<https://www.youtube.com/watch?v=v7R0FTYW3gA> ---playlist

# What is data factory

Azure Data Factory (ADF) is Microsoft’s cloud-based data integration service tailored for modern organizations. It empowers users to design, manage, and automate workflows that handle data movement and transformation tasks at an enterprise scale.

ADF stands out for its user-friendly, no-code interface, which allows both technical and non-technical users to [build data pipelines](https://www.datacamp.com/tutorial/introduction-to-data-pipelines-for-data-professionals) easily. Its extensive integration capabilities support [over 90 native connectors](https://learn.microsoft.com/en-us/fabric/data-factory/connector-overview), enabling data flow across diverse sources, including **on-premises** systems and **cloud-based** services.

[Azure Data Factory (ADF)](https://azure.microsoft.com/en-us/products/data-factory) is a cloud-based data integration service designed to orchestrate and automate data workflows.

## Features of Azure Data Factory(ADF)

### Data integration

Azure Data Factory supports integration with [**over 90 data sources**](https://learn.microsoft.com/en-us/fabric/data-factory/connector-overview), including cloud-based and on-premises systems. It includes support for [**SQL databases**](https://www.datacamp.com/blog/all-about-sql-the-essential-language-for-database-management), **[NoSQL systems](https://www.datacamp.com/tutorial/nosql-tutorial" \t "_blank)**, [**REST APIs**](https://www.datacamp.com/tutorial/graphql-vs-rest), and file-based data sources, allowing you to unify data workflows regardless of the source or format.

### No-code pipeline authoring

ADF’s drag-and-drop interface simplifies how users create data pipelines. With prebuilt templates, guided configuration wizards, and an intuitive visual editor, even users with no coding expertise can design comprehensive end-to-end workflows.

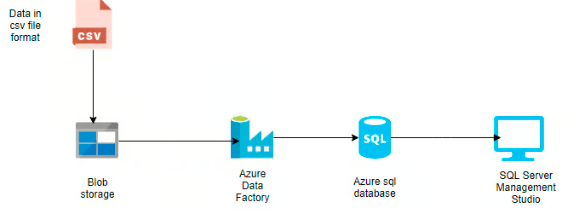
### Scheduling

Azure Data Factory’s scheduling tools offer workflow automation. Users can set up triggers based on specific conditions, such as a file’s arrival in cloud storage or scheduled time intervals. These scheduling options eliminate the need for manual interventions and ensure workflows are executed consistently and reliably.

## Core Components

### Pipelines

Pipelines are the backbone of Azure Data Factory. They represent data-driven workflows that define the steps required to move and transform data. Each pipeline serves as a container for one or more activities, executed sequentially or in parallel, to achieve the desired data flow.



### Activities

Activities are the functional building blocks of pipelines, each performing a specific operation. They are broadly categorized into:

**Data movement activities:** These activities facilitate data transfer between different storage systems. For example, the "Copy data" activity moves data from Azure Blob Storage to an Azure SQL Database.

**Data transformation activities:** These activities allow you to manipulate or process data. For instance, data flows or custom scripts can be used to transform data formats, aggregate values, or cleanse datasets.

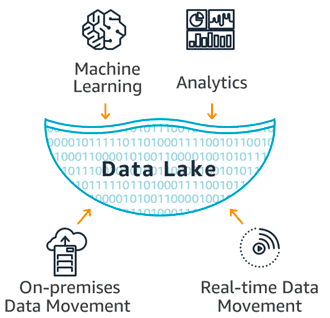
**Control flow activities:** These manage the logical flow of execution within pipelines. Examples include conditional branching, loops, and parallel execution, which provide flexibility in handling complex workflows.

# What is Azure data lake(ADL)

Data storage + Analytics

It is a one-of-a-kind solution for getting started with big data in the cloud. A data lake is a consolidated storage system developed to hold, manage, and safeguard a vast volume of structured, semi-structured, and unstructured data.  It has the capability to maintain data in its original format and process any type of data, regardless of its size.

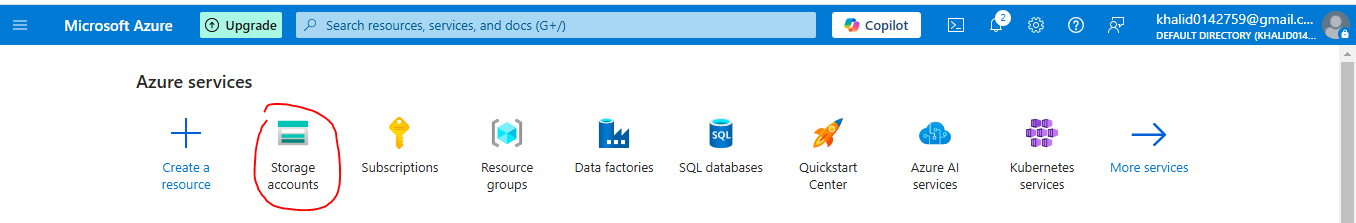
* A data lake is a central storage repository that carries **big data** from many sources of raw data in its native form until it is needed.
* It can store [**structured, unstructured data, or semi-structured**](https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/), which means data can be kept in a more flexible format for future use.
* A data lake is capable of storing and analyzing **petabyte-size** files and trillions of objects.
* It also develops massively parallel programs easily.
* It does all types of **analytics** and **processing** across platforms and languages.

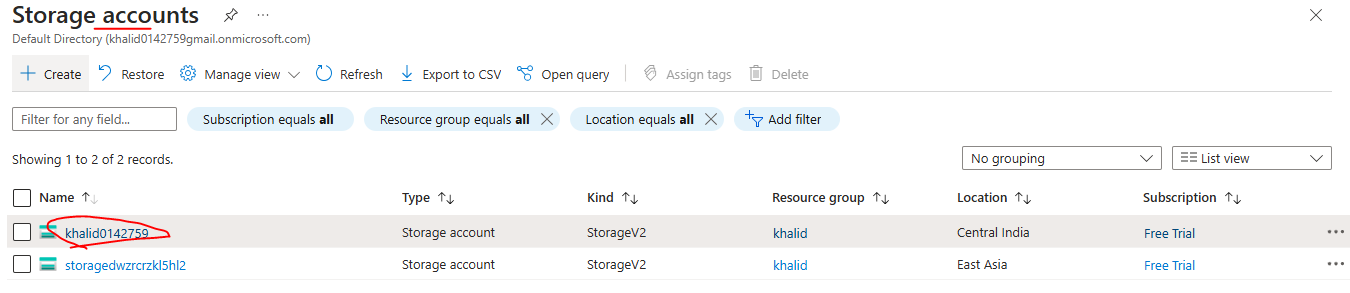


# Azure Blob storage

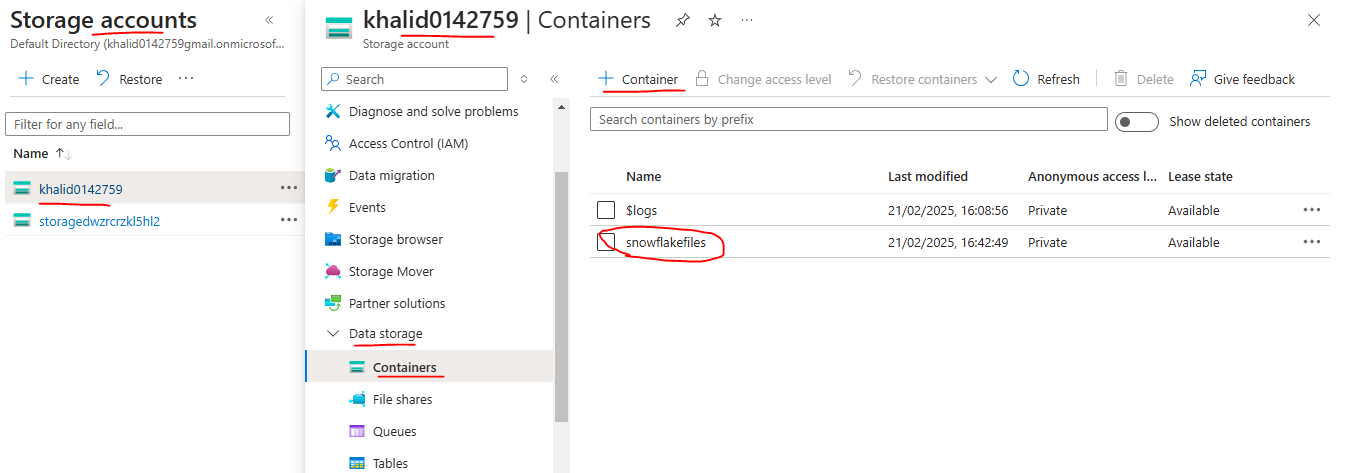
if you need a simple, scalable, and cost-effective storage solution for images, videos, documents, and backups.

**Step1**: Create Storage account





**Step2**: create container under “storage account”



### Create Storage integration in snowflake

CREATE OR REPLACE STORAGE INTEGRATION kk\_azure

TYPE = EXTERNAL\_STAGE

STORAGE\_PROVIDER = 'AZURE'

ENABLED = TRUE

AZURE\_TENANT\_ID = '061dd179-5037-4b7f-b244-9923cbf54048'

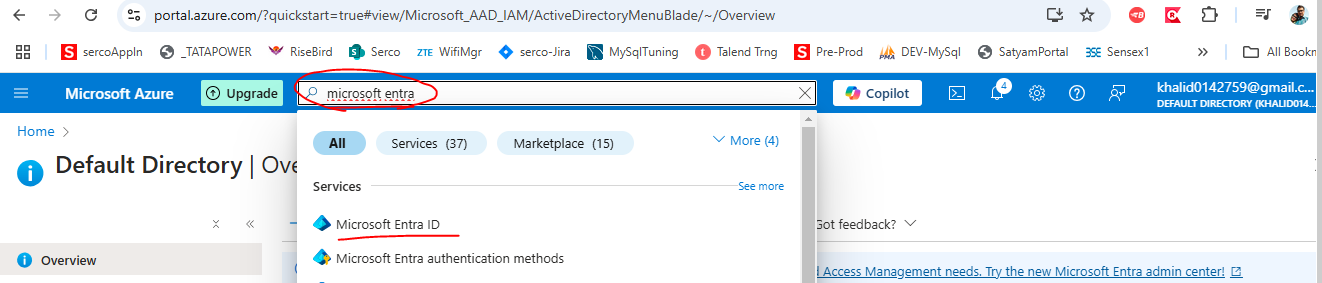
STORAGE\_ALLOWED\_LOCATIONS = ('azure://khalid0142759.blob.core.windows.net/snowflakefiles/')

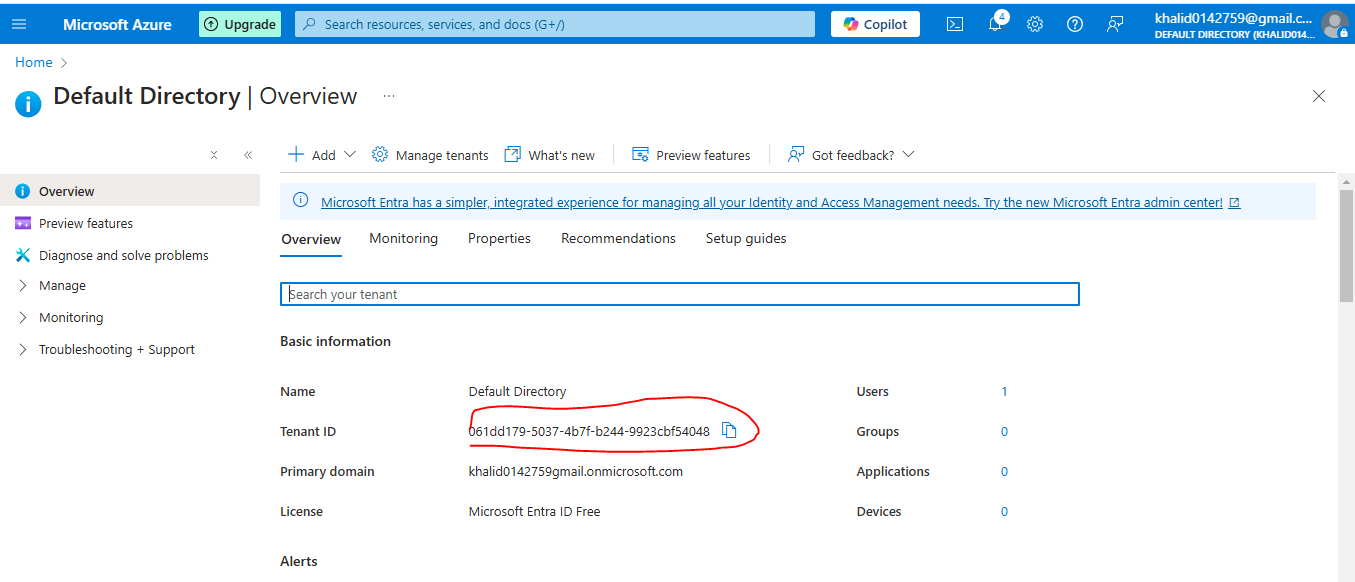
Grant permission to this azure integration

GRANT usage on INTEGRATION kk\_azure to role accountadmin

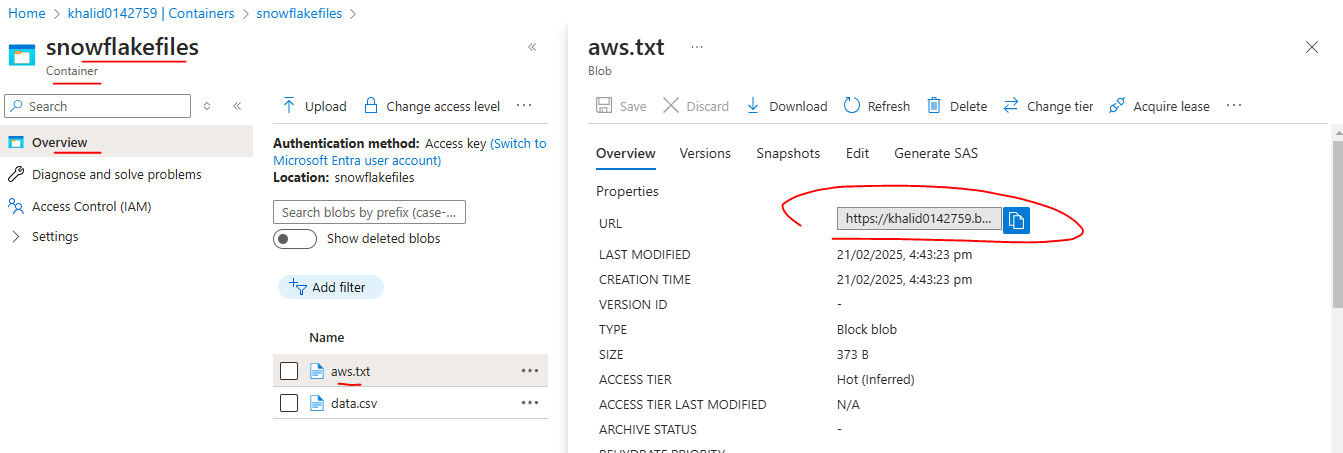
### Find tenant ID as below azure portal

Search “Microsoft entra id” as below

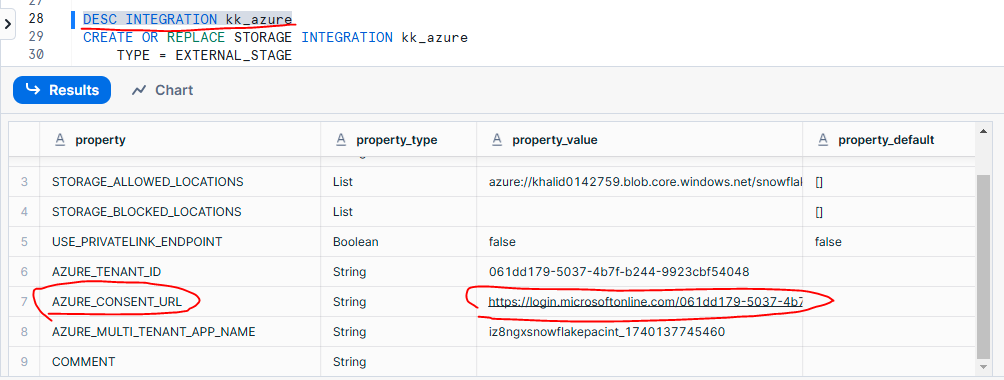




### Find container URL in azure portal as below



Describe storage integration in snowflake

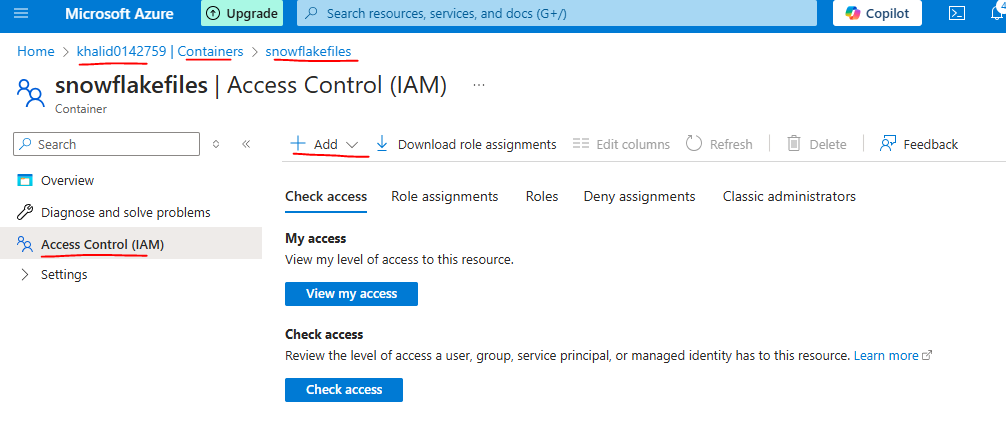


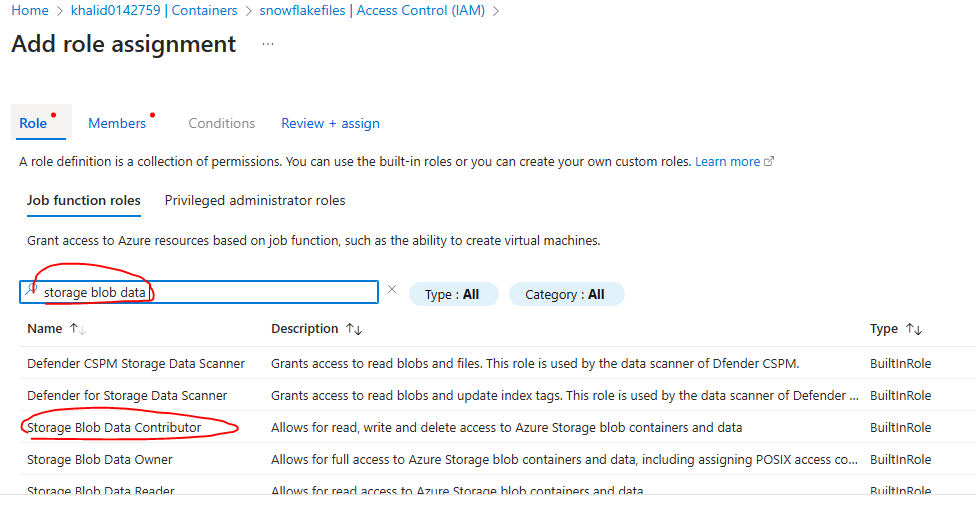
Click on “Azure\_Consent\_URL” link and accept consent as below

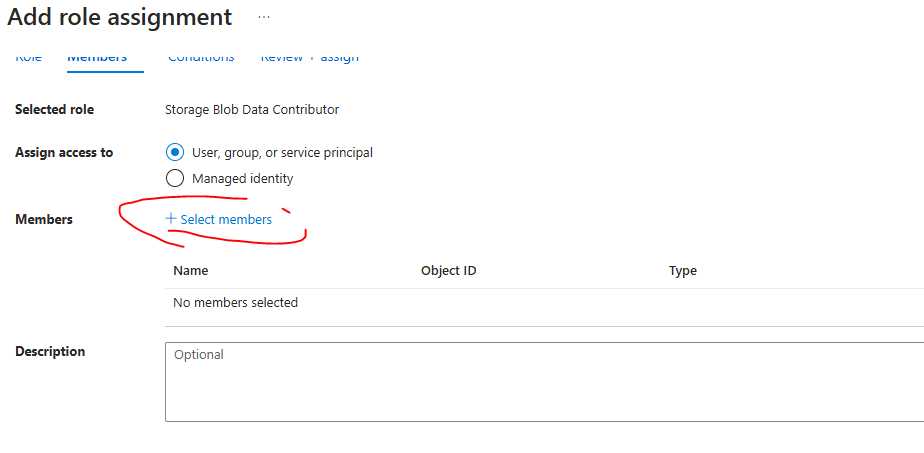
### Create IAM user in azure portal

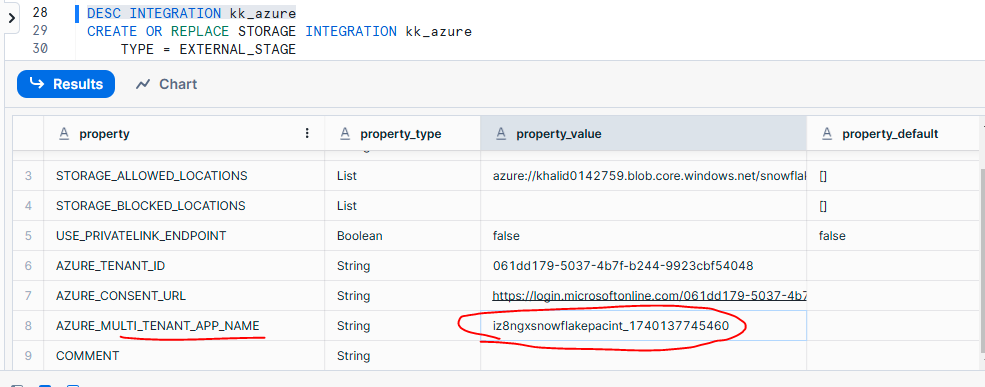
Step1: Click on Storage account

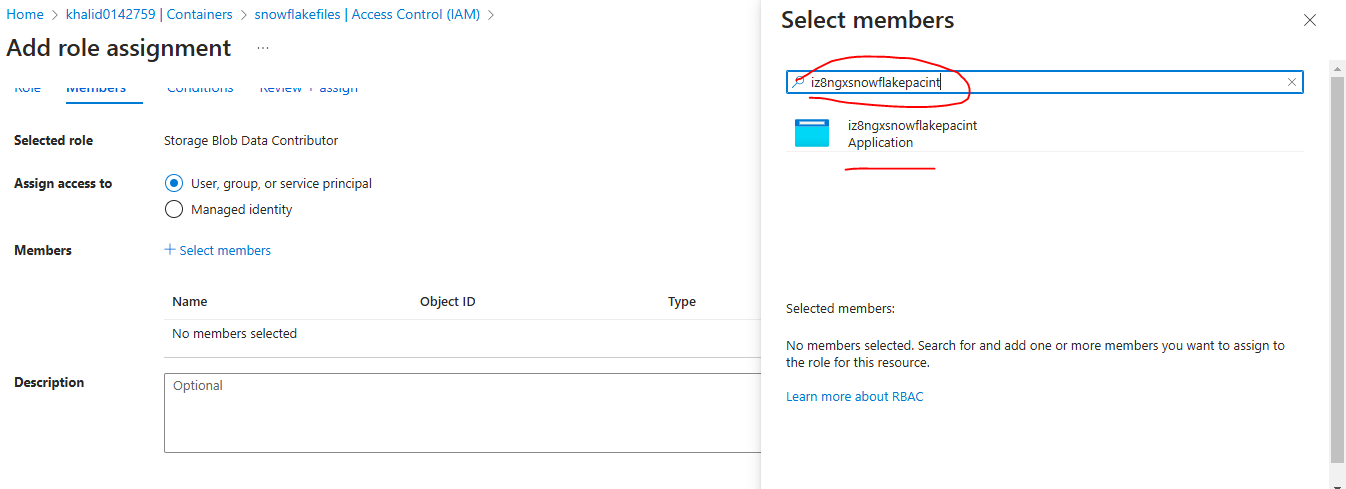
Step2: click on container

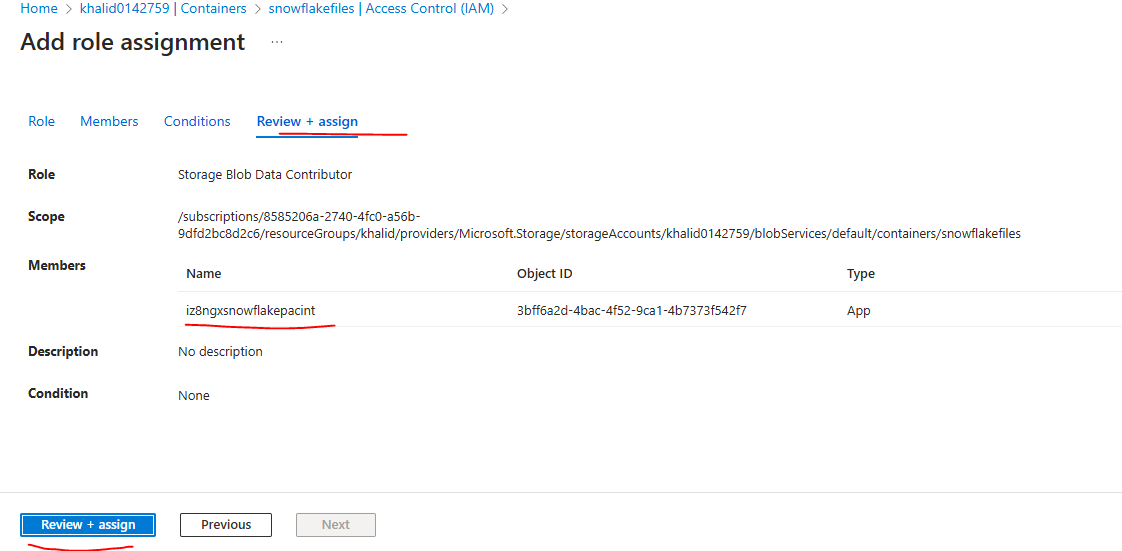


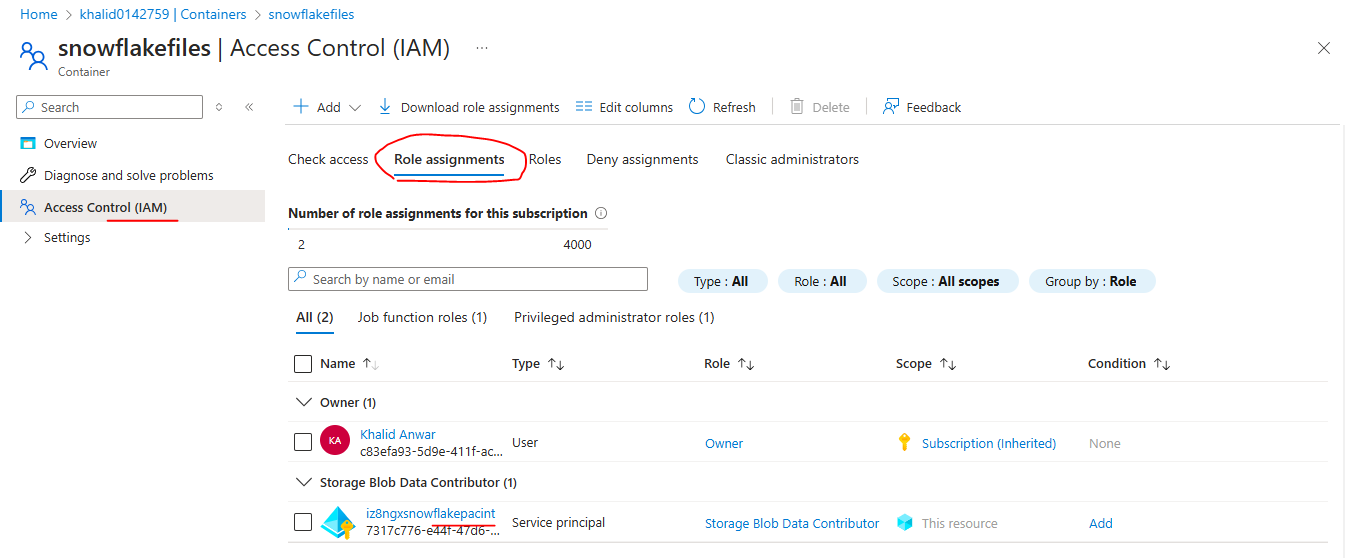












Create Stage in snowflake

CREATE OR REPLACE STAGE kk\_azure\_stage

storage\_integration = kk\_azure

url = 'azure://khalid0142759.blob.core.windows.net/snowflakefiles/'

file\_format = kk\_csv\_format;

listing all azure files

LIST @kk\_azure\_stage

### Load azure data(all files in container) into snowflake table

COPY INTO azureDate

FROM @kk\_azure\_stage/

FIle\_Format=kk\_csv\_format

## Difference between Blob storage and data lake storage

| **Feature** | **Azure Blob Storage** | **Azure Data Lake Storage (ADLS)** |
| --- | --- | --- |
| **Purpose** | General-purpose object storage (images, videos, backups, logs, etc.) | Optimized for big data analytics and hierarchical storage |
| **Storage Type** | Flat structure (containers & blobs) | Hierarchical namespace (folders & files) |
| **Use Case** | Storing unstructured data (e.g., documents, media, backups) | Big data workloads, analytics, machine learning |
| **Performance** | Optimized for high-speed data storage & retrieval | Optimized for large-scale analytical processing |
| **Access Control** | Role-based access control (RBAC) at the account and container level | Fine-grained access control using ACLs (Access Control Lists) |
| **Protocol Support** | REST API, SDKs, Azure CLI | Supports REST API, HDFS (Hadoop Distributed File System), and Azure SDKs |
| **File System Support** | Object-based storage with no hierarchical directory structure | Supports hierarchical namespace for file/folder operations |
| **Analytics Integration** | Requires external compute engines for big data processing | Natively integrates with Azure Synapse, HDInsight, Databricks, etc. |
| **Cost** | Lower cost for simple storage needs | Slightly higher cost due to additional features |

# Create Azure SQL database

<https://learn.microsoft.com/en-us/azure/azure-sql/database/free-offer?view=azuresql>