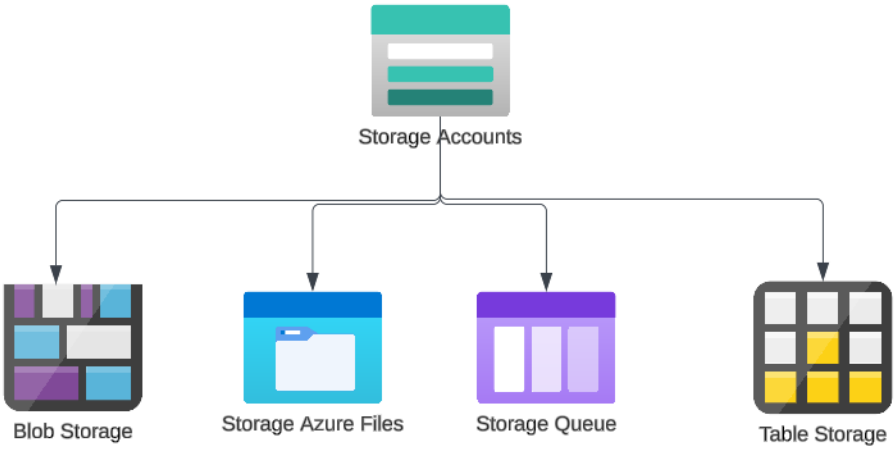
Design and implement data storage – Basics

We need different types of storage services

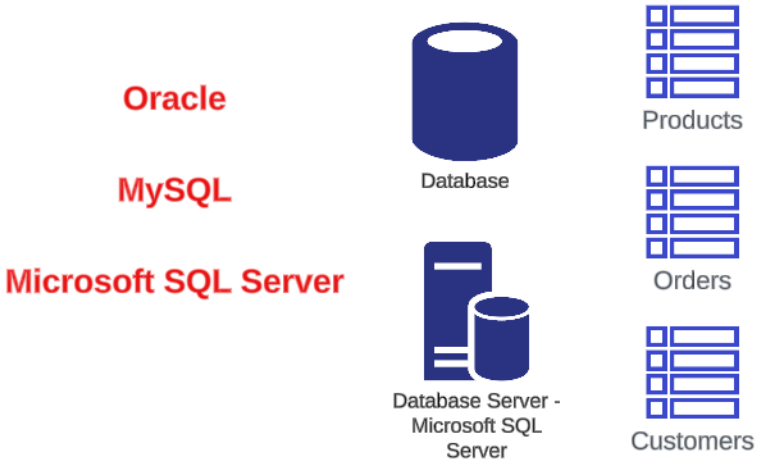
We wil start simple when it comes to using Azure services.

Azure Storage Accounts - This is storage on the Azure cloud for your blob objects, files, queues and tables.



We are going to focus on the Blob service. This is an object storage service on the cloud. This is used for storing objects such as videos, audio files, images.

Azure SQL Database



https://t.me/learningnets

We need to install the database software on some sort of compute hardware.

Virtual Machine

We can install the database engine on a virtual machine.

Database administrator responsibilities

1. Uptime of the database server
2. Database backups and restore
3. Patch installation at the operating system and database engine level.

Azure SQL

If the company does not want the burden of managing the underlying infrastruture, they can opt to use the Azure SQL Database service.

Virtual Machine

SQL Servers

Here the underlying server is managed for you. The database software will be in place. It also has features such as backup/restore and several other features.

SQL Databases

You can simply start hosting your databases. The Azure SQL database is the cloud version of Microsoft SQL Server.

https://t.me/learningnets

Design and implement data storage - Azure Synapse Analytics

What have we seen so far

Azure Data Lake Gen2 storage account

To this account, we can add files as objects. Here the files can be in different formats - CSV, Parquet, JSON.

We could visualize these files in PowerBI.

Normally in companies, you have different sources that can stream data in different formats and store it in a data lake.

Then we have transactional data in a database.

Companies want to have the capability to analyze and visualize data coming in from various sources.

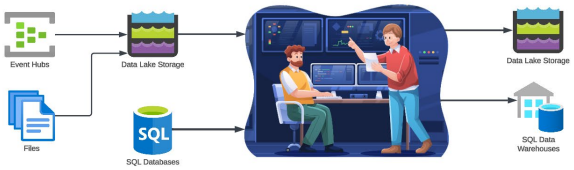
https://t.me/learningnets

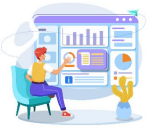
Companies want to have the capability to analyze and visualize data coming in from various sources.

In the end the business users should be able to visualize and report on the data.

Events coming in from Event Hub

Data coming in - Semi-structured, unstructured.





The Data Engineer's job is to make sure the data that comes in, is in a shape that can be used for analysis and visualization.

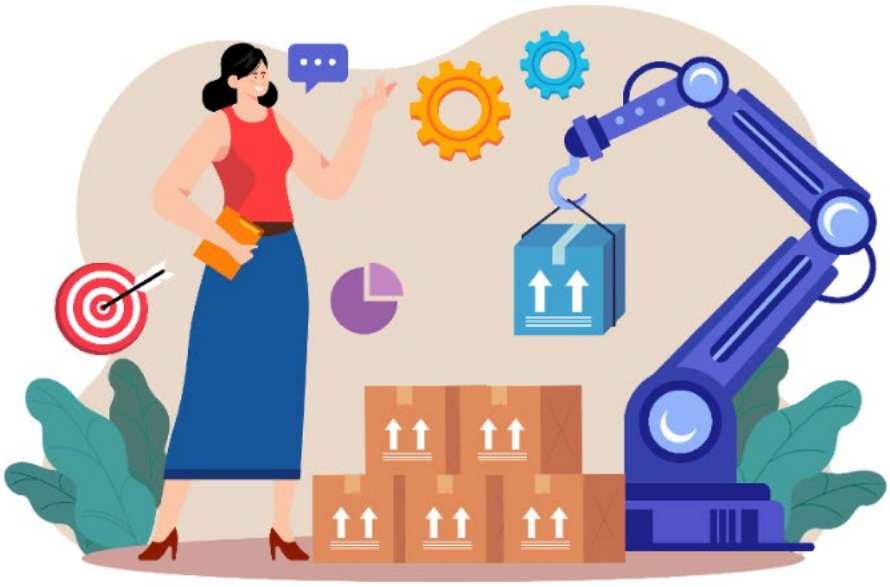
Data might need to be cleant, transformed and put in a target state and data store.

The transformed data could reside in the data lake itself or in a SQL data warehouse.

https://t.me/learningnets

A data warehouse

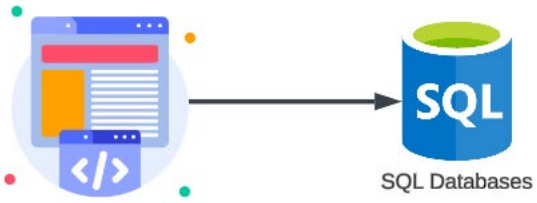
Using data warehouse



A data warehouse is a repository that can be used to store structured data.

You can use the SQL query language to work with the data in the data warehouse.

But why do we need to have a SQL data warehouse in the first place.



An Online Course platform as a web application. This uses a SQL database.

https://t.me/learningnets

Let's say that the orders placed for courses by students are being stored in the Order table.

Now the SQL database is part of an OLTP system - Online Transactional Processing system.

Here transactions are flowing into database almost every second.

Senior management or business wants to get some information based on the purchases being made.

How are we performing to-date?

What would be the forecast for the next quarter?

What's the most popular subject based on the courses being purchased? What's the current market trend?

SQL Databases

In order to provide answers to these questions, we would need to perform analysis over existing data and historical data.

https://t.me/learningnets

We might also need to extract data from other sources as well just to get the requirted answers.

Now we don't try to get these answers from the relational database. This is just used as part of our transactional system.

If we try to generate reports over the current database, it could have an impact and affect our application.

Also we might need to make sure that the data in the database is in a form that makes it easier to get the required insights that business wants.

Hence we normally take our data from the SQL database and transfer it over to a SQL data warehouse.

The engine that is normally used to host a SQL data warehouse is designed differently.

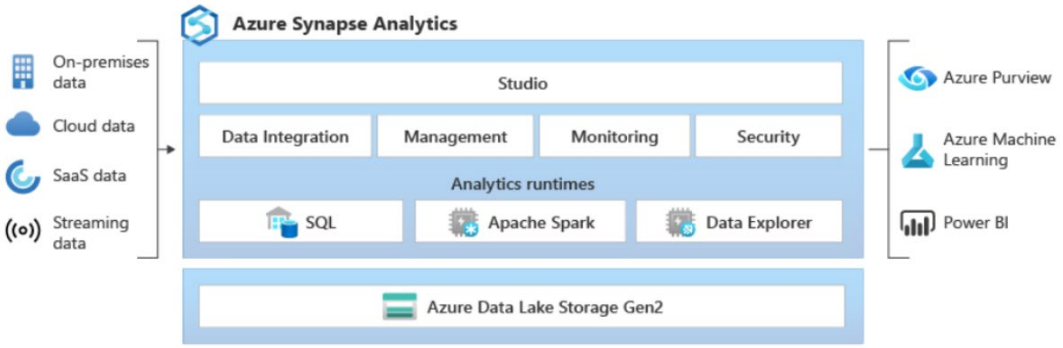
It's designed to host a lot of data. It's designed to be able to query over large data sets.

We would normally keep on taking our transactional data over time and transfer it to the data warehouse.

https://t.me/learningnets

Welcome to Azure Synapse Analytics

Enterprise Analytical service



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

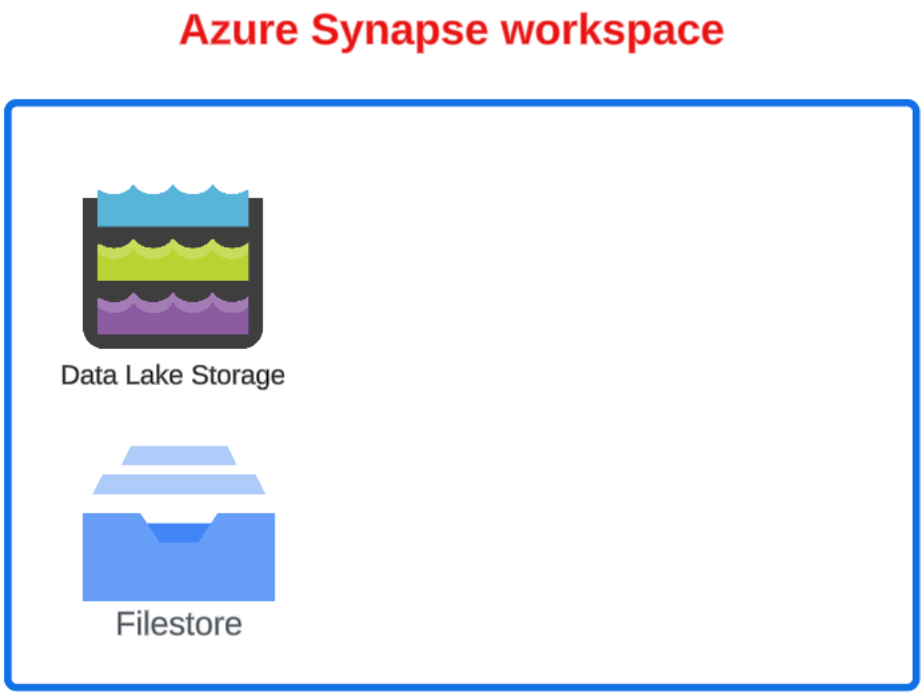
Synapse SQL - Here you can host your SQL data warehouse.

Apache Spark for Azure Synapse - You get access to Spark that assists you in the entire data engineering process.

Data Integration - You can Azure Data Factory like features to ingest your data.

https://t.me/learningnets

Lab - Let's create a Azure Synapse workspace

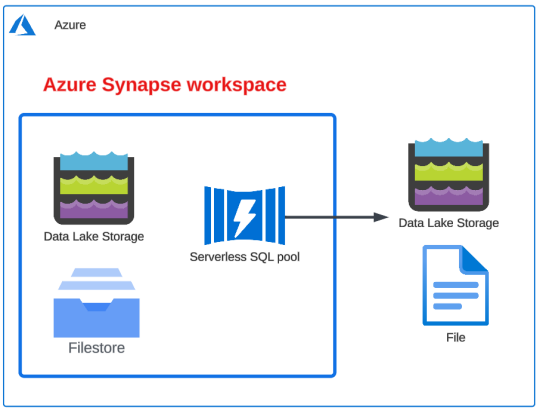


Within your Azure account, you will create an Azure Synapse workspace. This is a secure boundary for performing the various analytical operations on the cloud.

The workspace needs a file system on an Azure Data Lake Gen 2 storage account for temporary data.

https://t.me/learningnets

About the serverless SQL pool



As part of your Azure account, you could be having data that streaming in onto another Azure data lake gen 2 storage account.

Now let's say you have a CSV-based file and you want to analyze the data in the file.

You can use the built-in Serverless SQL pool to query the data that's in the Azure Data Lake storage account.

Your files - Delimited text, Parquet, Delta Lake.

You can use SQL-like queries to work with the data.

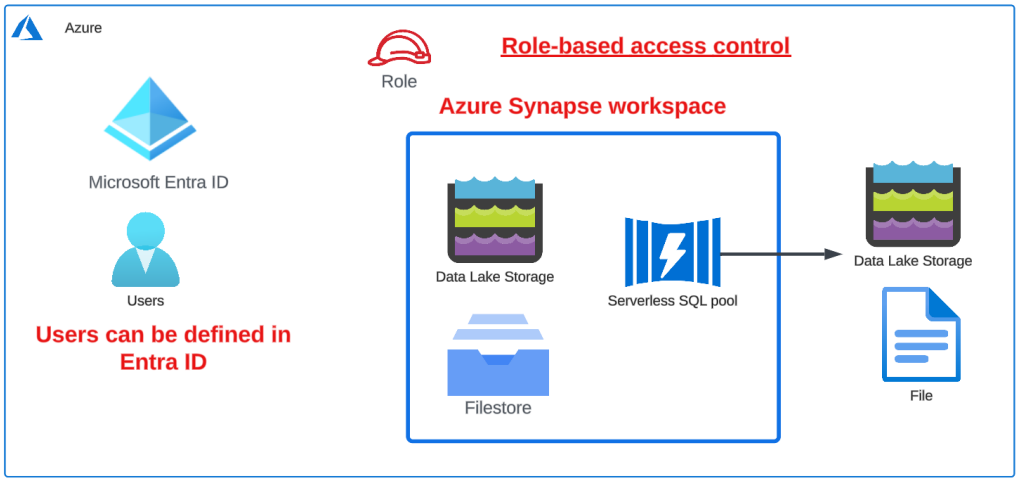
Here the serverless SQL pool has underlying compute that is managed for you.

The compute will manage the queries for you.

You are charged for the data that is processed by the queries.

Quick note on Microsoft Entra ID and permissions

Microsoft Entra ID is an identity provider that is used for authentication and authorization.



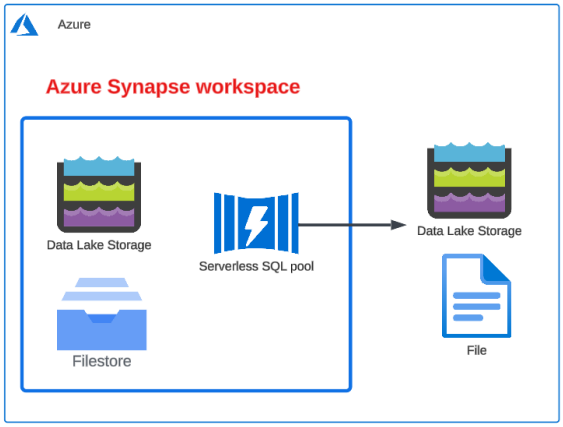
We can assign different roles to a user.

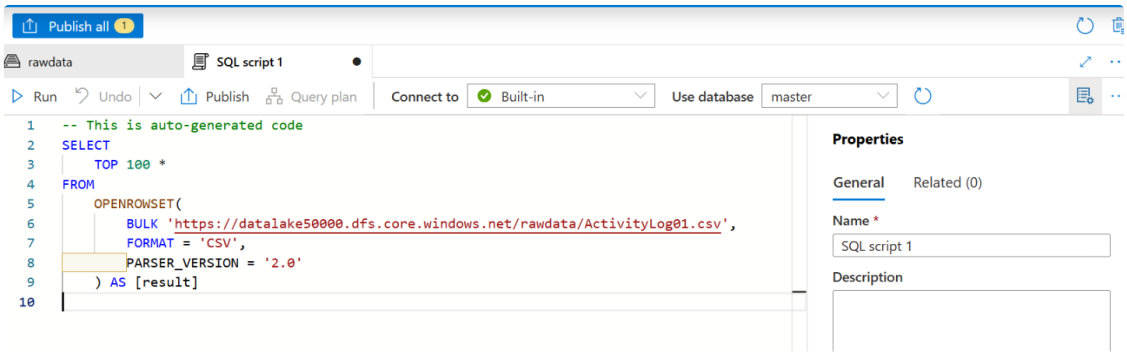
There are many in-built roles.

You can also define your own custom roles.

https://t.me/learningnets

Lab - Using External tables – CSV



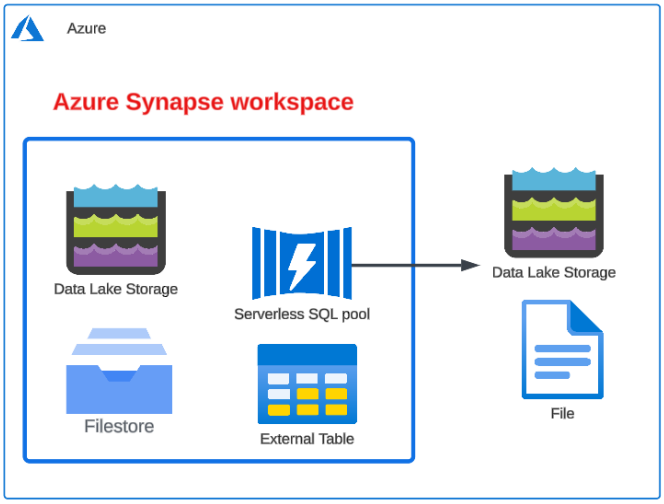


Currently we kind of executed an adhoc query using the context of the master database.

This is fine as a starting point. But data engineers would want to query the data in the data lake files via the use of normal table like structures.

https://t.me/learningnets

For this we can make use of External tables.



Here we create an external table definition that points to the data located in our Azure Data Lake account.

You can use external tables for both read and write operations.

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

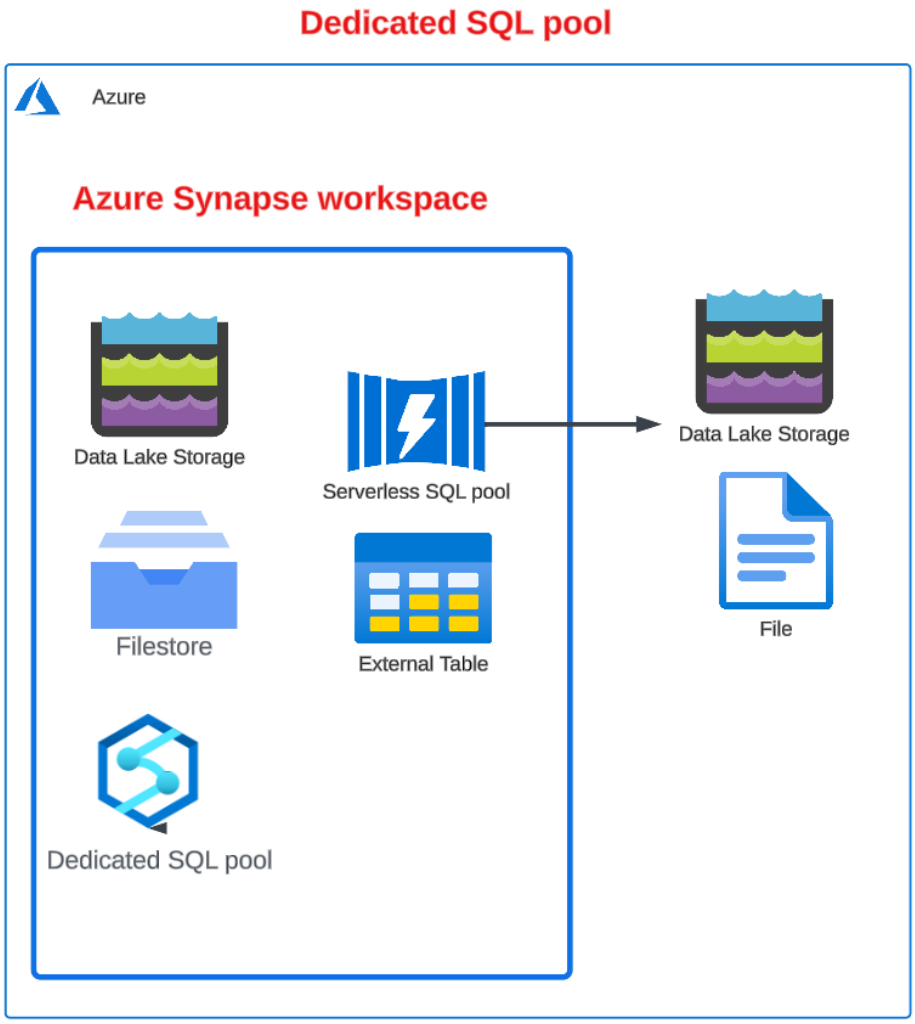
This provides the reference to external storage and the credentials to be used.

This defines the table schema.

This describes the format of the files.

https://t.me/learningnets

Using the dedicated SQL pool



You can host a SQL data warehouse with the help of the dedicated SQL pool.

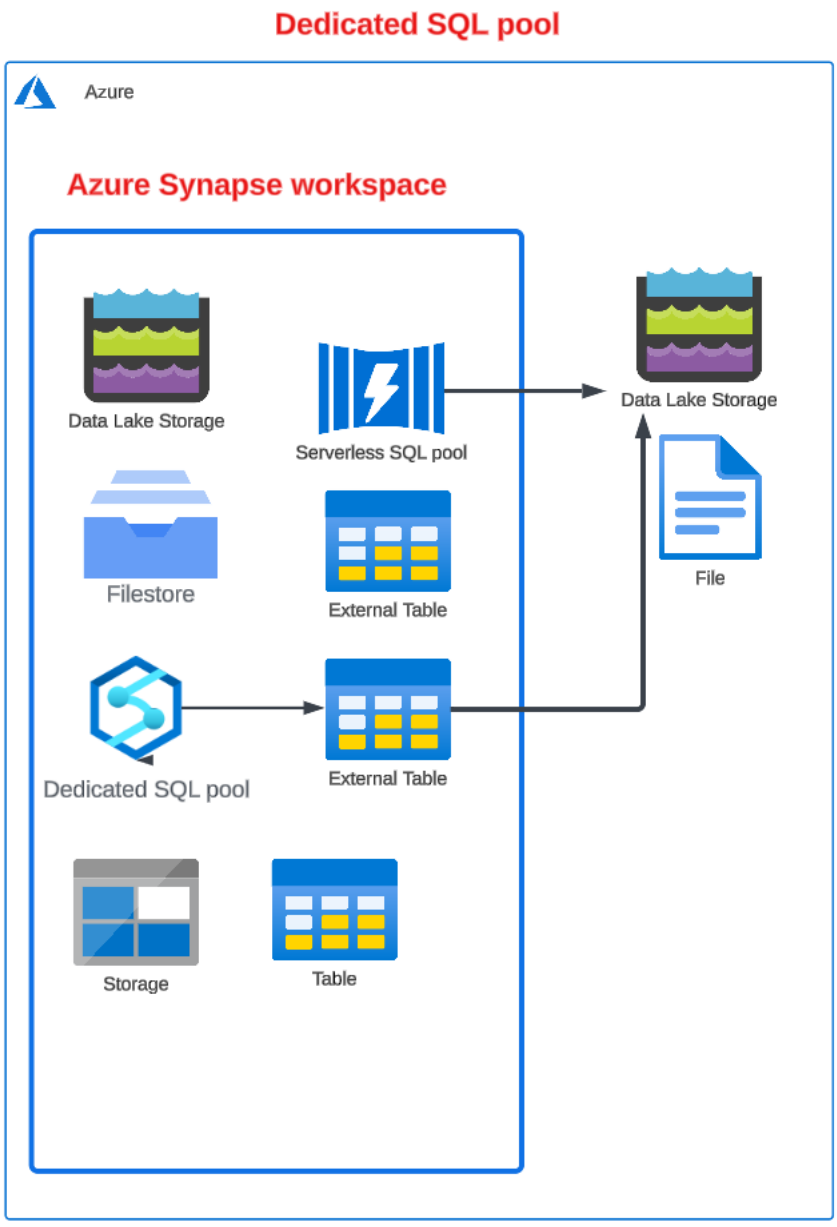
With the Serverless SQL pool, you can just define the table schema. The data itself resides in external storage.

But if you need to persist the data in actual tables and query them via SQL , we need to have a SQL data warehouse in place.

The data warehouse gets dedicated compute and storage. The data in the tables are stored in columar format which reduces data storage costs and improves the query performance.

https://t.me/learningnets

Lab - SQL Pool - External Tables



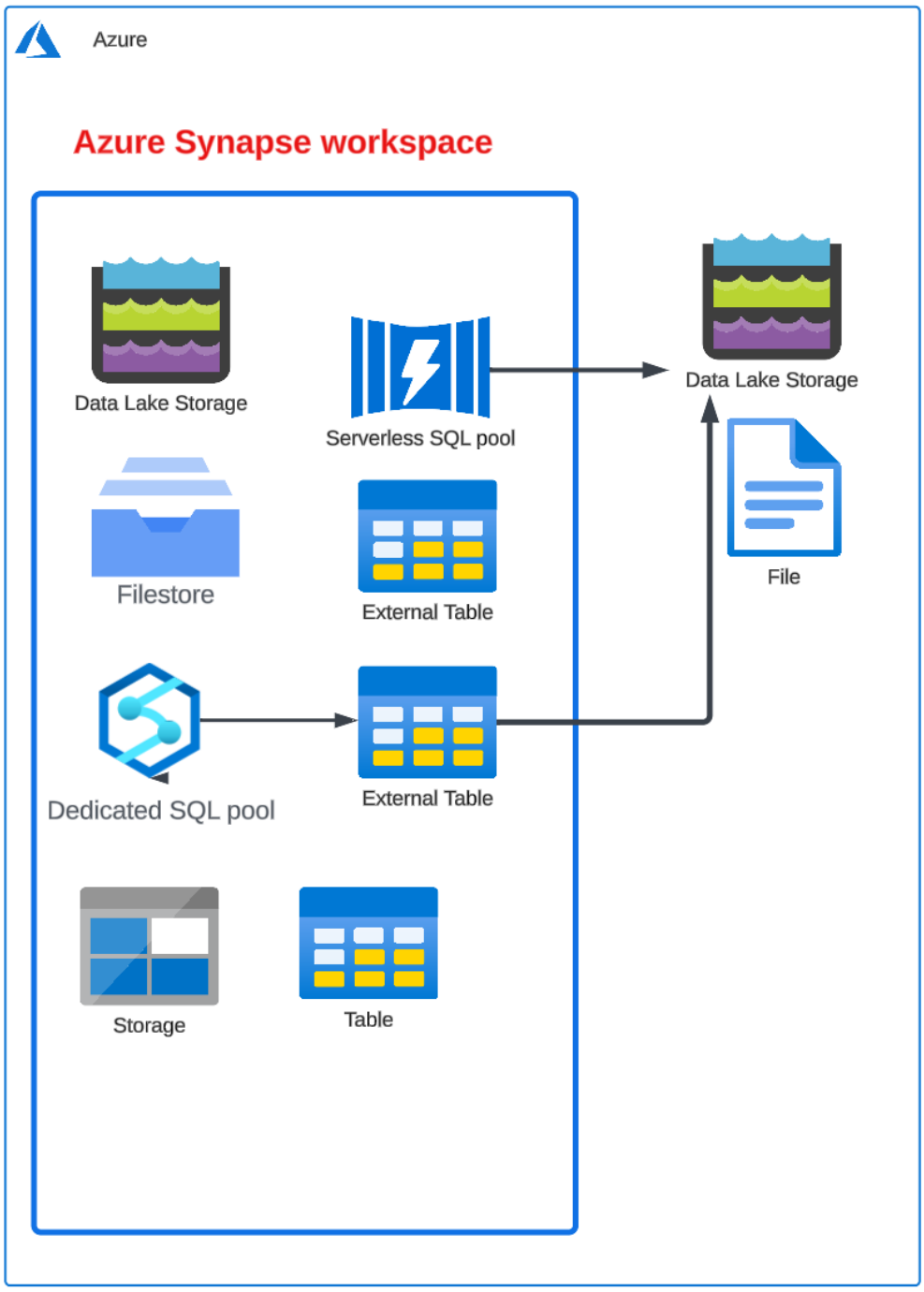
We are going to first see how to use external tables in the dedicated SQL pool.

Now why would you want to create an external table when you have a dedicated pool where you can define persistent tables with data.

One use case is wherein you need to bring the data from an Azure Data Lake account into the dedicated pool. You then create an external table to the data source. And then define a persistent table in the dedicated SQL pool and pull in the data via the use of the external table.

https://t.me/learningnets

Lab - Loading data into a SQL pool using Polybase

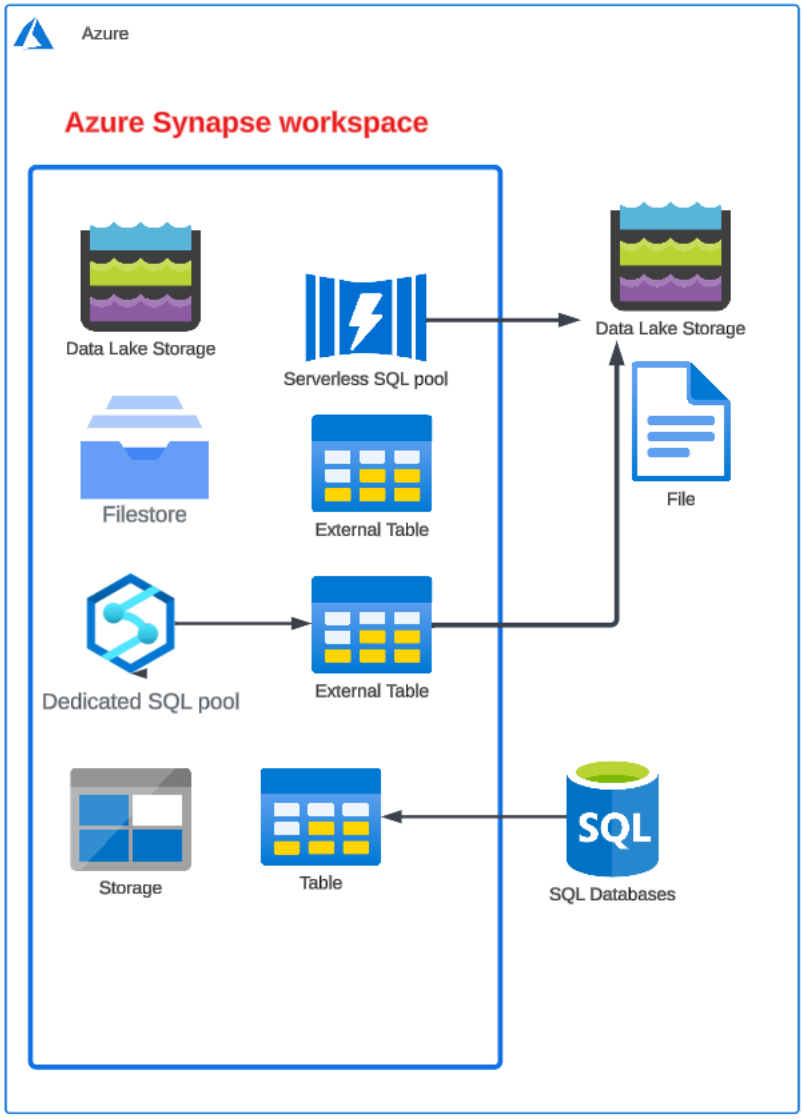


We now want to define a persistent table in the dedicated SQL pool and load the data via the use of an external table.

We are going to use Polybase which is a technology that allows us to load data from Azure Data Lake in a faster and much more efficient manner.

https://t.me/learningnets

Designing a data warehouse



We have seen how to transfer data to a table via the use of Integration pipelines in Azure Synapse.

A very simple example when it comes to storing data in a data warehouse.

But there is a design strategy that is followed when designing tables in a data warehouse.

https://t.me/learningnets

Just looking back as the usage of a SQL database as a backend for an application.

Now normally the application would add data to a table via the use of the INSERT SQL statement.

Here a row of data would be added to a table in the database.

But in a data warehouse things work a little differently.

Remember here data is used for analytical purposes.

Normally you don't insert rows of data one by one.

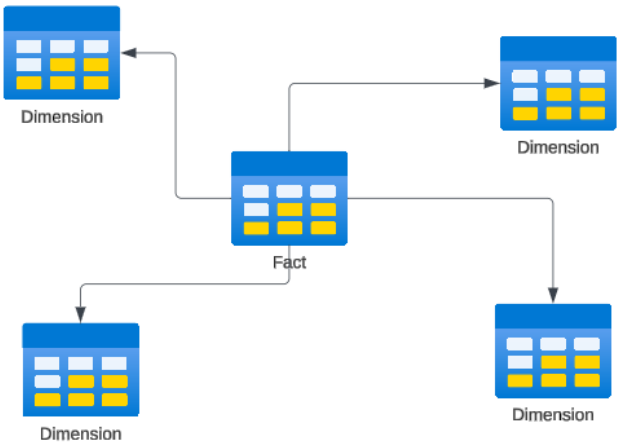
Instead you add and delete rows of data in bulk. This is because of the huge amount of data that is stored in the data warehouse.

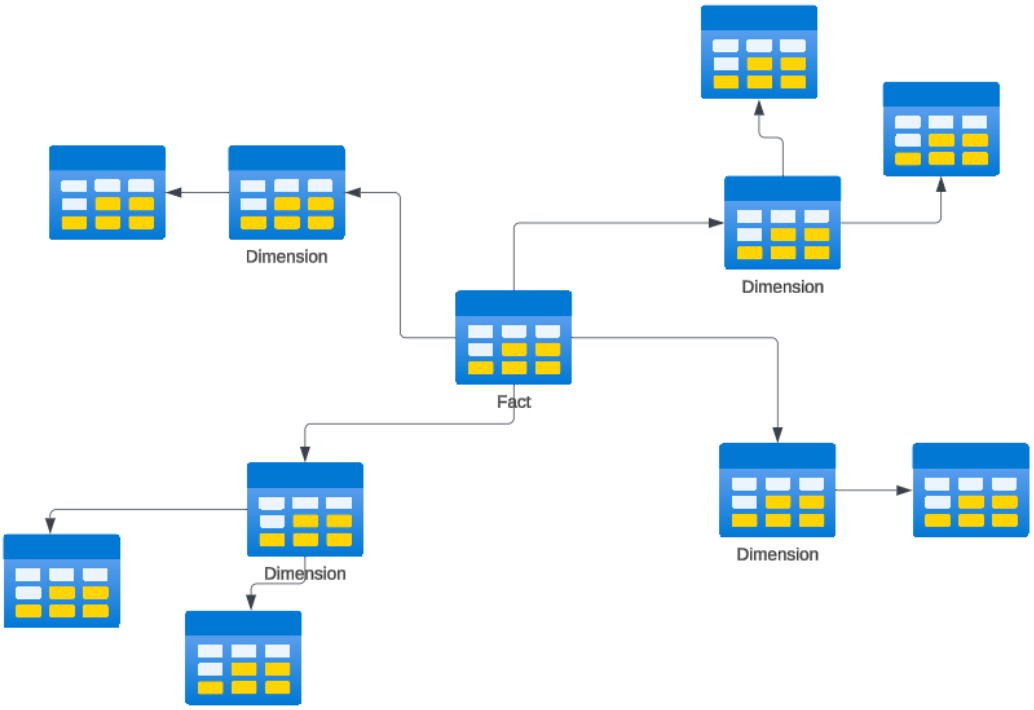
https://t.me/learningnets

Next is the design of the tables

The tables in a data warehouse are split into Fact and Dimension tables.

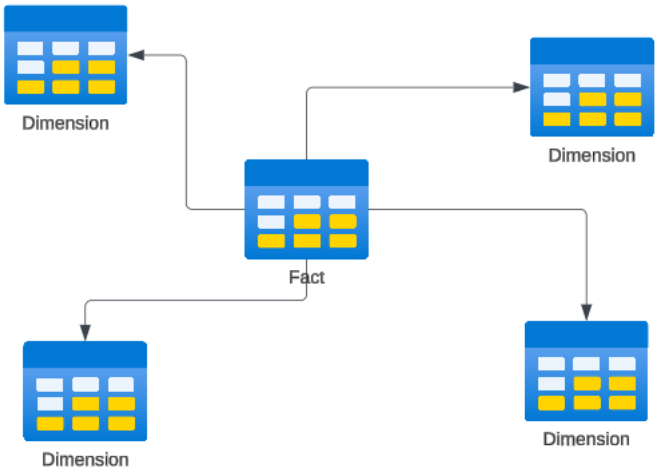
These tables either conform onto a Star or Snowflake schema.



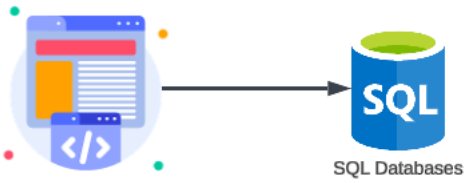


https://t.me/learningnets

Fact and Dimension Tables



A Fact table is meant to store quantitative data, that is data that can be measured.



So let's say that users are making purchases via an ecommerce platform.

The sales data for various products are being recorded in the OLTP SQL database.

Now the sales being recorded are quantitative in nature. You can take the data over time and store in a Fact-based table in the data warehouse.

https://t.me/learningnets

Dimension tables are used to used to present some context to the facts.

For example, based on the sales being made, you want to analyze what are the best selling products - Hence the product-related information would become a dimension.

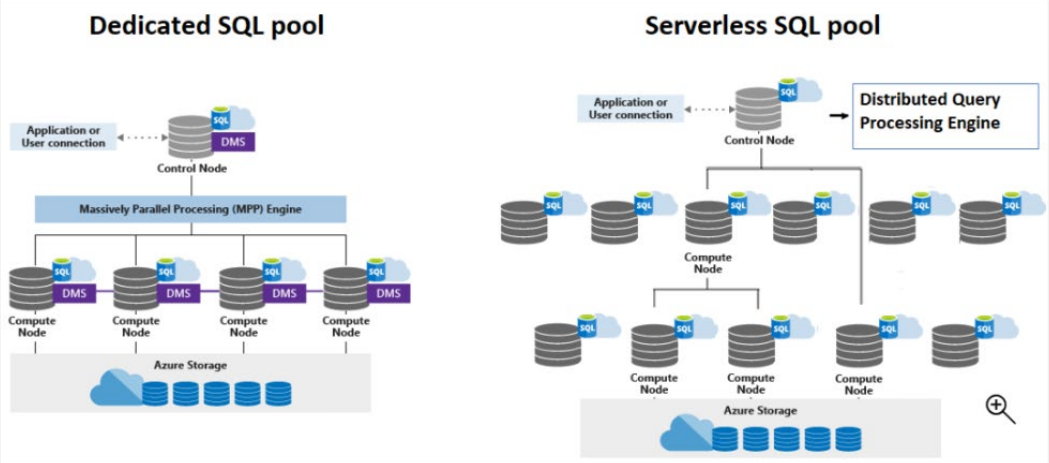
This is because the product information is giving you a view or insight into the sales data.

Or you want to look at the top regions where sales are being made based on the customer's location. So the customer's information can be another dimension to give some more context to the sales.

In this way , we can construct our Fact and Dimension tables.

Understanding Azure Synapse Architecture

Synapse SQL architecture



Reference - https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/overview-architecture

In Synapse SQL , the compute and storage are seperate so that each can be scaled seperately.

In the dedicated SQL pool , the compute power allocated to the pool is determined by a unit known as the data warehouse unit.

All queries are targeted towards the Control Node. And then the Control Node distributes the query for parallel processing across the compute nodes.

https://t.me/learningnets

Performance level	Compute nodes	Distributions per Compute node	Memory per data warehouse (GB)
DW100c	1	60	60
DW200c	1	60	120
DW300c	1	60	180
DW400c	1	60	240
DW500c	1	60	300
DW1000c	2	30	600
DW1500c	3	20	900
DW2000c	4	15	1200
DW2500c	5	12	1500
DW3000c	6	10	1800
DW5000c	10	6	3000
DW6000c	12	5	3600
DW7500c	15	4	4500
DW10000c	20	3	6000
DW15000c	30	2	9000
DW30000c	60	1	18000

Reference - https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/memory-concurrency-limits

So if you have 60 compute nodes , then you query will be split into 60 small queries and run in parallel.

All user data in Synapse SQL is stored in Azure Storage.

In Synapse SQL , the data in the dedicated SQL pool is sharded into distributions.

https://t.me/learningnets

Understanding table types

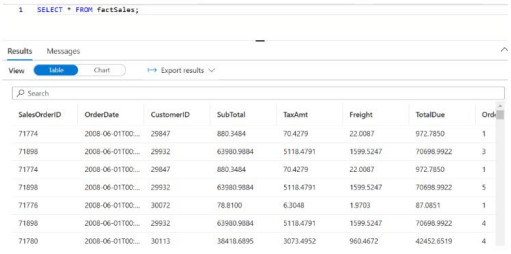
Hash-distributed tables

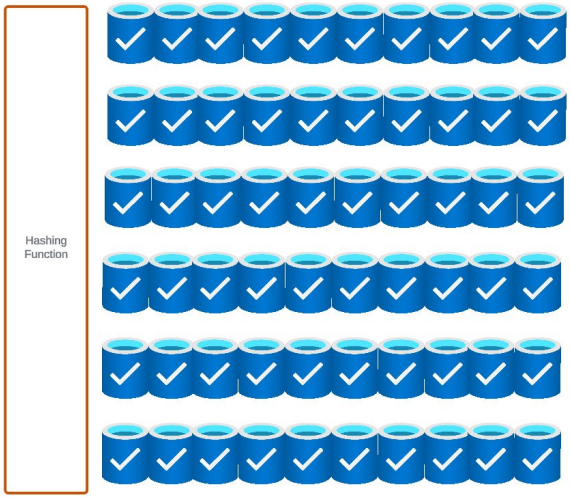
Here the dedicated SQL pool used a hash function to decide to which distribution to assign the row to.

As part of the table definition, you decide which column should be used as the distribution column.

This is ideal for your fact tables.

Let's say you decide that SalesOrderID needs to be the distribution column.





Round-Robin distributed tables

Here the data is evenly distributed across all distributions.

This type of table is effective when you need to load data into staging tables. It provides a fast way to load data.

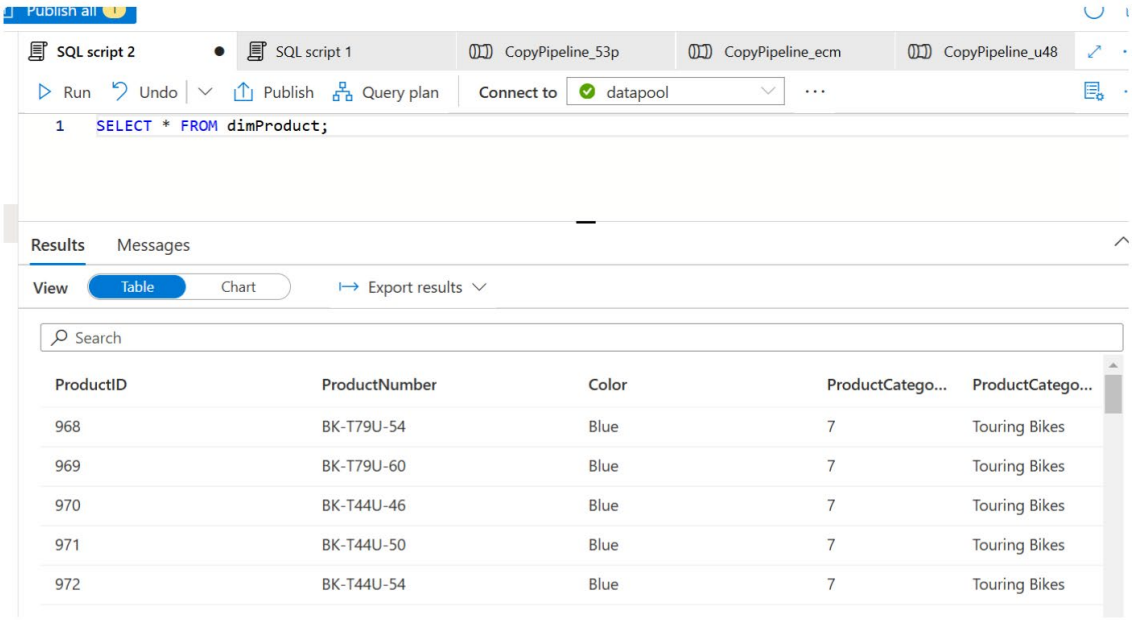
Replicated Tables

Here each compute node would cache a full copy of the table.

This is ideal for dimension tables. When you perform JOINS for your fact and dimension tables, if you have multiple compute nodes, there is a Data Movement service that ensures that data is avaiable on the compute node to fulfil the query operation.

https://t.me/learningnets

Lab - Surrogate keys for dimension tables



Now sometimes for your dimension table, you will want to have a column that helps to uniquely identify the rows in the dimension table.

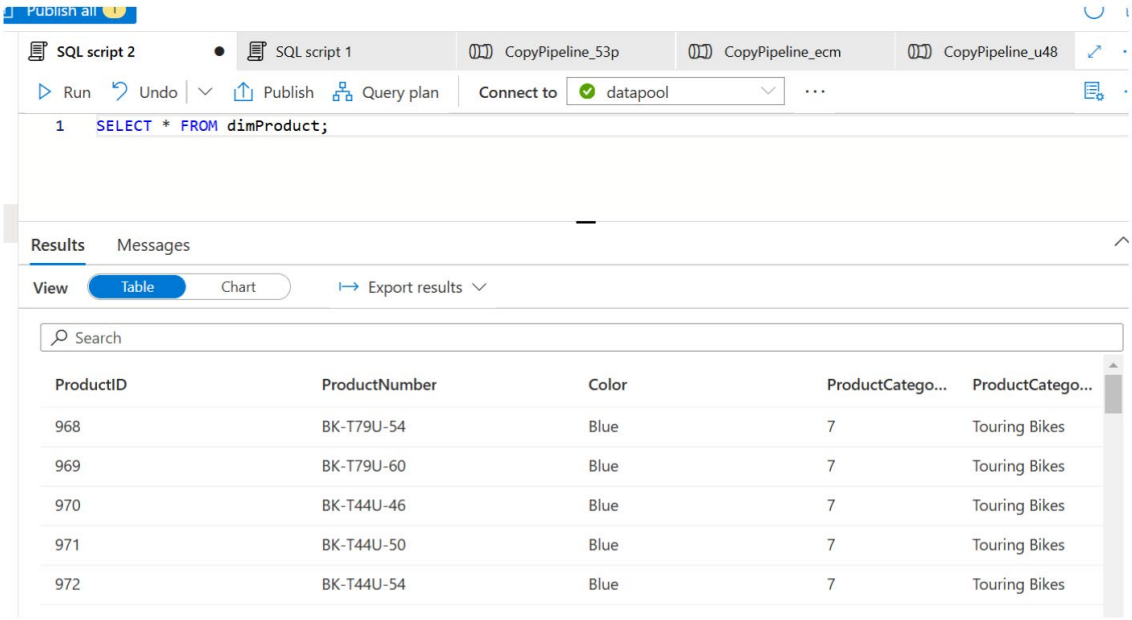
Here ProductID refers to the Business Key because this is taken from the source.

Sometimes data can be taken from multiple sources and they have the same ProductID. Hence we will not use this as the key to uniquely identify each row in the table.

We can define a surrogate key for the table. This could be a simple incrementing number.

https://t.me/learningnets

Slowly Changing dimensions



Now sometimes for your dimension table, you will want to have a column that helps to uniquely identify the rows in the dimension table.

Here ProductID refers to the Business Key because this is taken from the source.

Sometimes data can be taken from multiple sources and they have the same ProductID. Hence we will not use this as the key to uniquely identify each row in the table.

We can define a surrogate key for the table. This could be a simple incrementing number.

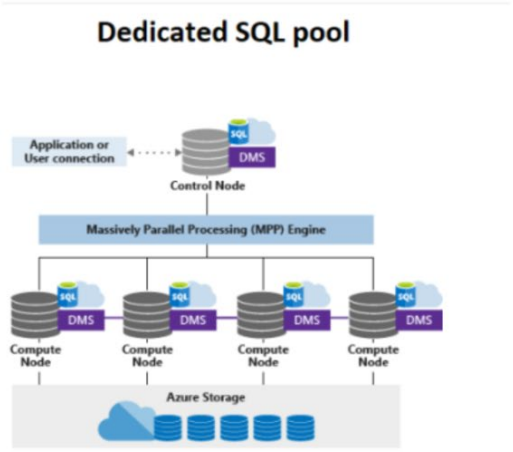
https://t.me/learningnets

Which Load Method to use

Copy command vs Polybase vs Bulk Insert

The Bulk Insert method is slower than the Copy command or Polybase

Polybase gives you better performance when you are loading large amounts of data



With Bulk Insert all of the commands go through the Control Node

You can use the Bulk insert command when you have less data to transfer

With Polybase, your data movement operations go through the compute nodes in parallel. This gives better loading times. And if you have more compute nodes, the data loading process becomes faster.

Polybase also remember allows you to create EXTERNAL tables. And then you can import data into the dedicated SQL pool , creating tables using the CREATE TABLE AS SELECT command.

Even the COPY command has high throughput and goes through the compute nodes in the dedicated SQL pool.

The COPY command is simple to execute, you don't need to define the EXTERNAL DATA SOURCE, EXTERNAL FILE FORMAT and EXTERNAL TABLE.

Design and Develop Data Processing - Azure Data Factory

https://t.me/learningnets

Extract, Transform and Load

Files

Media File

Data Lake Storage

Your data lake would consist of files - semi-structured or unstructured.

Initially all of the data would be in raw format. Especially if you have log files.

In the log files there would probably be a lot of data that you don't need.

SQL Database

Table

Then you probably also have data in your relational databases.

You might want to extract data from here that could be used for analysis purposes.

https://t.me/learningnets

Files

Data Lake Storage

Media File

Step 1 : Extract the data that is required

SQL Database

Table

The next step is to transform your data.

https://t.me/learningnets

Step 2 : Transform the data.

Take only the data values that are required.

Clean data - Remove any NULL or incorrect values.

Convert the data into a desired target format.

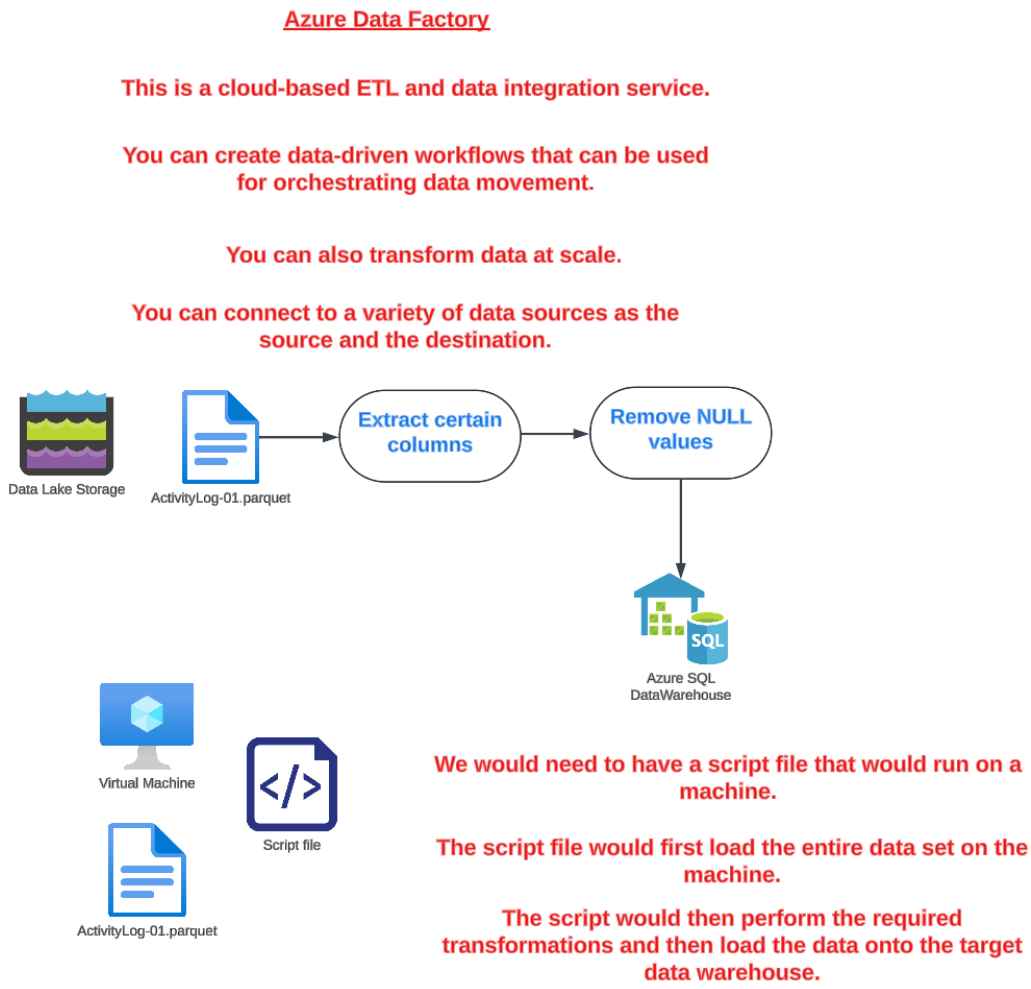
Step 3 : Load the data.

Data Lake Storage

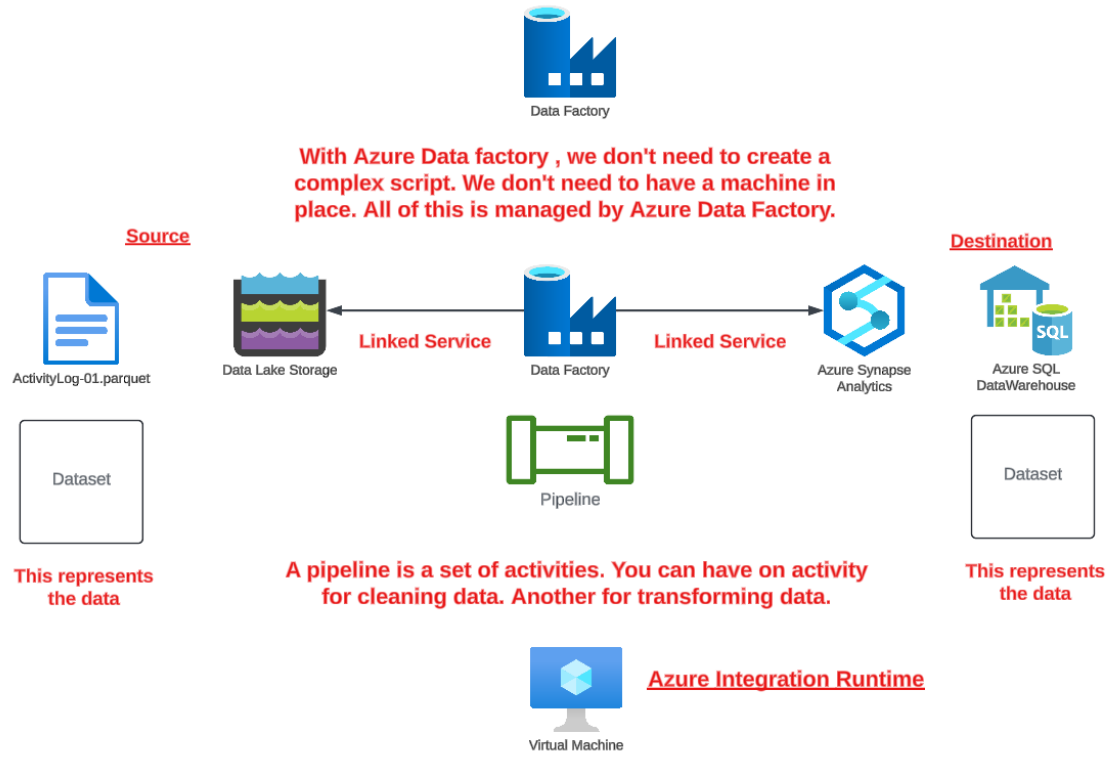
Azure SQL DataWarehouse

https://t.me/learningnets

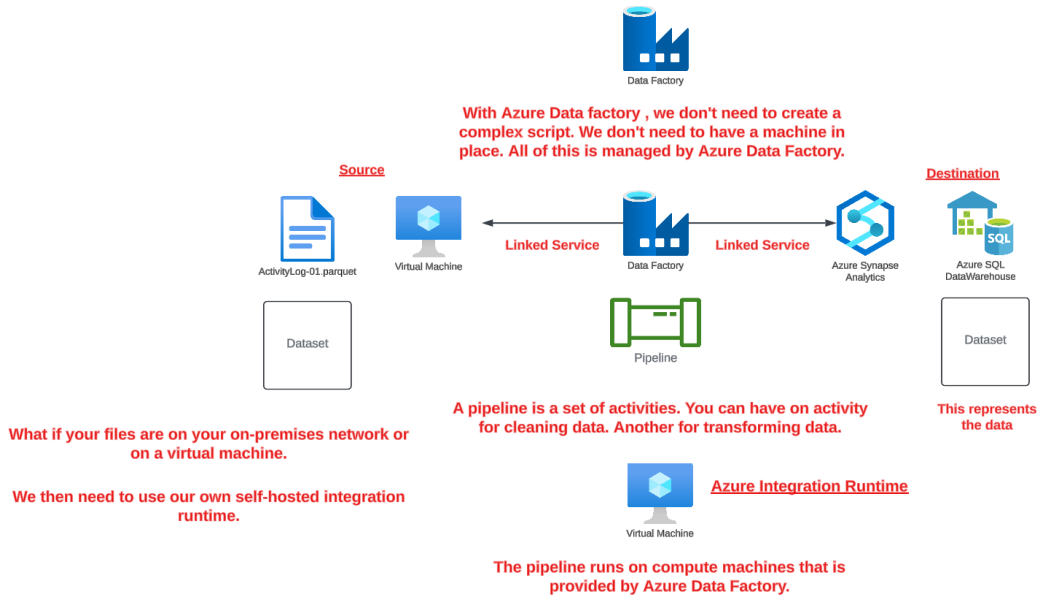
What is Azure Data Factory



https://t.me/learningnets



Self-Hosted Integration Runtime



https://t.me/learningnets

What are going to implement

Build a Windows Server 2022-based Azure virtual machine.

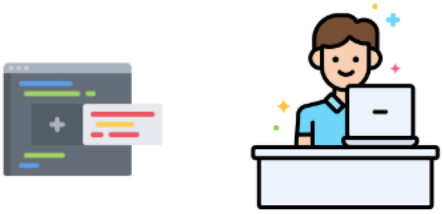
We will install a web server - Internet Information Services.

Install the Self-hosted runtime and register it with Azure Data Factory.

Copy Activity - Transfer the log data file from the web server onto Azure Data Lake.

Mapping Data Flow - Take the log data , extract the contents and transfer it to a table in Azure Synapse.

Azure Data Factory and Git



Git is a popular version control software. Developers normally used this software to manage different versions of their code.

You can integrate Azure Data Factory with a git-based repository for version control of pipelines.

https://t.me/learningnets

Design and Develop Data Processing - Azure Event Hubs ,Stream Analytics

Batch and Real-Time Processing

How do you want to process your data.

Your data could be coming in from multiple data streams.

The data could be in different formats.

The size of data needs to be considered.

The velocity of data needs to be considered.

There are different ways to process your data

Batch Processing

**Here multiple data records are collected at a time.
They are stored first. Then the records are processed as batches at regular time intervals.**

Sales transactions for an e-commerce site are being recorded on a daily basis to a database.

Job

Daily transaction data needs to be processed so that reports can be generated.

The processing can happen at the end of the day. This is when existing compute infrastructure can be used to process the data.

You have to ensure that your source data is not filled with errors. It could impact the batch process that runs on a daily basis.

If the nightly batch process takes time and if the data causes the batch process to fail. Then you will need to run the process again.

In a batch process , the latency is high. It takes time to see the final results.

https://t.me/learningnets

Stream Processing

Here the data is processed in real time. For example for an incoming stream of data , you process the data every 5 minutes.

Here the latency is less because you get results faster.

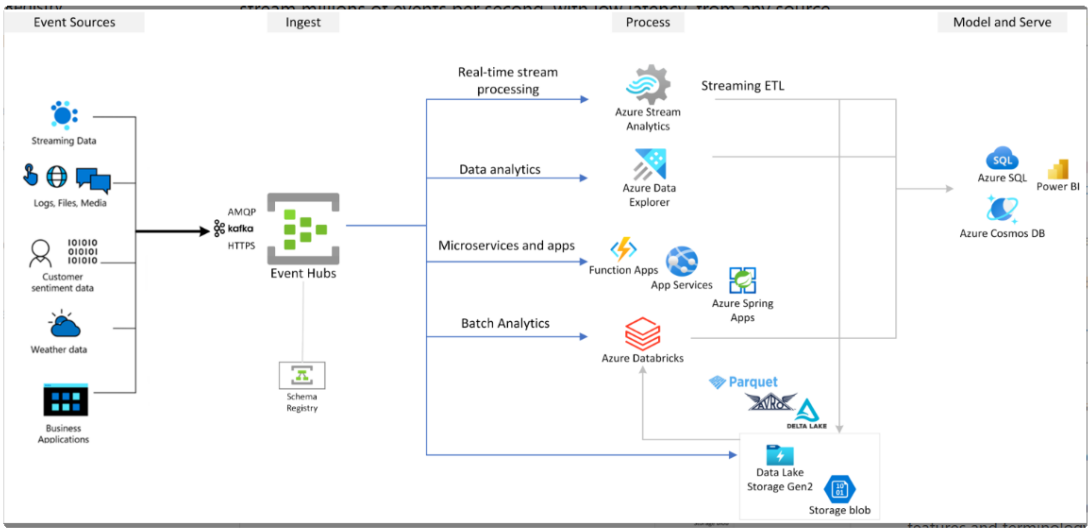
But you need to ensure that you have a processing system that is capable of ingesting and processing data at that speed.

What are Azure Event Hubs

Azure Event Hubs

This is a data streaming service that can stream millions of events per second.

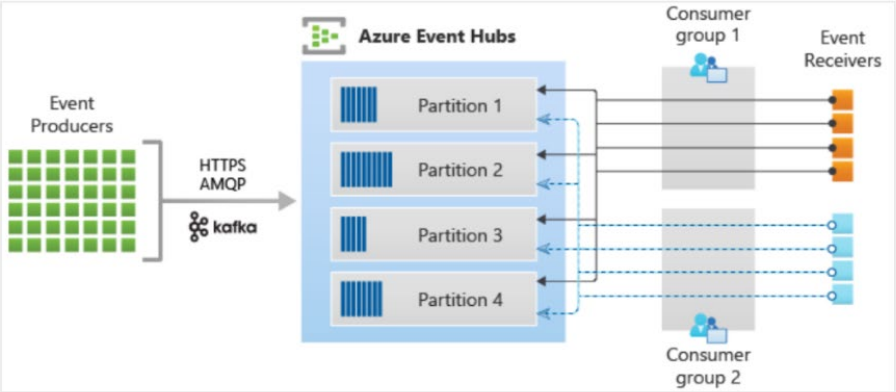
This can be from any source or destination.



Reference -
https://learn.microsoft.com/en-us/azure/event-hubs/event-hubs-about

https://t.me/learningnets

Event Hubs Architecture



Reference - https://learn.microsoft.com/en-us/azure/event-hubs/event-hubs-features

Event Hub namespace - This is a container for multiple Event Hubs.

As event data comes in, they are stored across multiple partitions.

Each Event can have the body of the event. Any user-defined properties. Metadata - offset in the partition, stream sequence number.

Having multiple partitions helps to increase the overall throughput of the system.

Each partition generally can sustain 1 MB/s throughput.

You can also use a partition key in your data events to map the events to a specific partition.

Event Publisher - This sends data to the Event hub.

Event Retention - Standard - 7 days, Premium , Dedicated - 90 days maximum.

Event Consumer- These consumes the events from Event Hub.

Consumer Group - This is a logical grouping of consumers that read data from the event hub.

Throughput capacity of the Event Hub is controlled via the number of throughput units you assign. These are prepurchased but billed per hour.

Ingress - 1 MB per second or 1000 events per second

Egress - 2 MB per second or 4096 events per second

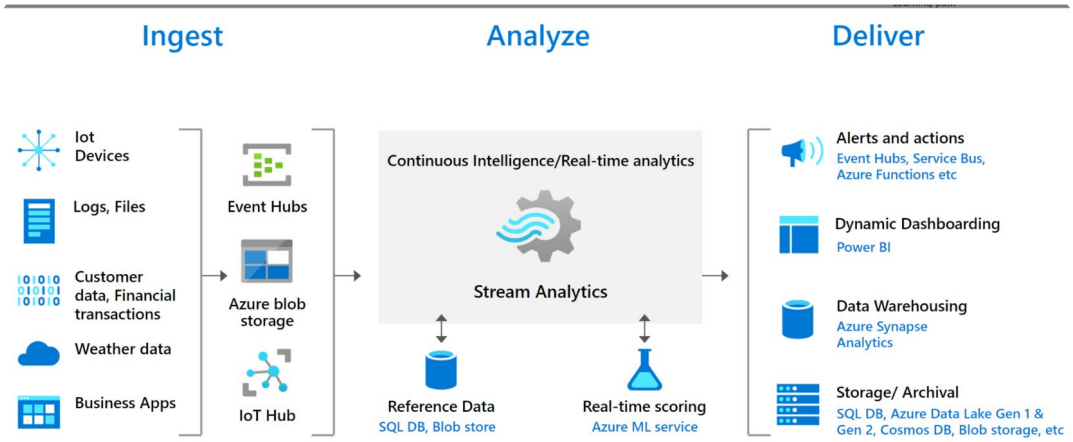
https://t.me/learningnets

What is Azure Stream Analytics

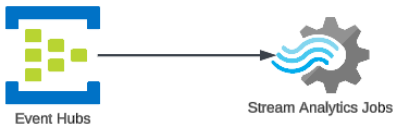
Azure Stream Analytics

This is a fully managed stream processing engine.

You can use this to process large amounts of data in real time.



About Windowing functions



A lot of data gets generated during the process of streaming

Most of the times , we want to aggregate the data over a span of time.

Windowing functions

Tumbling window - This is used to segment the data stream into distinct time segments. A function is performed against the data values in the window.

Here data values don't repeat or overlap.

An event can't belong to more than one time window.

Here events can be consolidated over time windows.

https://t.me/learningnets

Design and Develop Data Processing - Scala, Notebooks and Spark

Why Spark

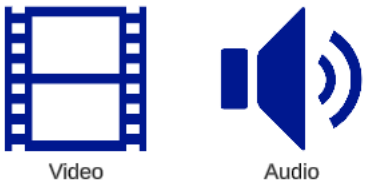
We have seen how to host a SQL data warehouse in Azure Synapse.

We can extract, process data from Azure Data Lake and send it across to a SQL data warehouse

We have seen how to send data onto Azure Event Hubs and process data via the use of Azure Stream Analytics.

But the fact that we need to consider is that a lot of source data is in semi-structured or unstructured in nature.

JSON, CSV



Binary data

Yes, now we do have a lot of cloud services to process data.

https://t.me/learningnets

Before the advent of the cloud, Hadoop was used to process Big data sets.

Then came Apache Spark which is a much more efficient engine when it comes to processing data.

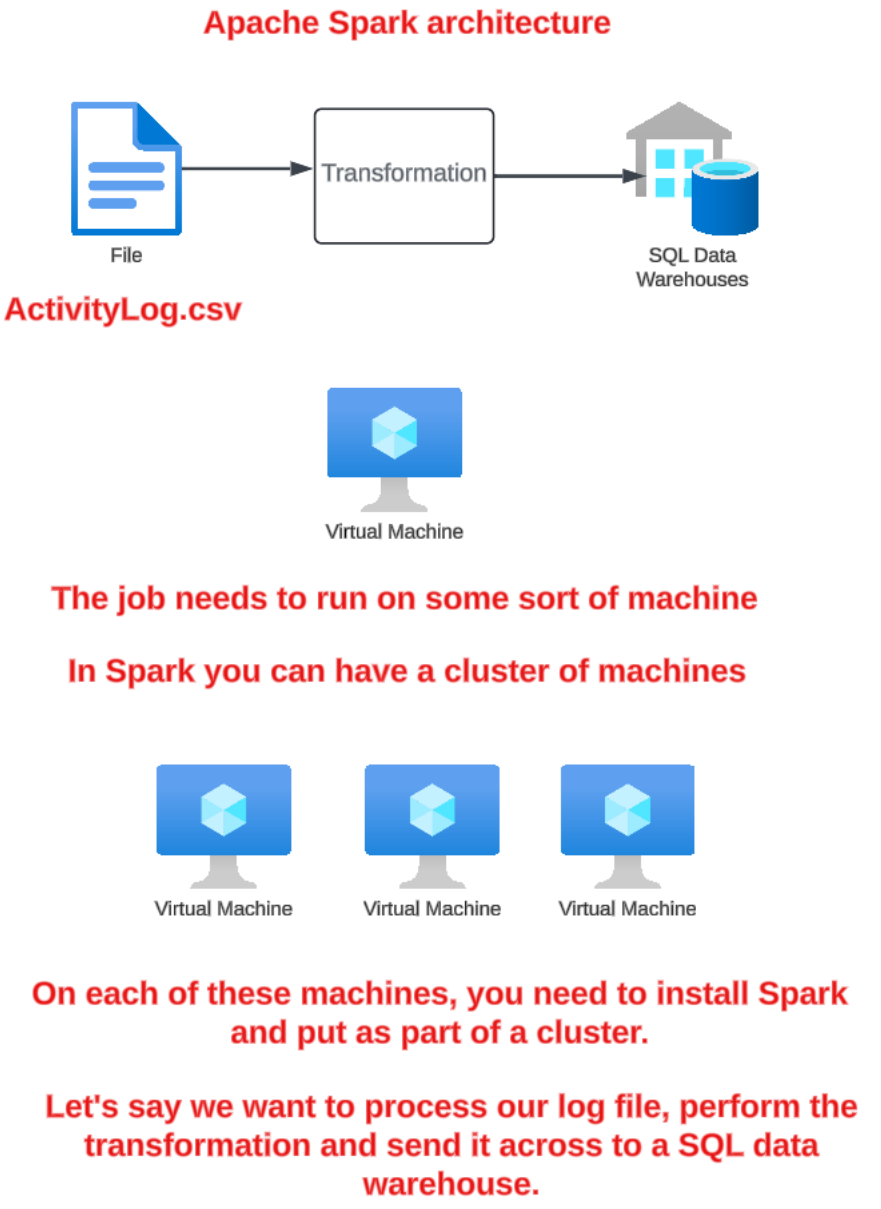
Here you can work with both Batch and streaming data.

You can also use a variety of programming languages to work with data - Python, Scala, Java, R - You are not constrained for just using SQL.

There are many libraries that were infused to give features when it comes to Data Science and Machine Learning.

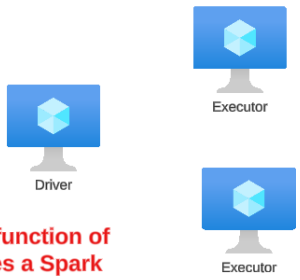
https://t.me/learningnets

Spark Architecture



https://t.me/learningnets

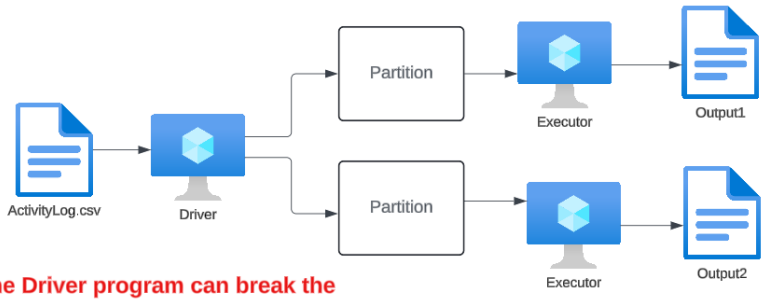
We can submit onto Spark.



The Executor runs the tasks and keeps the data in memory or on the disk.

The driver runs the main function of an application and creates a Spark context

Spark is efficient and fast in processing data because its all processed within memory.



The Driver program can break the file into multiple partitions. Then the task which is the program and the data can be sent to an executor.

Each executor can work in parallel with the data and provide the output accordingly.

https://t.me/learningnets

Installing Spark

Virtual Machine

We are going to create a Windows Server-based Azure virtual machine.

We will need to install Python and Java.

We will download Spark and WinUtils.

We need to setup Environment variables.

We can then run Spark.

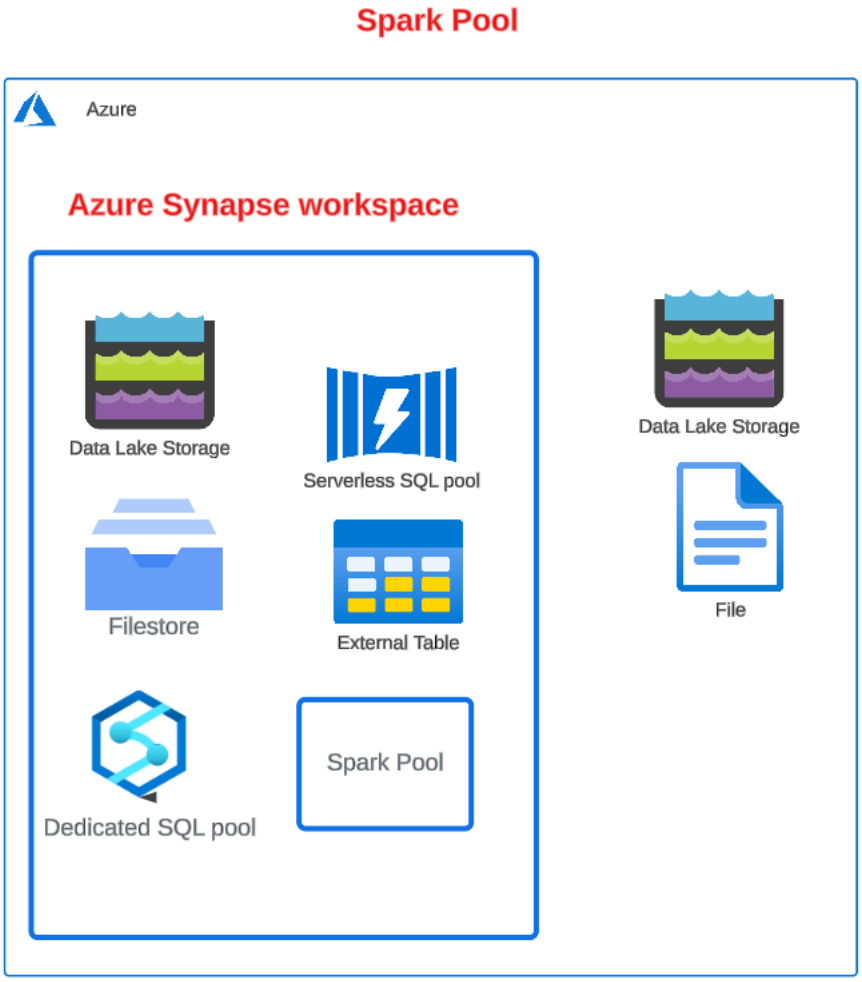
Notebook

We are going to use Notebooks - This is an interactive way when it comes to development.

https://t.me/learningnets

Design and Develop Data Processing - Azure Synapse and Azure Databricks

Azure Synapse - Spark Pool – Concepts

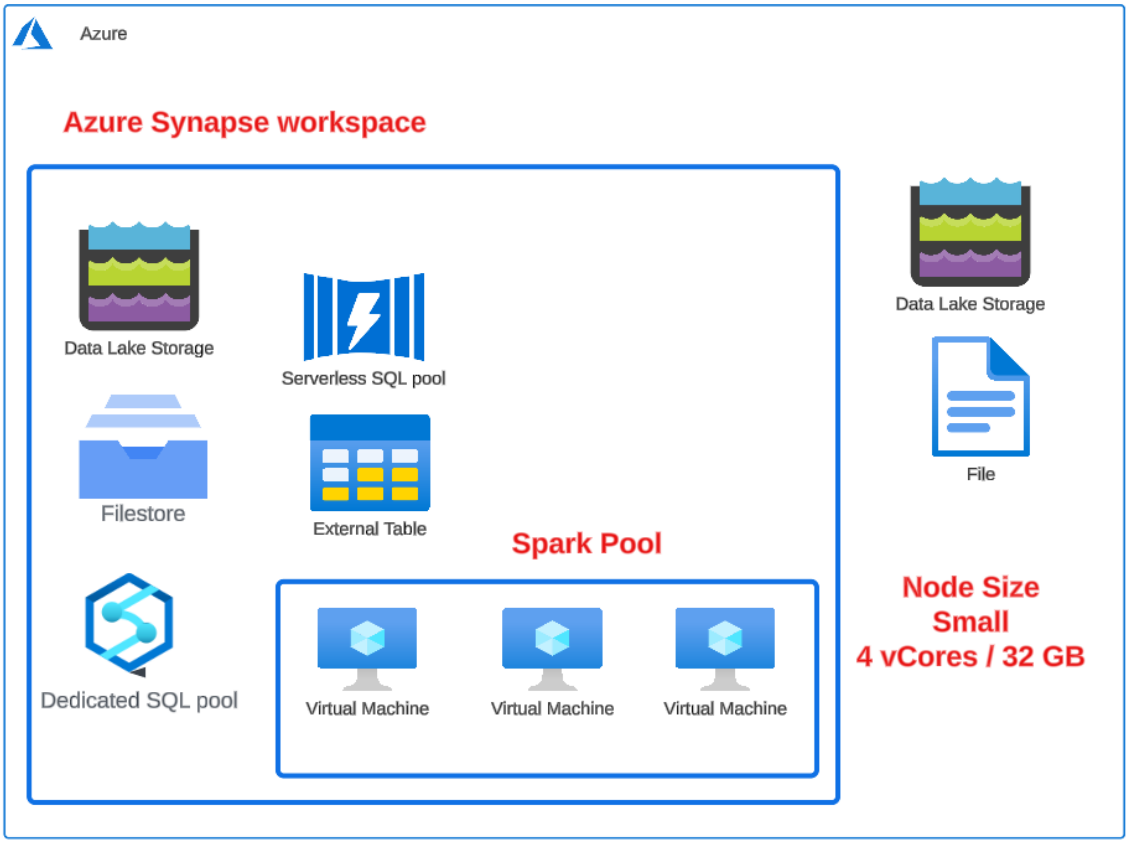


You can create a serverless Spark pool

When you define the spark pool, only the meta data is defined. There are no instances created, hence there is no cost associated with the pool.

When you create a session and run a spark job, that is the time the instances get created.

https://t.me/learningnets



https://t.me/learningnets

What is Azure Databricks

Databricks

Databricks as a company was founded by the original creators of Apache Spark.

It is a cloud platform that allows enterprises to build, scale and govern data and AI.

They brought about the data lakehouse which is a combination of a data lake and a data warehouse.

Here it tries to address the same issues we saw earlier on - Data keeps on streaming in onto a data lake. It becomes difficult to manage and govern the delta data. And then we have the expense of maintaining a data warehouse.

Databricks is available on cloud platforms such as Microsoft Azure.

Atomicity - Here transactions are encapsulated in a single unit.

Consistency - Changes are made in a predictable manner.

Isolation - Here transactions don't interfere with each other.

Durability - Transations are saved in the end.

Keep track of your data assets.

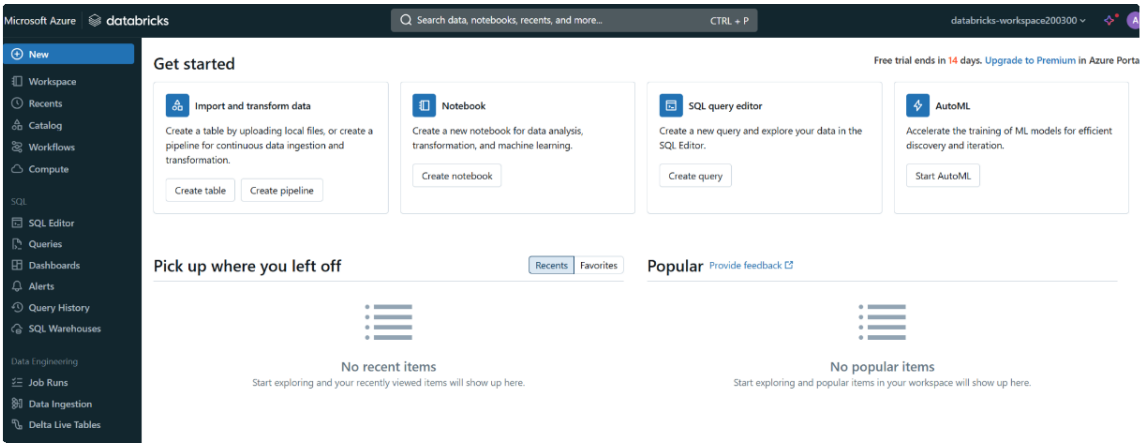
Data can be ingested in different formats.

Be able to have structure for your data.

https://t.me/learningnets

Concepts with Azure Databricks

In Azure Databricks, we first need to create a workspace. This is used for storing all Databricks related assets.



Then we need to create a compute cluster which will have Spark installed - This can be used to run our Notebooks.

You can create different types of clusters.

You can run your Notebooks on Interactive clusters - You can manually terminate the clusters.

You can also run jobs on job-based clusters. Here the cluster is terminated once the job is complete.

Databricks File system - This is an abstraction layer on top of blob storage.

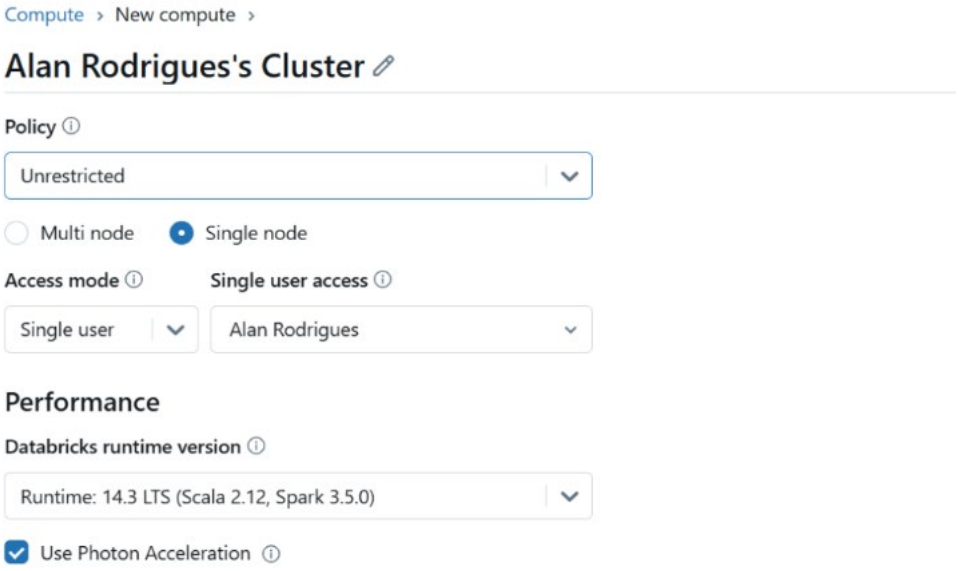
Database - Collection of objects such as tables and views.

Table - This is a representation of structured data.

https://t.me/learningnets

Lab - Creating a cluster

When creating a cluster we can choose a single or multi node



If you just need to run small jobs, we can make use of a single node.

In the single node, the driver and executor run on the same machine. All of the stderr, stdout and log4j log outputs are on the single node, on the driver log.

https://t.me/learningnets

Access Nodes

Access Mode	Visible to user	UC Support	Supported Languages	Notes
Single user	Always	Yes	Python, SQL, Scala, R	Can be assigned to and used by a single user. Referred to as **Assigned** access mode in some workspaces.
Shared	Always (**Premium plan required**)	Yes	Python (on Databricks Runtime 11.3 LTS and above), SQL, Scala (on Unity Catalog-enabled compute using Databricks Runtime 13.3 LTS and above)	Can be used by multiple users with data isolation among users.
No Isolation Shared	Admins can hide this access mode by enforcing user isolation in the admin settings page.	No	Python, SQL, Scala, R	There is a related account-level setting for No Isolation Shared compute.
Custom	Hidden (For all new compute)	No	Python, SQL, Scala, R	This option is shown only if you have existing compute without a specified access mode.

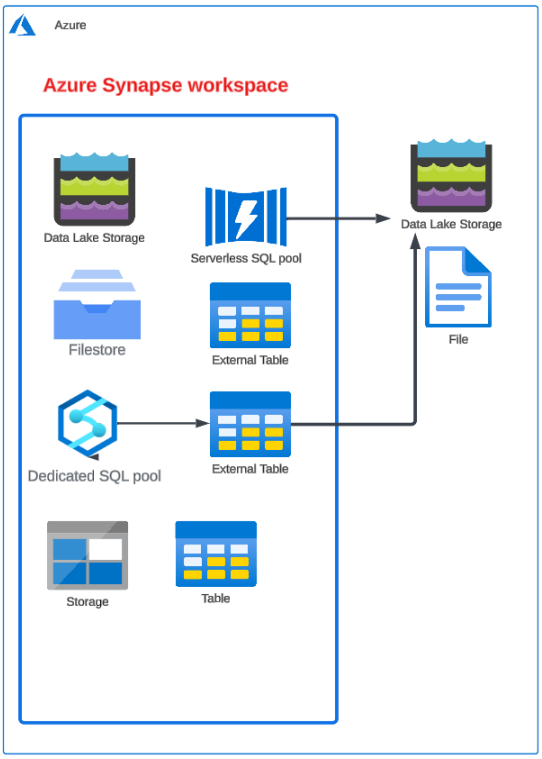
Reference - https://learn.microsoft.com/en-us/azure/databricks/compute/configure

https://t.me/learningnets

Design and Implement Data Security

Authorization for Azure Data Lake Gen2

Authorization when it comes to Azure Data Lake Gen2 account.



When it comes to Authorization, we can use the Account Key or the Shared Access Signature.

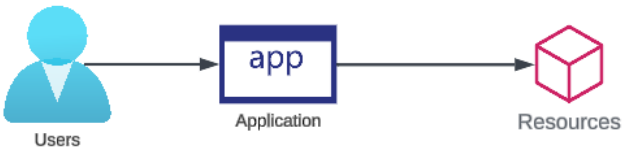
Users or applications can use these methods to authorize themselves to use the Azure Data Lake Gen2 account.

But we can also make use of Microsoft Entra ID as well.

Microsoft Entra ID

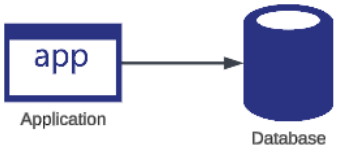
https://t.me/learningnets

Authentication and Authorization



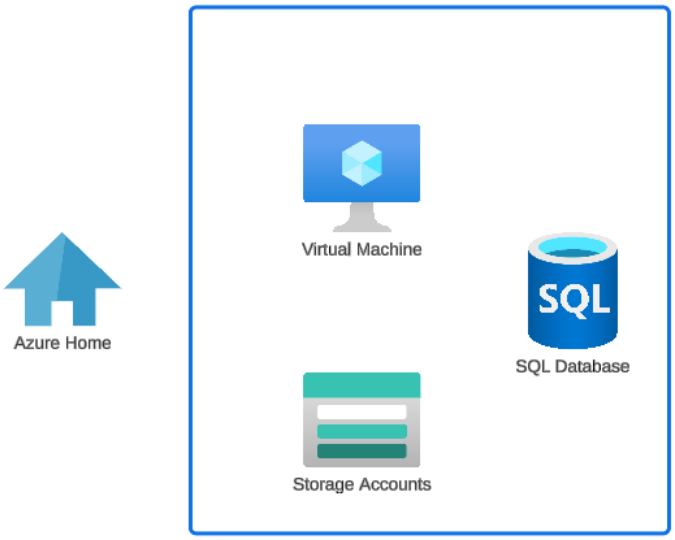
User must need to authenticate themselves - This can be done via a user name and password.

If the user needs to access a resource, the user needs to have the right permissions.



The application would need to maintain a data store for the user credentials and permissions.

But nowdays applications can offload the authentication and authorization to other third-party identity providers.



So far we have been working with Azure resources with our Azure Admin Account.

But in an organization, you want to have users who can access and manage resources.

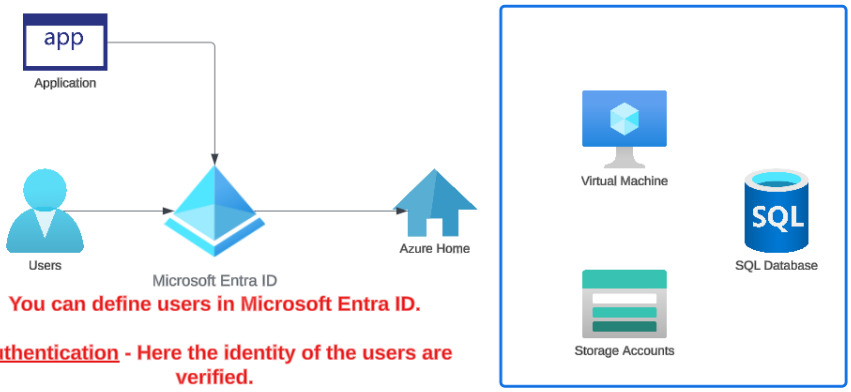
Who has permission to create resources. Who has permission to access resources.

https://t.me/learningnets

We need to create users and be able to assign permissions.

Microsoft Entra ID - This is a cloud-based identity and access management service. This identity service can be used for Azure, Microsoft 365 and even other Software-as-a-service applications.

Even Applications can be linked to identities and be given access accordingly.

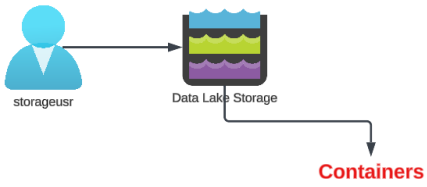


You can define users in Microsoft Entra ID.

Authentication - Here the identity of the users are verified.
Authorization - Here the permissions are checked for the users.

https://t.me/learningnets

Lab - Using Microsoft Entra ID - Using RBAC - Storage Blob Data Reader



Role	Description
Storage Blob Data Owner	Full access to the Blob containers and the data.
Storage Blob Data Contributor Role	Gives read, write and delete access to the Blob containers and data. But this does not give permissions to manage the Access Control lists.
Storage Blob Data Reader	Gives permission to read and list the Blob storage containers and blobs.

Name	Last modified	Anonymous access lev...	Lease state
$logs	6/23/2024, 12:21:34 PM	Private	Available
data	7/2/2024, 8:21:02 AM	Private	Available
insights-logs-networksecuritygroupflowevent	7/10/2024, 2:28:09 PM	Private	Available
logs	6/24/2024, 4:12:25 PM	Private	Available
staging	6/23/2024, 2:49:57 PM	Private	Available

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: data

Search blobs by prefix (case-sensitive) Show deleted objects

Name	Modified	Access tier	Archive status	Blob type
ActivityLog01.csv	7/18/2024, 8:01:08 AM	Hot (Inferred)		Block blob
ActivityLog01.parquet	7/18/2024, 3:10:58 PM	Hot (Inferred)		Block blob

https://t.me/learningnets

Lab - Using Access Control Lists

We can assign access levels to security principals for files and directories.

Reference - https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

	File	Directory
Read (R)	Can read the contents of a file	Requires **Read** and **Execute** to list the contents of the directory
Write (W)	Can write or append to a file	Requires **Write** and **Execute** to create child items in a directory
Execute (X)	Does not mean anything in the context of Data Lake Storage Gen2	Required to traverse the child items of a directory

Numeric form	Short form	What it means
7	`RWX`	Read + Write + Execute
5	`R-X`	Read + Execute
4	`R--`	Read
0	`---`	No permissions

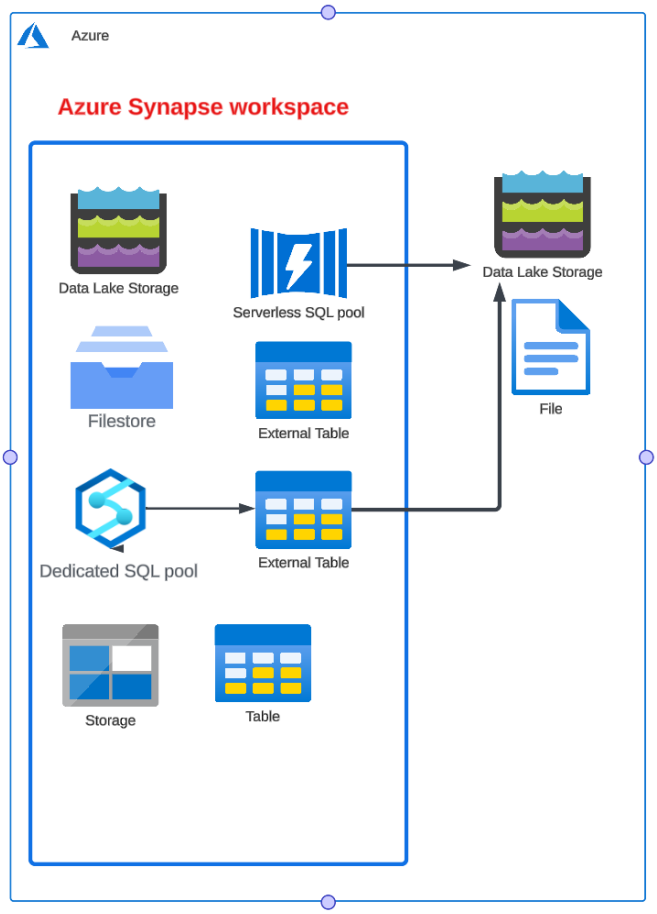
https://t.me/learningnets

Azure Synapse - Dedicated Pool Encryption

When it comes to storage of data for your dedicated SQL pool, remember in the end that the data is stored on disks in an Azure Data center.

There are several security protocols for the Azure Data center, but you can still enable the encryption of the underlying data when it is stored on the underlying disks in the Azure data center.

This is known as Transparent Data Encryption. Here the encryption occurs in the background, applications don't need to enable anything to use the encryption feature.



https://t.me/learningnets

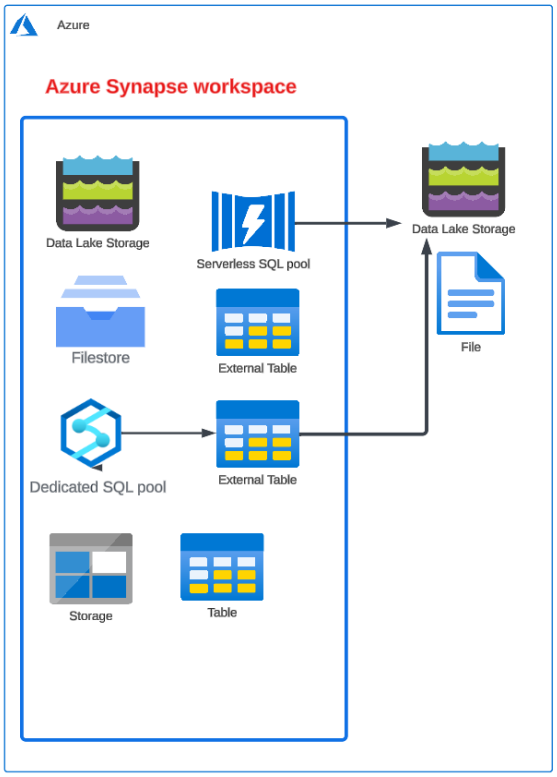
Azure Synapse Workspace Encryption

For the Dedicated SQL Pool we could enable Transparent Data Encryption.

Here the data is encrypted at rest.

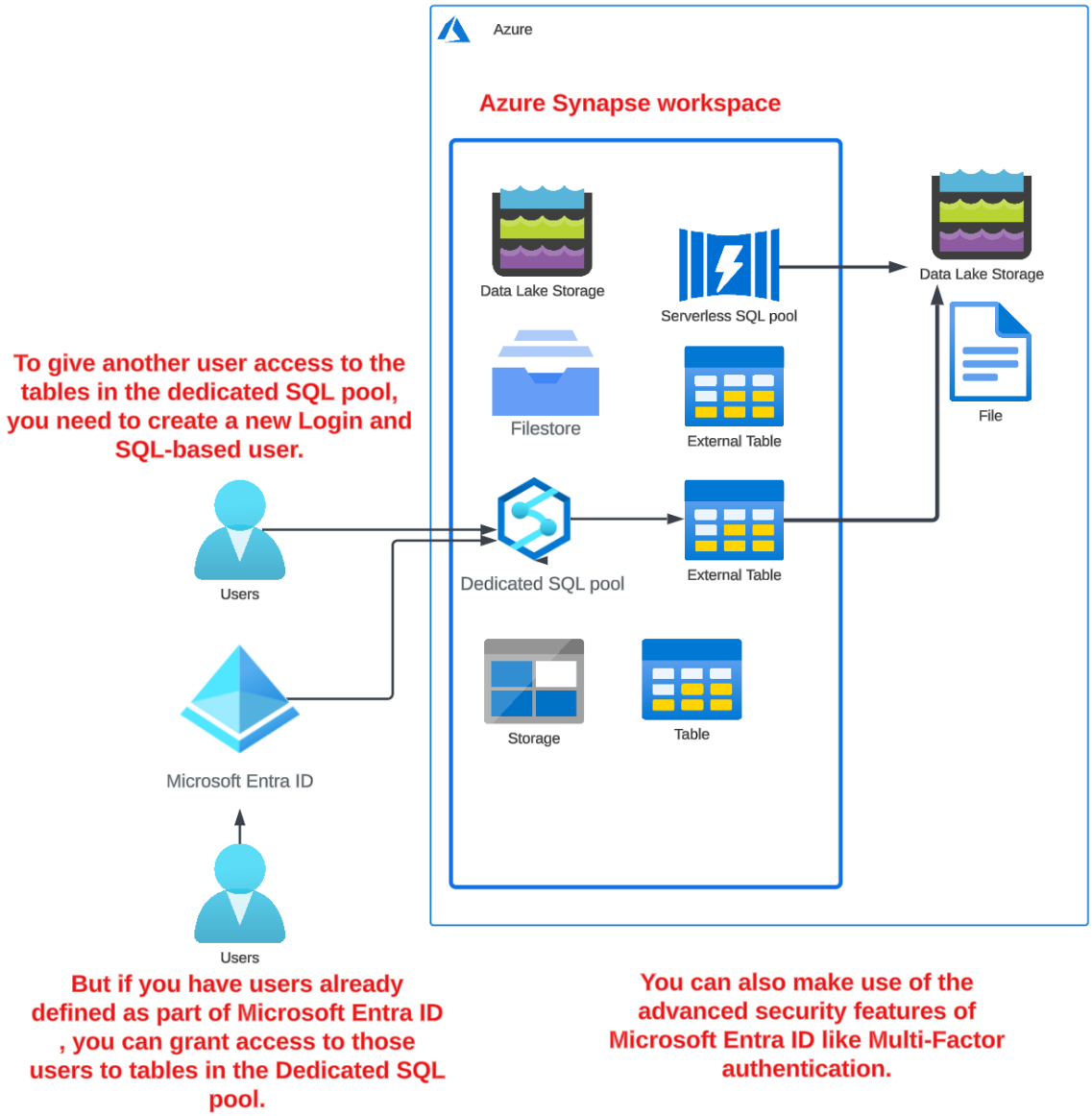
We can also enable encryption for the entire Azure Synapse workspace.

Here we can use another Azure service - Azure Key vault to manage the encryption key used for encryption purposes.



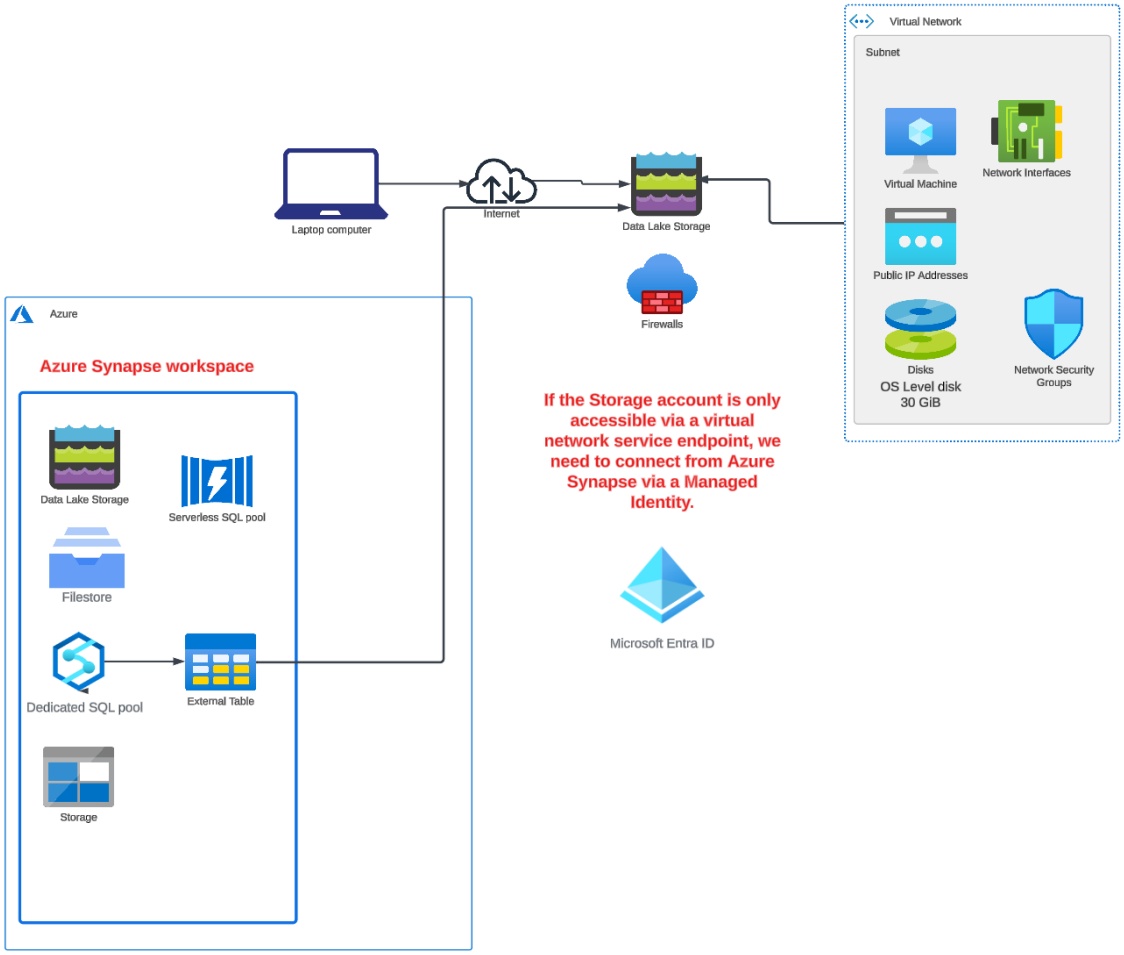
https://t.me/learningnets

Azure Synapse - Microsoft Entra ID



https://t.me/learningnets

Azure Storage Accounts - Network and Firewall



https://t.me/learningnets

About Managed Identities

Managed Identities

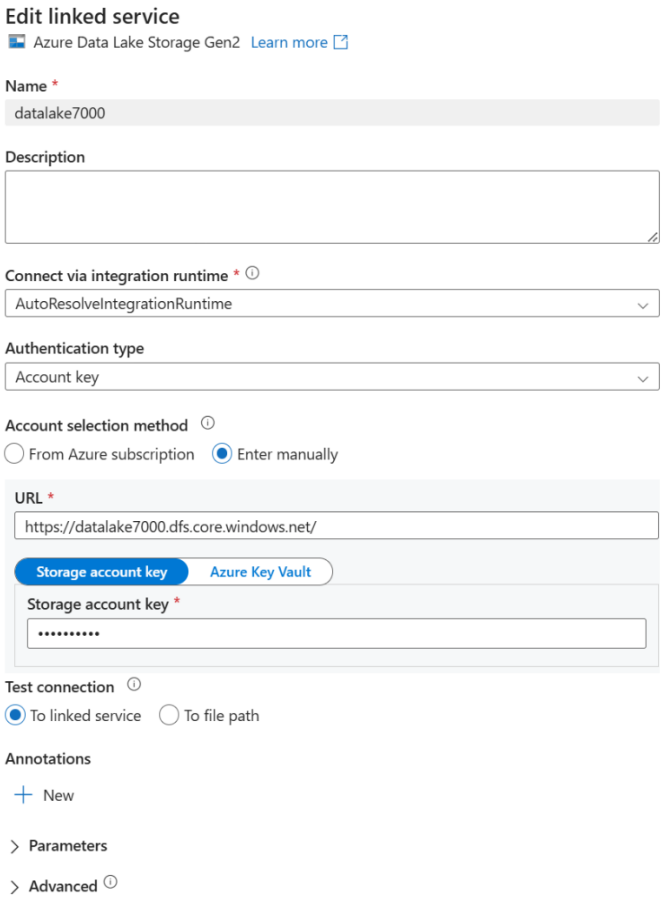
Provides a more secure way to access resources on Azure.

Linked Service

Here Azure Data Factory uses the Account Key to connect to the storage account.

This is secure , but you can also use a managed identity.

Instead of depending on the Account Key which in the end is like using a password for authorization purposes.



https://t.me/learningnets

You can enable the managed identity for Azure Data factory

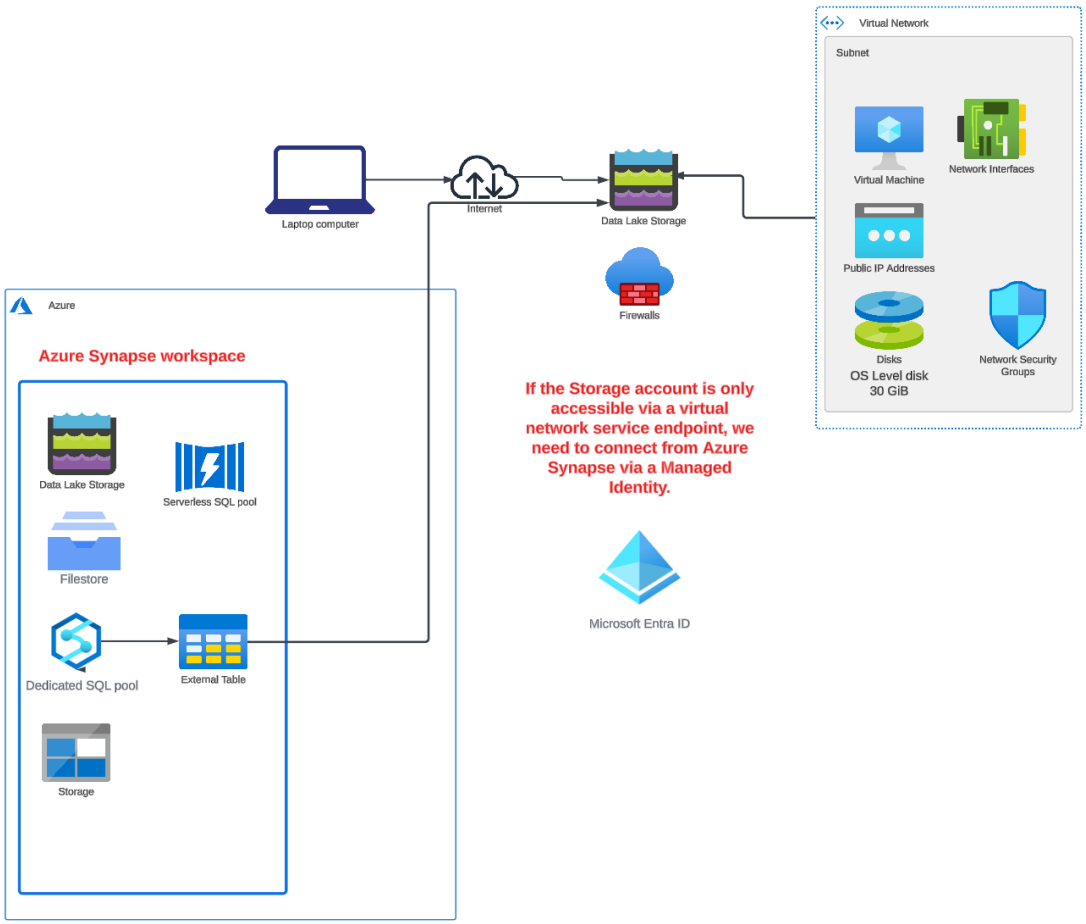
Microsoft Entra ID

This creates a security principal in Microsoft Entra ID.

The linked service can use the Managed Identity for authorization purposes.

https://t.me/learningnets

Azure Synapse - Managed Identity connectivity



https://t.me/learningnets

Azure Data Factory – Encryption

We need to use an Encryption key to encrypt the data in Azure Data Factory.

1. **Azure Key Vault - Soft delete and purge protection is enabled.**

2. **Provide access to the Azure Key Vault for the Azure Data Factory resource.**

3. **Create an encryption key in the key vault.**

4. **Enable encryption in Azure Data Factory.**

Monitor and optimize data storage and data processing

https://t.me/learningnets

Microsoft Purview

Welcome to the Microsoft Purview portal

Microsoft Purview brings together solutions across data governance, data security, and compliance so that you can govern and secure your data wherever it lives.

Supported cloud platforms:

Microsoft 365 Microsoft Azure Microsoft Fabric

Other cloud platforms

Ideal when you want to discover data assets that a company has.

Data Lake Storage

Azure Synapse Analytics

Helps to maintain a data map - Here you can scan your data stores, extract the metadata and understand the data assests that you have.

You can use the Data Catalog to browse for assets based on the metadata.

You can also get an entire data lineage of your data assests , the various stages that your data can go through.

https://t.me/learningnets

Best practices for data storage - Azure Data Lake

https://t.me/learningnets

You have a wide variety of file formats available - CSV, JSON, XML, Avro, Parquet, ORC(Optimized Row Columnar)

Avro, Parquet and ORC are all compressed. Hence file sizes are smaller. The schema is embedded in each file. Avro uses a row format and Parquet and ORC use column format.

Parquet and ORC file formats are good when there are more read operations and there is focus to get a subset of columns in the records.

Avro file format is ideal when there are more write-based transactions and there is a need to retrieve multiple rows that need to fetch the entire row information.

Directory structure

IoT structure

In IoT workloads, there can be a great deal of data being ingested that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout:

- *{Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/*

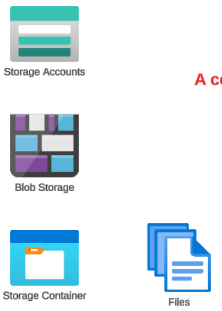
Batch job structure

- *{Region}/{SubjectMatter(s)}/In/{yyyy}/{mm}/{dd}/{hh}/*
- *{Region}/{SubjectMatter(s)}/Out/{yyyy}/{mm}/{dd}/{hh}/*
- *{Region}/{SubjectMatter(s)}/Bad/{yyyy}/{mm}/{dd}/{hh}/*

Reference - https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices

https://t.me/learningnets

Azure Data Lake Gen2 - Access tiers

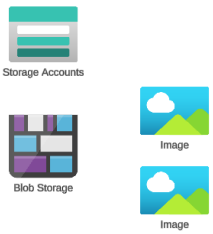


A company can look towards millions of objects in an Azure Storage Account.

Data storage prices pay-as-you-go	Premium	Hot	Cool	Cold	Archive
First 50 terabyte (TB) / month	$0.15 per GB	$0.018 per GB	$0.01 per GB	$0.0036 per GB	$0.00099 per GB
Next 450 TB / month	$0.15 per GB	$0.0173 per GB	$0.01 per GB	$0.0036 per GB	$0.00099 per GB
Over 500 TB / month	$0.15 per GB	$0.0166 per GB	$0.01 per GB	$0.0036 per GB	$0.00099 per GB

A company would want to monitor their storage costs.

An this can especially be the case if objects are not being used.

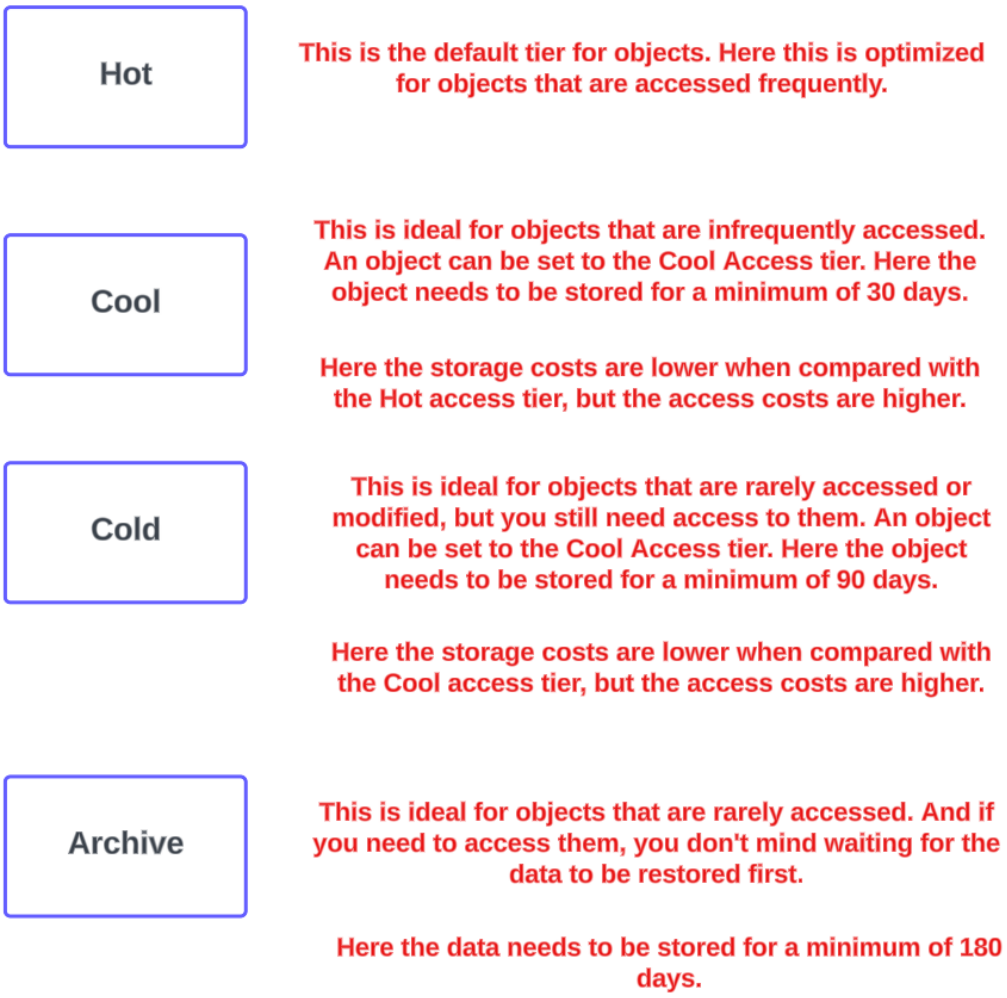


A thousand images have been uploaded on a particular day. During the first week the images are being used regularly.

But after a week the images are not being accessed. Should be still pay the same when it comes to storage costs.

We can use Access tiers to help in this regard.

https://t.me/learningnets



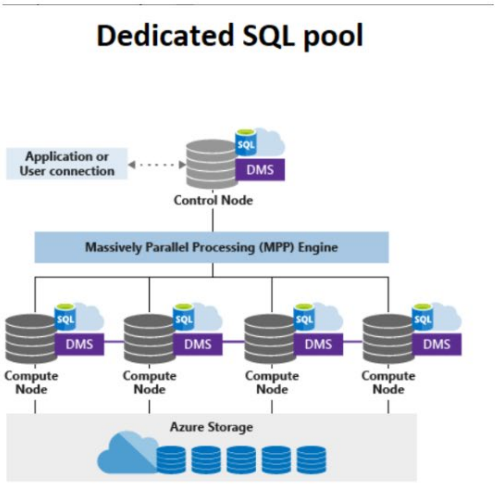
Azure Data Factory Logging

https://t.me/learningnets

In the Monitor section for Azure Data Factory you can view the pipeline runs, see the status when it comes to the Activities. All of this information is retained for 45 days.

If you want to retain this information for a longer time, you can use a Log Analytics workspace. Here you can direct the logs from an Azure Data Factory resource.

Azure Synapse - Result set caching



We can enable result set caching for the dedicated SQL pool.

If there are frequent requests for the same query that returns the same results, then instead of the compute nodes formulating the results again, the result can just be taken from the cache.

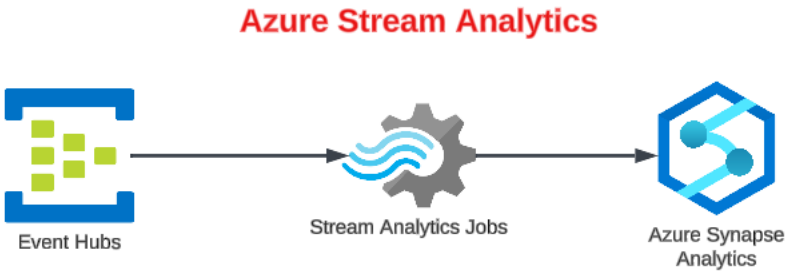
The following types of queries don't get cached

1. Queries with built-in functions - e.g DateTime.Now() , GetDate()
2. Queries that use user-defined functions
3. Queries that use tables with row level security.
4. Queries that return data with row size larger than 64 KB.
5. Queries that return a large data set - greater than 10GB.

You can then use the view - sys.dm_pdw_exec_requests view along with the request id to see if the SQL request used the cache.

https://t.me/learningnets

Azure Stream Analytics – Optimization



The streaming units allocated to a job represent the compute resources allocated to the job.

The streaming job is highly dependent on memory , because the data is processed in memory.

If the job runs out of memory, then the job will start to fail.

You can monitor the metric - SU (Memory) % Utilization - If this is going beyond 80% , then you might need to increase the number of streaming units.

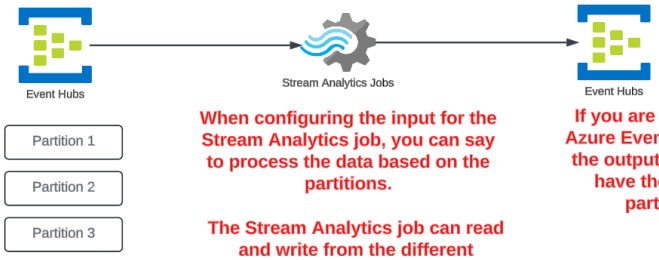
Other metrics to monitor - Backlogged Events - If this is a non-zero vaue for regular intervals it means that the job cannot keep up with the streaming events. You might need to scale up the Stream Analytics job.

https://t.me/learningnets

You can also partition the events to achieve better throughput

Data can be ingested across multiple partitions in Azure Event Hub.

Your application sending events can decide which property of the data set can be used to partition the events.



When configuring the input for the Stream Analytics job, you can say to process the data based on the partitions.

The Stream Analytics job can read and write from the different partitions in parallel.

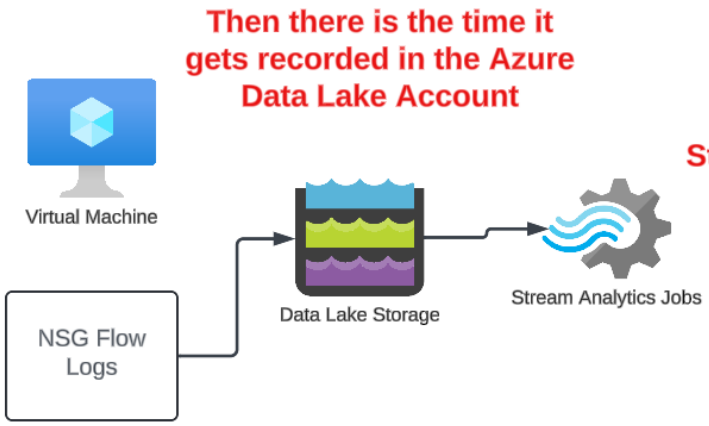
If you are writing to an Output Azure Event Hub, when defining the output, make sure that you have the same number of partitions defined.

Query	Max SUs for the job
• The query contains one step. • The step isn't partitioned.	1 SU V2
• The input data stream is partitioned by 16. • The query contains one step. • The step is partitioned.	16 SU V2 (1 * 16 partitions)
• The query contains two steps. • Neither of the steps is partitioned.	1 SU V2
• The input data stream is partitioned by 3. • The query contains two steps. The input step is partitioned and the second step isn't. • The **SELECT** statement reads from the partitioned input.	4 SU V2s (3 for partitioned steps + 1 for nonpartitioned steps

Reference - https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

https://t.me/learningnets

Azure Stream Analytics - The importance of time



By default Azure Stream Analytics processed the events based on the arrival time. For example , the time when the Blob is uploaded onto Azure Blob Storage or the time when the event arrives into Azure Event Hub.

But you can also process your events based on the application time in the event payload. Here you need to use the TIMESTAMP BY clause.

Watermark delay - This is an event timer that is maintained by Azure Stream Analytics that notes the point at which events have been ingested.

Normally streaming data is unbounded in nature. Here the data never stops. So Azure Stream Analytics uses the watermark delay to see if events are not coming in.

Because of the depedency on time, you can have late arriving events or early arriving events.

Why would such aspects occur.

https://t.me/learningnets

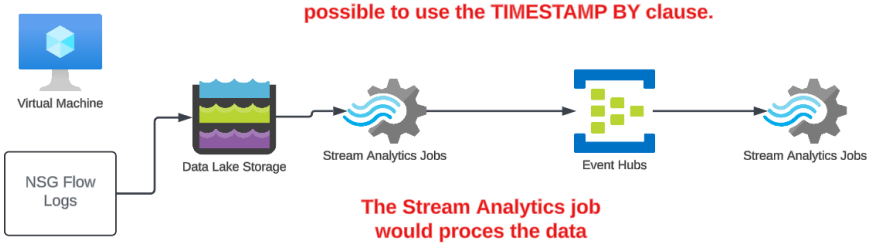
There could be a slight variation on the system clock used by the application and the ingestion service.

There is network latency.

Maybe the application has stopped working and not generating events.

The Azure Event Hub has too many partitions and data is not arriving at all partitions.

When defining steps in your SQL query, its not possible to use the TIMESTAMP BY clause.



Then we can define another Stream Analytics job to process the data.

The Stream Analytics job would proces the data based on the NSG Flow Logs , based on the steps and load the data into another Azure Event Hub.

https://t.me/learningnets