A diagram of data processing

AI-generated content may be incorrect.

Data set

A screenshot of a computer

AI-generated content may be incorrect.

Fact tables: returns and sales

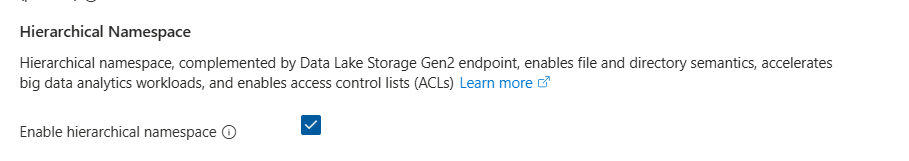
🡪 we have to build dimensions

Procedure: -

Create a resource group:

* Tags are used to categorized

Storage account: -



* This will create Data lake other wise BLOB storage account will be created
* In data lake we can create directories inside it but BLOB cannot provide

Azure Data Factory: -

-----> in the storage account create containers in the data source like as in medilion architecture. Bronze, silver and gold

A screenshot of a computer

AI-generated content may be incorrect.

* Launch studio in azure data factory
* Create pipeline
* A computer screen shot of a computer

  AI-generated content may be incorrect.
* Here sink is destination
* Link service is a connection, it needs to get data from github and push to data lake

A diagram of a data processing process

AI-generated content may be incorrect.

* Create linked services

A screenshot of a computer

AI-generated content may be incorrect.

* Create for http and data lake, click on source create a httplinked and sink create a datalake one give bronze path

A screenshot of a computer

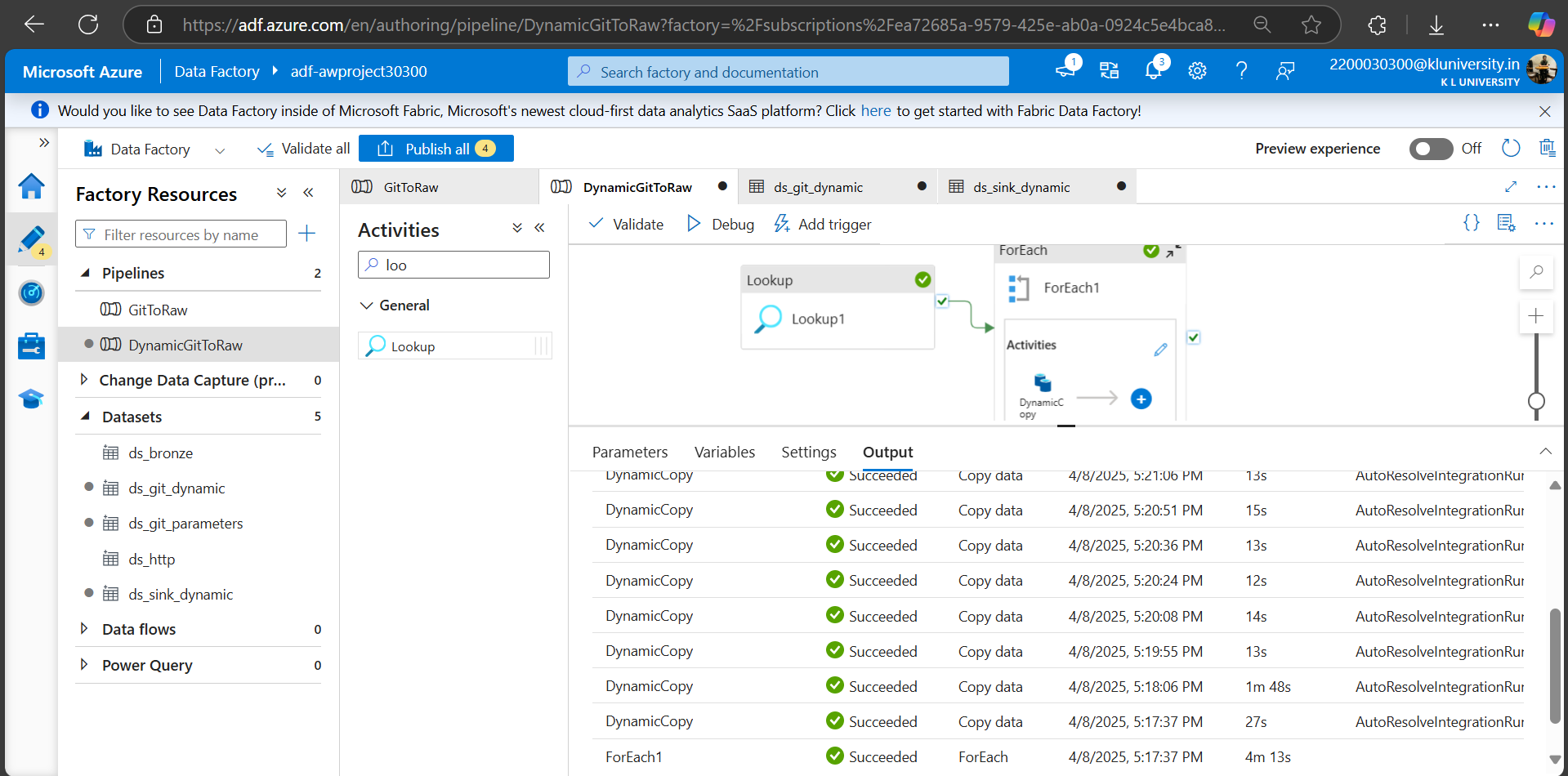
AI-generated content may be incorrect.

* Hectic process to create for each file, now we fetch data dynamically.

1. We have relative url file
2. Folder changes
3. File

* We will use for-each activity
* Create a json file
* A screenshot of a computer

  AI-generated content may be incorrect.
* Add all the relative urls
* Just create another container in data lake and add this json file in it
* Go to data factory add a lookup in dynamic pipeline and upload this file in it
* In the same pipeline we have a for each loop just drag the COPYGITDATA item.
* Go to activity of for each just add the parameters in source and sink tabs
* Boom! Just debug it mann!



PHASE -2

🡺now we have to transform the data using databricks

🡪To deal with big data data bricks is used

🡪it is used for transformation layer

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Create cluster

🡪it is auto managed we don’t need to manage

🡪 the data bricks needs key to access data lake, can be created using ENTRA ID

A black and white diagram of a key

AI-generated content may be incorrect.

* Open entra id register an app and copy credential
* Open data lake then go to IAM just get secret key id
* Now open data bricks, start the compute engine (virtual machine)

PYSPARK

* In the workspace just add this below code and give the credentials which copied in previous steps  
  spark.conf.set("fs.azure.account.auth.type.<storage-account>.dfs.core.windows.net", "OAuth")  
  spark.conf.set("fs.azure.account.oauth.provider.type.<storage-account>.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")  
  spark.conf.set("fs.azure.account.oauth2.client.id.<storage-account>.dfs.core.windows.net", "<application-id>")  
  spark.conf.set("fs.azure.account.oauth2.client.secret.<storage-account>.dfs.core.windows.net", service\_credential)  
  spark.conf.set("fs.azure.account.oauth2.client.endpoint.<storage-account>.dfs.core.windows.net", "https://login.microsoftonline.com/<directory-id>/oauth2/token")
* df = spark.read.format("csv").option("header", "true").option("inferSchema",True).load('abfss://bronze-container@awstorage30300.dfs.core.windows.net/AdventureWorks\_Calendar')
* inferschema: default the data in csv it takes as text data, we are telling spark to analyze the data on its own to decide it.
* Load:
  + abfss : azure blob storage

Adventure Calender

* df\_cal = df\_cal.withColumn('Month',month(col('Date')))\
* .withColumn('Year',year(col('Date')))
* df\_cal.display()
* transformation to columnar data

modes:

* append: to apply union (merge data)
* overwrite: replace existing data
* error: write data that already existed
* ignore: if data is there in folder it does not thrown any error, it not write data

df\_cal.write.format('parquet')\

    .mode('append')\

    .option("path","abfss://silver-container@awstorage30300.dfs.core.windows.net/AdventureWorks\_Calendar")\

.save()

AdventureWorks – Customers

df\_cus = df\_cus.withColumn('fullName',concat\_ws(' ',col('Prefix'),col('FirstName'),col('LastName')))

Products table:

Split: - to fetch desired index

Sales Table

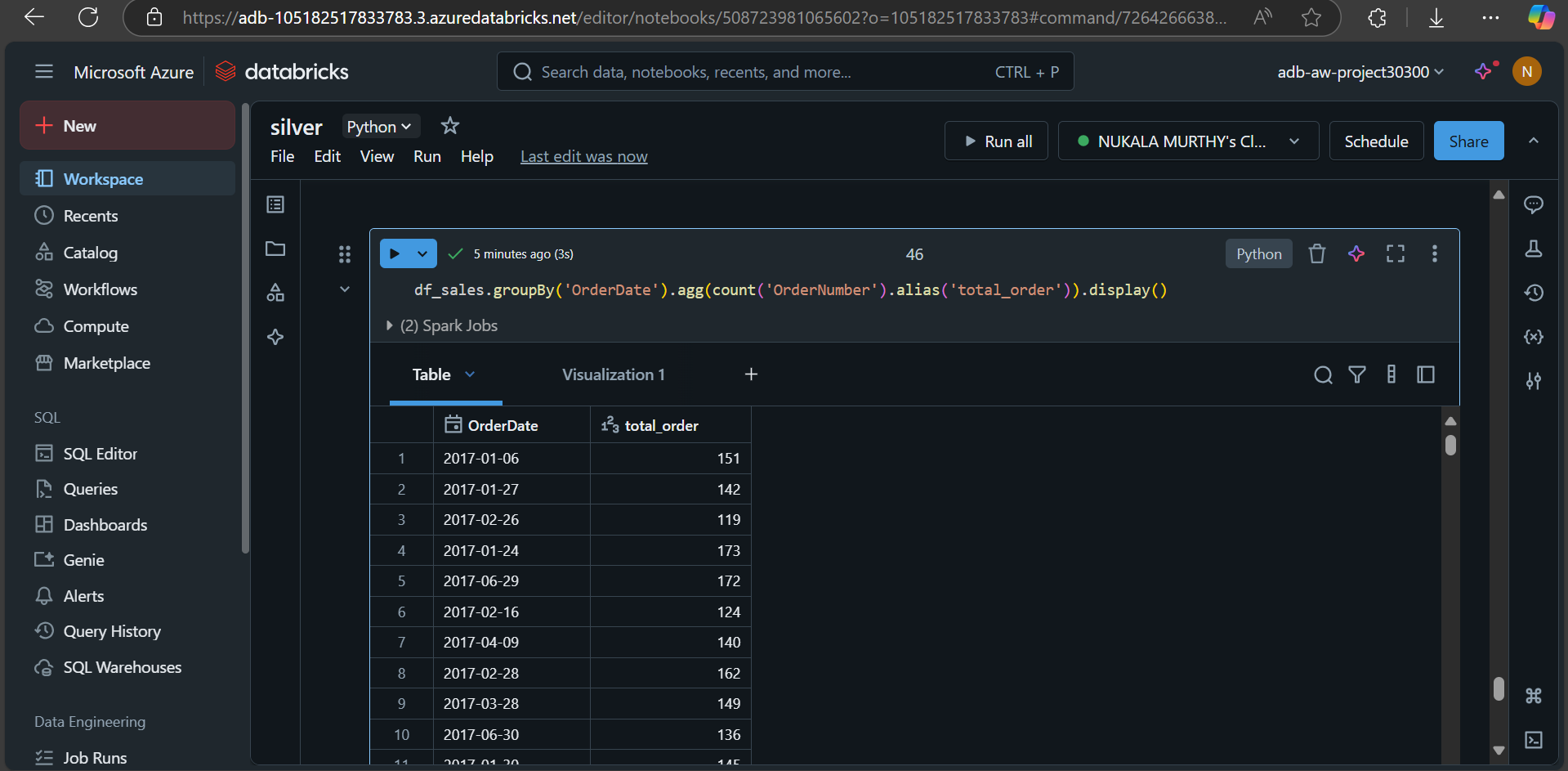
df\_sales = df\_sales.withColumn('StockDate',to\_timestamp('StockDate'))

df\_sales = df\_sales.withColumn('OrderNumber',regexp\_replace(col('OrderNumber'),'S','T'))

df\_sales = df\_sales.withColumn('multiply',col('OrderLineItem')\*col('OrderQuantity'))

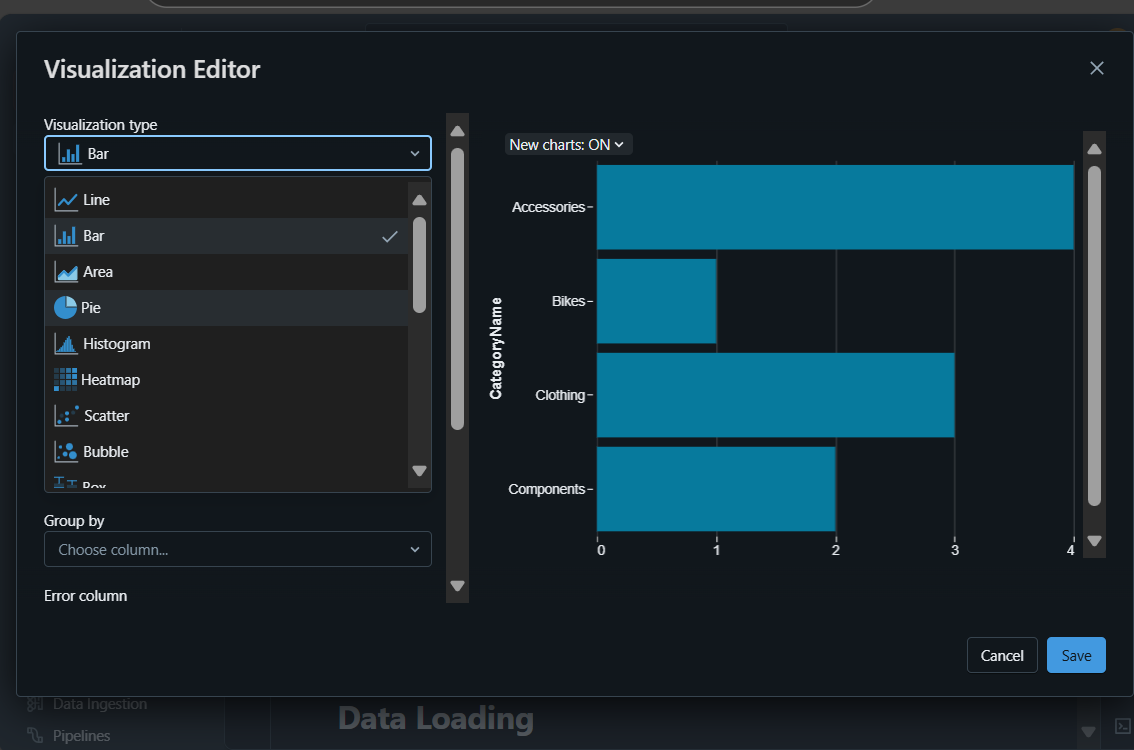
BIG DATA ANALYTICS USING PY SPARK

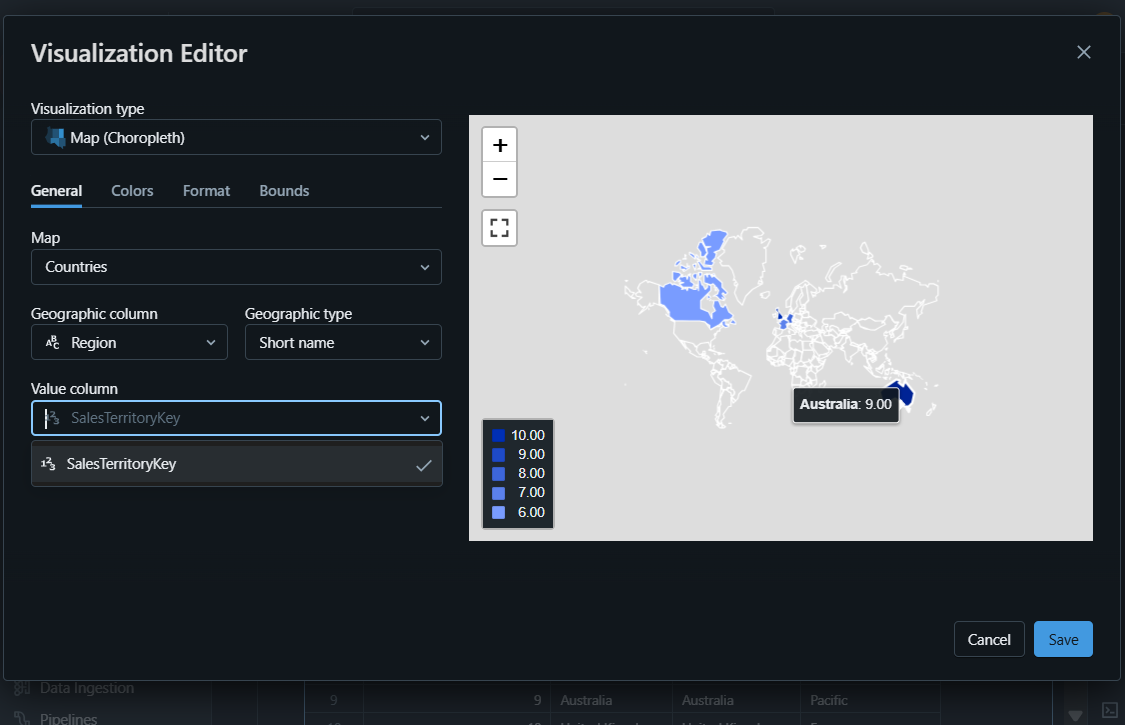
df\_sales.groupBy('OrderDate').agg(count('OrderNumber').alias('total\_order')).display()



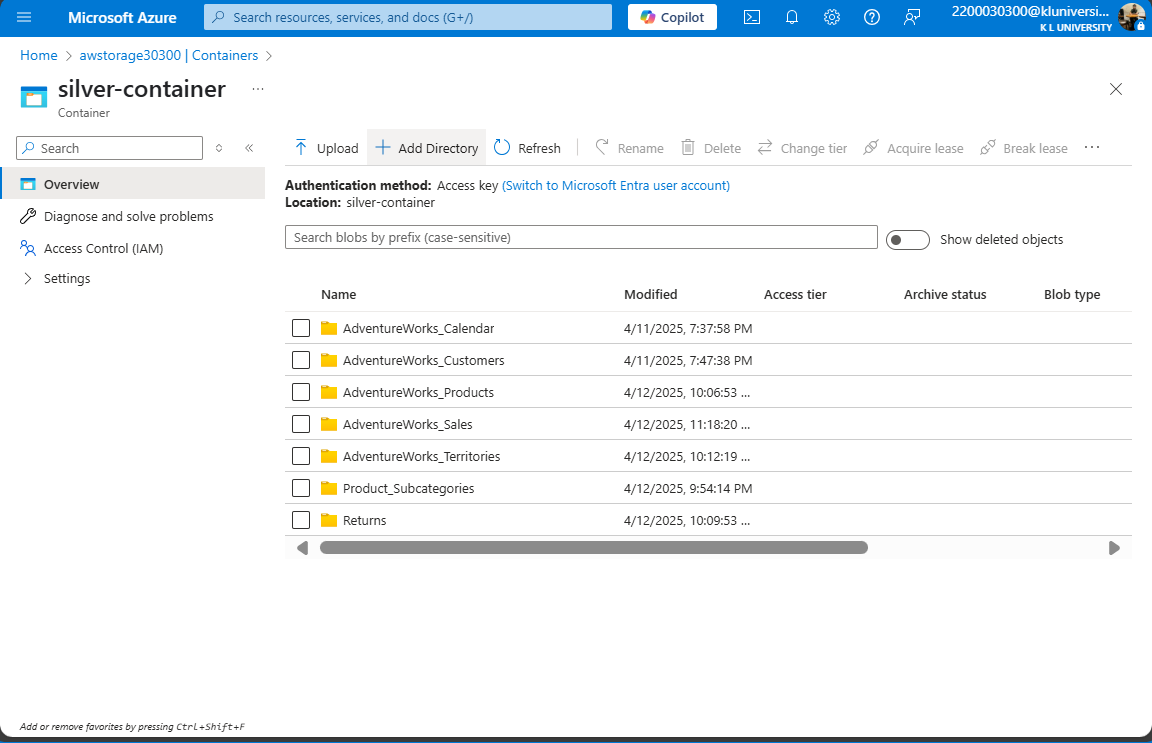
A screenshot of a computer

AI-generated content may be incorrect.





Done



PHASE 3- GOLD LAYER

🡪synapse analytics is a unifed platform in which it combines

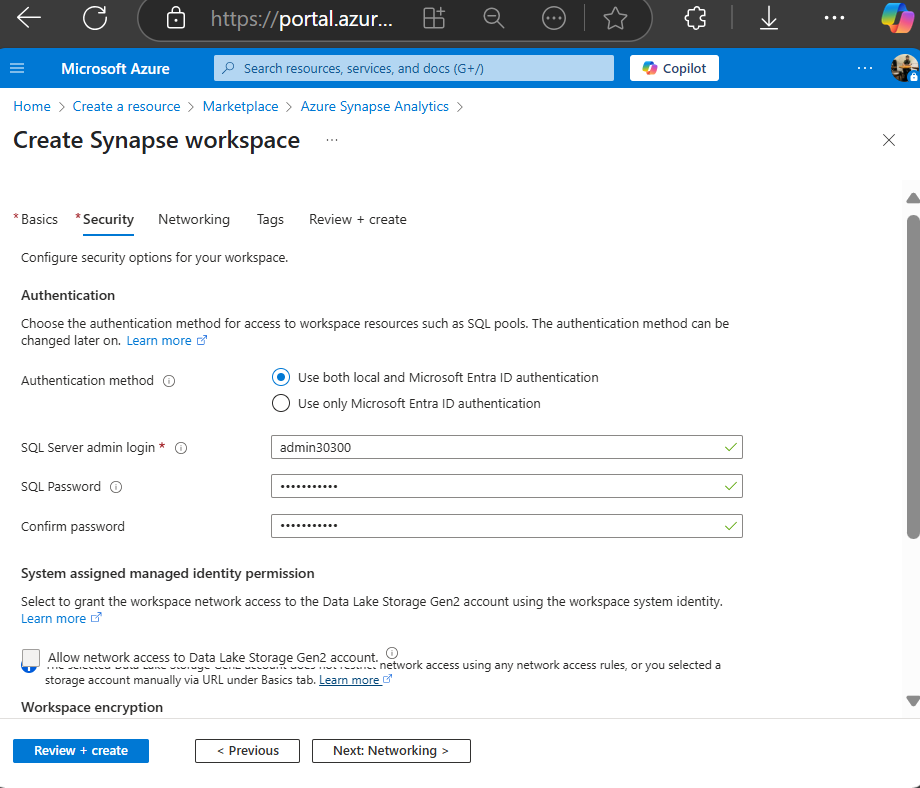
Azure data factory + Spark (Big Data Analytics) + Warehousing

🡪Synapse needs permission to access data Lake, MANAGED IDENTITY or SYSTEM MANAGED IDENTITY

A screenshot of a computer

AI-generated content may be incorrect.

Don’t use the same data lake that has used in previous layers.



Note: - All regions are not working (US East, south india), Central US is working fine

🡪To give permission for synapse to access Data lake

* go to data lake then click on IAM add a role assignment of managed identity

pass: mypassword

A computer screen shot of a computer

AI-generated content may be incorrect.

-- come to synapse

Dedicated: it is a traditional way to store data, in which the data actually stored in mysql and postgresql

Serverless: no need to manage the server

Data resides in the data lake, in csv format

LAKE HOUSE, gives privilege to apply all database concepts to data lake  
A screenshot of a computer

AI-generated content may be incorrect.

A diagram of data flow

AI-generated content may be incorrect.

🡪 To Pull Data From the silver layer we can use openrowset()

A screenshot of a computer

AI-generated content may be incorrect.

Get the link from data lake ie, go to the file and copy link then change ‘blob’ to dfs

IAM access role assignment: Storage Blob Data Contributor

Allows for read, write and delete access to Azure Storage blob containers and dat

🡺 to create external tables in synapse, these steps are requires

Create credential

External data source

External file format