
1 Read ?
In this case we cached the result of the select(..).
We never even get to the part of the lineage that involves the shuffle, let alone Stage #1.
Instead, we pick up with the cache and resume execution from there:
Step Transformation
1 Read skipped
2 Select skipped
3 Filter skipped
4a GroupBy 1/2 skipped
4b shuffle write skipped
4c shuffle read skipped
4d GroupBy 2/2 skipped
5a cache read -
5b Select -
6 Filter -
7 Write
https://databricks.com/blog/2015/04/28/project-tungsten-bringing-spark-closer-to-bare-metal.html



- It serializes the work to make the work sequential, thereby lowering CPU and RAM cost.
- It allows Spark to work on various activities simultaneously using multiple nodes.
- It allows Azure to distribute the load to the required number of processors to optimize the load.

Report Error

Q. 37

Scenario: You are working on a project with a 3rd party vendor to build a website for a customer. The image assets that will be used on the website are stored in an Azure Storage account that is held in your subscription. You want to give read access to this data for a limited period of time.

What security option would be the best option to use?

- ○ CORS Support
- ○ Private Link
- ◉ Shared Access Signatures

Explanation:- A shared access signature is a string that contains a security token that can be attached to a URI. Use a shared access signature to delegate access to storage objects and specify constraints, such as the permissions and the time range of access.
Shared Access Signatures (SAS)
Access keys are the easiest approach to authenticating access to a storage account. However they provide full access to anything in the storage account, similar to a root password on a computer.
Storage accounts offer a separate authentication mechanism called shared access signatures that support expiration and limited permissions for scenarios where you need to grant limited access. You should use this approach when you are allowing other users to read and write data to your storage account. There are links to our documentation on this advanced topic at the end of the module.
https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

- ○ Storage Account

Report Error

Q. 38 What happens if the command option ("checkpointLocation", pointer-to-checkpoint directory) is not specified in Structured Streaming?

- ○ It will not be possible to create more than one streaming query that uses the same streaming source since they will conflict.
- ◉ When the streaming job stops, all state data around the streaming job is lost, and upon restart, the job must start from scratch.

Explanation:- Setting the checkpointLocation is required for many sinks used in Structured Streaming. For those sinks where this setting is optional, keep in mind that when you do not set this value, you risk losing your place in the stream.
https://www.waitingforcode.com/apache-spark-structured-streaming/checkpoint-storage-structured-streaming/read

- ○ The streaming job will function as expected since the checkpointLocation option does not exist.
- ○ When the streaming job stops, all state around the streaming job dumped to a default location, and upon restart, the job must start from aggregated data rather than tuned specific data.

Report Error

Q. 39 Which is an element of a Spark Pool in Azure Synapse Analytics?

- [x] Spark Instance

Explanation:- The definition of a Spark pool is that, when instantiated, it is used to create a Spark instance that processes data.
Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications. Apache Spark in Azure Synapse Analytics is one of Microsoft's implementations of Apache Spark in the cloud.
Azure Synapse makes it easy to create and configure Spark capabilities in Azure. Azure Synapse provides a different implementation of these Spark capabilities that are documented here.
Spark pools
A serverless Apache Spark pool is created in the Azure portal. It's the definition of a Spark pool that, when instantiated, is used to create a Spark instance that processes data. When a Spark pool is created, it exists only as metadata, and no resources are consumed, running, or charged for. A Spark pool has a series of properties that control the characteristics of a Spark instance. These characteristics include but aren't limited to name, size, scaling behaviour, time to live.
As there's no dollar or resource cost associated with creating Spark pools, any number can be created with any number of different configurations. Permissions can also be applied to Spark pools allowing users only to have access to some and not others.
A best practice is to create smaller Spark pools that may be used for development and debugging and then larger ones for running production workloads.
Spark instances
Spark instances are created when you connect to a Spark pool, create a session, and run a job. As multiple users may have access to a single Spark pool, a new Spark instance is created for each user that connects.
When you submit a second job, if there is capacity in the pool, the existing Spark instance also has capacity. Then, the existing instance will process the job. Otherwise, if capacity is available at the pool level, then a new Spark instance will be created.
https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-concepts

- [ ] Spark Console
- [ ] Databricks
- [ ] HDI

Report Error

Q. 40 Which statement about the Azure Databricks Data Plane is true?

- [x] The Data Plane is hosted within the client subscription and is where all data is processed and stored.

Explanation:- All data is processed by clusters hosted within the client Azure subscription and data is stored within Azure Blob storage and any connected Azure services within this portion of the platform architecture.
https://docs.microsoft.com/en-us/azure/key-vault/general/security-overview

- [ ] The Data Plane is hosted within a Microsoft-managed subscription.
- [ ] The Data Plane contains the Cluster Manager and coordinates data processing jobs.
- [ ] The Data Plane is where you manage Key Vault itself and it is the interface used to create and delete vaults.

Report Error

Q. 41

Synapse Analytics removes the barrier of setting up multiple different services for Spark or SQL.

The interoperability between Spark and SQL helps you achieve as follows:
• A shared Hive-compatible metadata system enables you to define tables on files in the data lake such that it can be consumed by either Spark or Hive.
• Both SQL and Spark can directly explore, and analyze Parquet, CSV, TSV, and JSON files stored in the data lake.
• The enablement of fast scalable load and unload for data transferring between SQL and Spark databases.

The Azure Synapse Apache Spark to Synapse SQL connector is designed to efficiently transfer data between serverless Apache Spark pools and SQL pools in Azure Synapse.

Which of the following are valid use cases for Apache Spark and SQL integration within Synapse analytics? (Select all that apply)

- [x] Dealing with different type of analytics

Explanation:- Synapse Analytics removes the barrier of setting up multiple different services for Spark or SQL. Therefore, it removes the traditional thinking about these technologies. It enables you to use both technologies within one platform, which allowed you to switch between Spark or SQL based on the needs and expertise you have in-house.
A spark orientated data engineer can now easily communicate with a SQL based data engineer and communicate together on the same platform.
The interoperability between Spark and SQL helps you achieve as follows:
• A shared Hive-compatible metadata system enables you to define tables on files in the data lake such that it can be consumed by either Spark or Hive.
• Both SQL and Spark can directly explore, and analyze Parquet, CSV, TSV, and JSON files stored in the data lake.
• The enablement of fast scalable load and unload for data transferring between SQL and Spark databases.
The question might raise as how would that SQL and Spark integration then work.
That's when the Azure Synapse Apache Spark to Synapse SQL connector comes in place. It is designed to efficiently transfer data between serverless Apache Spark pools (preview) and SQL pools in Azure Synapse. However, at the moment, the Azure Synapse Apache Spark to Synapse SQL connector works on dedicated SQL pools only, it doesn't work with serverless SQL pools.
In the commonly used existing approach, you often see the use of the JDBC. The JDBC would open the connection. Then, filters and projections would be applied and spark would read the data serially. Given two distributed systems such as Spark and SQL pools, JDBC could become a bottleneck with serial data transfer.

Therefore the New Approach we would take is JDBC and PolyBase. First, the JDBC issues CETAS and send filters and projections. Then filters and projections would be applied and the DataWarehouse exports the data in parallel. Spark reads the data in parallel all based on the user provisioned workspace default data lake storage.
The Azure Synapse Apache Spark Pool to Synapse SQL connector would then be a data source implementation for apache spark where the ADLS Gen 2 is used as well as PolyBase in the dedicated SQL Pools to transfer data between the Spark instance and SQL pool efficiently.
The use cases for Apache Spark and SQL integration within Synapse analytics are as following:
• Dealing with different type of analytics
• Scalability
• Big data computational powers
• Flexibility in the use of Spark and SQL languages and frameworks
Since Apache Spark is integrated in Synapse Analytics, there is more to that than giving use for the big data analytics framework Apache Spark enables. When you deploy a synapse cluster, ADLS Gen2 capacity that can store Spark SQL Tables is provisioned with it.
If you use Spark SQL Tables, you might know that these tables can be queried from a SQL-server-based T-SQL language without you having to use commands like CREATE EXTERNAL TABLE. Within synapse analytics, these queries integrate natively with data files that are stored in an Apache Parquet format.
The other thing to take in mind is that beyond the capabilities mentioned above, the Azure Synapse Studio experience gives you an integrated notebook experience. Within this notebook experience, you can attach a SQL or Spark pool, and develop and execute, for example, transformation pipelines using Python, Scala, and native Spark SQL.
So, let's say you would like to write to a SQL pool after you've performed engineering tasks in spark. You can reference the SQL Pool data as a source for joining with Spark Dataframes that can contain data from other files. When you decide to use the Azure Synapse Apache Spark to Synapse SQL connector, you're now able to efficiently transfer data between the Spark and SQL Pools.
The Azure Synapse Apache Spark pool to Synapse SQL connector is a data source implementation for Apache Spark. It uses the Azure Data Lake Storage Gen2 and PolyBase in SQL pools to efficiently transfer data between the Spark cluster and the Synapse SQL instance.
https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/synapse-spark-sql-pool-import-export

- [x] Big data computational powers

Explanation:- Synapse Analytics removes the barrier of setting up multiple different services for Spark or SQL. Therefore, it removes the traditional thinking about these technologies. It enables you to use both technologies within one platform, which allowed you to switch between Spark or SQL based on the needs and expertise you have in-house.

A spark orientated data engineer can now easily communicate with a SQL based data engineer and communicate together on the same platform.
The interoperability between Spark and SQL helps you achieve as follows:
• A shared Hive-compatible metadata system enables you to define tables on files in the data lake such that it can be consumed by either Spark or Hive.
• Both SQL and Spark can directly explore, and analyze Parquet, CSV, TSV, and JSON files stored in the data lake.
• The enablement of fast scalable load and unload for data transferring between SQL and Spark databases.
The question might raise as how would that SQL and Spark integration then work.
That's when the Azure Synapse Apache Spark to Synapse SQL connector comes in place. It is designed to efficiently transfer data between serverless Apache Spark pools (preview) and SQL pools in Azure Synapse. However, at the moment, the Azure Synapse Apache Spark to Synapse SQL connector works on dedicated SQL pools only, it doesn't work with serverless SQL pools.
In the commonly used existing approach, you often see the use of the JDBC. The JDBC would open the connection. Then, filters and projections would be applied and spark would read the data serially. Given two distributed systems such as Spark and SQL pools, JDBC could become a bottleneck with serial data transfer.
Therefore the New Approach we would take is JDBC and PolyBase. First, the JDBC issues CETAS and send filters and projections. Then filters and projections would be applied and the DataWarehouse exports the data in parallel. Spark reads the data in parallel all based on the user provisioned workspace default data lake storage.
The Azure Synapse Apache Spark Pool to Synapse SQL connector would then be a data source implementation for apache spark where the ADLS Gen 2 is used as well as PolyBase in the dedicated SQL Pools to transfer data between the Spark instance and SQL pool efficiently.
The use cases for Apache Spark and SQL integration within Synapse analytics are as following:
• Dealing with different type of analytics
• Scalability
• Big data computational powers
• Flexibility in the use of Spark and SQL languages and frameworks
Since Apache Spark is integrated in Synapse Analytics, there is more to that than giving use for the big data analytics framework Apache Spark enables. When you deploy a synapse cluster, ADLS Gen2 capacity that can store Spark SQL Tables is provisioned with it.
If you use Spark SQL Tables, you might know that these tables can be queried from a SQL-server-based T-SQL language without you having to use commands like CREATE EXTERNAL TABLE. Within synapse analytics, these queries integrate natively with data files that are stored in an Apache Parquet format.
The other thing to take in mind is that beyond the capabilities mentioned above, the Azure Synapse Studio experience gives you an integrated notebook experience. Within this notebook experience, you can attach a SQL or Spark pool, and develop and execute, for example, transformation pipelines using Python, Scala, and native Spark SQL.

So, let's say you would like to write to a SQL pool after you've performed engineering tasks in spark. You can reference the SQL Pool data as a source for joining with Spark Dataframes that can contain data from other files. When you decide to use the Azure Synapse Apache Spark to Synapse SQL connector, you're now able to efficiently transfer data between the Spark and SQL Pools.
The Azure Synapse Apache Spark pool to Synapse SQL connector is a data source implementation for Apache Spark. It uses the Azure Data Lake Storage Gen2 and PolyBase in SQL pools to efficiently transfer data between the Spark cluster and the Synapse SQL instance.
https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/synapse-spark-sql-pool-import-export

- [x] Scalability

Explanation:- Synapse Analytics removes the barrier of setting up multiple different services for Spark or SQL. Therefore, it removes the traditional thinking about these technologies. It enables you to use both technologies within one platform, which allowed you to switch between Spark or SQL based on the needs and expertise you have in-house.
A spark orientated data engineer can now easily communicate with a SQL based data engineer and communicate together on the same platform.
The interoperability between Spark and SQL helps you achieve as follows:
• A shared Hive-compatible metadata system enables you to define tables on files in the data lake such that it can be consumed by either Spark or Hive.
• Both SQL and Spark can directly explore, and analyze Parquet, CSV, TSV, and JSON files stored in the data lake.
• The enablement of fast scalable load and unload for data transferring between SQL and Spark databases.
The question might raise as how would that SQL and Spark integration then work.
That's when the Azure Synapse Apache Spark to Synapse SQL connector comes in place. It is designed to efficiently transfer data between serverless Apache Spark pools (preview) and SQL pools in Azure Synapse. However, at the moment, the Azure Synapse Apache Spark to Synapse SQL connector works on dedicated SQL pools only, it doesn't work with serverless SQL pools.
In the commonly used existing approach, you often see the use of the JDBC. The JDBC would open the connection. Then, filters and projections would be applied and spark would read the data serially. Given two distributed systems such as Spark and SQL pools, JDBC could become a bottleneck with serial data transfer.
Therefore the New Approach we would take is JDBC and PolyBase. First, the JDBC issues CETAS and send filters and projections. Then filters and projections would be applied and the DataWarehouse exports the data in parallel. Spark reads the data in parallel all based on the user provisioned workspace default data lake storage.
The Azure Synapse Apache Spark Pool to Synapse SQL connector would then be a data source implementation for apache spark where the ADLS Gen 2 is used as well as PolyBase in the dedicated SQL Pools to transfer data between the Spark instance and SQL pool efficiently.
The use cases for Apache Spark and SQL integration within Synapse analytics are as following:

• Dealing with different type of analytics
• Scalability
• Big data computational powers
• Flexibility in the use of Spark and SQL languages and frameworks
Since Apache Spark is integrated in Synapse Analytics, there is more to that than giving use for the big data analytics framework Apache Spark enables. When you deploy a synapse cluster, ADLS Gen2 capacity that can store Spark SQL Tables is provisioned with it.
If you use Spark SQL Tables, you might know that these tables can be queried from a SQL-server-based T-SQL language without you having to use commands like CREATE EXTERNAL TABLE. Within synapse analytics, these queries integrate natively with data files that are stored in an Apache Parquet format.
The other thing to take in mind is that beyond the capabilities mentioned above, the Azure Synapse Studio experience gives you an integrated notebook experience. Within this notebook experience, you can attach a SQL or Spark pool, and develop and execute, for example, transformation pipelines using Python, Scala, and native Spark SQL.
So, let's say you would like to write to a SQL pool after you've performed engineering tasks in spark. You can reference the SQL Pool data as a source for joining with Spark Dataframes that can contain data from other files. When you decide to use the Azure Synapse Apache Spark to Synapse SQL connector, you're now able to efficiently transfer data between the Spark and SQL Pools.
The Azure Synapse Apache Spark pool to Synapse SQL connector is a data source implementation for Apache Spark. It uses the Azure Data Lake Storage Gen2 and PolyBase in SQL pools to efficiently transfer data between the Spark cluster and the Synapse SQL instance.
https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/synapse-spark-sql-pool-import-export

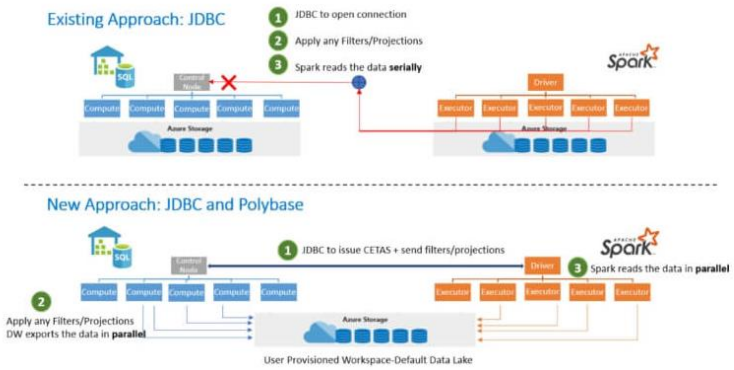
- [ ] VNet and On-prem sync

- [x] Flexibility in the use of Spark and SQL languages and frameworks

Explanation:- Synapse Analytics removes the barrier of setting up multiple different services for Spark or SQL. Therefore, it removes the traditional thinking about these technologies. It enables you to use both technologies within one platform, which allowed you to switch between Spark or SQL based on the needs and expertise you have in-house.
A spark orientated data engineer can now easily communicate with a SQL based data engineer and communicate together on the same platform.
The interoperability between Spark and SQL helps you achieve as follows:
• A shared Hive-compatible metadata system enables you to define tables on files in the data lake such that it can be consumed by either Spark or Hive.
• Both SQL and Spark can directly explore, and analyze Parquet, CSV, TSV, and JSON files stored in the data lake.

• The enablement of fast scalable load and unload for data transferring between SQL and Spark databases.
The question might raise as how would that SQL and Spark integration then work.
That's when the Azure Synapse Apache Spark to Synapse SQL connector comes in place. It is designed to efficiently transfer data between serverless Apache Spark pools (preview) and SQL pools in Azure Synapse. However, at the moment, the Azure Synapse Apache Spark to Synapse SQL connector works on dedicated SQL pools only, it doesn't work with serverless SQL pools.
In the commonly used existing approach, you often see the use of the JDBC. The JDBC would open the connection. Then, filters and projections would be applied and spark would read the data serially. Given two distributed systems such as Spark and SQL pools, JDBC could become a bottleneck with serial data transfer.
Therefore the New Approach we would take is JDBC and PolyBase. First, the JDBC issues CETAS and send filters and projections. Then filters and projections would be applied and the DataWarehouse exports the data in parallel. Spark reads the data in parallel all based on the user provisioned workspace default data lake storage.
The Azure Synapse Apache Spark Pool to Synapse SQL connector would then be a data source implementation for apache spark where the ADLS Gen 2 is used as well as PolyBase in the dedicated SQL Pools to transfer data between the Spark instance and SQL pool efficiently.
The use cases for Apache Spark and SQL integration within Synapse analytics are as following:
• Dealing with different type of analytics
• Scalability
• Big data computational powers
• Flexibility in the use of Spark and SQL languages and frameworks
Since Apache Spark is integrated in Synapse Analytics, there is more to that than giving use for the big data analytics framework Apache Spark enables. When you deploy a synapse cluster, ADLS Gen2 capacity that can store Spark SQL Tables is provisioned with it.
If you use Spark SQL Tables, you might know that these tables can be queried from a SQL-server-based T-SQL language without you having to use commands like CREATE EXTERNAL TABLE. Within synapse analytics, these queries integrate natively with data files that are stored in an Apache Parquet format.
The other thing to take in mind is that beyond the capabilities mentioned above, the Azure Synapse Studio experience gives you an integrated notebook experience. Within this notebook experience, you can attach a SQL or Spark pool, and develop and execute, for example, transformation pipelines using Python, Scala, and native Spark SQL.
So, let's say you would like to write to a SQL pool after you've performed engineering tasks in spark. You can reference the SQL Pool data as a source for joining with Spark Dataframes that can contain data from other files. When you decide to use the Azure Synapse Apache Spark to Synapse SQL connector, you're now able to efficiently transfer data between the Spark and SQL Pools.

The Azure Synapse Apache Spark pool to Synapse SQL connector is a data source implementation for Apache Spark. It uses the Azure Data Lake Storage Gen2 and PolyBase in SQL pools to efficiently transfer data between the Spark cluster and the Synapse SQL instance.
https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/synapse-spark-sql-pool-import-export



Report Error

Q. 42

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 done to improve high availability of the real-time data processing solution. Which of the following should you propose as the best solution?

- ○ Set Data Lake Storage to use geo-redundant storage (GRS).
- ○ Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- ○ Deploy a High Concurrency Databricks cluster.
- ◉ Deploy identical Azure Stream Analytics jobs to paired regions in Azure.

Explanation:- The best solution to move forward is to deploy identical Azure Stream Analytics jobs to paired regions in Azure. The application development team should 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.
Guarantee Stream Analytics job reliability during service updates
Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.
How do Azure paired regions address this concern?
Stream Analytics guarantees jobs in paired regions are updated in separate batches. As a result there is a sufficient time gap between the updates to identify potential issues and remediate them.

With the exception of Central India (whose paired region, South India, does not have Stream Analytics presence), the deployment of an update to Stream Analytics would not occur at the same time in a set of paired regions. Deployments in multiple regions in the same group may occur at the same time.
The article on availability and paired regions has the most up-to-date information on which regions are paired.
It is recommended to deploy identical jobs to both paired regions. You should then monitor these jobs to get notified when something unexpected happens. If one of these jobs ends up in a Failed state after a Stream Analytics service update, you can contact customer support to help identify the root cause. You should also fail over any downstream consumers to the healthy job output.
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability

Q. 43

Scenario: Honest Eddie's Car Dealership is an establishment in South Carolina USA, which is dedicated to the purchase and sale of cars and light trucks. Currently the IT team is trying to determine whether they should use a star schema or a snowflake schema.

What is the difference between a star schema and a snowflake schema?

- ○ All dimensions in a star schema are normalized while all dimensions in a snowflake schema join directly to the fact table (denormalized).
- ◉ All dimensions in a star schema join directly to the fact table (denormalized) while some dimension tables in a snowflake schema are normalized.

Explanation:-

A star schema is highly denormalized so that the fact table joins directly to dimension; a snowflake schema normalizes some dimensions into multiple tables such as DimProduct, DimProductSubcategory, and DimProductCategory.

Star schema is a mature modelling approach widely adopted by relational data warehouses. It requires modellers to classify their model tables as either dimension or fact.

Dimension tables

Dimension tables describe business entities—the things you model. Entities can include products, people, places, and concepts including time itself. The most consistent table you'll find in a star schema is a date dimension table. A dimension table contains a key column (or columns) that acts as a unique identifier, and descriptive columns.

Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table. Instead, the fact table and the dimension table can share a customer ID. A query can join the two tables to associate a customer's profile and transactions.

Fact tables

Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc. A fact table contains dimension key columns that relate to dimension tables, and numeric measure columns. The dimension key columns determine the dimensionality of a fact table, while the dimension key values determine the granularity of a fact table. For example, consider a fact table designed to store sale targets that has two dimension key columns Date and ProductKey. It's easy to understand that the table has two dimensions. The granularity, however, can't be determined without considering the dimension key values. In this example, consider that the values stored in the Date column are the first day of each month. In this case, the granularity is at month-product level.

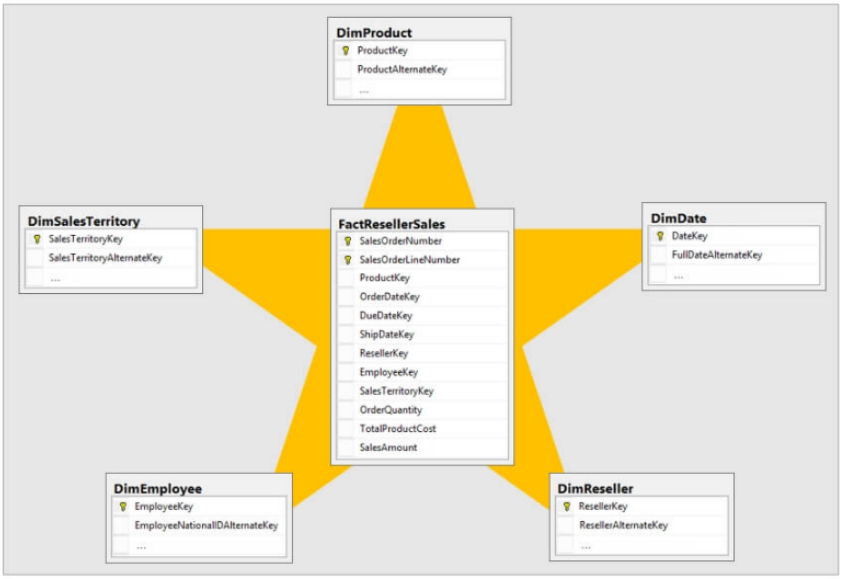
Generally, dimension tables contain a relatively small number of rows. Fact tables, on the other hand, can contain a very large number of rows and continue to grow over time.

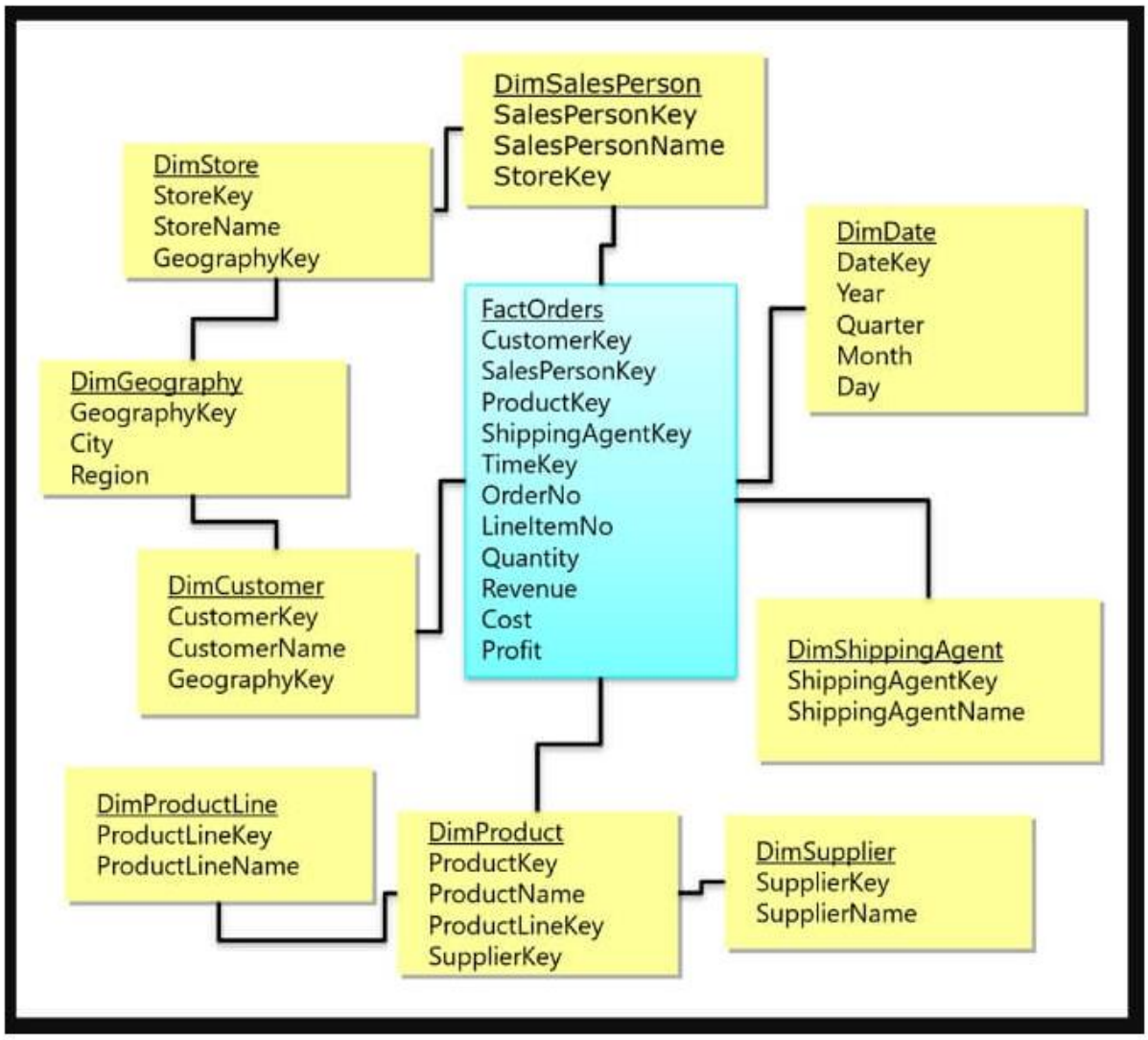
Below is an example star schema, where the fact table is in the middle, surrounded by dimension tables:

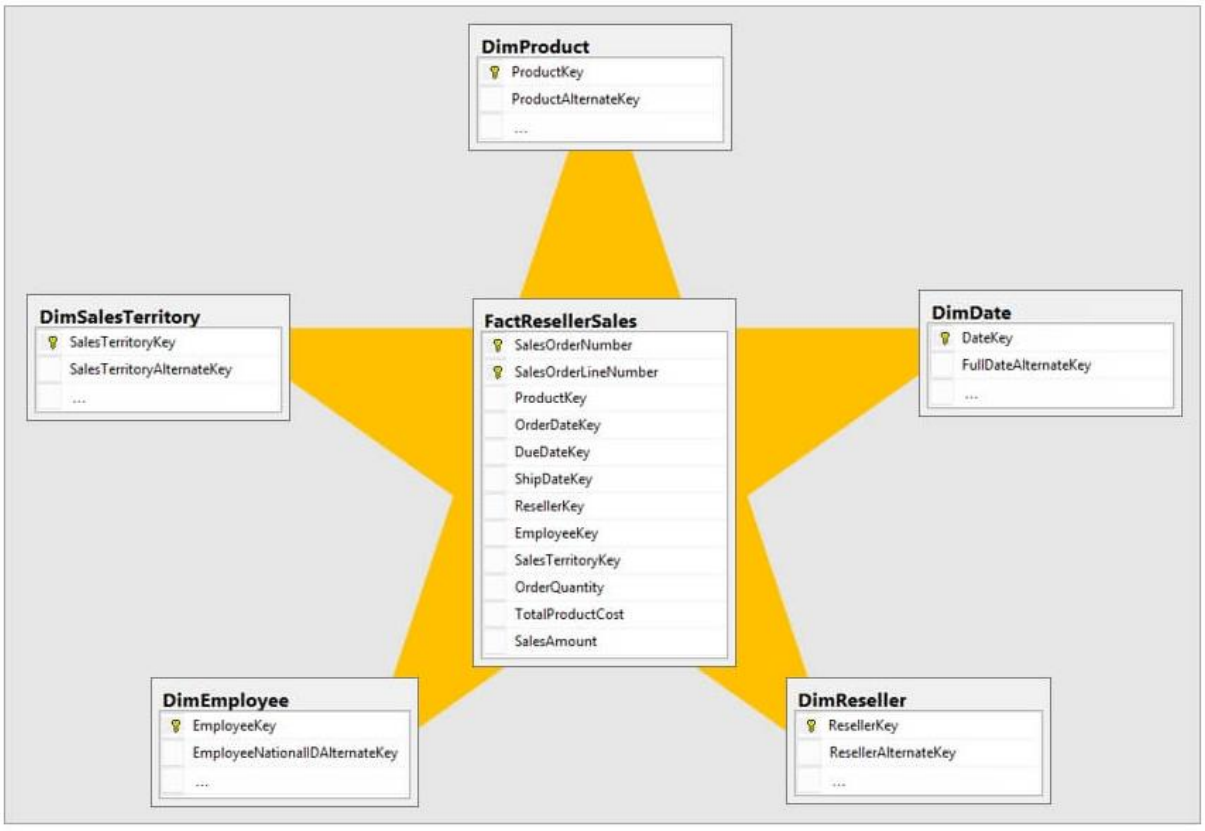
https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

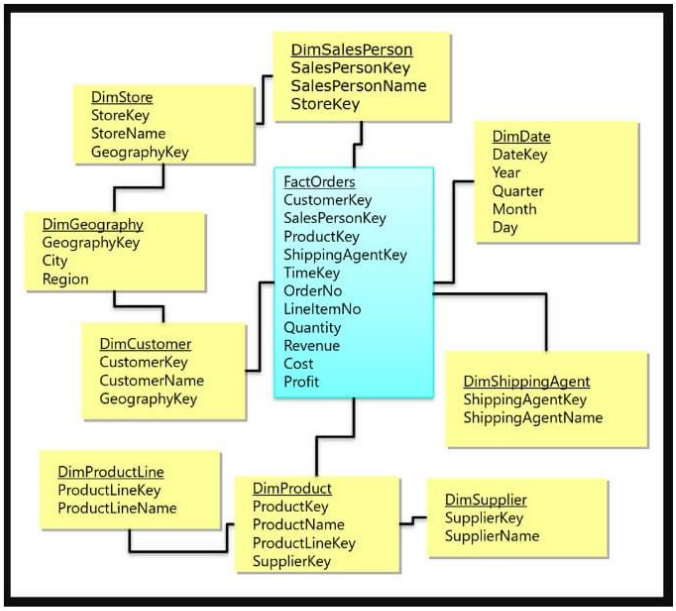
A snowflake schema is a set of normalized tables for a single business entity. For example, Adventure Works classifies products by category and subcategory. Categories are assigned to subcategories, and products are in turn assigned to subcategories. In the Adventure Works relational data warehouse, the product dimension is normalized and stored in three related tables: DimProductCategory, DimProductSubcategory, and DimProduct.

The snowflake schema is a variation of the star schema. You add normalized dimension tables to a star schema to create a snowflake pattern. In the following diagram, you see the yellow dimension tables surrounding the blue fact table. Notice that many of the dimension tables relate to one another in order to normalize the business entities:




DimSalesPerson
SalesPersonKey
SalesPersonName
StoreKey
DimStore
StoreKey
StoreName
GeographyKey
DimDate
DateKey
Year
Quarter
Month
Day
FactOrders
CustomerKey
SalesPersonKey
ProductKey
ShippingAgentKey
TimeKey
OrderNo
LineItemNo
Quantity
Revenue
Cost
Profit
DimGeography
GeographyKey
City
Region
DimCustomer
CustomerKey
CustomerName
GeographyKey
DimShippingAgent
ShippingAgentKey
ShippingAgentName
DimProductLine
ProductLineKey
ProductLineName
DimProduct
ProductKey
ProductName
ProductLineKey
SupplierKey
DimSupplier
SupplierKey
SupplierName


DimProduct
ProductKey
ProductAlternateKey
...
DimSalesTerritory
SalesTerritoryKey
SalesTerritoryAlternateKey
...
FactResellerSales
SalesOrderNumber
SalesOrderLineNumber
ProductKey
OrderDateKey
DueDateKey
ShipDateKey
ResellerKey
EmployeeKey
SalesTerritoryKey
OrderQuantity
TotalProductCost
SalesAmount
DimDate
DateKey
FullDateAlternateKey
...
DimEmployee
EmployeeKey
EmployeeNationalIDAlternateKey
...
DimReseller
ResellerKey
ResellerAlternateKey
...



- ○ A star schema uses surrogate keys while a snowflake schema uses business keys.
- ○ A star schema has one fact table while a snowflake schema has multiple fact tables.

Report Error

Q. 44

Azure HDInsight is a low-cost cloud solution which provides technologies to help you ingest, process, and analyze big data.

Which of the following are supported in the HDInsight solution? (Select all that apply)

- ○ Storm
- ◉ All of these

Explanation:- Azure HDInsight provides technologies to help you ingest, process, and analyze big data. It supports batch processing, data warehousing, IoT, and data science.
Key features
HDInsight is a low-cost cloud solution. HDInsight supports the latest open-source projects from the Apache Hadoop and Spark ecosystems. It includes Apache Hadoop, Spark, Kafka, HBase, Storm, and Interactive Query.
• Hadoop includes Apache Hive, HBase, Spark, and Kafka. Hadoop stores data in a file system (HDFS). Spark stores data in memory. This difference in storage makes Spark about 100 times faster.
• HBase is a NoSQL database built on Hadoop. It's commonly used for search engines. HBase offers automatic failover.
• Storm is a distributed real-time streamlining analytics solution.
• Kafka is an open-source platform that's used to compose data pipelines. It offers message queue functionality, which allows users to publish or subscribe to real-time data streams.
Ingesting data
As a data engineer, use Hive to run ETL operations on the data you're ingesting. Or orchestrate Hive queries in Azure Data Factory.
https://azure.microsoft.com/en-us/services/hdinsight/#features

- ○ Hbase
- ○ Spark
- ○ Interactive Query
- ○ Hadoop

Report Error

Q. 45

Scenario: Queen Consolidated was overtaken by Raymond Carson Palmer and rebranded as Palmer Technologies. Now that Ray is overseeing the operations at Palmer, Ray has decided to implement on-premises Microsoft SQL Server pipelines by using a custom solution.
Currently, you are in a meeting with the IT team and discussing a project to pull data from SQL Server and migrate it to Azure Blob storage.

Required:
• The process must orchestrate and manage the data lifecycle.
• The process must configure Azure Data Factory to connect to the on-premises SQL Server database.

Ray and the IT team have put together a list of actions they think need to be performed to meet the needs of the project, but they are not sure on the order to execute. Below is a list of the actions they are considering.

Proposed Actions:
a. Create an Azure Data Factory resource.
b. Install and configure Azure Data Factory Integration Runtime
c. Create a virtual private network (VPN) connection from on-prem to MS Azure.
d. Create a database master key on SQL Server.
e. Backup the database and send it to Azure Blob storage.
f. Configure the on-prem SQL Server instance with an integration runtime.

As you are the Azure SME, Ray and the team look to you for direction on selecting the required items and putting them in the proper order.

Which of the below contains the correct items in the correct sequence to meet the requirements?

- ○ c,d,a,b,f
- ◉ a,b,f

Explanation:- Step 1: Create an Azure Data Factory
The instructions for creating a new Azure Data Factory and a resource group in the Azure portal are provided Create an Azure Data Factory. Name the new ADF instance adfdsp and name the resource group created adfdsprg.
Step 2: Install and configure Azure Data Factory Integration Runtime
The Integration Runtime is a customer-managed data integration infrastructure used by Azure Data Factory to provide data integration capabilities across different network environments. This runtime was formerly called "Data Management Gateway".
To set up, follow the instructions for creating a pipeline
Step 3: Configure the on-prem SQL Server instance with an integration runtime.
Create linked services to connect to the data resources. A linked service defines the information needed for Azure Data Factory to connect to a data resource. We have three resources in this scenario for which linked services are needed:
1. On-premises SQL Server
2. Azure Blob Storage
3. Azure SQL Database
https://docs.microsoft.com/pt-pt/azure/machine-learning/team-data-science-process/move-sql-azure-adf
It's not necessary to Create a virtual private network (VPN) connection from on-premises to Microsoft Azure - all communication from IR to ADF is over HTTPS, ? VPN is not a Required item.
Encryption in transit
All data transfers are via secure channel HTTPS and TLS over TCP to prevent man-in-the-middle attacks during communication with Azure services.
You can also use IPSec VPN or Azure ExpressRoute to further secure the communication channel between your on-premises network and Azure.
Azure Virtual Network is a logical representation of your network in the cloud. You can connect an on-premises network to your virtual network by setting up IPSec VPN (site-to-site) or ExpressRoute (private peering).
The following table summarizes the network and self-hosted integration runtime configuration recommendations based on different combinations of source and destination locations for hybrid data movement.
https://docs.microsoft.com/en-us/azure/data-factory/data-movement-security-considerations

It's not necessary to Create a virtual private network (VPN) connection from on-premises to Microsoft Azure - all communication from IR to ADF is over HTTPS, ? VPN is not a Required item.
Encryption in transit
All data transfers are via secure channel HTTPS and TLS over TCP to prevent man-in-the-middle attacks during communication with Azure services.
You can also use IPSec VPN or Azure ExpressRoute to further secure the communication channel between your on-premises network and Azure.
Azure Virtual Network is a logical representation of your network in the cloud. You can connect an on-premises network to your virtual network by setting up IPSec VPN (site-to-site) or ExpressRoute (private peering).
The following table summarizes the network and self-hosted integration runtime configuration recommendations based on different combinations of source and destination locations for hybrid data movement.
https://docs.microsoft.com/en-us/azure/data-factory/data-movement-security-considerations
Move data from a SQL Server database to the SQL Database with Azure Data Factory
AZure Data Factory is a fully managed cloud-based data integration service that orchestrates and automates the movement and transformation of data. The key concept in the ADF model is the pipeline. A pipeline is a logical grouping of Activities, each of which defines the actions to be performed on the data contained in the Data Sets. The linked services are used to define the information necessary for the Data Factory to connect to the data resources.
With the ADF, existing data processing services can be composed of highly available data pipelines and managed in the cloud. These data pipelines can be programmed to ingest, prepare, transform, analyze and publish data, and the ADF manages and orchestrates the complex data and processing dependencies. Solutions can be quickly built and deployed in the cloud, connecting an increasing number of data sources on-premises and in the cloud.
Consider using the ADF:
• when data needs to be continuously migrated in a hybrid scenario that accesses both on-premises and cloud resources
• when data needs transformation or has business logic added to it when it is migrated.
The ADF allows scheduling and monitoring of jobs using simple JSON scripts that manage the movement of data on a periodic basis. ADF also has other capabilities, such as supporting complex operations. For more information about the ADF, see the documentation at Azure Data Factory (ADF).
The set
We created an ADF pipeline that comprises two data migration activities. Together, they move data daily between a SQL Server database and the Azure SQL Database. The two activities are:
• Copy data from a SQL Server database to an Azure Blob storage account.
• Copy the data from the Azure Blob storage account to the Azure SQL Database.
Upload the data to your instance of SQL Server
Create an Azure Data Factory

Create an Azure Data Factory
Instructions for creating a new Azure Data Factory and resource group on the Azure portal are provided Create an Azure Data Factory. Name the new adf instance adfdsp and name the resource group created adfdsprg.
Install and configure Azure data factory integration time
Integration Runtime is a customer-managed data integration infrastructure used by Azure Data Factory to provide data integration capabilities in different network environments. This uptime was previously called "Data Management Gateway".
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/move-sql-azure-adf

Source	Destination	Network configuration	Integration runtime setup
On-premises	Virtual machines and cloud services deployed in virtual networks	IPSec VPN (point-to-site or site-to-site)	The self-hosted integration runtime should be installed on an Azure virtual machine in the virtual network.
On-premises	Virtual machines and cloud services deployed in virtual networks	ExpressRoute (private peering)	The self-hosted integration runtime should be installed on an Azure virtual machine in the virtual network.
On-premises	Azure-based services that have a public endpoint	ExpressRoute (Microsoft peering)	The self-hosted integration runtime can be installed on-premises or on an Azure virtual machine.

- ○ c,a,b,f
- ○ d,c,e,b

Report Error

Q. 46

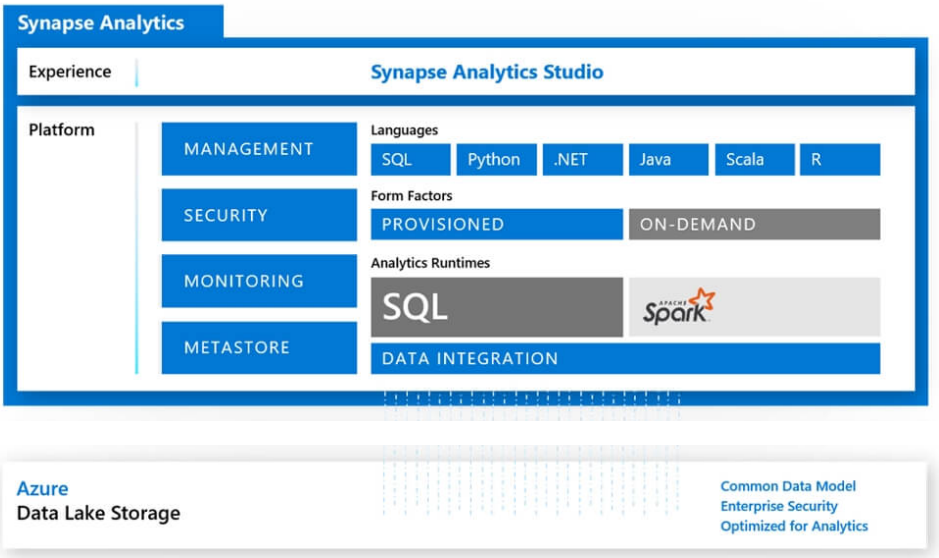
Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

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 [?], you can create and schedule data-driven workflows that can ingest data from disparate data stores.

- ○ Azure Synapse SQL
- ○ Azure Synapse Link
- ◉ Azure Synapse Pipelines

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



- ○ Apache Spark for Azure Synapse
- ○ Azure Cosmos DB

Report Error

Q. 47 How many drivers does a Cluster have?

- ○ Configurable between one and eight
- ○ Configurable between one and ten
- ○ Two, running in parallel
- ◉ Only one

Explanation:- A Cluster has one and only one driver.
Cluster node type
A cluster consists of one driver node and worker nodes.
You can pick separate cloud provider instance types for the driver and worker nodes, although by default the driver node uses the same instance type as the worker node. Different families of instance types fit different use cases, such as memory-intensive or compute-intensive workloads.
https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

Report Error

Q. 48 What should be done when a connector in data factory is not supported in mapping data flow in order to transform data from one of these sources? (Select all that apply)

- [x] Ingest the data into a supported source using the copy activity.

Explanation:- If a connector in Data factory is not supported, create a copy activity of the source data into a supported data source in mapping dataflow and continue the transformations from there.
Integrate with more data stores
Azure Data Factory can reach a very broad set of data stores. If you need to move data to/from a data store that is not in the Azure Data Factory built-in connector list, here are some extensible options:
• For database and data warehouse, usually you can find a corresponding ODBC driver, with which you can use generic ODBC connector.
• For SaaS applications:
• If it provides RESTful APIs, you can use generic REST connector.
• If it has OData feed, you can use generic OData connector.
• If it provides SOAP APIs, you can use generic HTTP connector.
• If it has ODBC driver, you can use generic ODBC connector.
• For others, check if you can load data to or expose data as any ADF supported data stores, e.g. Azure Blob/File/FTP/SFTP/etc, then let ADF pick up from there. You can invoke custom data loading mechanism via Azure Function, Custom activity, Databricks/HDInsight, Web activity, etc.

- [ ] Use an aggregate transformation in Dataflow.

- [ ] Use a group by activity in Dataflow.

- [x] Use a generic ODBC connector.

Explanation:- If a connector in Data factory is not supported, create a copy activity of the source data into a supported data source in mapping dataflow and continue the transformations from there.
Integrate with more data stores
Azure Data Factory can reach a very broad set of data stores. If you need to move data to/from a data store that is not in the Azure Data Factory built-in connector list, here are some extensible options:
• For database and data warehouse, usually you can find a corresponding ODBC driver, with which you can use generic ODBC connector.
• For SaaS applications:
• If it provides RESTful APIs, you can use generic REST connector.
• If it has OData feed, you can use generic OData connector.
• If it provides SOAP APIs, you can use generic HTTP connector.
• If it has ODBC driver, you can use generic ODBC connector.
• For others, check if you can load data to or expose data as any ADF supported data stores, e.g. Azure Blob/File/FTP/SFTP/etc, then let ADF pick up from there. You can invoke custom data loading mechanism via Azure Function, Custom activity, Databricks/HDInsight, Web activity, etc.

- [x] Use a generic REST connector.

Explanation:- If a connector in Data factory is not supported, create a copy activity of the source data into a supported data source in mapping dataflow and continue the transformations from there.
Integrate with more data stores
Azure Data Factory can reach a very broad set of data stores. If you need to move data to/from a data store that is not in the Azure Data Factory built-in connector list, here are some extensible options:
• For database and data warehouse, usually you can find a corresponding ODBC driver, with which you can use generic ODBC connector.
• For SaaS applications:
• If it provides RESTful APIs, you can use generic REST connector.
• If it has OData feed, you can use generic OData connector.
• If it provides SOAP APIs, you can use generic HTTP connector.
• If it has ODBC driver, you can use generic ODBC connector.
• For others, check if you can load data to or expose data as any ADF supported data stores, e.g. Azure Blob/File/FTP/SFTP/etc, then let ADF pick up from there. You can invoke custom data loading mechanism via Azure Function, Custom activity, Databricks/HDInsight, Web activity, etc.

Q. 49

Scenario: You are working in a department which requires preparation of data for ad hoc data exploration and analysis based on market fluctuations. The Department Head has tasked you with determining the most effective resource model in Azure Synapse Analytics to employ.

Which of the following should you choose?

- ○ Pipelines
- ◉ Serverless

Explanation:- Serverless SQL pool is a pay per query service that doesn't require you to pick the right size. The system automatically adjusts based on your requirements, freeing you up from managing your infrastructure and picking the right size for your solution.
The serverless resource model is the ideal resource model in this scenario as it makes use of the resources when required.
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/resource-consumption-models

- ○ IoT Central
- ○ Databricks
- ○ Dedicated

Report Error

Q. 50

Scenario: You are working on a project and your team is moving data from an Azure Data Lake Gen2 store to Azure Synapse Analytics. The team is planning to do a data copy activity and you are discussing with integration runtime to use.

Which Azure Data Factory integration runtime should be used in a data copy activity?

- ○ Linked Services
- ○ Activities
- ○ Azure-SSIS
- ○ Self-hosted
- ○ Datasets
- ◉ Azure

Explanation:- When moving data between Azure data platform technologies, the Azure Integration runtime is used when copying data between two Azure data platform.
Integration runtime types
Data Factory offers three types of Integration Runtime, and you should choose the type that best serve the data integration capabilities and network environment needs you are looking for. These three types are:
• Azure
• Self-hosted
• Azure-SSIS
You can explicitly define the Integration Runtime setting in the connectVia property, if this is not defined, then the default Integration Runtime is used with the property set to Auto-Resolve.
The following describes the capabilities and network support for each of the integration runtime types:
IR type: Azure
Public network: Data Flow Data movement Activity dispatch
Private network: --
IR type: Self-hosted
Public network: Data movement Activity dispatch
Private network: Data movement Activity dispatch
IR type: Azure-SSIS
Public network: SSIS package execution
Private network: SSIS package execution
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

Report Error

Q. 51

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

Roles such as business stakeholders, business analysts, and business intelligence developers are well known and valuable.
Which of the available roles is best described by:

"Works with artificial intelligence services such as Cognitive Services, Cognitive Search, and Bot Framework. Cognitive Services includes Computer Vision, Text Analytics, Bing Search, and Language Understanding (LUIS).

Rather than creating models, they apply the pre-built capabilities of Cognitive Services APIs. Part of their job is to embed these capabilities within a new or existing application or bot. They rely on the expertise of data engineers to store information that's generated from artificial intelligence."

- [x] AI Engineer

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

- [ ] Data Engineer
- [ ] Project Manager
- [ ] Data Scientist
- [ ] Solution Architects
- [ ] RPA Developers

Report Error

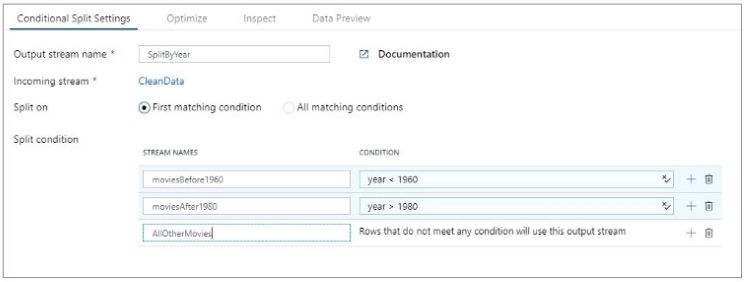
Q. 52

Scenario: While working on a project using Azure Data Factory, you are routing data rows to different streams based on matching conditions.

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

- Optimize
- Lookup
- Inspect
- Select
- Conditional Split

Explanation:- Conditional Split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language.
The Split on setting determines whether the row of data flows to the first matching stream or every stream it matches to.
Use the data flow expression builder to enter an expression for the split condition. To add a new condition, click on the plus icon in an existing row. A default stream can be added as well for rows that don't match any condition.
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split



Report Error

Q. 53 What steps are required to authorize Azure DevOps to connect to and deploy notebooks to a staging or production Azure Databricks workspace?

- None of the listed options.
- In the production or staging Azure Databricks workspace, enable Git integration to Azure DevOps, then link to the Azure DevOps source code repo.
- Create an Azure Active Directory application, copy the application ID, then use that as the Databricks bearer token in the Databricks Notebooks Deployment step of the Release pipeline.
- Create a new Access Token within the user settings in the production Azure Databricks workspace, then use the token as the Databricks bearer token in the Databricks Notebooks Deployment step of the Release pipeline.

Explanation:- To authorize Azure DevOps to connect to and deploy notebooks to a staging or production Azure Databricks workspace, create an Azure Active Directory application, copy the application ID, then use that as the Databricks bearer token in the Databricks Notebooks Deployment step of the Release pipeline.
The Access Token allows you to grant access to resources within an Azure Databricks workspace without passing in user credentials.
https://social.technet.microsoft.com/wiki/contents/articles/53094.azure-devops-integrate-with-an-azure-subscription-or-management-group.aspx

Report Error

Q. 54 Which transformation in the Mapping Data Flow is used to routes data rows to different streams based on matching conditions?

- ○ Alter row
- ○ Multiple inputs/outputs
- ○ Lookup
- ◉ Conditional Split

Explanation:- A Conditional Split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language.
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split
Transforming data using Mapping Data Flow
Mapping Data Flows provide an environment for building a wide range of data transformations visually without the need to use code. The resulting data flows that are created are subsequently executed on scaled-out Apache Spark clusters that are automatically provisioned when you execute the Mapping Data Flow. Mapping Data Flows also provides the capability to monitor the execution of the transformations so that you can view how the transformations are progressing, or to understand any errors that may occur
Mapping Data Flows provides a number of different transformations types that enable you to modify data. They are broken down into the following categories:
• Schema modifier transformations
• Row modifier transformations
• Multiple inputs/outputs transformations
Below is a list of transformations that is available in the Mapping Data Flows:
Name & Category: Aggregate - Schema modifier
Description: Define different types of aggregations such as SUM, MIN, MAX, and COUNT grouped by existing or computed columns.
Name & Category: Alter row - Row modifier

Description: Set insert, delete, update, and upsert policies on rows. You can add one-to-many conditions as expressions. These conditions should be specified in order of priority, as each row will be marked with the policy corresponding to the first-matching expression. Each of those conditions can result in a row (or rows) being inserted, updated, deleted, or upserted. Alter Row can produce both DDL & DML actions against your database.
Name & Category: Conditional split - Multiple inputs/outputs
Description: Route rows of data to different streams based on matching conditions.
Name & Category: Derived column - Schema modifier
Description: Generate new columns or modify existing fields using the data flow expression language.
Name & Category: Exists - Multiple inputs/outputs
Description: Check whether your data exists in another source or stream.
Name & Category: Filter - Row modifier
Description: Filter a row based upon a condition.
Name & Category: Flatten - Schema modifier
Description: Take array values inside hierarchical structures such as JSON and unroll them into individual rows.
Name & Category: Join - Multiple inputs/outputs
Description: Combine data from two sources or streams
Name & Category: Lookup - Multiple inputs/outputs
Description: Enables you to reference data from another source.
Name & Category: New branch - Multiple inputs/outputs
Description: Apply multiple sets of operations and transformations against the same data stream.
Name & Category: Pivot - Schema modifier
Description: An aggregation where one or more grouping columns has distinct row values transformed into individual columns.
Name & Category: Select - Schema modifier
Description: Alias columns and stream names, and drop or reorder columns.
Name & Category: Sink – N/A
Description: A final destination for your data
Name & Category: Sort - Row modifier
Description: Sort incoming rows on the current data stream
Name & Category: Source – N/A
Description: A data source for the data flow
Name & Category: Surrogate key - Schema modifier

Description: Pivot columns into row values.
Name & Category: Window - Schema modifier
Description: Define window-based aggregations of columns in your data streams.
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

- ○ Derived column
- ○ Select

Report Error

Q. 55

Before we can query our data using Azure Synapse Analytics using Azure Synapse Link, we must first create the container that is going to hold our data at the same time enabling it to have an analytical store.

Correct or Incorrect : Enabling analytical store is only available at the time of creating a container however it can be deactivated or reactivated at anytime thereafter.

- ◉ Incorrect

Explanation:- Before we can query our data using Azure Synapse Analytics using Azure Synapse Link, we must first create the container that is going to hold our data at the same time enabling it to have an analytical store.
Enabling analytical store is only available at the time of creating a container and cannot be completely disabled without deleting the container. Setting the default analytical store TTL value to 0 or null effectively disables the analytical store by no longer synchronize new items to it from the transactional store and deleting items already synchronized from the analytical store.
https://docs.microsoft.com/en-us/azure/cosmos-db/configure-synapse-link

- ○ Correct

Report Error

Q. 56

Scenario: You have been assigned to a new project and your first task is to initialize the Blob Storage client library within an application.

Which of the following can be used to do this?

- ○ The Azure Storage account datacentre and location identifiers.
- ◉ The Azure Storage account connection string.

Explanation:- A storage account connection string contains all the information needed to connect to Blob storage, most importantly the account name and the account key.
https://docs.microsoft.com/en-us/azure/storage/common/storage-configure-connection-string

- ○ A globally-unique identifier (GUID) that represents the application.
- ○ An Azure username and password.

Report Error

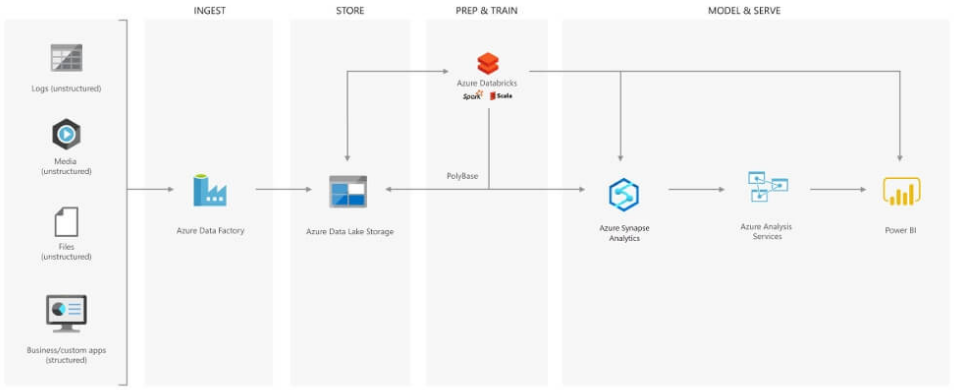
Q. 57

Scenario: You are a Data Engineering consultant for a Avengers Security. In the past, they've created an on-premises business intelligence solution that used a Microsoft SQL Server Database Engine, SQL Server Integration Services, SQL Server Analysis Services, and SQL Server Reporting Services to provide historical reports. They tried using the Analysis Services Data Mining component to create a predictive analytics solution to predict the buying behaviour of customers. While this approach worked well with low volumes of data, it couldn't scale after more than a gigabyte of data was collected. Furthermore, they were never able to deal with the JSON data that a third-party application generated when a customer used the feedback module of the point of sale (POS) application.

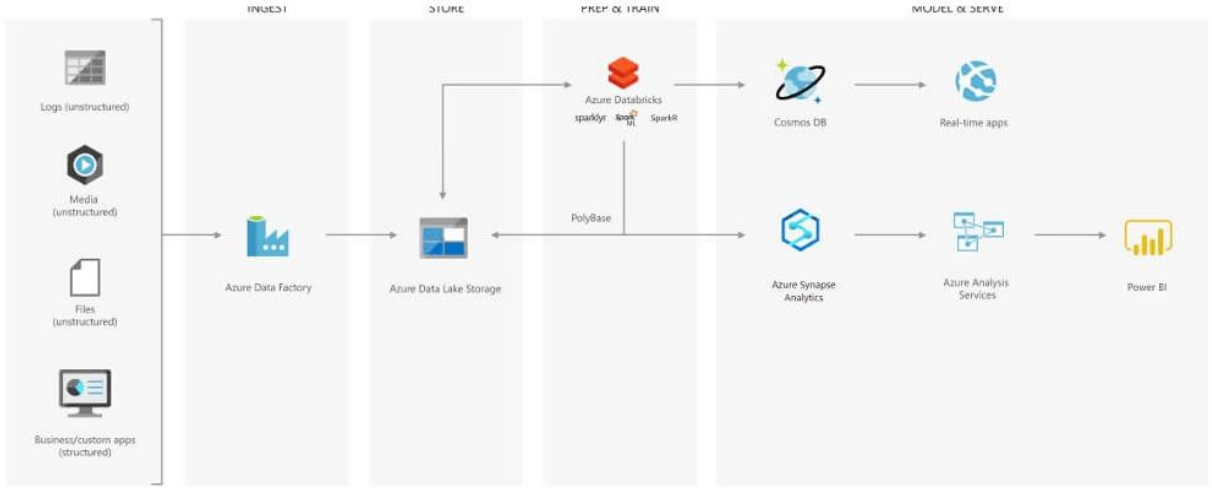
The company has turned to you for help with creating an architecture that can scale with the data needs that are required to create a predictive model and to handle the JSON data so that it's integrated into the BI solution.

Review the following architecture designs.

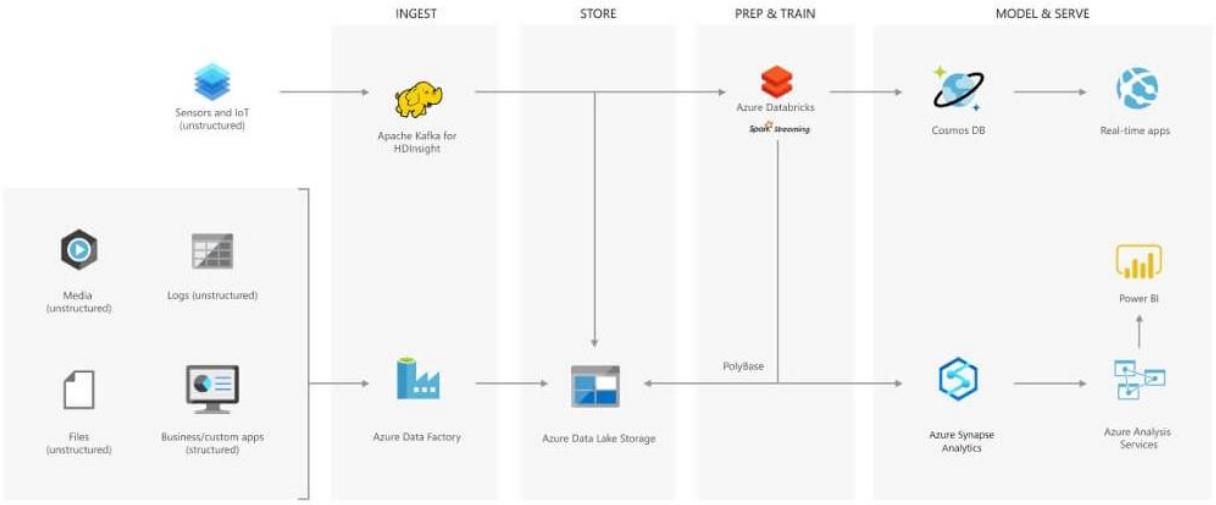
Design A:



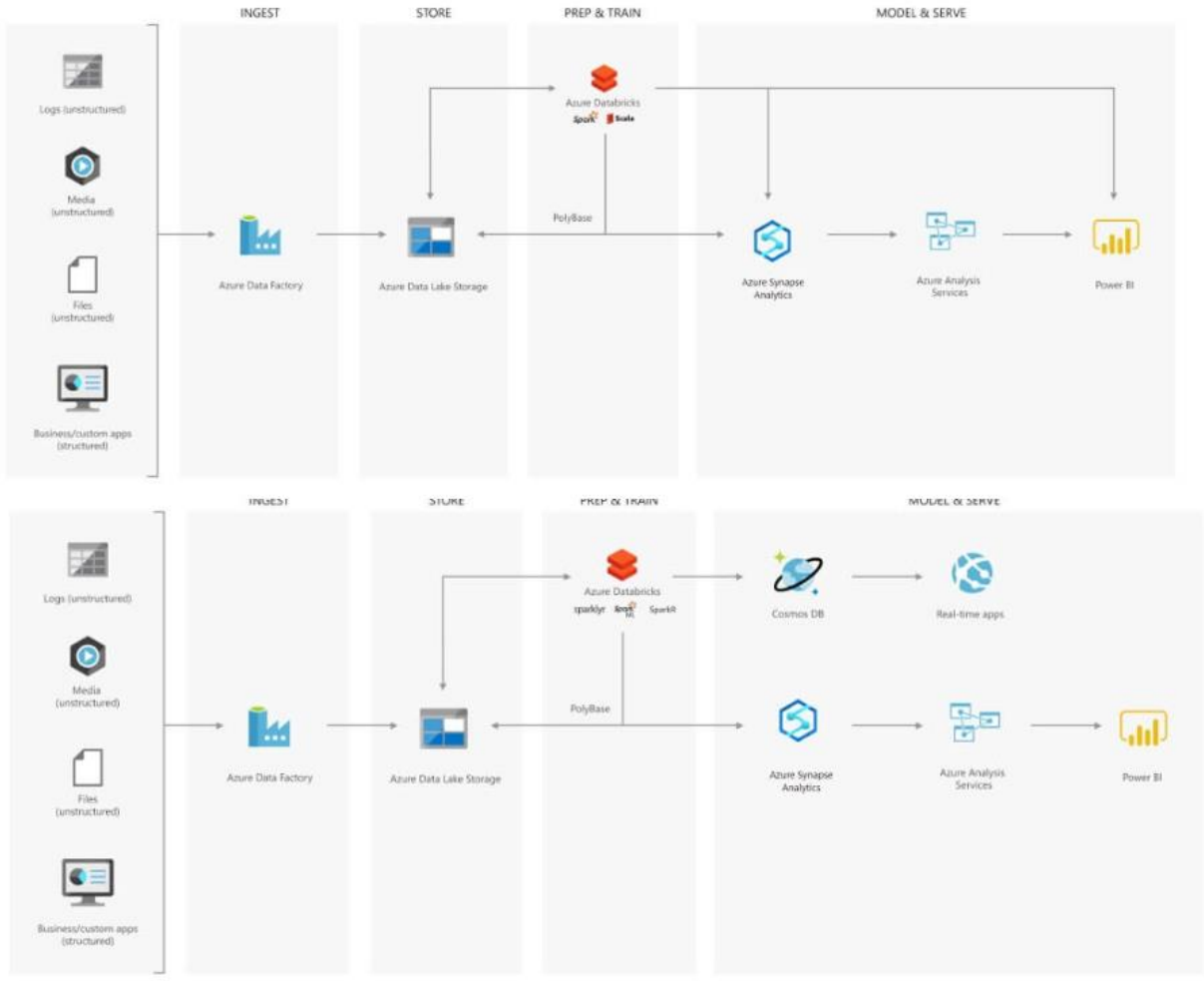
Design B:

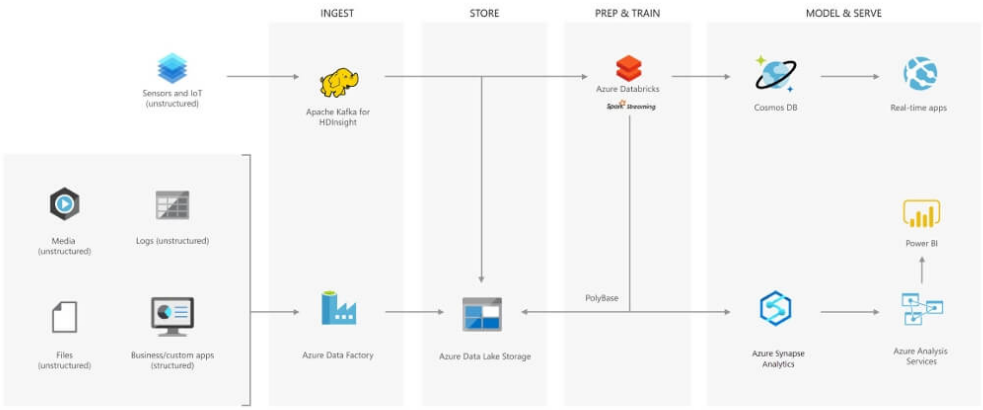


Design C:


INGEST
STORE
PREP & TRAIN
MODEL & SERVE
Sensors and IoT (unstructured)
Apache Kafka for HDInsight
Azure Databricks
Spark Streaming
Cosmos DB
Real-time apps
Media (unstructured)
Logs (unstructured)
Power BI
PolyBase
Files (unstructured)
Business/custom apps (structured)
Azure Data Factory
Azure Data Lake Storage
Azure Synapse Analytics
Azure Analysis Services

Which architecture would be best suited for the need?





Design A

Explanation:-

Creating a modern data warehouse

Imagine you're a Data Engineering consultant for a Avengers Security. In the past, they've created an on-premises business intelligence solution that used a Microsoft SQL Server Database Engine, SQL Server Integration Services, SQL Server Analysis Services, and SQL Server Reporting Services to provide historical reports. They tried using the Analysis Services Data Mining component to create a predictive analytics solution to predict the buying behaviour of customers. While this approach worked well with low volumes of data, it couldn't scale after more than a gigabyte of data was collected. Furthermore, they were never able to deal with the JSON data that a third-party application generated when a customer used the feedback module of the point of sale (POS) application.

The company has turned to you for help with creating an architecture that can scale with the data needs that are required to create a predictive model and to handle the JSON data so that it's integrated into the BI solution. You suggest the following architecture:

The architecture uses Azure Data Lake Storage at the centre of the solution for a modern data warehouse. Integration Services is replaced by Azure Data Factory to ingest data into the Data Lake from a business application. This is the source for the predictive model that is built into Azure Databricks. PolyBase is used to transfer the historical data into a big data relational format that is held in Azure Synapse Analytics, which also stores the results of the trained model from Databricks. Azure Analysis Services provides the caching capability for SQL Data Warehouse to service many users and to present the data through Power BI reports.

Advanced analytics for big data

In this second use case, Azure Data Lake Storage plays an important role in providing a large-scale data store. Your skills are needed by Hydra Corporation, which is a global seller of bicycles and cycling components through a chain of resellers and on the internet. As their customers browse the product catalogue on their websites and add items to their baskets, a recommendation engine that is built into Azure Databricks recommends other products. They need to make sure that the results of their recommendation engine can scale globally. The recommendations are based on the web log files that are stored on the web servers and transferred to the Azure Databricks model hourly. The response time for the recommendation should be less than 1 ms. You propose the following architecture:

Real-time analytical solutions

To perform real-time analytical solutions, the ingestion phase of the architecture is changed for processing big data solutions. In this architecture, note the introduction of Apache Kafka for Azure HDInsight to ingest streaming data from an Internet of Things (IoT) device, although this could be replaced with Azure IoT Hub and Azure Stream Analytics. The key point is that the data is persisted in Data Lake Storage Gen2 to service other parts of the solution.

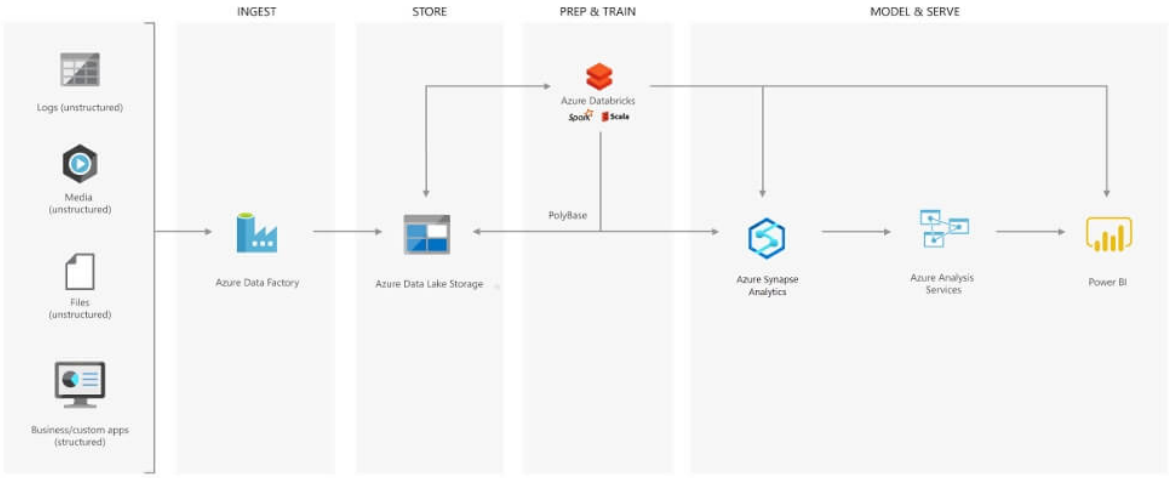
In this use case, you are a Data Engineer for HAMMER Industries, an organization that is working with a transport company to monitor the fleet of Heavy Goods Vehicles (HGV) that drive around Europe. Each HGV is equipped with sensor hardware that will continuously report metric data on the temperature, the speed, and the oil and brake solution levels of an HGV. When the engine is turned off, the sensor also outputs a file with summary information about a trip, including the mileage and elevation of a trip. A trip is a period in which the HGV engine is turned on and off.

Both the real-time data and batch data is processed in a machine learning model to predict a maintenance schedule for each of the HGVs. This data is made available to the downstream application that third-party garage companies can use if an HGV breaks down anywhere in Europe. In addition, historical reports about the HGV should be visually presented to users. As a result, the following architecture is proposed:

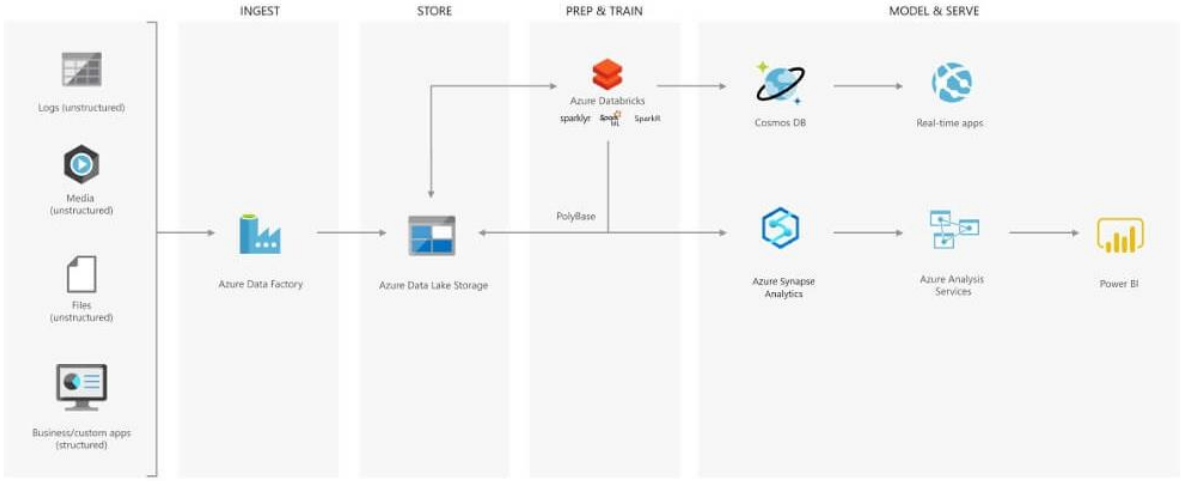
In this architecture, there are two ingestion streams. Azure Data Factory ingests the summary files that are generated when the HGV engine is turned off. Apache Kafka provides the real-time ingestion engine for the telemetry data. Both data streams are stored in Azure Data Lake Store for use in the future, but they are also passed on to other technologies to meet business needs. Both streaming and batch data are provided to the predictive model in Azure Databricks, and the results are published to Azure Cosmos DB to be used by the third-party garages. PolyBase transfers data from the Data Lake Store into SQL Data Warehouse where Azure Analysis Services creates the HGV reports by using Power BI.

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

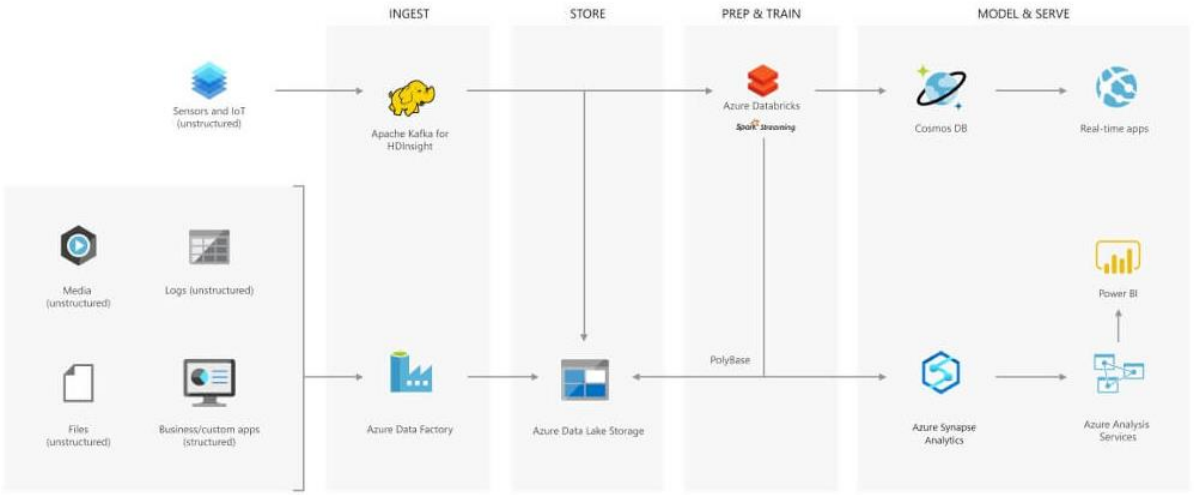
Modern data warehouse



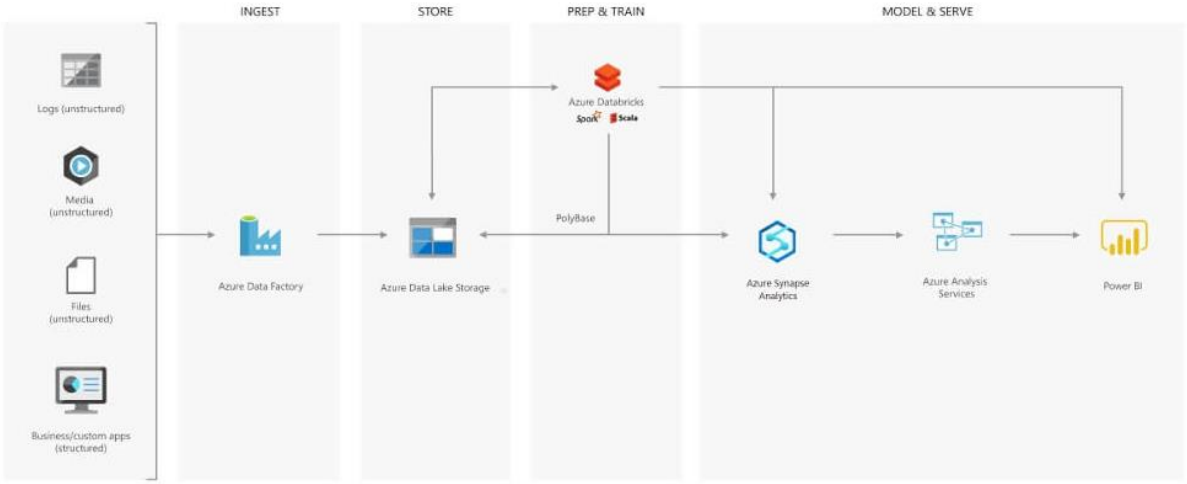
Advanced analytics on big data



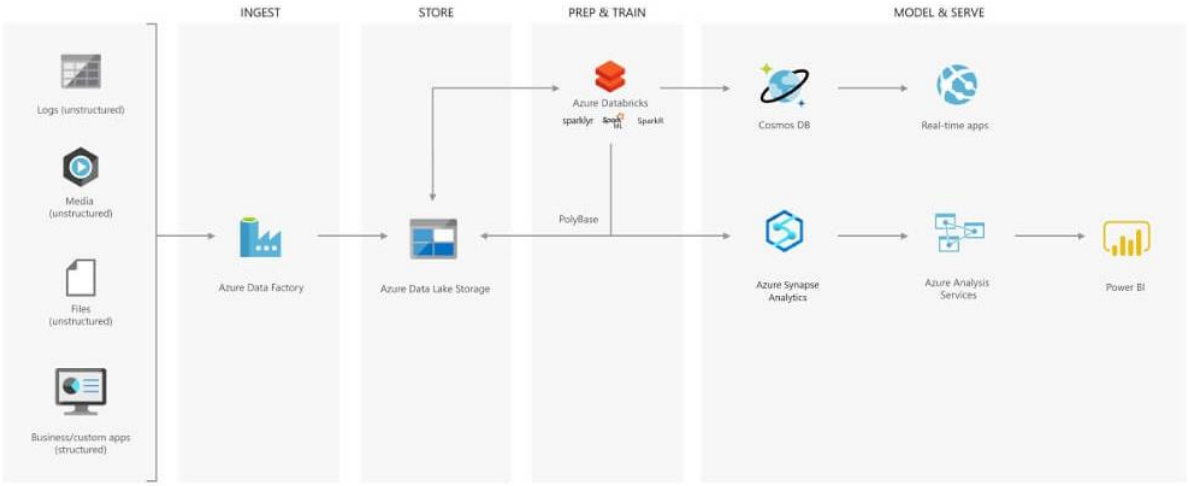
Real-time analytics



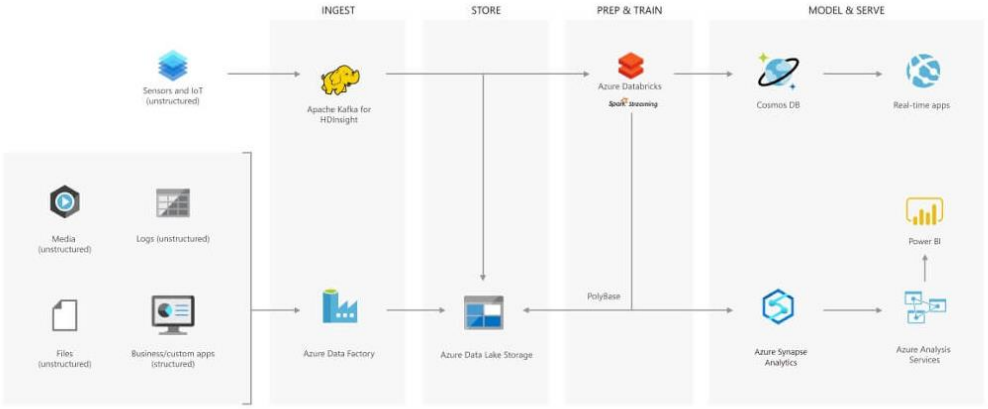
Modern data warehouse



Advanced analytics on big data



Real-time analytics



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

Report Error

Q. 58

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

An Azure integration runtime is capable of which of the following? (Select all that apply)

- [x] Dispatching transform activities in public network utilizing platforms such as Databricks Notebook/ Jar/ Python activity, HDInsight Hive activity and more.

Explanation:- In Data Factory, an activity defines the action to be performed. A linked service defines a target data store or a compute service. An integration runtime provides the bridge between the activity and linked services.
Azure integration runtime
An Azure integration runtime is capable of:
• Running Data Flows in Azure
• Running Copy Activity between cloud data stores
• Dispatching the following transform activities in public network: Databricks Notebook/ Jar/ Python activity, HDInsight Hive activity, HDInsight Pig activity, HDInsight MapReduce activity, HDInsight Spark activity, HDInsight Streaming activity, Machine Learning Batch Execution activity, Machine Learning Update Resource activities, Stored Procedure activity, Data Lake Analytics U-SQL activity, .NET custom activity, Web activity, Lookup activity, and Get Metadata activity.
You can set a certain location of an Azure IR, in which case the data movement or activity dispatch will happen in that specific region. If you choose to use the auto-resolve Azure IR which is the default, ADF will make a best effort to automatically detect your sink and source data store to choose the best location either in the same region if available or the closest one in the same geography for the Copy Activity. For anything else, it will use the IR in the Data Factory region. Azure Integration Runtime also has support for virtual networks.
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

- [x] Running Copy Activity between cloud data stores

Explanation:- In Data Factory, an activity defines the action to be performed. A linked service defines a target data store or a compute service. An integration runtime provides the bridge between the activity and linked services.
Azure integration runtime
An Azure integration runtime is capable of:
• Running Data Flows in Azure
• Running Copy Activity between cloud data stores
• Dispatching the following transform activities in public network: Databricks Notebook/ Jar/ Python activity, HDInsight Hive activity, HDInsight Pig activity, HDInsight MapReduce activity, HDInsight Spark activity, HDInsight Streaming activity, Machine Learning Batch Execution activity, Machine Learning Update Resource activities, Stored Procedure activity, Data Lake Analytics U-SQL activity, .NET custom activity, Web activity, Lookup activity, and Get Metadata activity.
You can set a certain location of an Azure IR, in which case the data movement or activity dispatch will happen in that specific region. If you choose to use the auto-resolve Azure IR which is the default, ADF will make a best effort to automatically detect your sink and source data store to choose the best location either in the same region if available or the closest one in the same geography for the Copy Activity. For anything else, it will use the IR in the Data Factory region. Azure Integration Runtime also has support for virtual networks.
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

- [x] Running Data Flows in Azure

Explanation:- In Data Factory, an activity defines the action to be performed. A linked service defines a target data store or a compute service. An integration runtime provides the bridge between the activity and linked services.
Azure integration runtime
An Azure integration runtime is capable of:
• Running Data Flows in Azure
• Running Copy Activity between cloud data stores
• Dispatching the following transform activities in public network: Databricks Notebook/ Jar/ Python activity, HDInsight Hive activity, HDInsight Pig activity, HDInsight MapReduce activity, HDInsight Spark activity, HDInsight Streaming activity, Machine Learning Batch Execution activity, Machine Learning Update Resource activities, Stored Procedure activity, Data Lake Analytics U-SQL activity, .NET custom activity, Web activity, Lookup activity, and Get Metadata activity.
You can set a certain location of an Azure IR, in which case the data movement or activity dispatch will happen in that specific region. If you choose to use the auto-resolve Azure IR which is the default, ADF will make a best effort to automatically detect your sink and source data store to choose the best location either in the same region if available or the closest one in the same geography for the Copy Activity. For anything else, it will use the IR in the Data Factory region. Azure Integration Runtime also has support for virtual networks.
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

- [ ] Triggering batch movement of ETL data on a dynamic schedule for most analytics solutions.
- [ ] None of the listed options.
- [ ] All the listed options.

Report Error

Q. 59

With the Azure-SSIS integration runtime installed and SQL Server Data Tools (SSDT) you have the capability to deploy and manage SSIS packages that you create in the cloud. For some packages, you may be able to rebuild them by redeploying them in the Azure-SSIS runtime. However, there may be some SSIS packages that already exist within your environment that may not be compatible.

You can use the [?] to perform an assessment of the SSIS packages that exist and identify any compatibility issues with them.

- ○ Azure Lab Services
- ○ Azure SQL Server Management Studio
- ○ Azure ARM templates
- ○ Azure SQL Server Upgrade Advisor
- ◉ Azure Data Migration Assistant

Explanation:- With the Azure-SSIS integration runtime installed and SQL Server Data Tools (SSDT) you have the capability to deploy and manage SSIS packages that you create in the cloud. For some packages, you may be able to rebuild them by redeploying them in the Azure-SSIS runtime. However, there may be some SSIS packages that already exist within your environment that may not be compatible? How should you deal with them?
Perform assessments of your SSIS packages.
When you migrate your database workloads from SQL Server on premises to Azure SQL database services, you may have to migrate SSIS packages as well. The first step required is to perform an assessment of you current SSIS packages to make sure that they are compatible in Azure. Fortunately, you can use the Data Migration Assistant (DMA) to perform an assessment of the SSIS packages that exist and identify any compatibility issues with them. The Data Migration Assistant has two main categories of information:
• Migration blockers: Issues that prevent your existing SSIS packages to run on Azure-SSIS Integration Runtime environments.

• Information issues: SSIS features within your packages that are only partially supported, or are deprecated. Regardless of which category of information you receive, the Data Migration Assistant will perform the assessment on a batch of SSIS packages and provide guidance and potential mitigation steps that you can use to address the blockers and issues that are raised.
Perform a migration of your packages
Before migrating, you must know which Azure SQL database service you are migrating to. This can include migrating to Azure SQL Managed Instance (MI), or Azure SQL Database. Furthermore, when migrating SSIS packages. you have to consider the location of the SSIS packages that you are migrating, as this can impact how you migrate the packages, and which tool you will need to use. There are four types of storage including:
• SSIS Catalog (also known as SSISDB)
• File System
• MSDB database in SQL Server
• SSIS Package store
Based on this information, you can use the following table as a basis for understanding the tools you can use to perform migration assessments, and to perform the migration itself.
Microsoft Data Migration Assistant
The Data Migration Assistant helps you upgrade to a modern data platform by detecting compatibility issues that can impact database functionality in your new version of SQL Server or Azure SQL Database. DMA recommends performance and reliability improvements for your target environment and allows you to move your schema, data, and objects from your source server to your target server.
This tool can be helpful to you in identifying any issues that can affect a migration to an Azure SQL data platform. The DMA can run assessment projects that will identify any blocking issues or unsupported features that are currently in use with your on-premises SQL Server. It can also help you understand the new features in the target SQL Server platform that the database can benefit from after a migration. The DMA can also perform migration projects that can migrate an on-premises SQL Server instance to a modern SQL Server instance hosted on-premises or on an Azure virtual machine (VM) that is accessible from your on-premises network.
The Data Migration Assistant replaces all previous versions of SQL Server Upgrade Advisor and should be used for upgrades for most SQL Server versions.
https://www.sqlshack.com/move-local-ssis-packages-to-azure-data-factory/

Source: SQL Server + SQL Agent		Destination: Azure SQL MI + MI Agent		Destination: Azure SQL DB + ADF	
Storage Types	**Package Assessment**	**Package Migration**	**Job Migration**	**Package Migration**	**Job Migration**
SSISDB	• Data Migration Assistant tool • SQL Server Data Tools	• Migrate the SSISDB to SSISDB using the Database Migration Service (DMS)	• Migrate SQL Server Agent Jobs to Managed Instance (MI) Agent Using PowerShell, T-SQL, or C# script. • Recreate in the Managed Instance (MI) Agent via SQL Server Management Studio (SSMS)	• Deploy to the SSISDB via SQL Server Data Tools (SSDT) or SQL Server Management Studio (SSMS)	• Migrate SQL Server Agent Jobs to Azure Data Factory (ADF) using PowerShell, T-SQL, or C# scripts • Recreate in Azure Data Factory (ADF) using SQL Server Management Studio (SSMS) or the Azure Data Factory (ADF) portal
File Systems	• Data Migration Assistant tool • SQL Server Data Tools	• Deploy to file shares, or Azure Files using dtinstall, or dtutil, or by a manual copy • Keep in file systems and access via Vnet, or Self-Hosted Integration Runtime (IR)	• Migrate SQL Server Agent Jobs to Managed Instance (MI) Agent Using PowerShell, T-SQL, or C# script. • Recreate in the Managed Instance (MI) Agent via SQL Server Management Studio (SSMS)	• Deploy to file shares, or Azure Files using dtinstall, dtutil, or by a manual copy • Keep in file systems and access via Vnet, or Self-Hosted Integration Runtime (IR)	• Migrate SQL Server Agent Jobs to Azure Data Factory (ADF) using SQL Server Management Studio (SSMS) • Recreate in Azure Data Factory (ADF) using SQL Server Management Studio (SSMS) or the Azure Data Factory (ADF) portal
MSDB	• Data Migration Assistant tool • SQL Server Data Tools	• Export to file systems, file shares, or Azure Files via SQL Server Management Studio (SSMS) or dtutil • Import and export to the Package store, or MSDB via SQL Server Management Studio (SSMS) or dtutil	• Migrate SQL Server Agent Jobs to Azure Data Factory (ADF) using PowerShell, T-SQL, or C# scripts • Recreate in Azure Data Factory (ADF) using SQL Server Management Studio (SSMS) or the Azure Data Factory (ADF) portal	• Export to file systems, file shares, or Azure Files using SQL Server Management Studio (SSMS) or dtutil	• Migrate SQL Server Agent Jobs to Azure Data Factory (ADF) using PowerShell, T-SQL, or C# scripts • Recreate in Azure Data Factory (ADF) using SQL Server Management Studio (SSMS) or the Azure Data Factory (ADF) portal
SSIS Package Store	• Data Migration Assistant tool • SQL Server Data Tools	• Export to file systems, file shares, or Azure Files via SQL Server Management Studio (SSMS) or dtutil • Import and export to the Package store, or MSDB via SQL Server Management Studio (SSMS) or dtutil	• Migrate SQL Server Agent Jobs to Azure Data Factory (ADF) using PowerShell, T-SQL, or C# scripts • Recreate in Azure Data Factory (ADF) using SQL Server Management Studio (SSMS) or the Azure Data Factory (ADF) portal	• Export to file systems, file shares, or Azure Files using SQL Server Management Studio (SSMS) or dtutil	• Migrate SQL Server Agent Jobs to Azure Data Factory (ADF) using PowerShell, T-SQL, or C# scripts • Recreate in Azure Data Factory (ADF) using SQL Server Management Studio (SSMS) or the Azure Data Factory (ADF) portal

- ○ Azure Advisor

Report Error

Q. 60

Scenario: Data loads at your company have increased the processing time for on-premises data warehousing descriptive analytic solutions. You have been tasked with looking into a cloud-based alternative to reduce processing time and release business intelligence reports faster. Your boss wants you to first consider scaling up on-premises servers but you discover this approach would reach its physical limits shortly.

The new solution must be on a petabyte scale that doesn't involve complex installations and configurations.

Which of the following would best suit the need?

- ○ Azure Cosmos DB
- ○ Azure On-prem Solution
- ○ Azure DataNow
- ○ Azure Table Storage
- ◉ Azure Synapse Analytics

Explanation:- Azure Synapse Analytics is a cloud-based data platform that brings together enterprise data warehousing and Big Data analytics. It can process massive amounts of data and answer complex business questions with limitless scale.
When to use Azure Synapse Analytics
The SQL Pools capability of Azure Synapse Analytics can meet the scenario needs.
The volume and variety of data that is being generated are providing opportunities to perform different types of analysis on the data. This can include techniques such as exploratory data analysis to identify initial patterns or meaning in the data. It can also include conducting predictive analytics for forecasting, or segmenting data. The Big Data Analytics capability of Azure Synapse Analytics will accommodate this.
Key features

SQL Pools uses massively parallel processing (MPP) to quickly run queries across petabytes of data. Because the storage is separated from the compute nodes, you can scale the compute nodes independently to meet any demand at any time.
In Azure Synapse Analytics, the Data Movement Service (DMS) coordinates and transports data between compute nodes as necessary. But you can use a replicated table to reduce data movement and improve performance. Azure Synapse Analytics supports three types of distributed tables: hash, round-robin and replicated. Use these tables to tune performance.
Importantly, Azure Synapse Analytics can also pause and resume the compute layer. This means you pay only for the computation you use. This capability is useful in data warehousing.
https://docs.microsoft.com/en-us/azure/synapse-analytics/overview-what-is

- ○ Azure Stream Analytics

Report Error