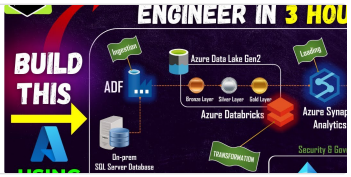
Notion: Azure end-to-end Data Engineering project

An End to End Azure Data Engineering Real Time Project Demo | Get Hired as an Azure Data Engineer

#azuredataengineer #endtoendproject #azuredataengineeringproject #azureintamil #azuredatafactory #azuredatabricks #azuresynapseanalytics #azuredatalake #datalake #powerbi #keyvault

https://www.youtube.com/watch?v=iQ41WqhHglk&t=4806s



- Source: **On-prem SQL server database**
  - For some reason I was unable to load the bak file from my desktop to had to move it to local disk D: then it loaded

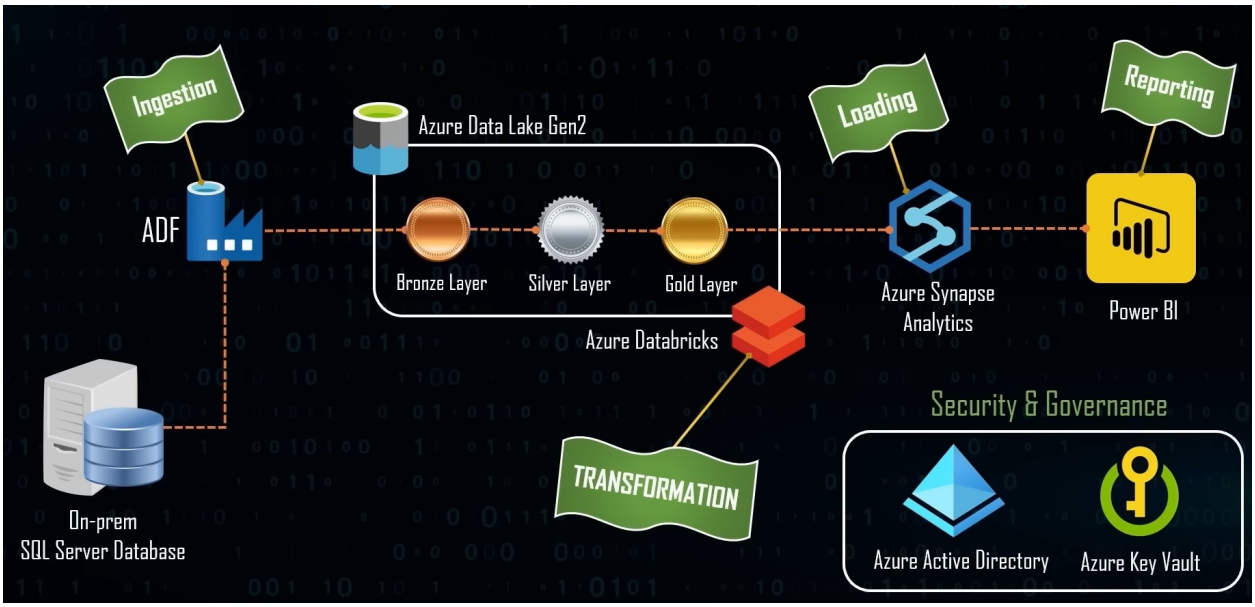
How to Import a .BAK File into a Database in SQL Server Step by Step

This article describes the detailed process to import a .bak file into SQL Server and how to backup a complete SQL Server database.

https://www.ubackup.com/enterprise-backup/sql-server-import-bak.amp.html

Extraction: **Azure data factory**

- Extract this data and move it to **Azure Data Lake Gen2**
  - In **Azure Data Lake Gen2** we have a lake house architecture
    - Bronze Layer: Exact copy of the source data → If something goes wrong you can go back to the bronze layer
    - Silver Layer: Do some transformations using Databricks and then load the data to the silver layer (changing datatypes, column names, etc) → minimal transformation
      - Now when we move data from on-prem to bronze, we make parquet format files
      - But now when we move data from bronze to silver, then we move it into DELTA file format, which is a format developed on top of parquet format by DATABRICKS
      - This means that the delta format has all the features of parquret and also some new features of its own such as:
        - Track Version History
        - Track schema changes
    - Gold Layer: Using Databricks, this is the final clean data, most curated data (maybe some aggregations)
- Azure Synapse Analytics: load the data after transformations to it
- Then use Power BI for reporting



Challenges that I faced:

1. Azure Key Vault was not correctly configured the first time I tried to create a vault with all default settings, But it was showing me RBAC error, which is the role-based access control error. So then I read the documentation and referred to the YouTube video, which helped me reconfigure the Azure key vault from scratch.

   What is Azure role-based access control (Azure RBAC)?
   Get an overview of Azure role-based access control (Azure RBAC). Use role assignments to control access to Azure resources.
   https://learn.microsoft.com/en-us/azure/role-based-access-control/overview?WT.mc_id=Portal-Microsoft_Azure_KeyVault

   This video was super helpful and helped me configure the VAULT

   Azure Key Vault Tutorial : Step-By-Step-Demo | Secret, Key, Certificates
   In this Video you ll learn about Azure Key vault one of very critical service you use for application security, Disk Encryption, and storing of the certificates.
   There is a step by step Demo for how to configure and secure access to your key vault. Also you can
   https://www.youtube.com/watch?v=xchSkmHDL0c

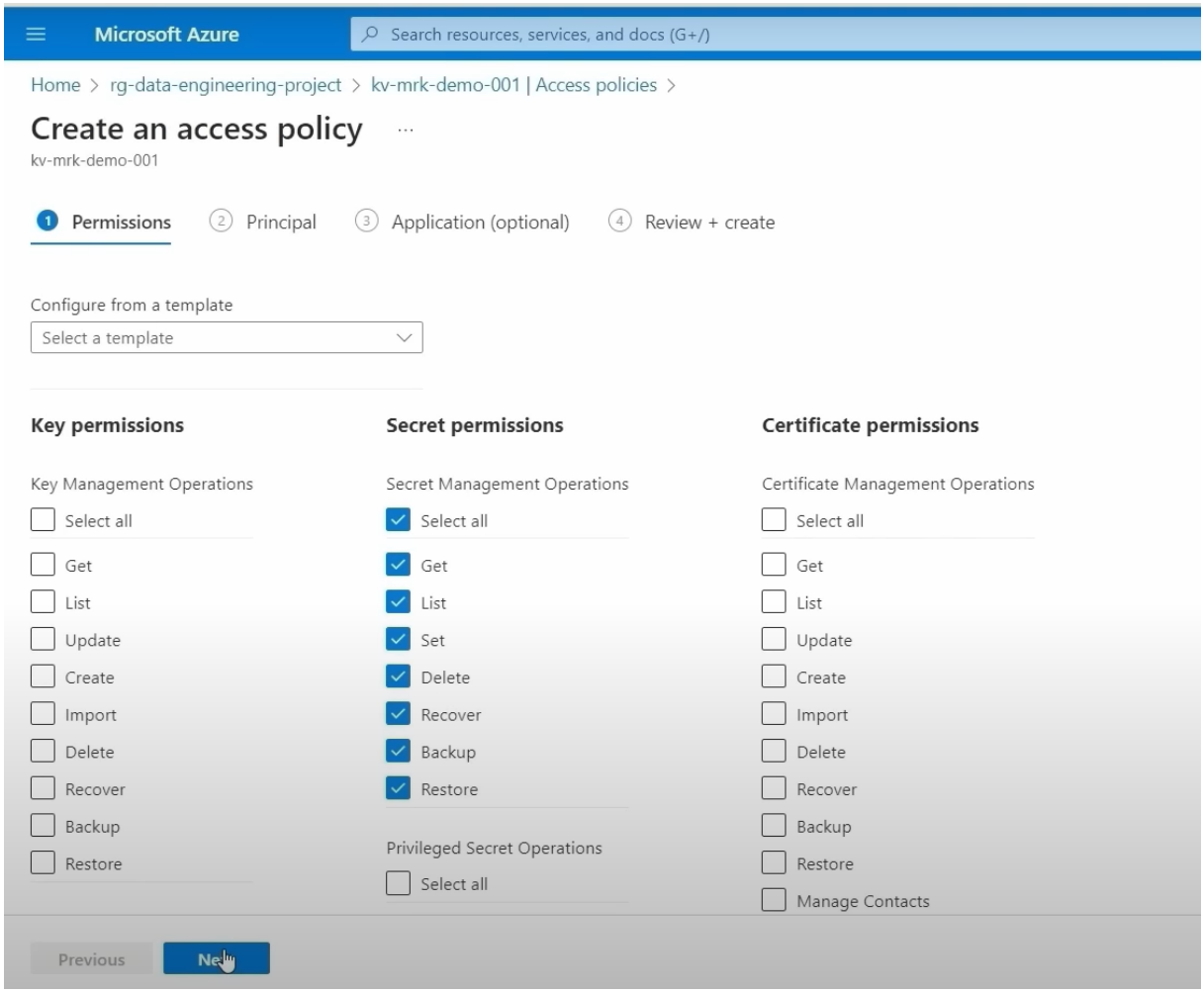
2. The second challenge that I faced was that I had successfully configured the Microsoft Integration Runtime configuration manager part. but my SQL Server database was not able to give its remote access to Azure Data Factory, so I had to follow along another youtube tutorial which helped me configure the firewall settings (enabling TCP port 1433) as well as the settings for user for a new user creation. and enabling the remote access for this database.

How to Configure Remote Access and Connect to a Remote SQL Server 2019? | MilesWeb

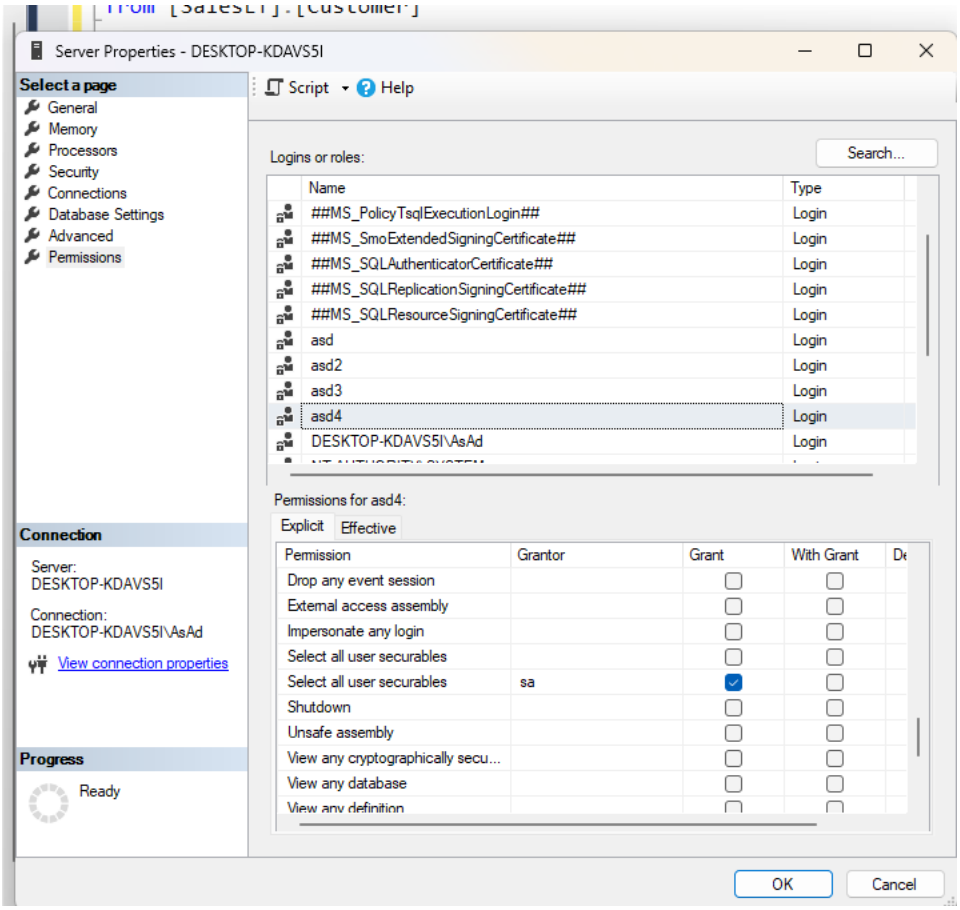
Here's an in-depth guide on how to configure remote access and connect to a remote SQL server 2019:
https://www.milesweb.com/hosting-faqs/configure-remote-access-connect-remote-sql-server2019/

https://www.youtube.com/watch?v=lJ_WRSN_wD0

3. The access policy in Azure Key Vault had to be configured as well.
    a. So basically I had to create an access policy where I gave Azure Data Factory access to the “Secrets” saved in the data vault.



4. So when I have successfully tried to import data into Azure Data factory, then I would click on get the table it would not show no tables over there. So then I had to go back to SQL Server and add this permission to my user.

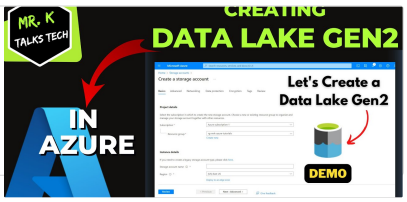


5. When creating SINK for ADF copy it asks in error message to disable soft delete for blobs

6. Configure Azure DataLake gen2 and Create a container (BRONZE) in Azure DataLake gen2 and connect to it

7. Creating an Azure Data Lake Gen2- Storage Account | Beginners Tutorials
#azuredatalake #azuretutorials #azuretutorialforbeginners #azurestorage #adlsgen2

In this Video, I have explained about how to create a Azure Data Lake Gen 2 (Storage Account) from
https://www.youtube.com/watch?v=B1FgexgPcqg&list=PLrG_BXEk3kXxv0lEASoJRTHuRq_DUqrjR&index=7



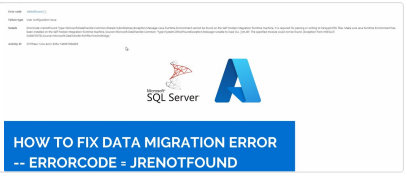
7. When I was debugging the ADF pipeline it was showing this error:

Data Migration Error -- errorcode = jrenotfound

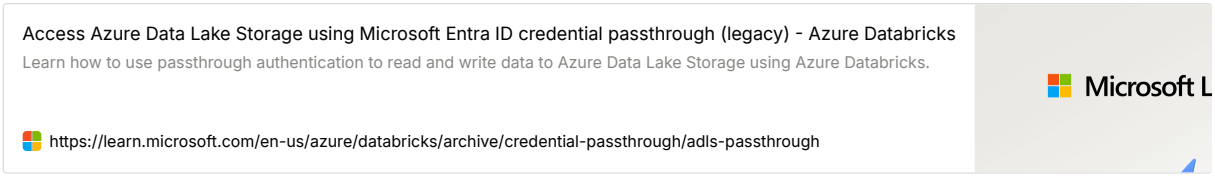
So to fix that I had to install JRE on my LOCAL machine as java runtime environment is necessary for converting SQL table to parquet

How to Fix Data Migration Error -- errorcode = jrenotfound
In this video, I describe how you can fix one of the common errors that pop up when migrating data from sql to azure datalake:

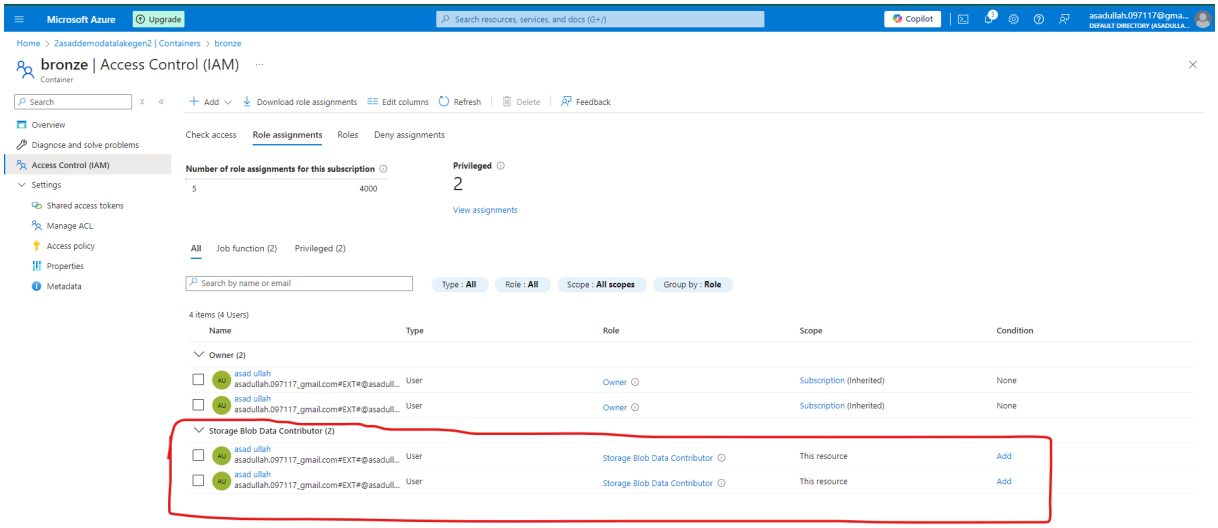
https://www.youtube.com/watch?v=5qRsrYQp0YU



8. Now first stage is completed, the data is loaded as is to the bronze container in Azure Data Lake gen2, so now we need to mount the Azure Storage to Azure DataBricks for that we will copy the config from here

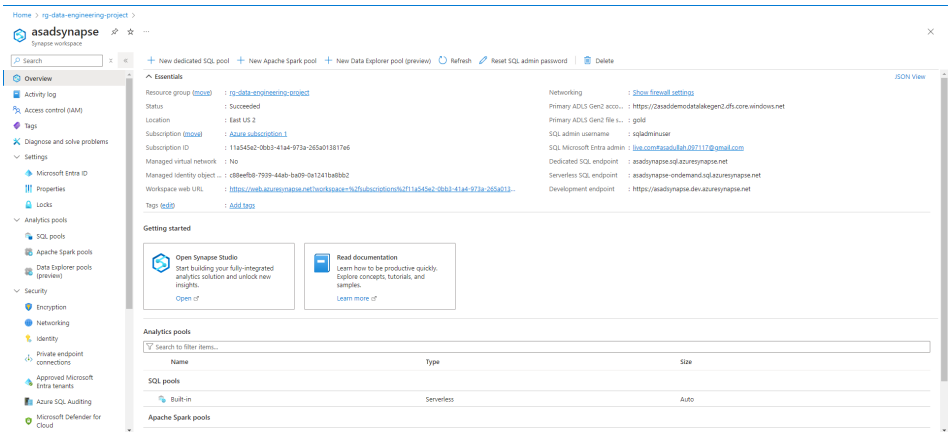


9. Although my user/id/email is the Owner of this blob but I still had to add a storage blob data contributor role to access this Data lake storage inside Databricks after mounting.



10. Need to do the same point 9 for Silver and Gold containers

11. create an azure synapse analytics workspace like this: select existing Azure Datalake gen2 account and a container. I have selected GOLD as default.
    a. Create SQL username and password as well



12. For this project we had selected serverless SQL database in Azure synapse, but here are the differences bw serverless and dedicated
    a.

    Here's a tabular comparison of **Dedicated SQL Pool** vs. **Serverless SQL Pool** in Azure Synapse Analytics:

Feature	Dedicated SQL Pool	Serverless SQL Pool
Data Storage	Data is stored internally in the SQL pool after loading.	Queries external data sources like Azure Data Lake directly.
Provisioning	Provisioned, requires manual allocation of resources (DWUs).	No provisioning, on-demand resource allocation.
Scaling	Manually scaled by adjusting Data Warehousing Units (DWUs).	Automatically scales based on query requirements.
Cost Model	Fixed cost based on provisioned compute and storage, even if idle.	Pay-per-query, charged based on the amount of data processed (per TB).
Use Case	Best for persistent, large-scale data warehouses.	Ideal for ad-hoc queries and exploratory analysis.
Data Types	Structured data (e.g., relational tables).	Can query structured, semi-structured, and unstructured data.
Performance	High performance for large, complex queries and heavy workloads.	Suitable for light to medium workloads, less optimized for heavy queries.
Pausing	Can be manually paused to stop compute costs.	No need to pause, only billed when querying.
Billing Model	Billed based on provisioned capacity (DWUs) and storage usage.	Billed only when queries are executed, based on data processed.
Management	Requires active management for scaling and resource optimization.	No active management required, auto-scales per query.
ETL/ELT Workloads	Requires loading data before querying (ETL/ELT processes).	No need to load data, queries external data directly.