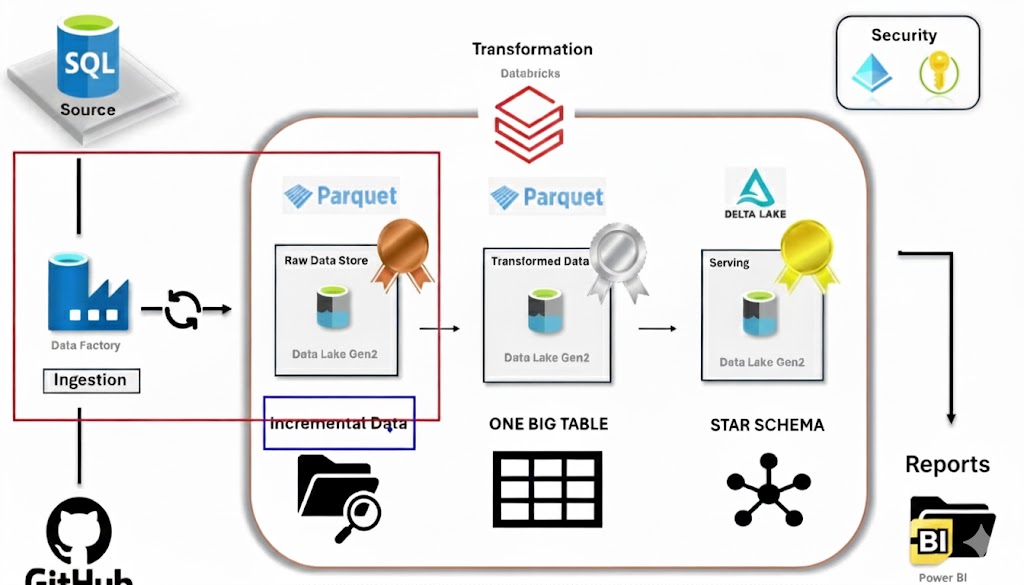


**End-to-End Azure Data Engineering Pipeline for Sales Data**



**Name: SALAPALA LOKESH REDDY**

**Agenda**

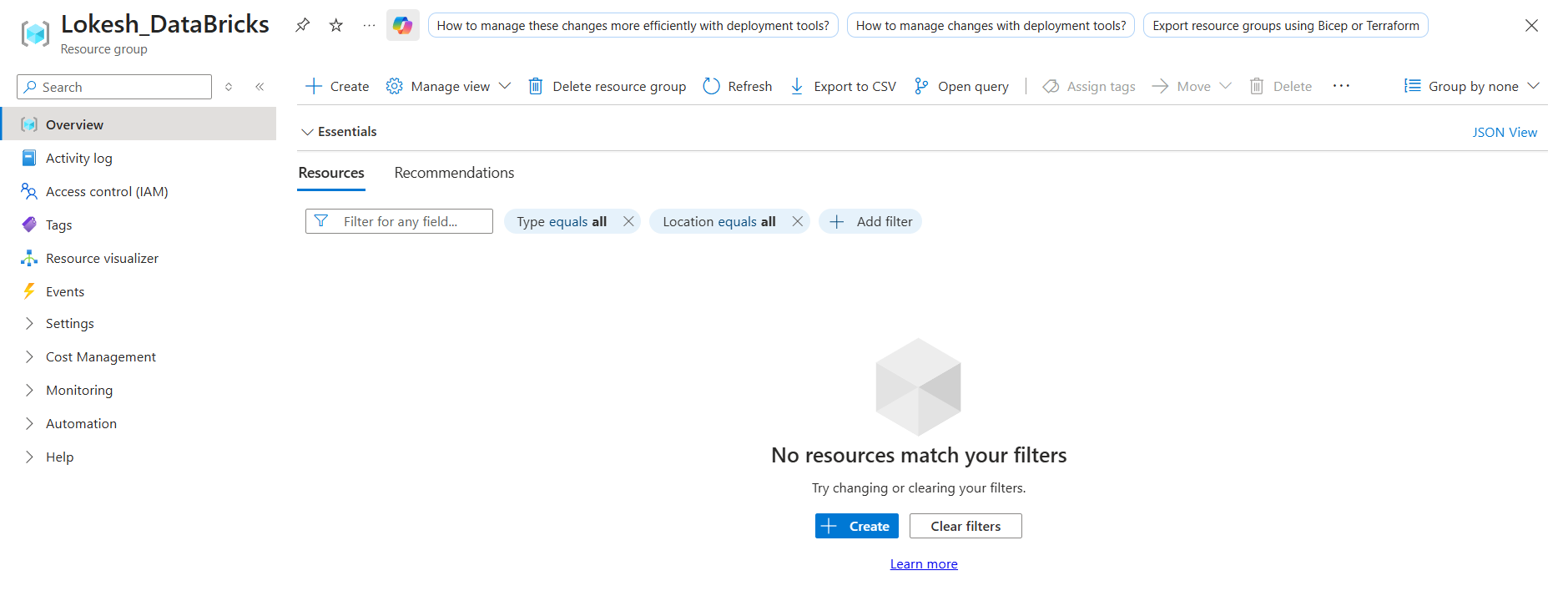
****

**Step 1: Collect the Data source for initial loading**



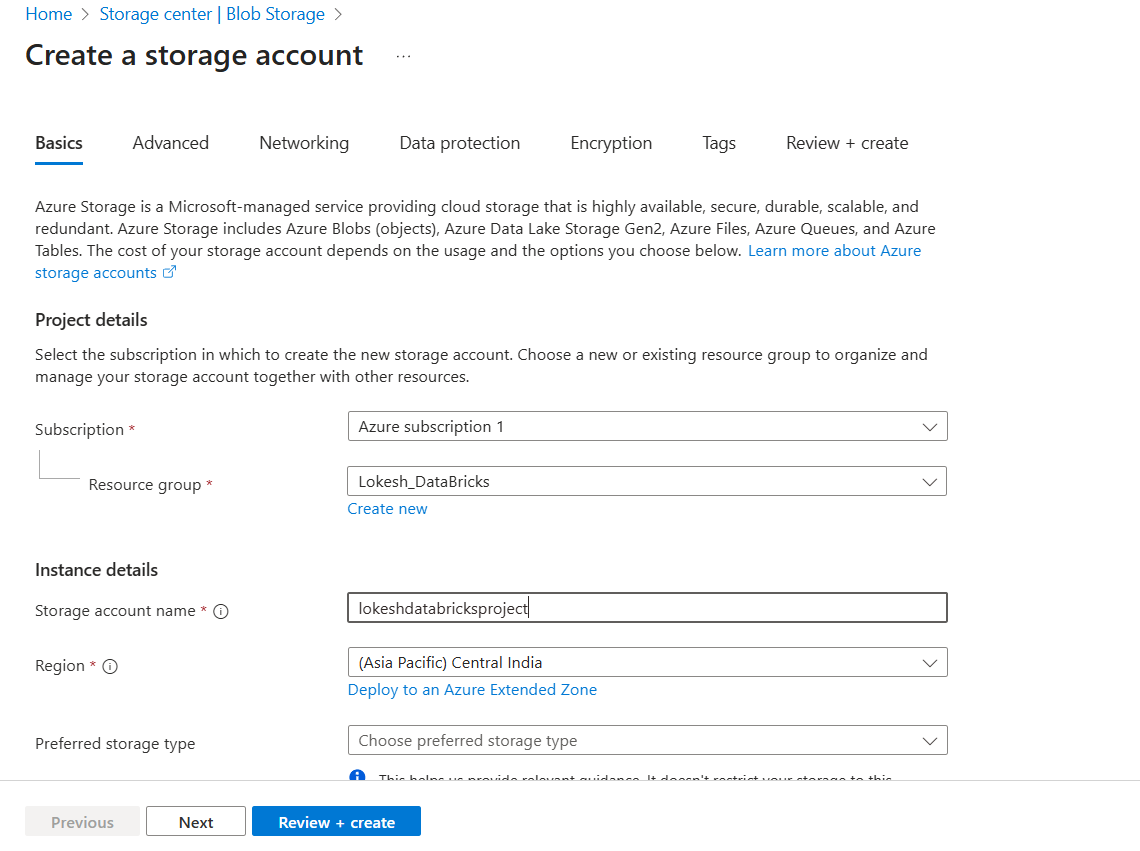
**Step 2: Create the Azure account**

**Step 3: Create the Resource Group**

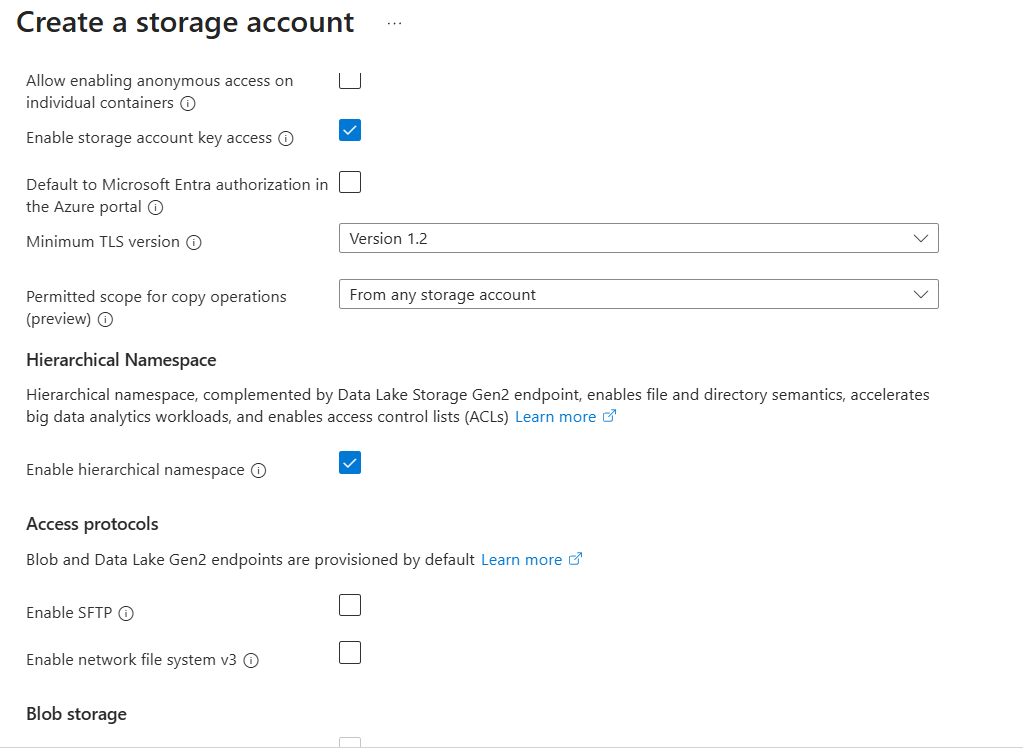
****

**Step 4: Create the Azure Data Lake**

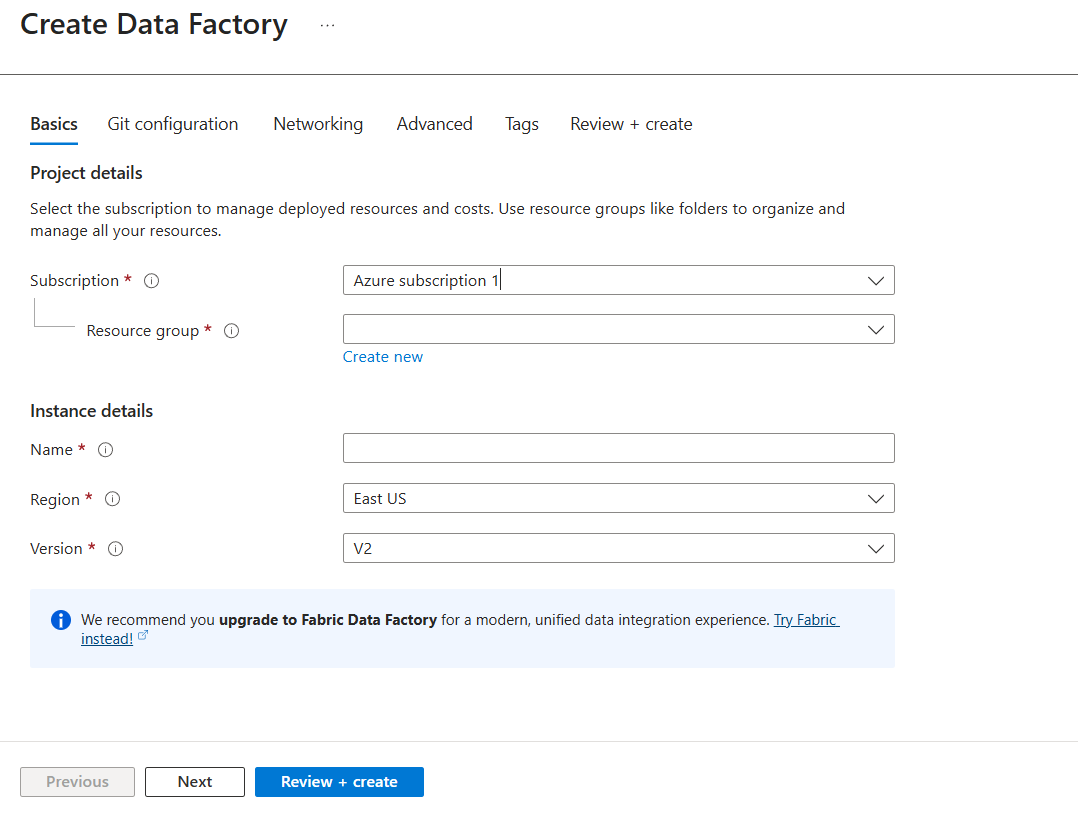
**Go to storage accounts 🡪 Provide the Resource group name, Storage account name which is unique**

****

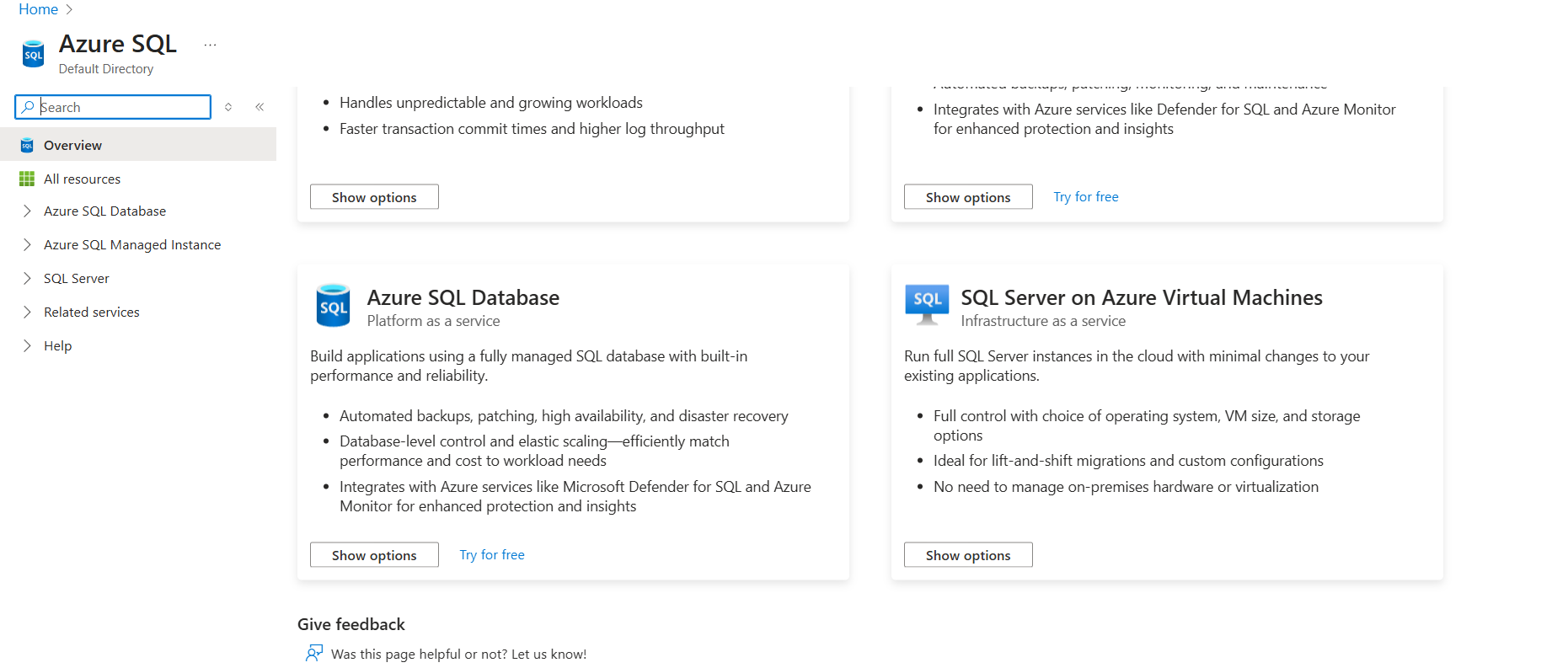
**Select the redundancy as LRS 🡪 Click next🡪 Click check box to enable hierarchical namespace so it creates the datalake**

****

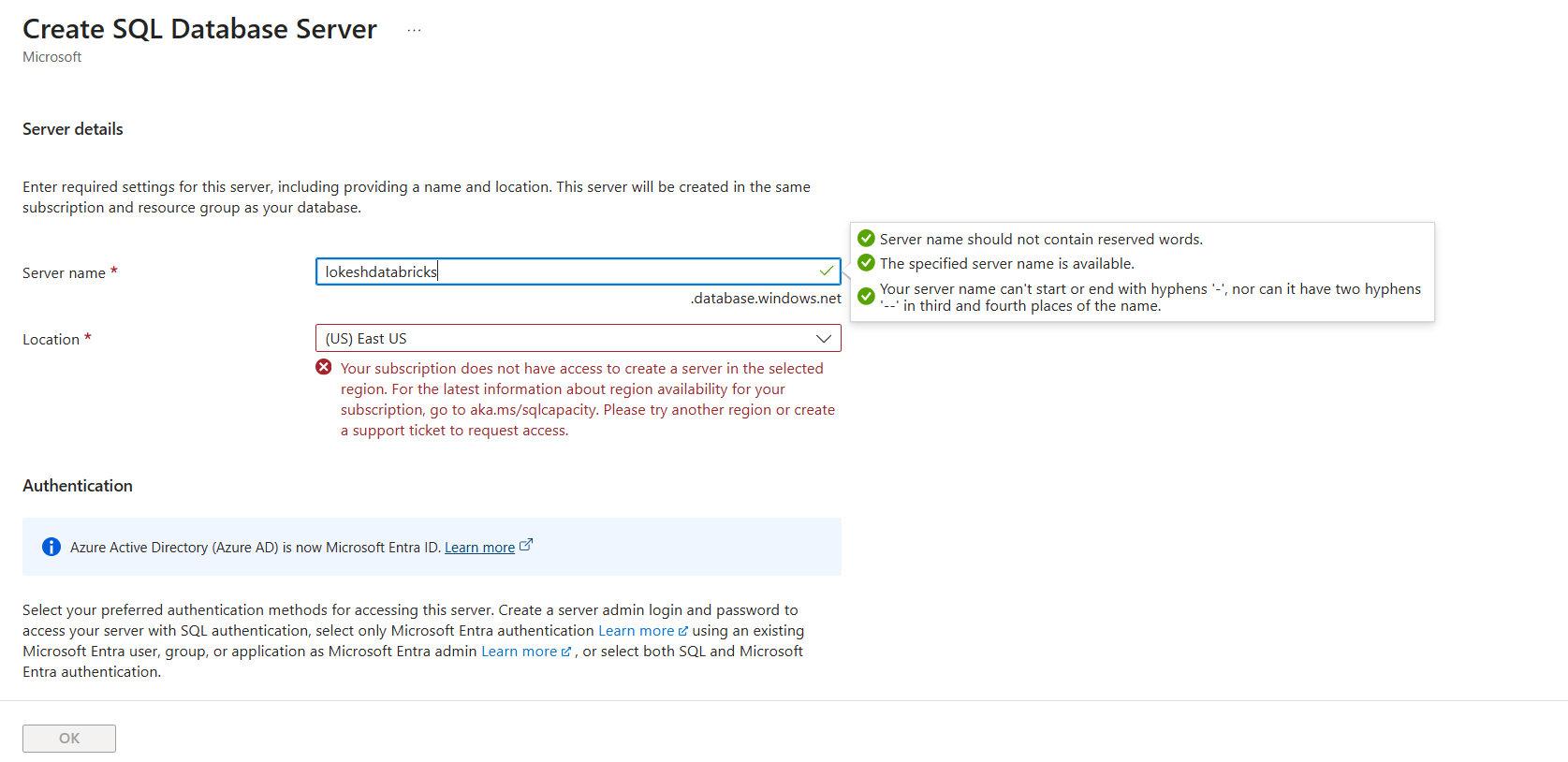
**Step 5: Create the Azure Data Factory 🡪 Provide RG name and Unique name for Data factory**

****

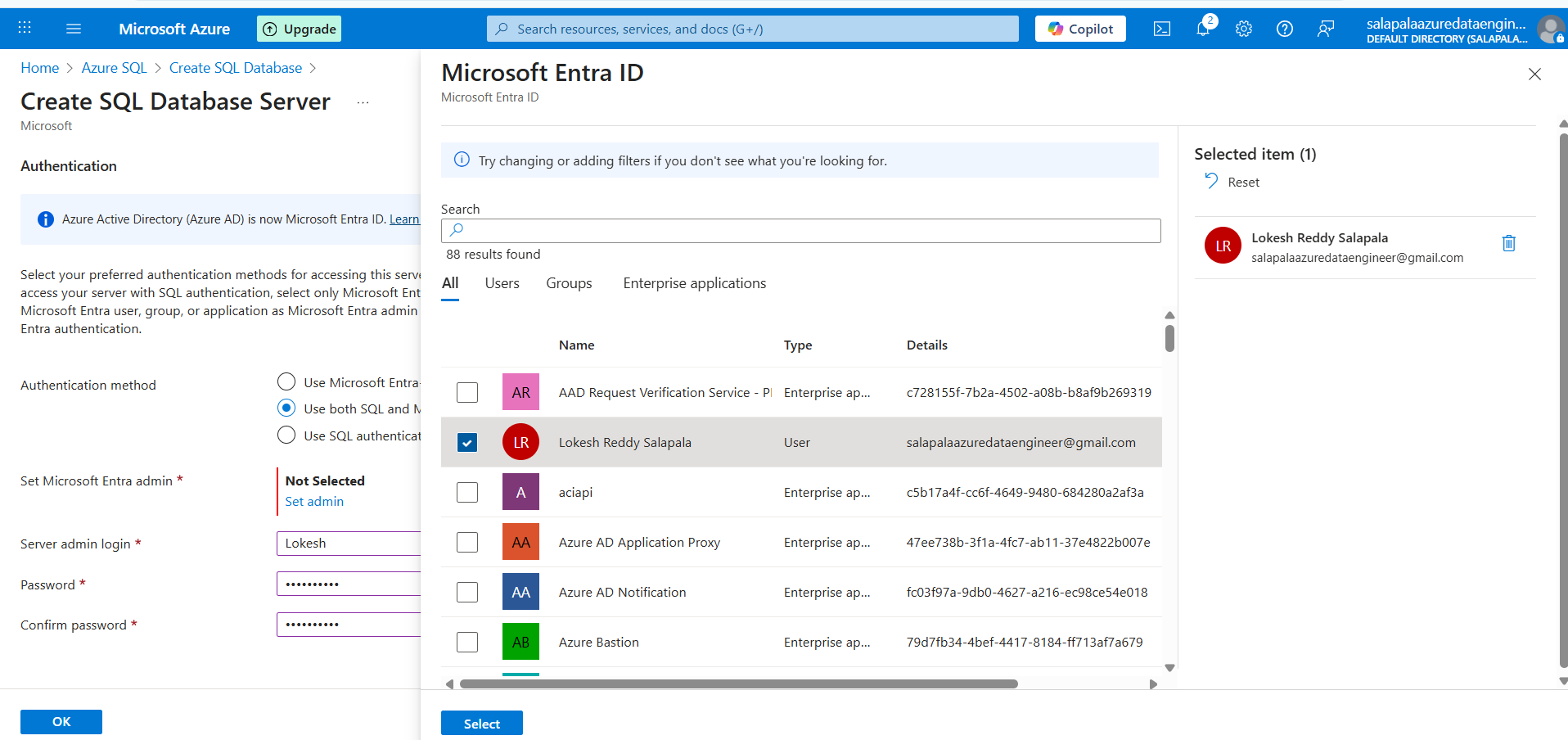
**Step 6: Create Azure SQL**

****

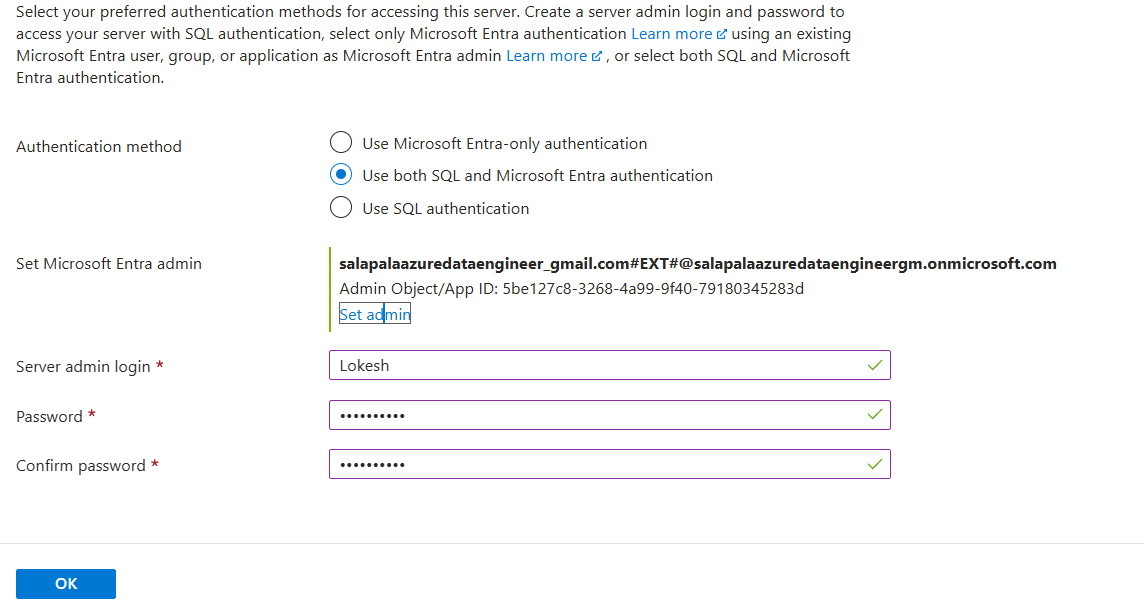
**Select the Azure SQL Database 🡪 for free trail account East US location is not supported we need to check the supported location**

****

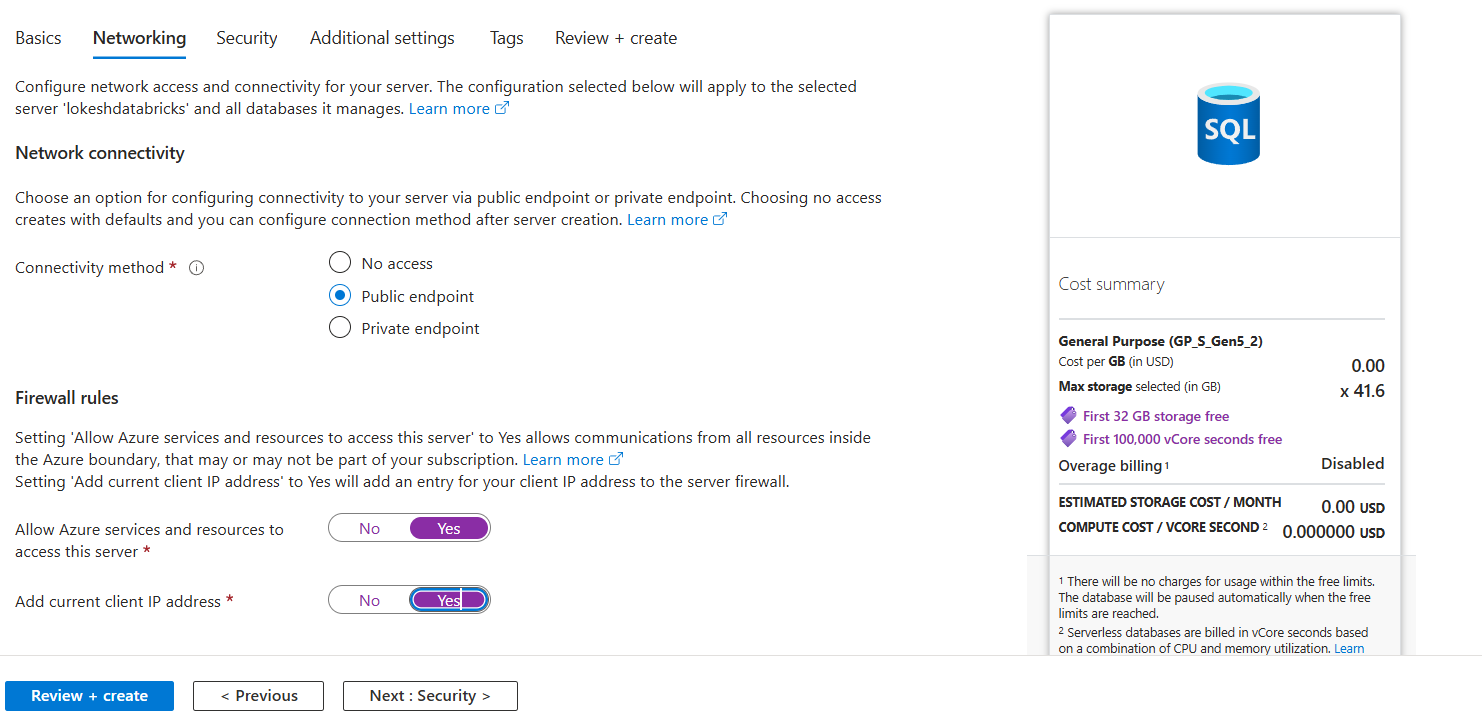
**In set Admin 🡪 select the both SQL and Microsoft Entra ID 🡪 select the User as our ID**

****

**Provide the username and password to login the SQL Database**

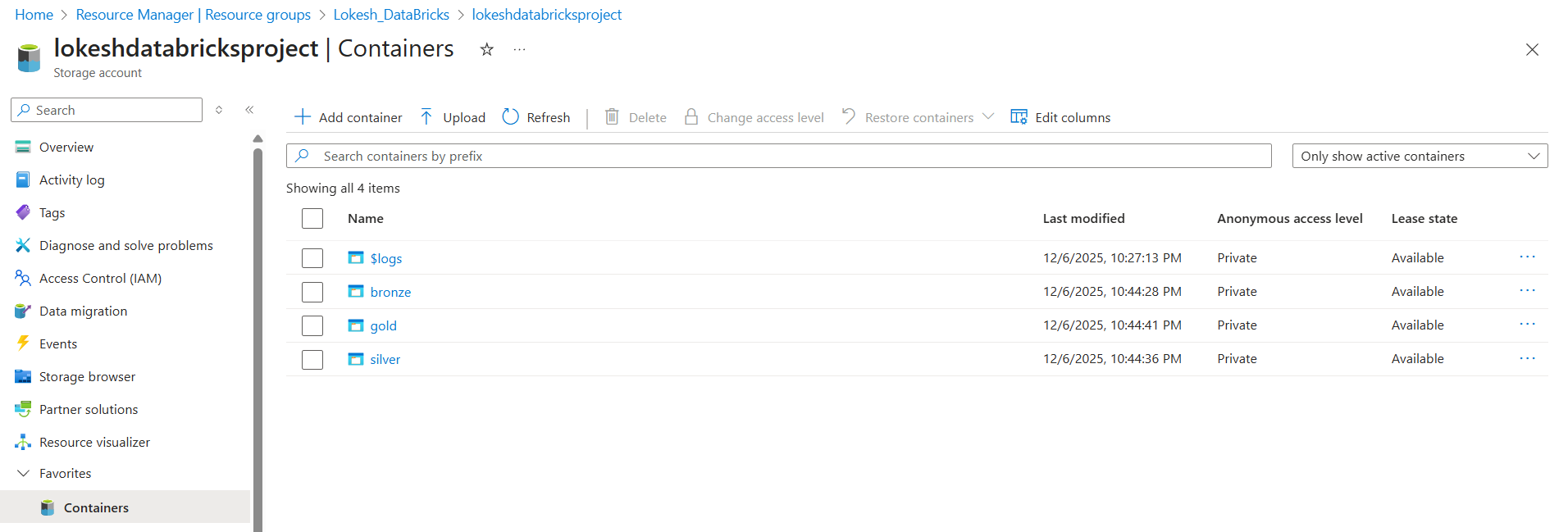
****

**Select the public endpoint and enable the firewall rules**

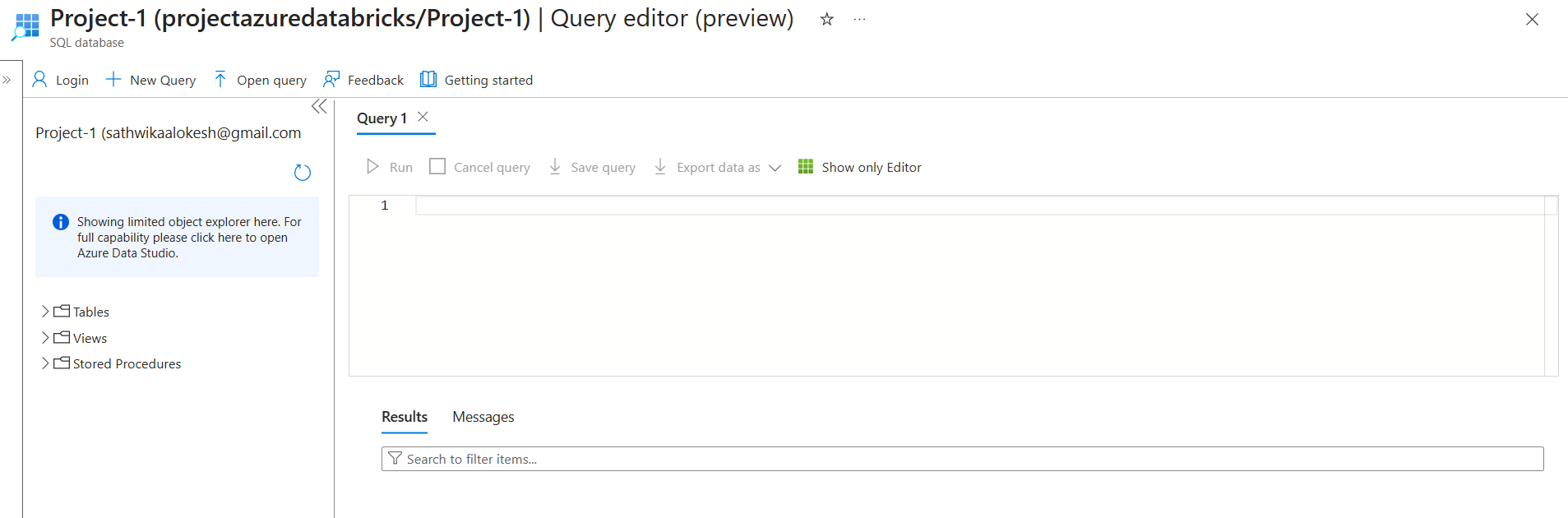
****

**Click on review and Create**

**Step 7: Go to Azure Date Lake 🡪 Create the containers as below, create the**

****

**Step 8: Login to the SQL Database using Querry Editor**

****

**Step 9: Create the Table to store the data in rows and columns**

**CREATE TABLE source\_cars\_data**

**(**

**Branch\_ID Varchar(2000),**

**Dealer\_ID Varchar(2000),**

**Model\_ID Varchar(200),**

**Revenue BIGINT,**

**Units\_Sold BIGINT,**

**Date\_ID Varchar(200),**

**Day INT,**

**Month INT,**

**Year INT,**

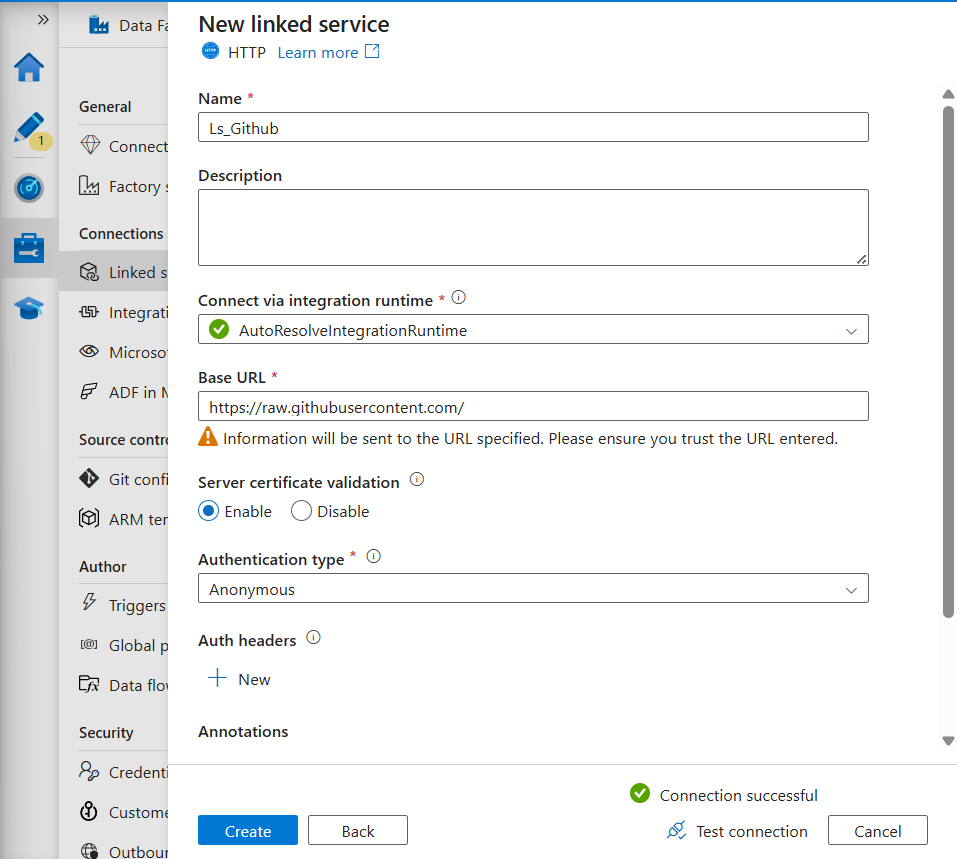
**BranchName Varchar(2000),**

**DealerName Varchar(2000)**

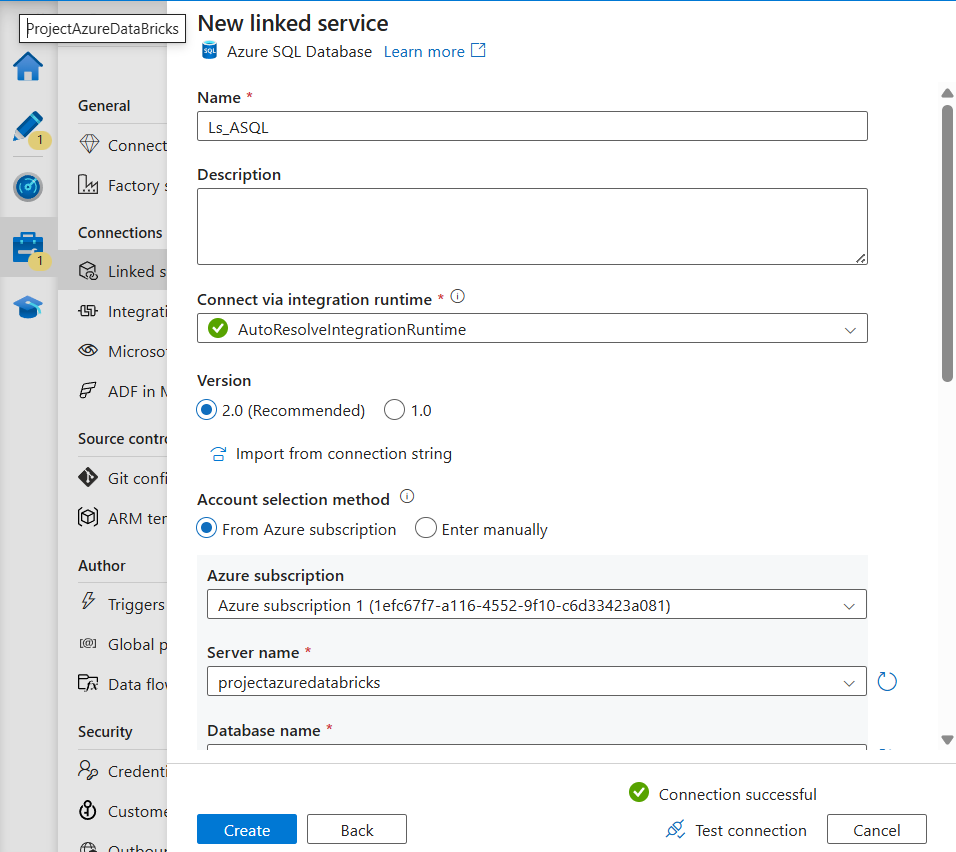
**)**

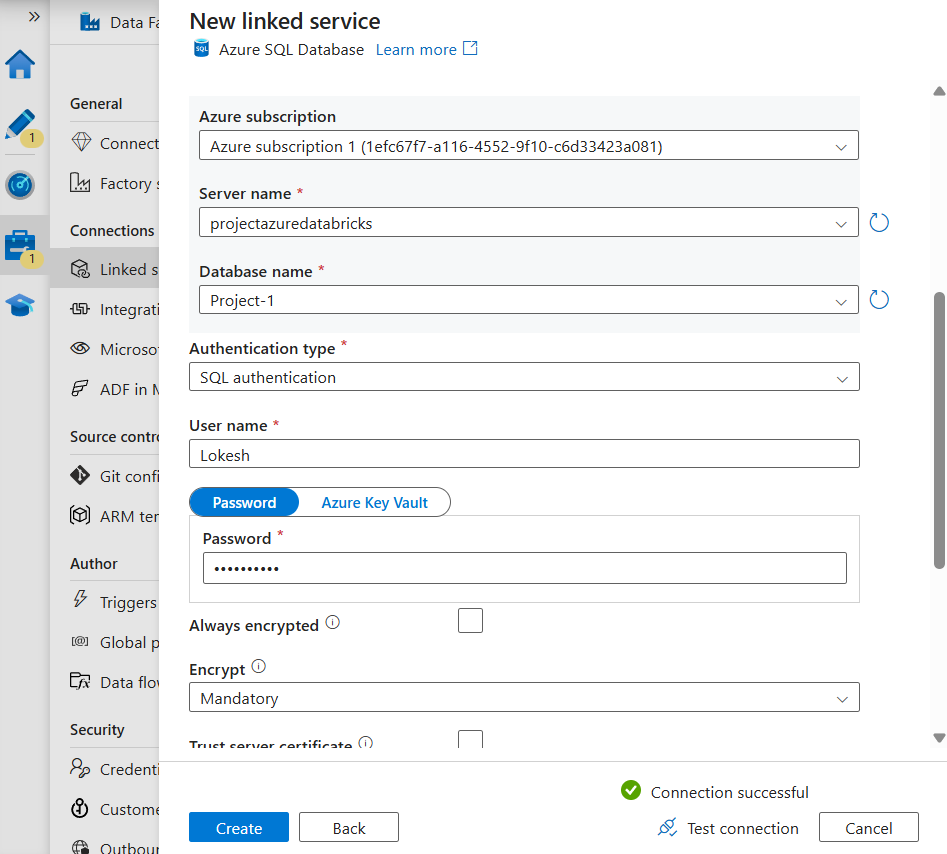
**Step 10: Go to Azure Data factory 🡪 Click on crate new pipeline🡪 Select the Copy data activity**

**then Click on Manage 🡪 Create Linked Service with Http 🡪 Go to GitHub and open the raw data source file and click on the base URL and paste as below**

****

**Step 11: Create new linked service with Azure SQL database 🡪 Select the subscription, server name, database name, Authentication type as SQL and provide username and password and test the connection**

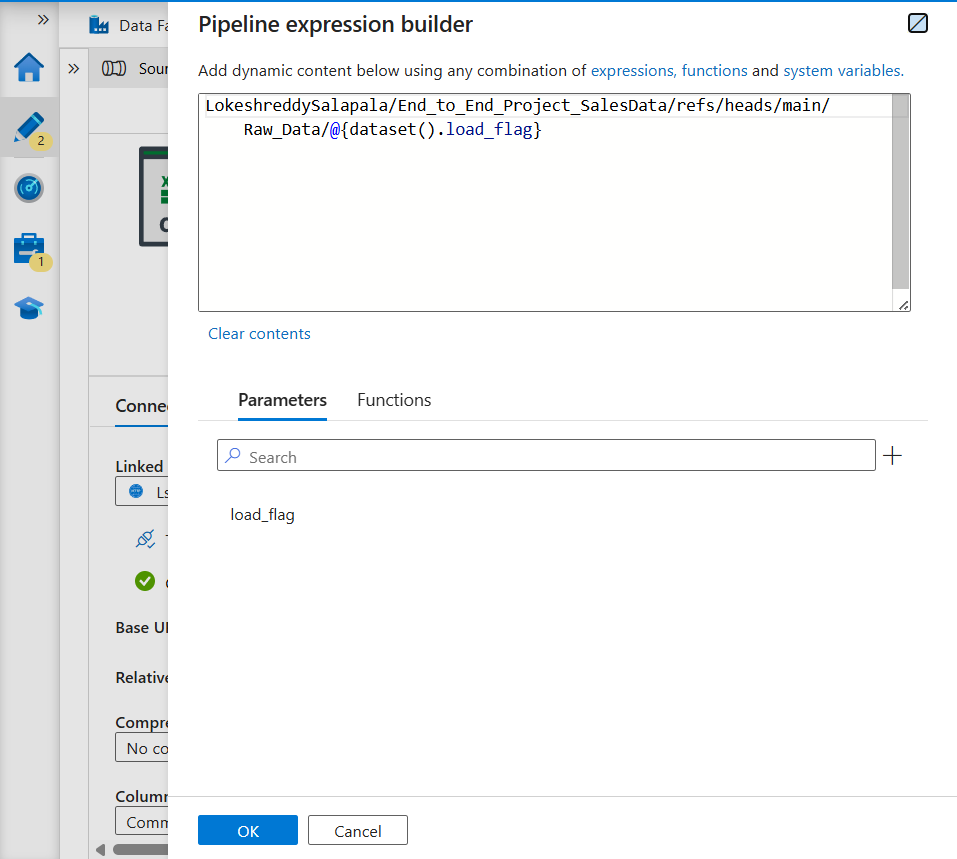
****

****

**Step 12: Goto pipeline 🡪 Source🡪create new Dataset with HTTP and dataset set type as Delimited text🡪 Select the Linked service and provide the raw data url**

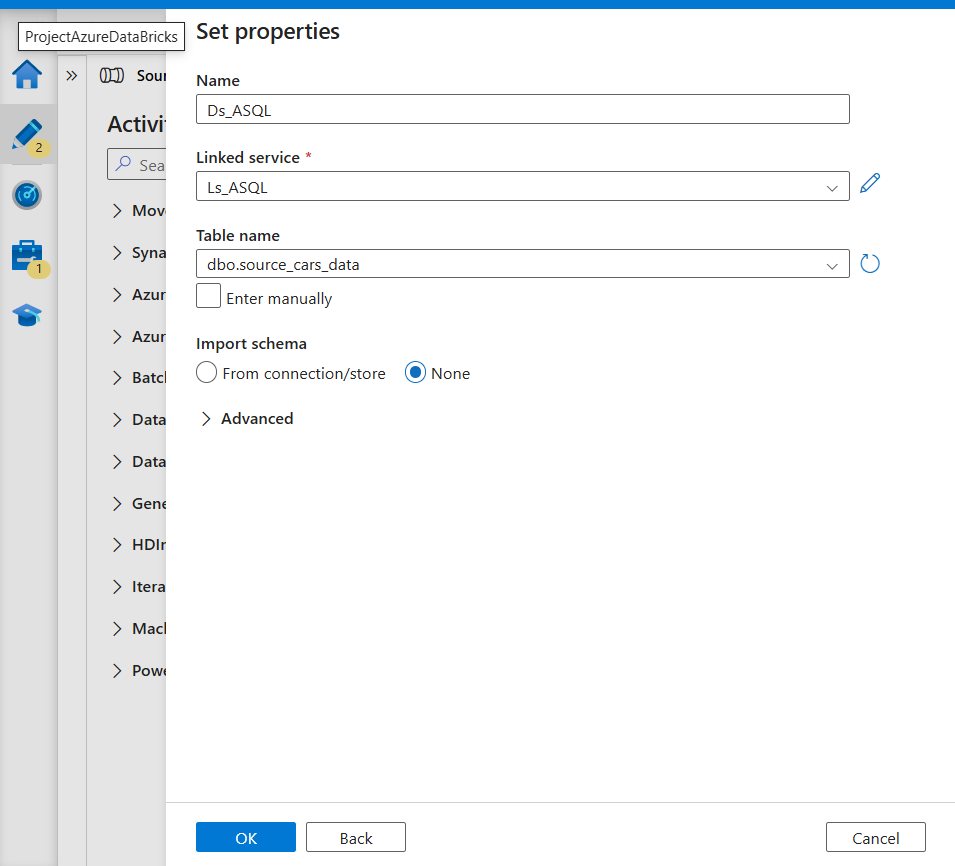
****

**Click on open dataset and click on parameter 🡪 give as load\_flag 🡪 Click on connection 🡪 add dynamic content 🡪 provide the path of GitHub url as below and click on ok**

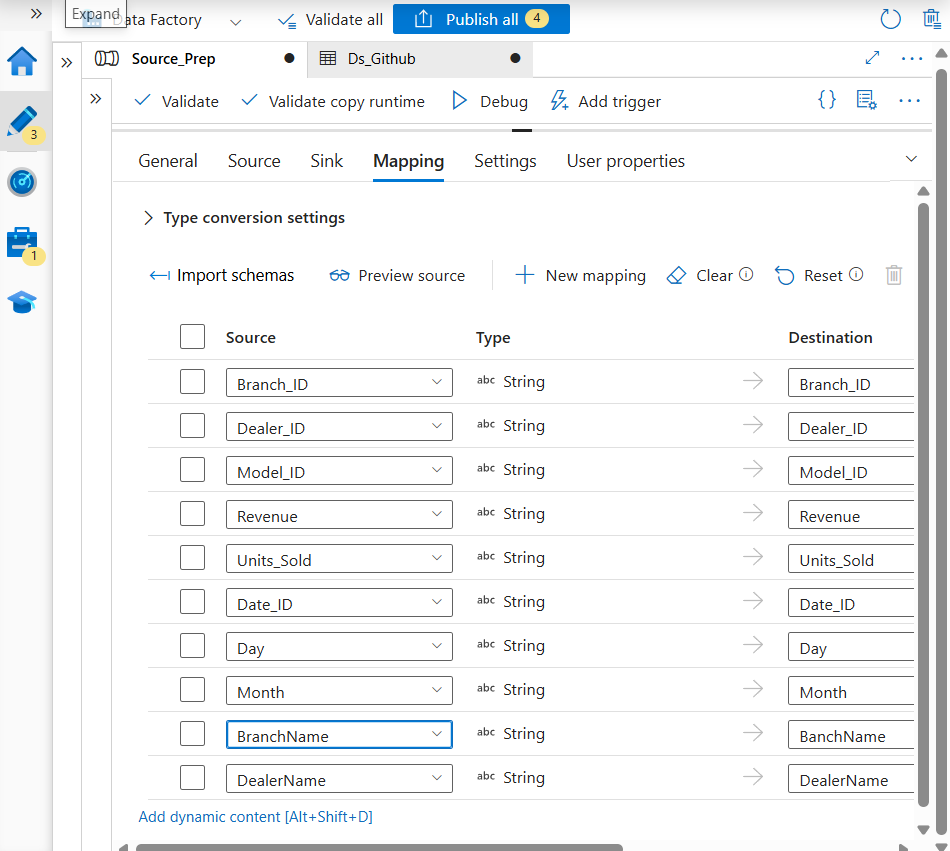
****

**Go to pipeline and preview the data**

**Step 13: Click on sink in pipeline 🡪 Create new dataset with Azure SQL Database 🡪 Click on enter manually and provide schema only and select None**

****

**Step 14: Select the mapping and import schemas 🡪 Map the correct src and destinations**

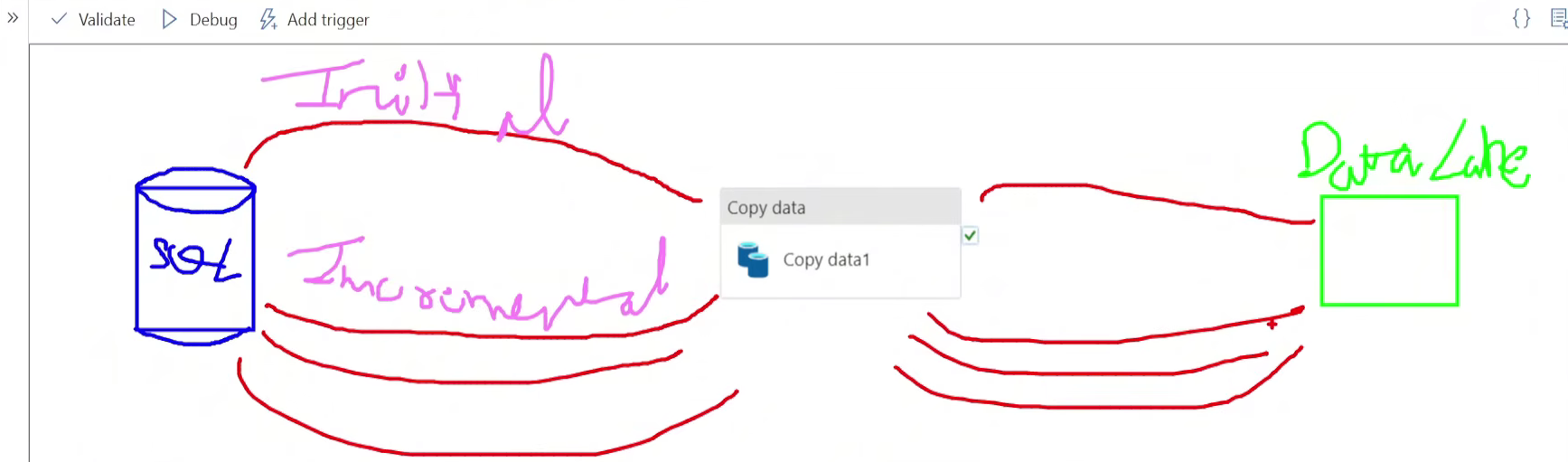
****

**Click on Debug 🡪 check the status successful or not**

**to validate 🡪go to azure SQL and query the select \* from < table name >**

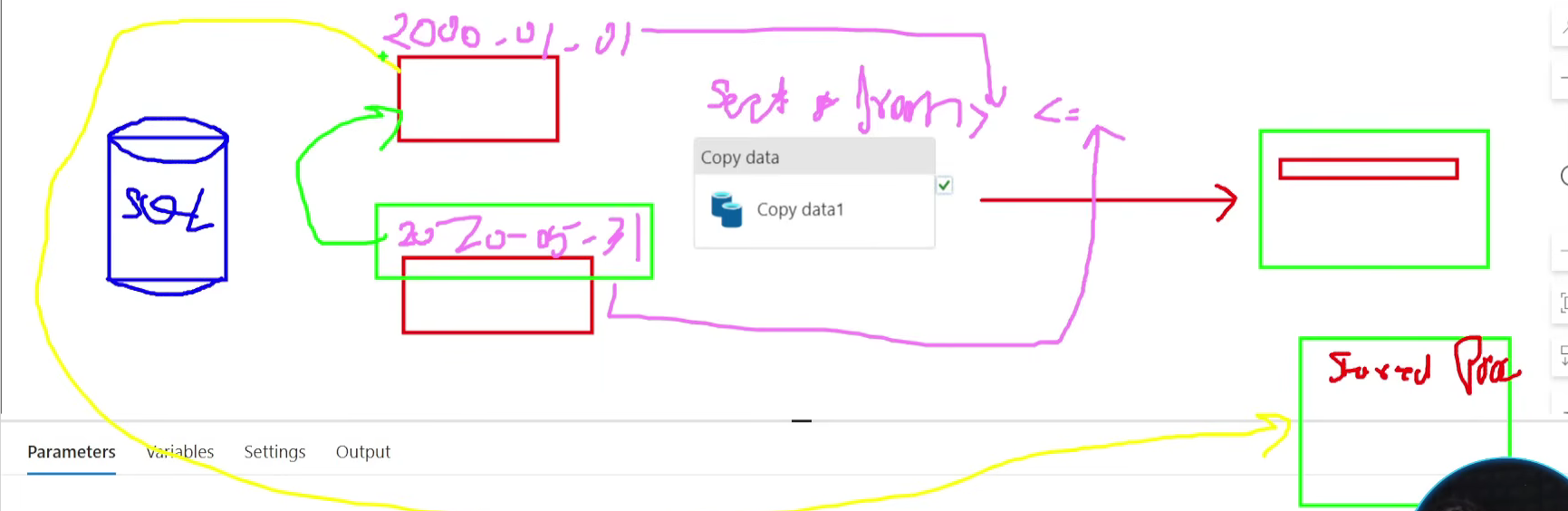
**Till here we have completed the collecting the raw\_data from GitHub and copied to the ASQL into the table we created**

**We need two pipelines one will be initial loading pipeline and the next pipeline will be incremental load pipeline**

****

**Flow**

**Here the last date for initial pipeline will be 2000-01-01 and the SQL will be select \* from date > initial date and date <= present current date and the next incremental load will be present date. Here the initial date we get it from stored procedure for initial date**

****

**Step 15: Create the watermark table which stores the date value**

**create table water\_table**

**(**

**last\_load Varchar(2000)**

**)**

**select \* from water\_table**

**select min (DATE\_ID ) from [dbo].[source\_cars\_data]**

**o/p 🡪 'DT00001'**

**So here we are inserting the values -1 in watermark table**

**insert into water\_table values ('DT00000')**

**select count(\*) from [dbo].[source\_cars\_data] where DATE\_ID > 'DT00000'**

**Step 16: Create the stored procedure**

**CREATE PROCEDURE UpdateWatermarkTable**

**@lastload Varchar(200)**

**AS**

**BEGIN**

**-- Start the transaction**

**BEGIN TRANSACTION;**

**-- Update the incremental column in the table**

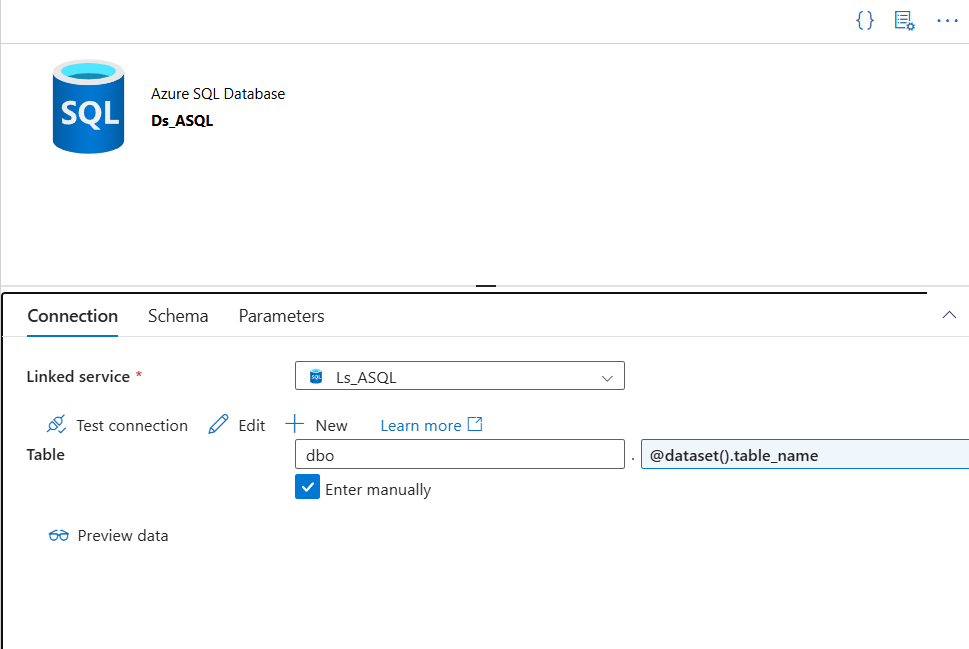
**UPDATE water\_table**

**SET  last\_load = @lastload**

**COMMIT TRANSACTION**

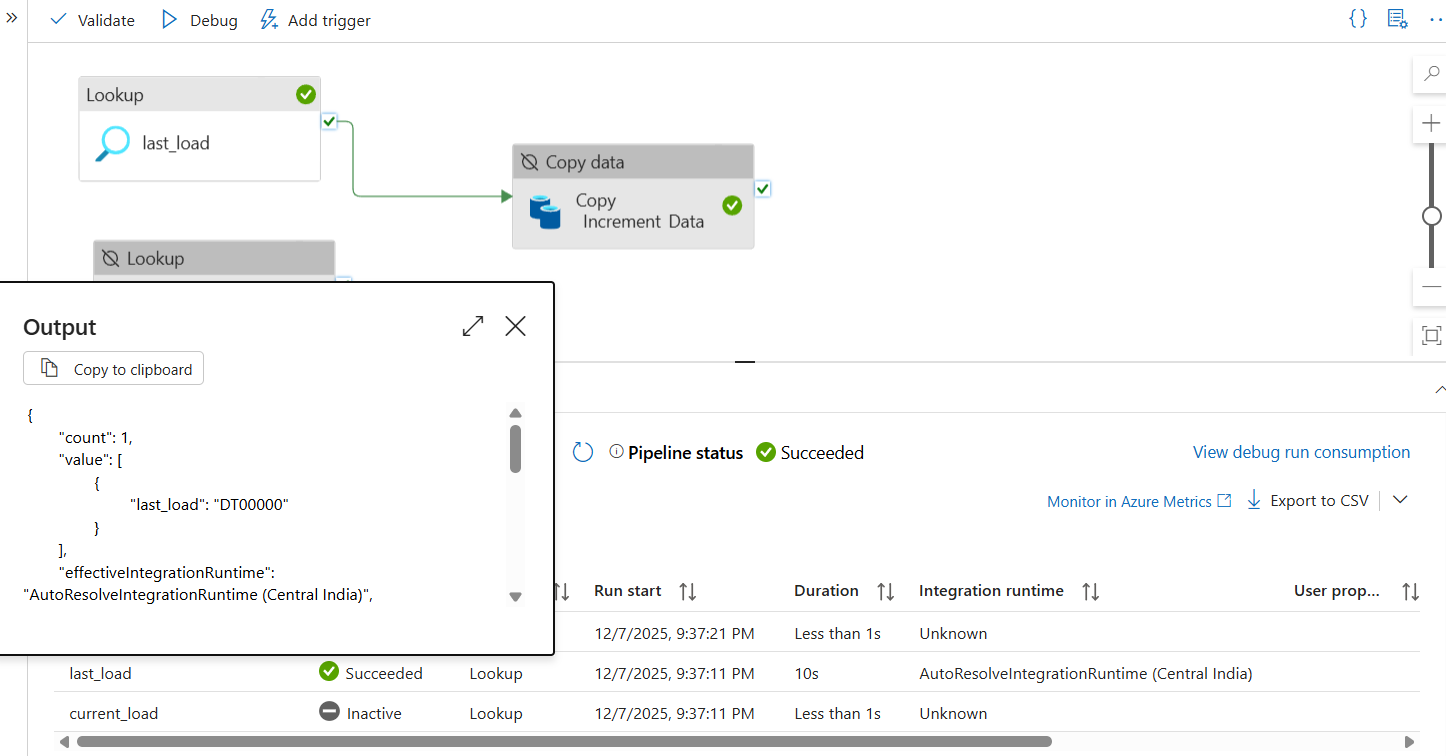
**END;**

**Step 17: Create new pipeline as incremental\_data\_pipeline 🡪 Drag the 2 lookup activity**

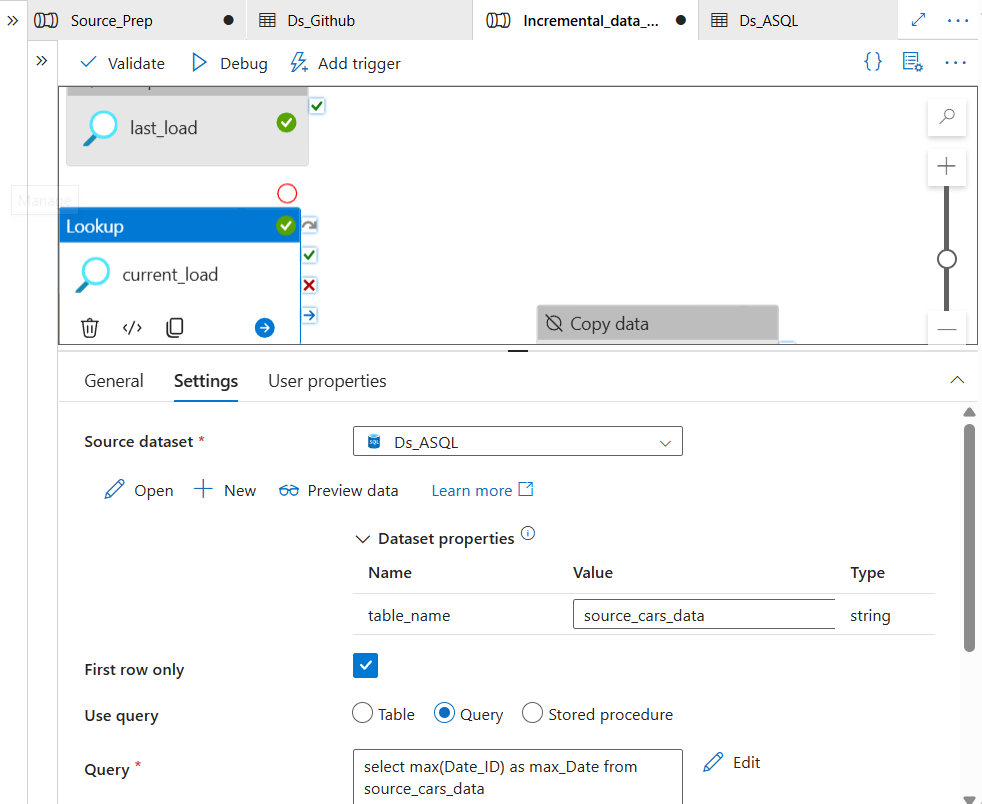
****

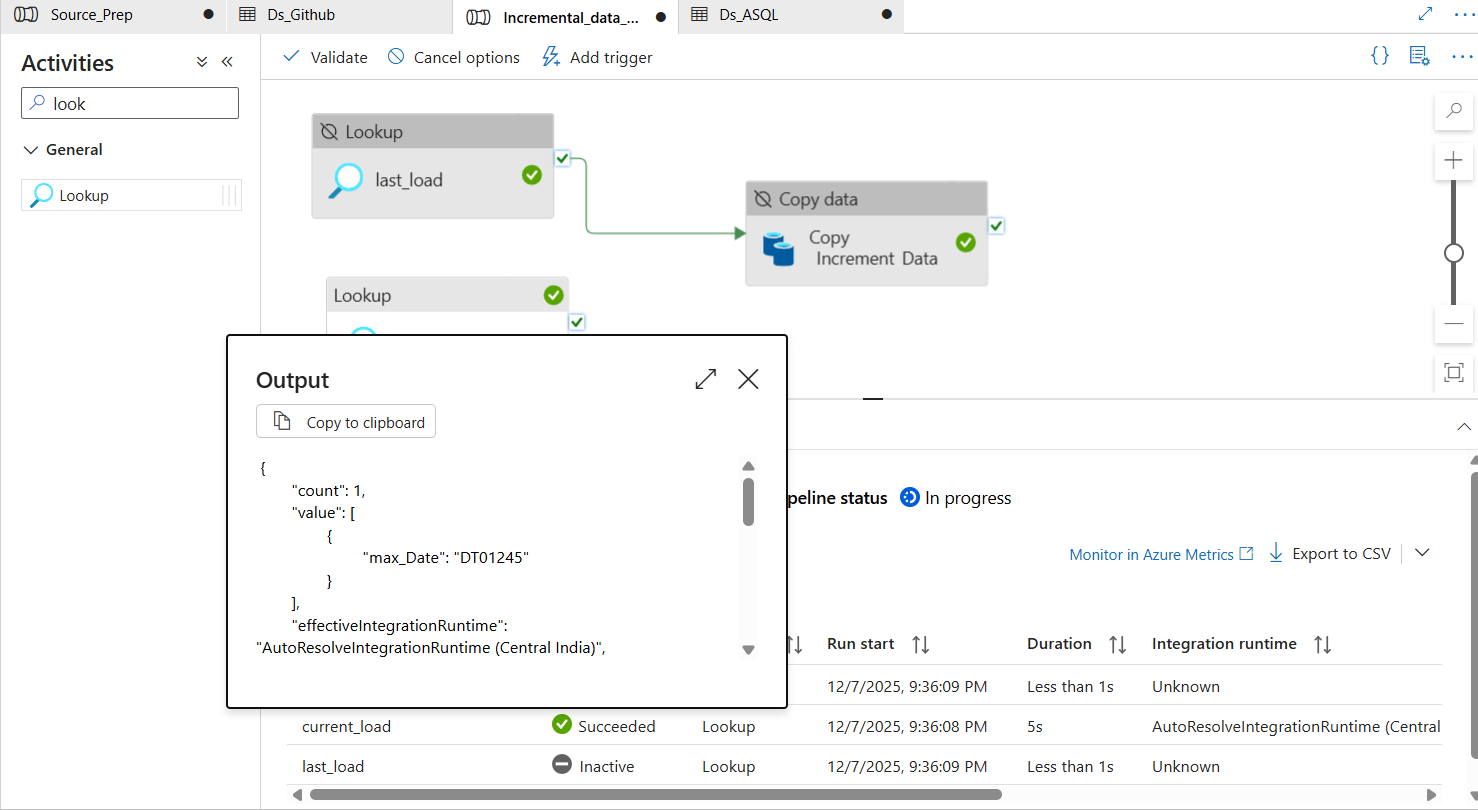
**last\_load will be used for initial load of data🡪 Here we need to provide the table name as watermark table🡪 select the query option and unselect the first row 🡪 select the dynamic content 🡪 give select \* from watermark\_table**

****

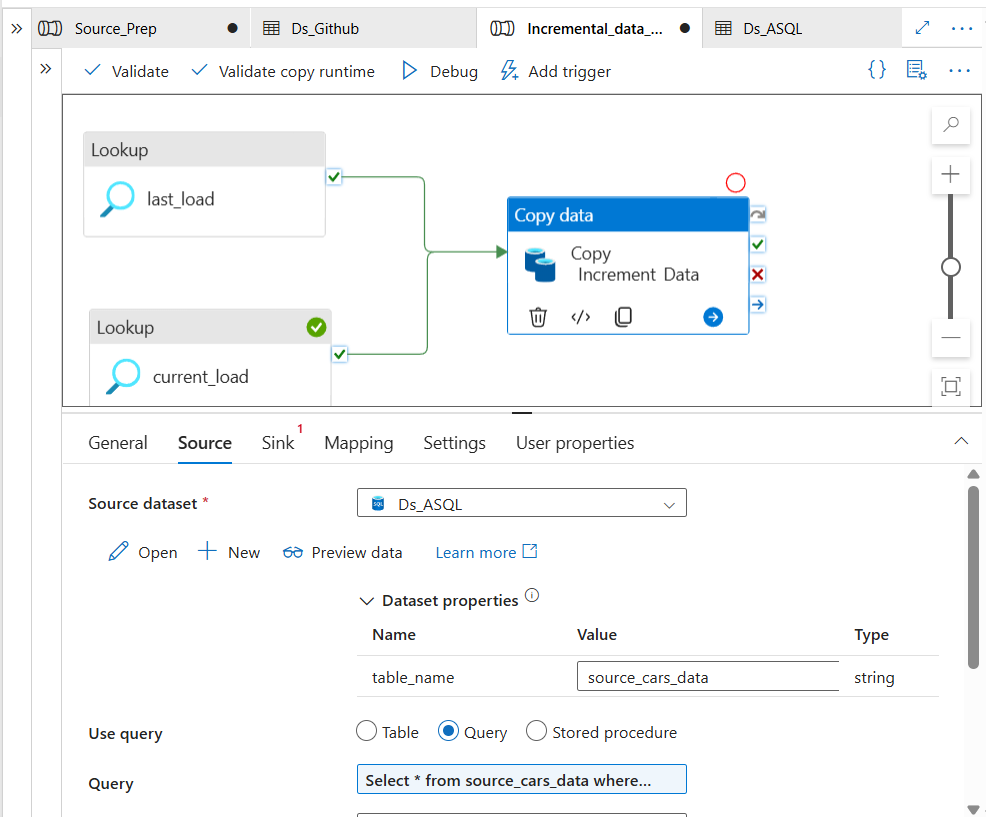
****

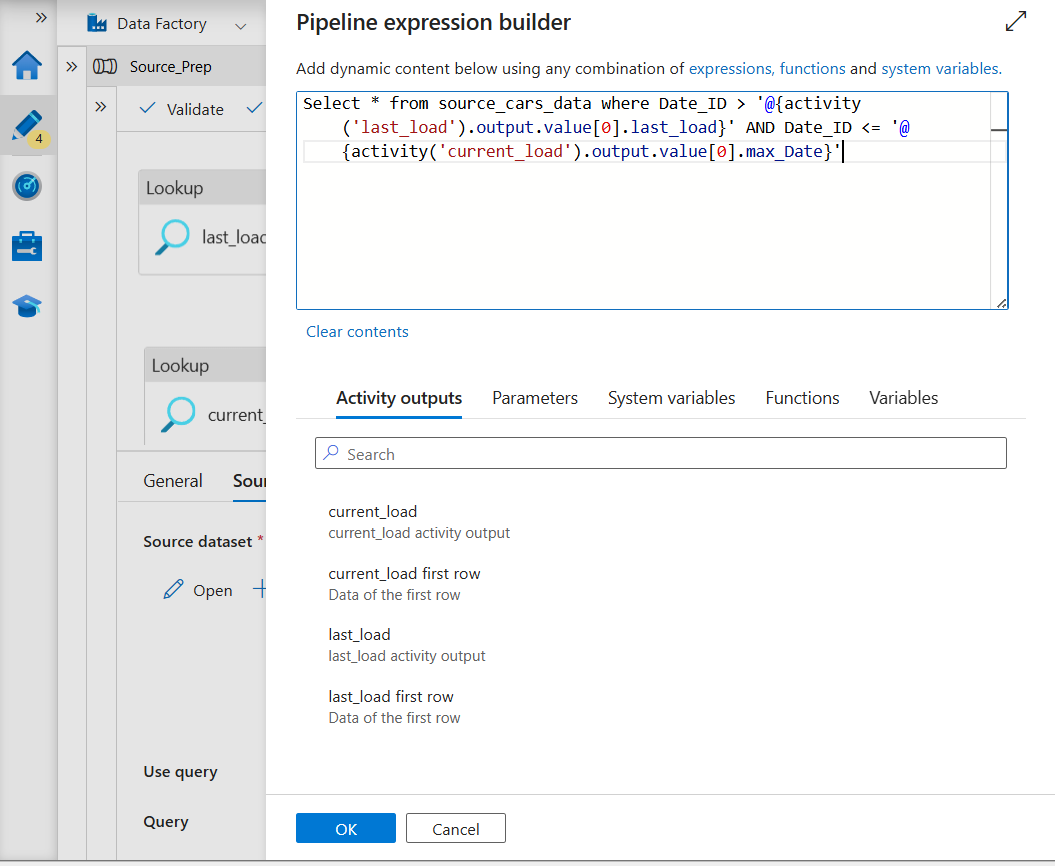
**Second lookup is for incremental load 🡪 provide the table name as source data table and provide the query as below**

****

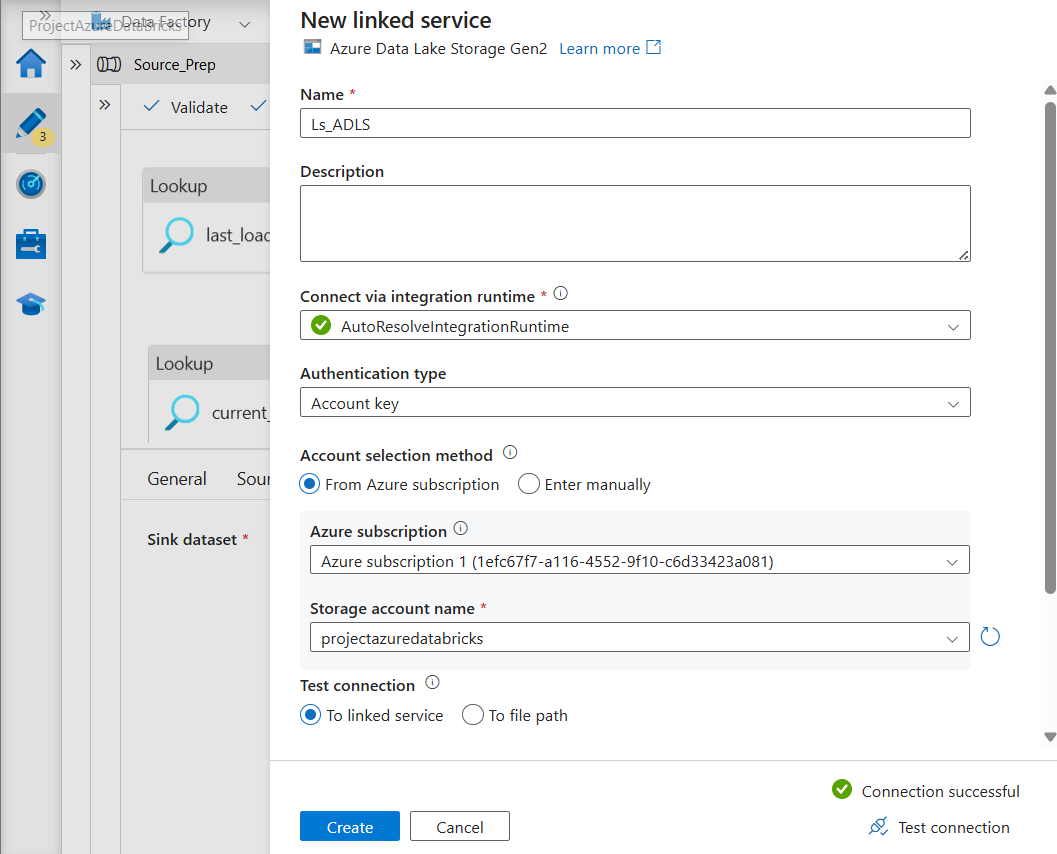
****

**Step 18: In Copy Data activity 🡪 Source**

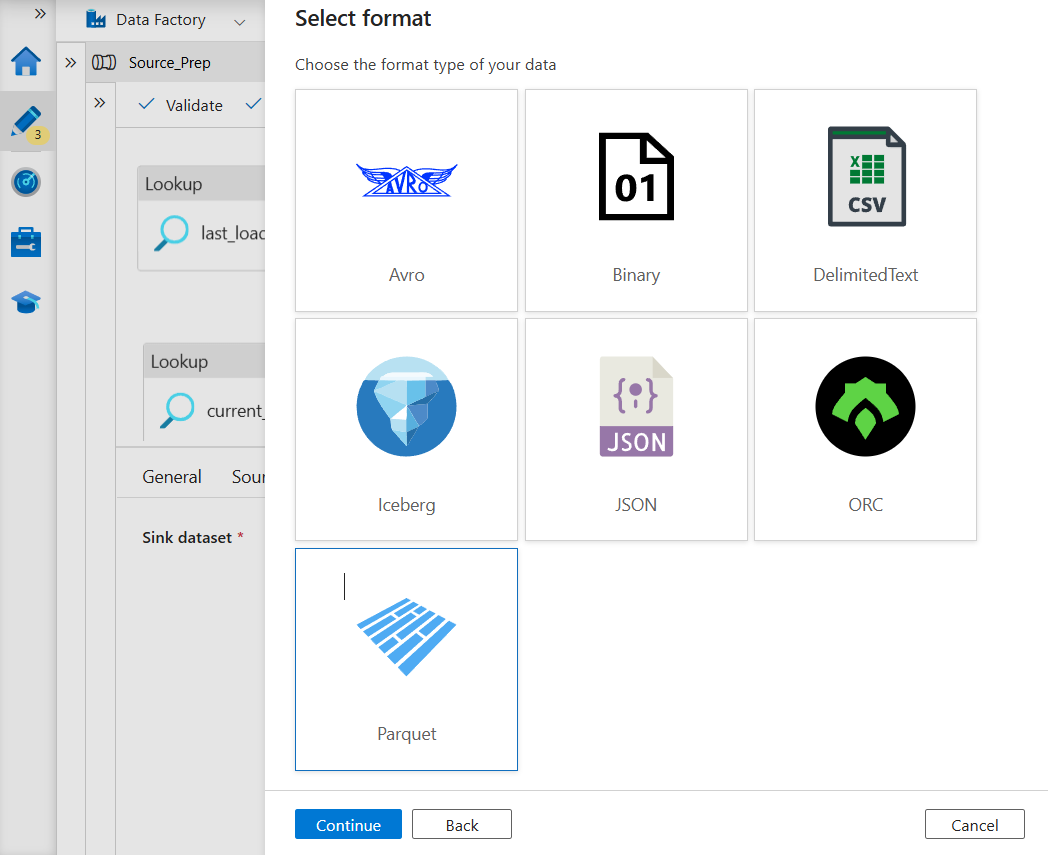
****

****

**In Sink 🡪 Create the Linked Service with ADLS Gen2 as below**

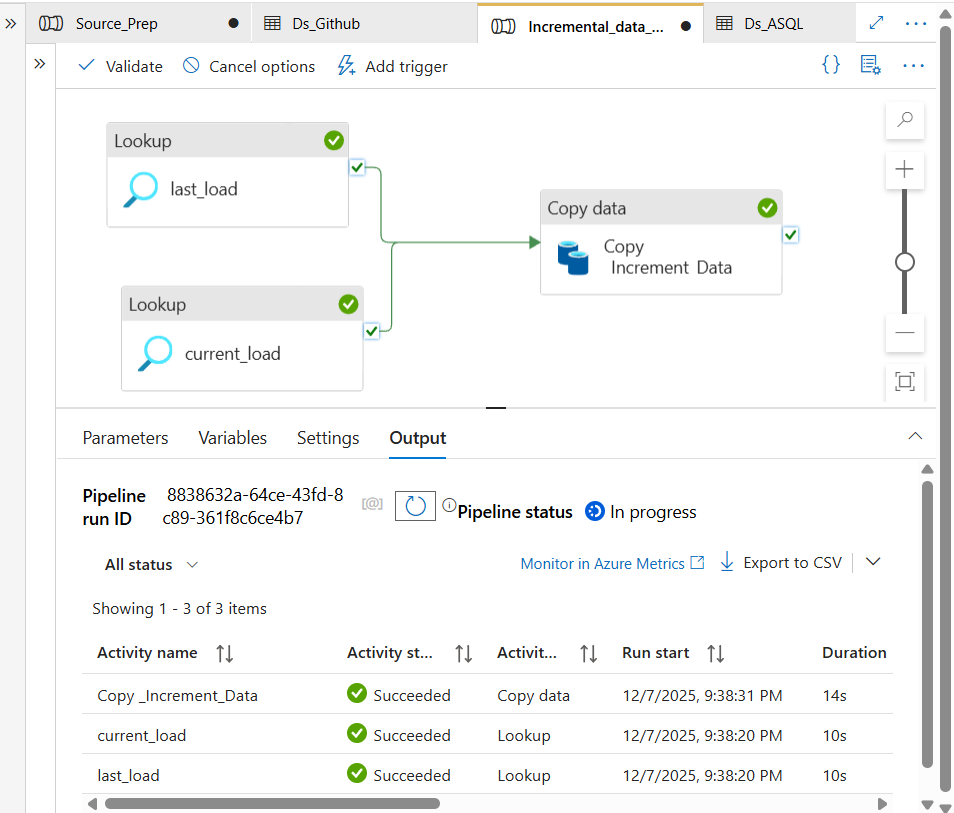
****

**Create the new Dataset with ADLS Gen2 🡪 File format as parquet**

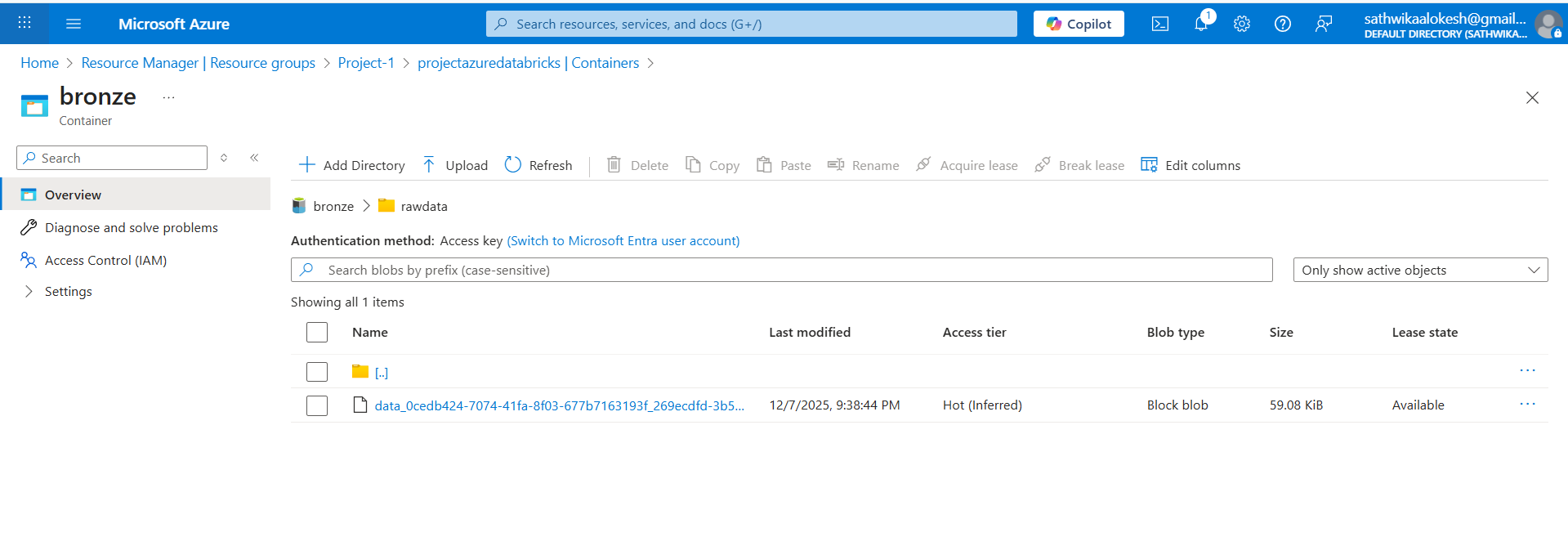
****

****

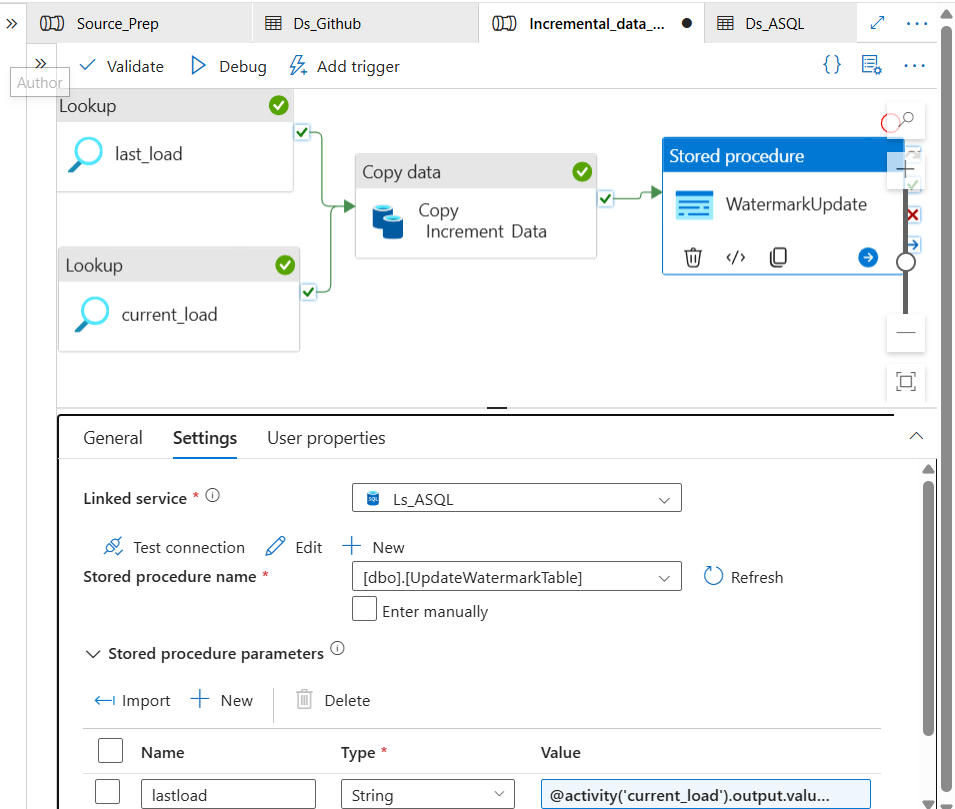
**Click on ok 🡪 Click on Debug**

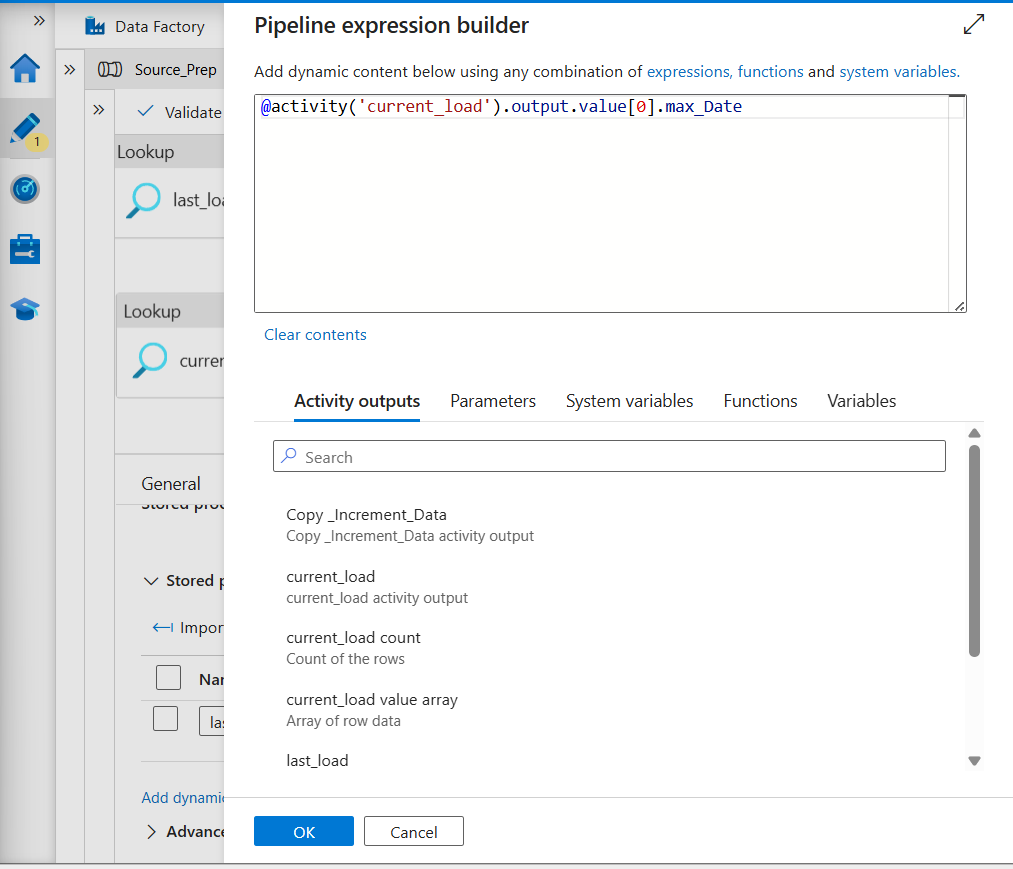
****

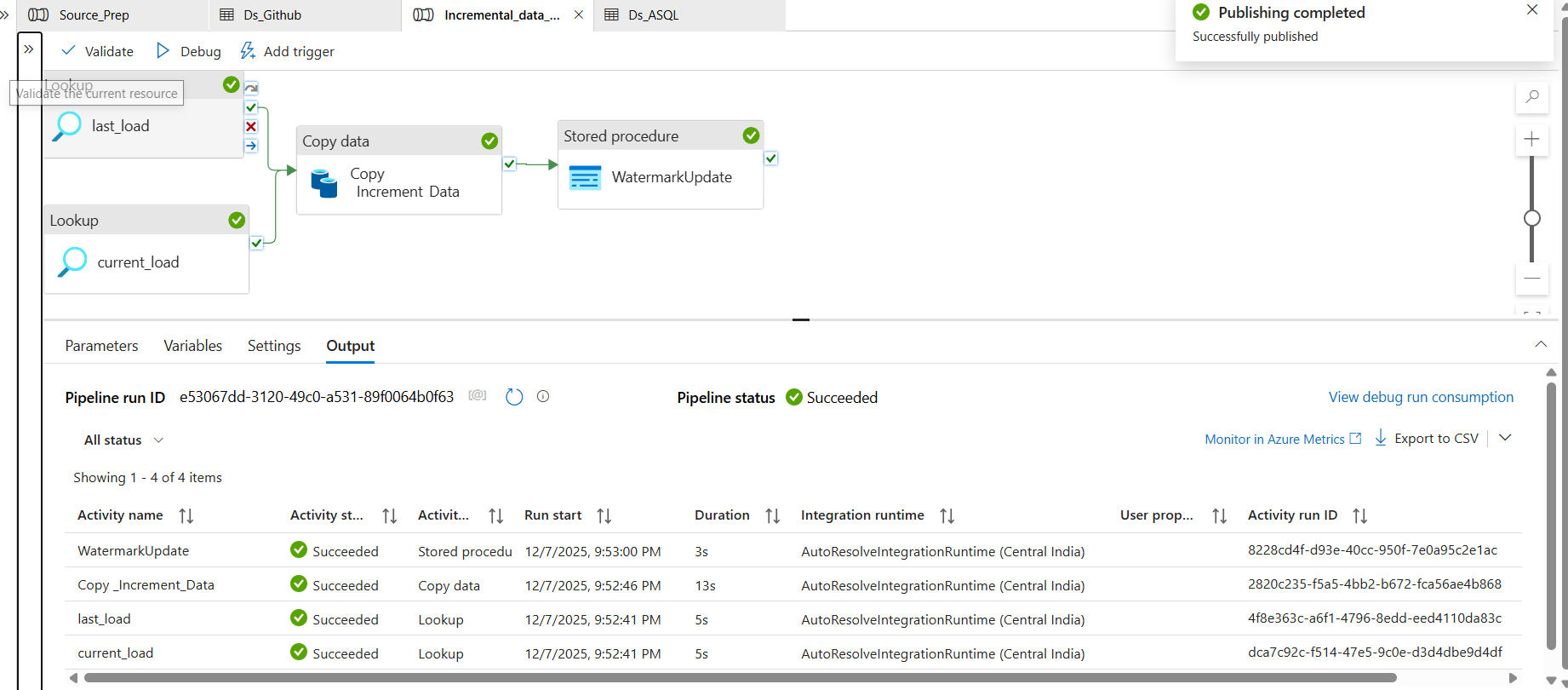
**Validate it in ADLS**

****

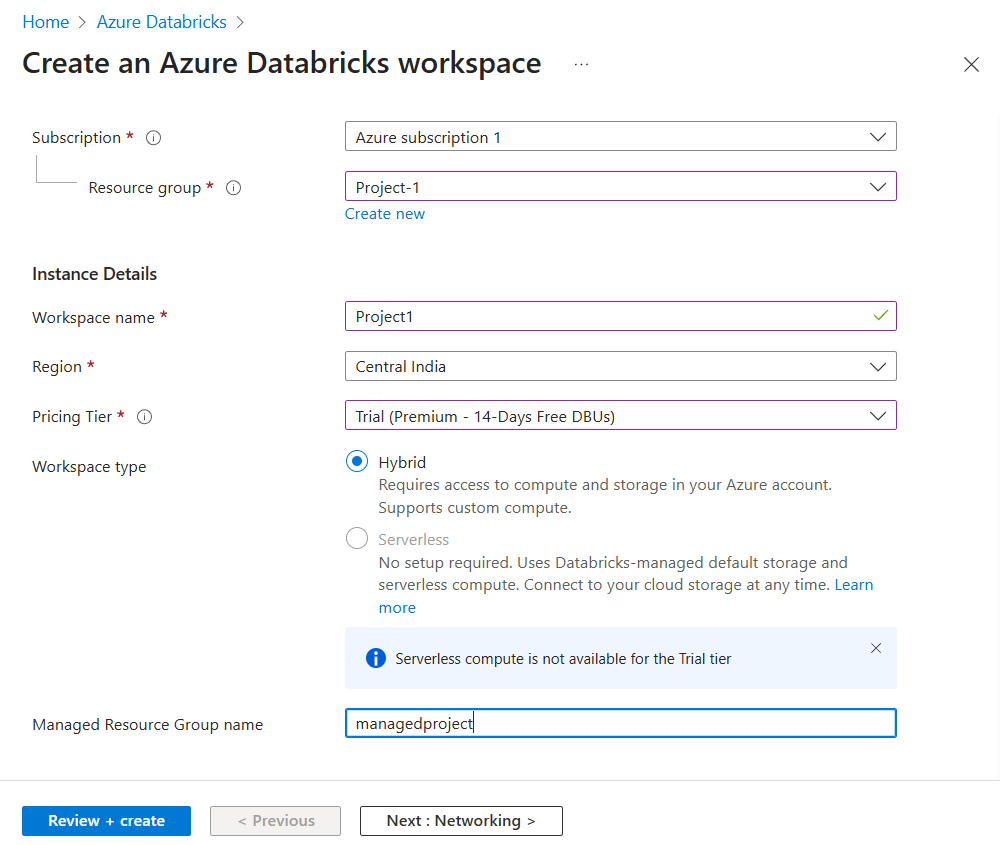
**Step 19: Once done Add the stored procedure to the copy data activity**

****

****

****

**Step 20: Create the Azure Data Bricks 🡪 Select pricing tag as free trail and provide the manages resource group**

****

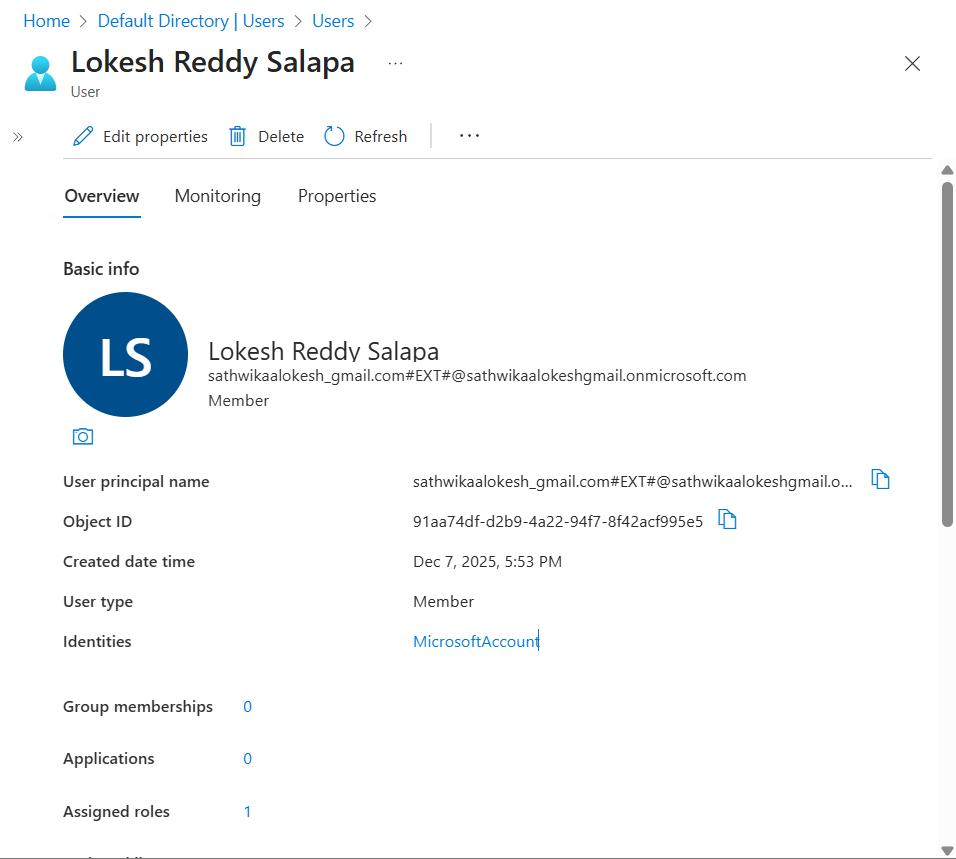
**review and create 🡪 once deployed launch the workspace**

**Once the Databricks workspace is launched follow the below steps**

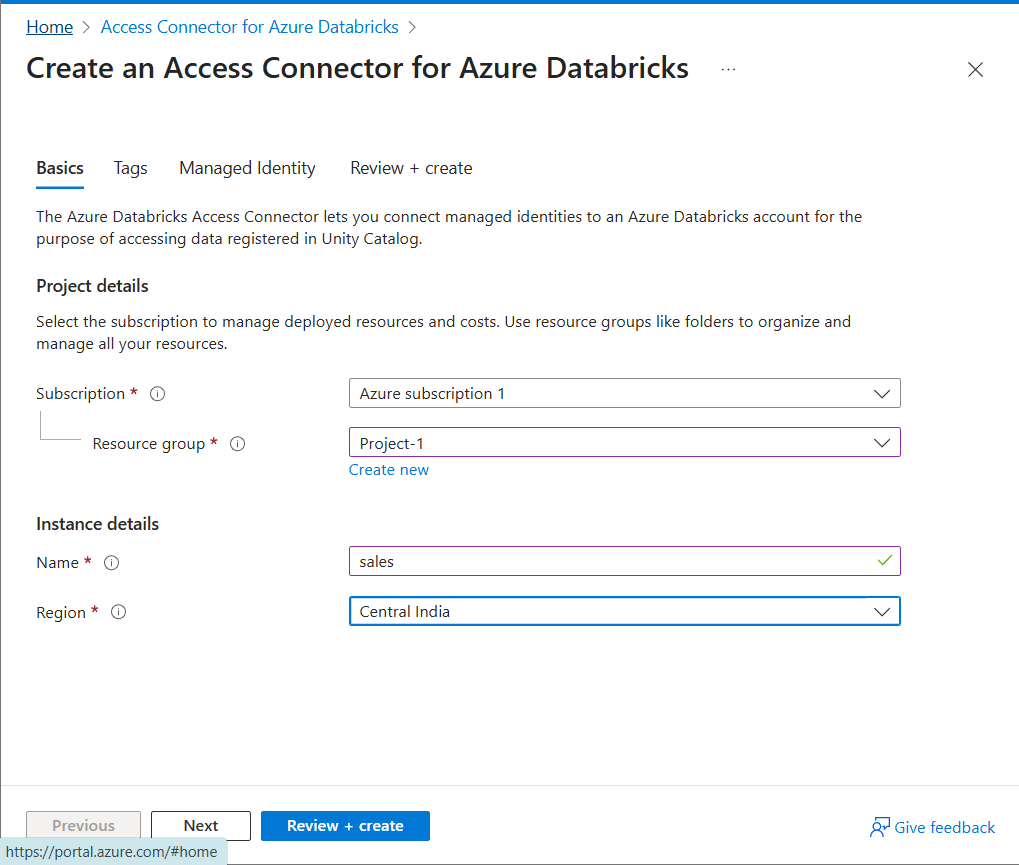
**Step 21: Go to Manage account 🡪 If manage account doesnot exist 🡪 Go to account.azuredatabricks.net in chrome 🡪 In Azure portal go to Entra ID 🡪 Users 🡪 copy the User principal name and give the username in accounts.azuredatabricks.in 🡪 click sign in by providing required steps🡪 once the loggedin**

**In the metastore Data bricks 🡪user management 🡪 create the group 🡪 add the users mainly add the default mail id of azure portal and add the group names as Admins 🡪 then go to catalog 🡪 add the meta admin as admins as below**

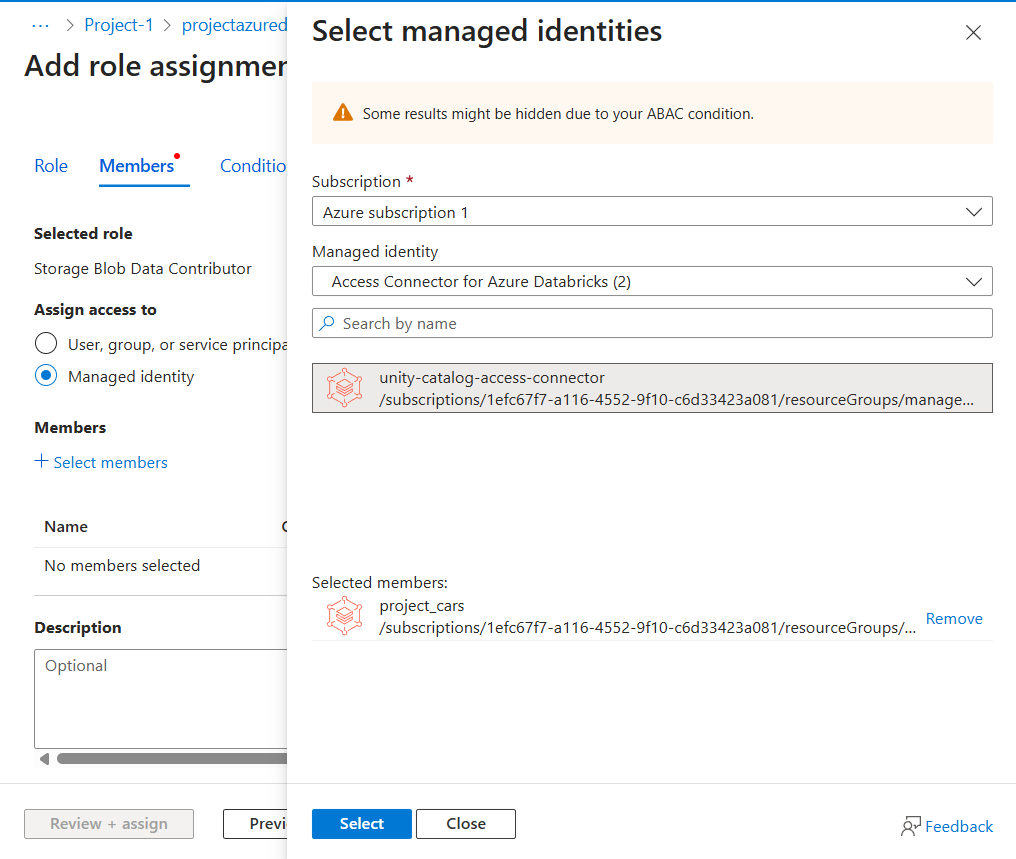
****

****

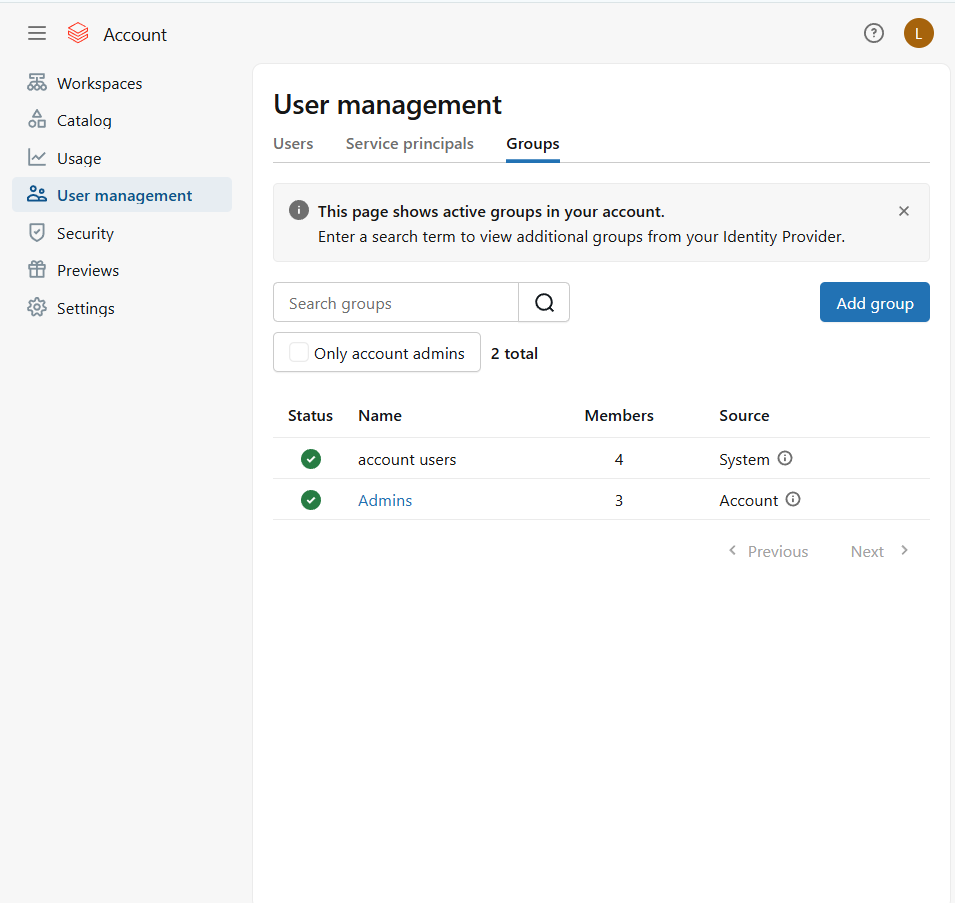
**Step 22: Create the access connector in azure portal**

****

**Step 23: Go to storage account🡪 select the ADLS 🡪 IAM 🡪add role assignment 🡪 search Storage Blob Data Contributor**

****

**Review and assign**

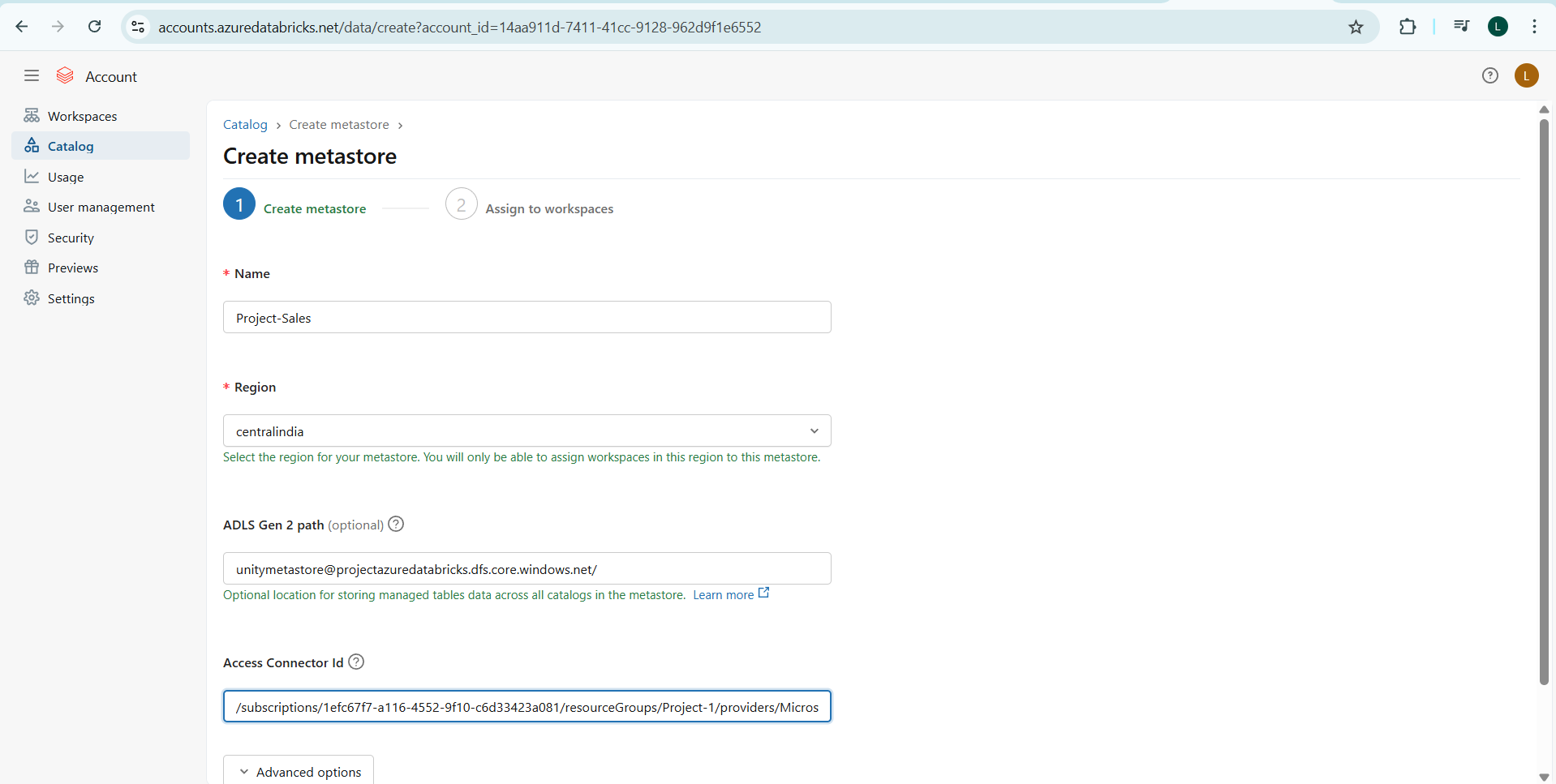


**Step 24: Create the meta store🡪**

**Create the container as unitymetastore in ADLS**

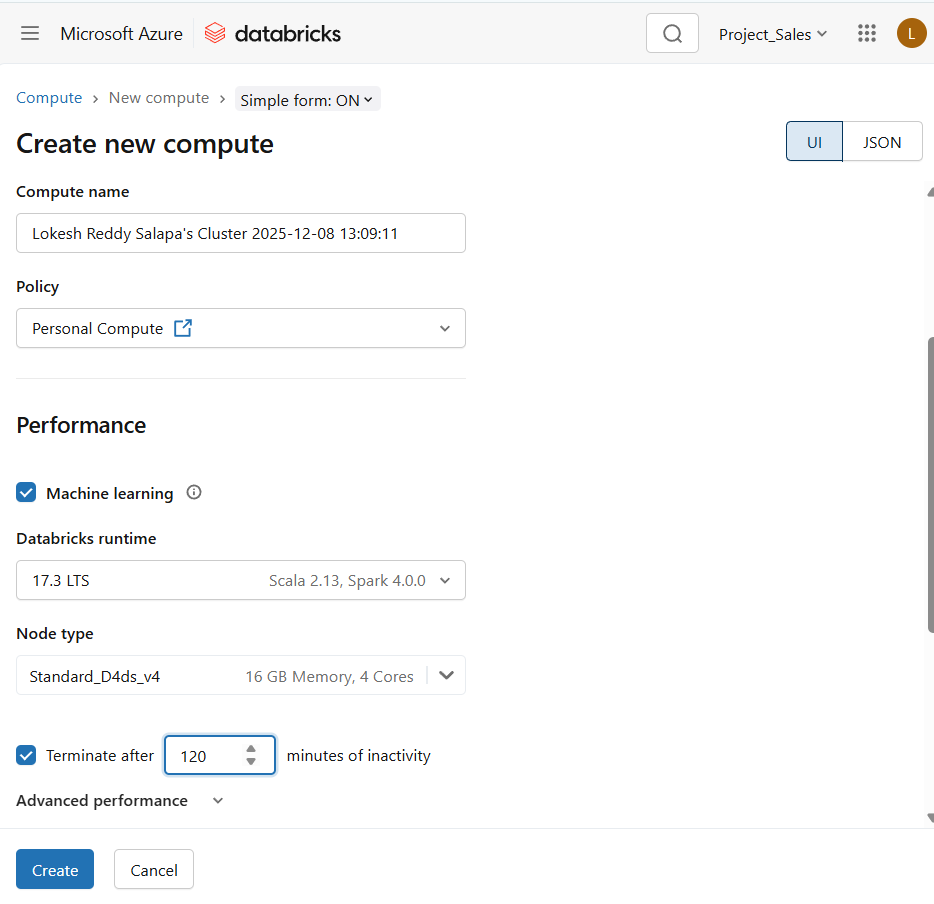
**Get the access connector resource ID as connector ID and ADLS path**

**unitymetastore@projectazuredatabricks.dfs.core.windows.net/**

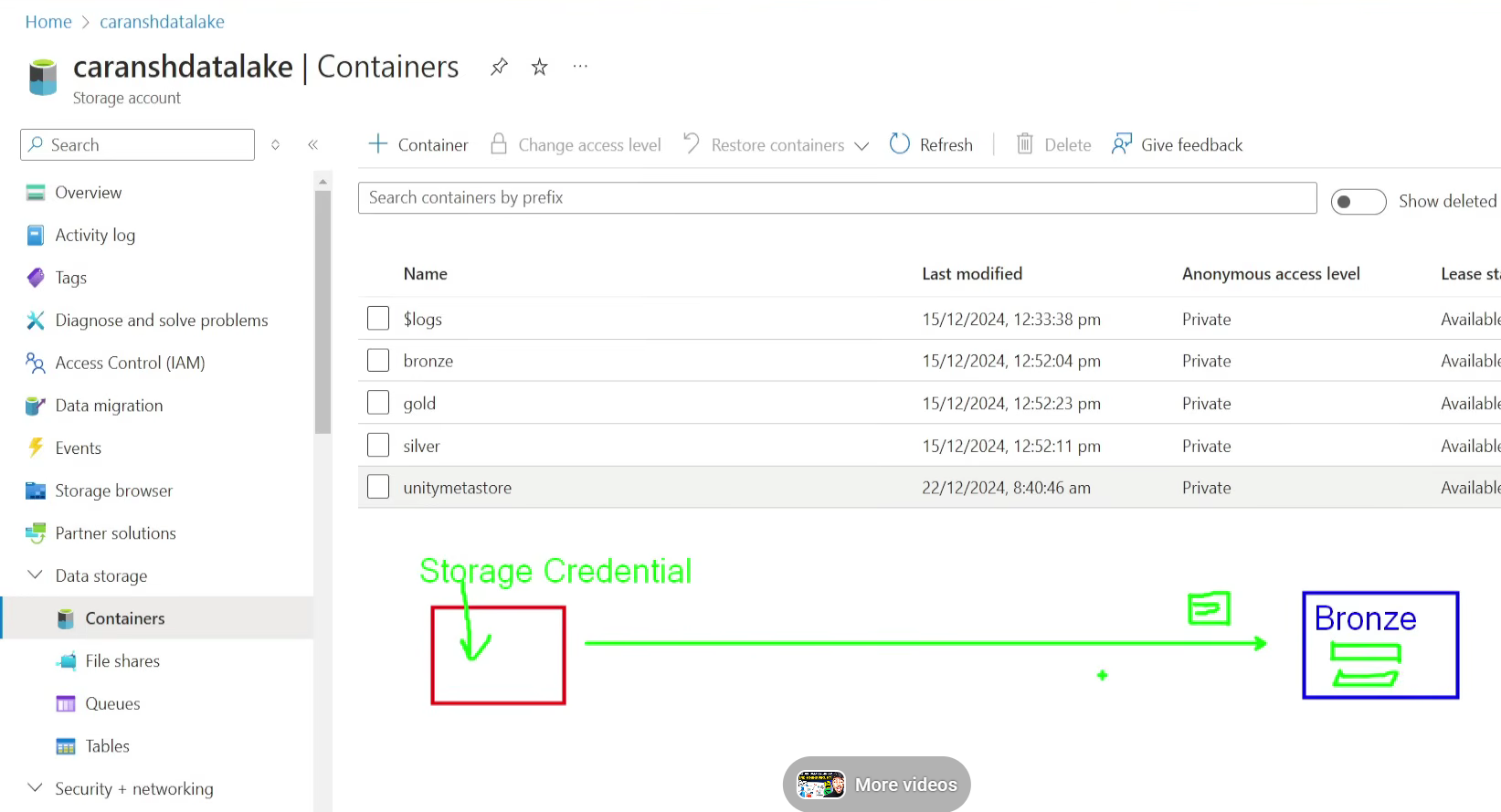
****

**Step 25: Once the meta store is created 🡪 go to main databricks account and refresh the page**

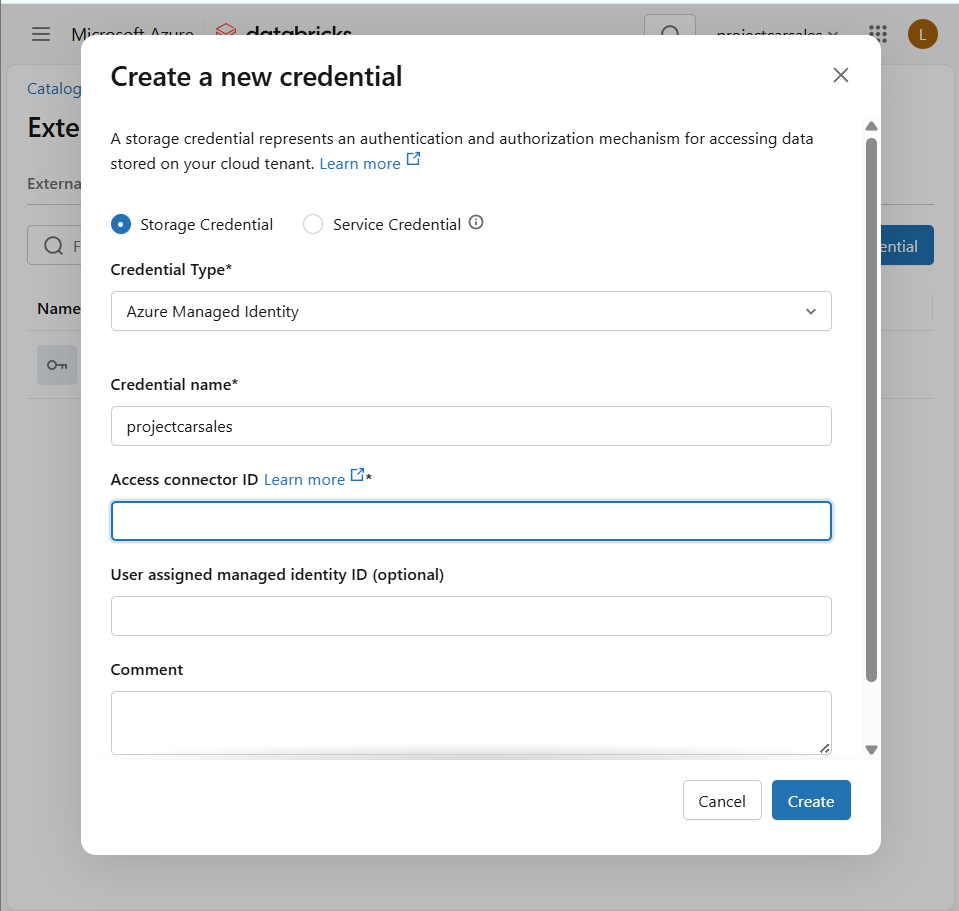
**Then Click on compute and create the compute cluster**

****

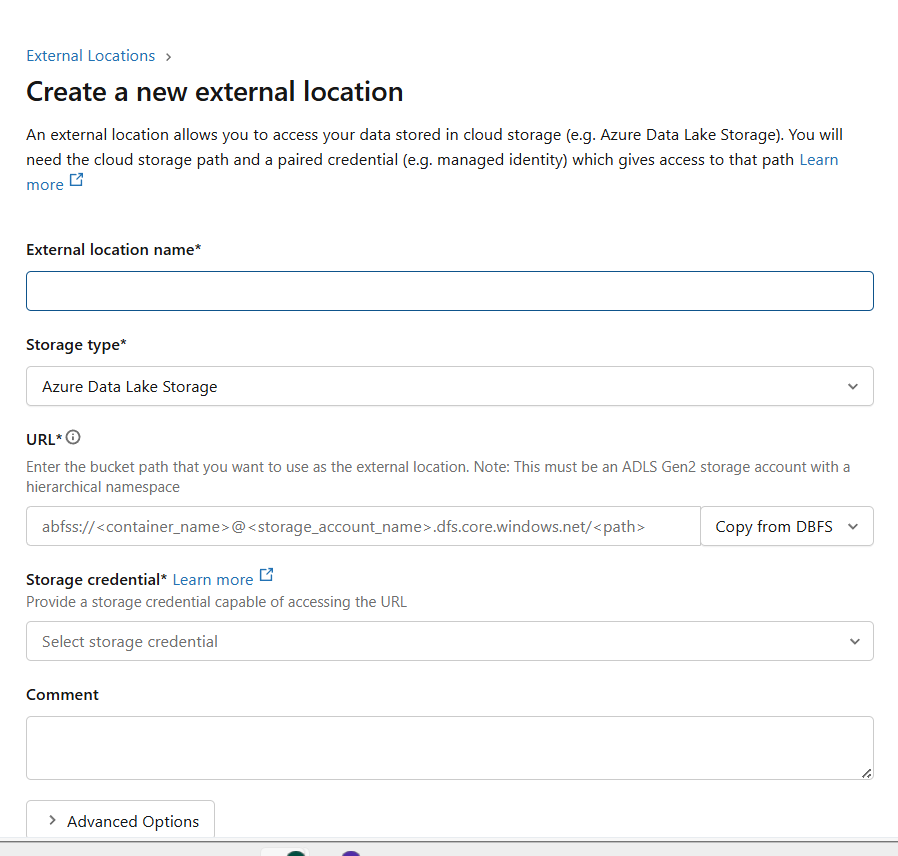
**Once it is created, Click on we need to create the 3 external location for Bronze, Silver and Gold to store the data. So for that we need to create the Storage Credentials**

****

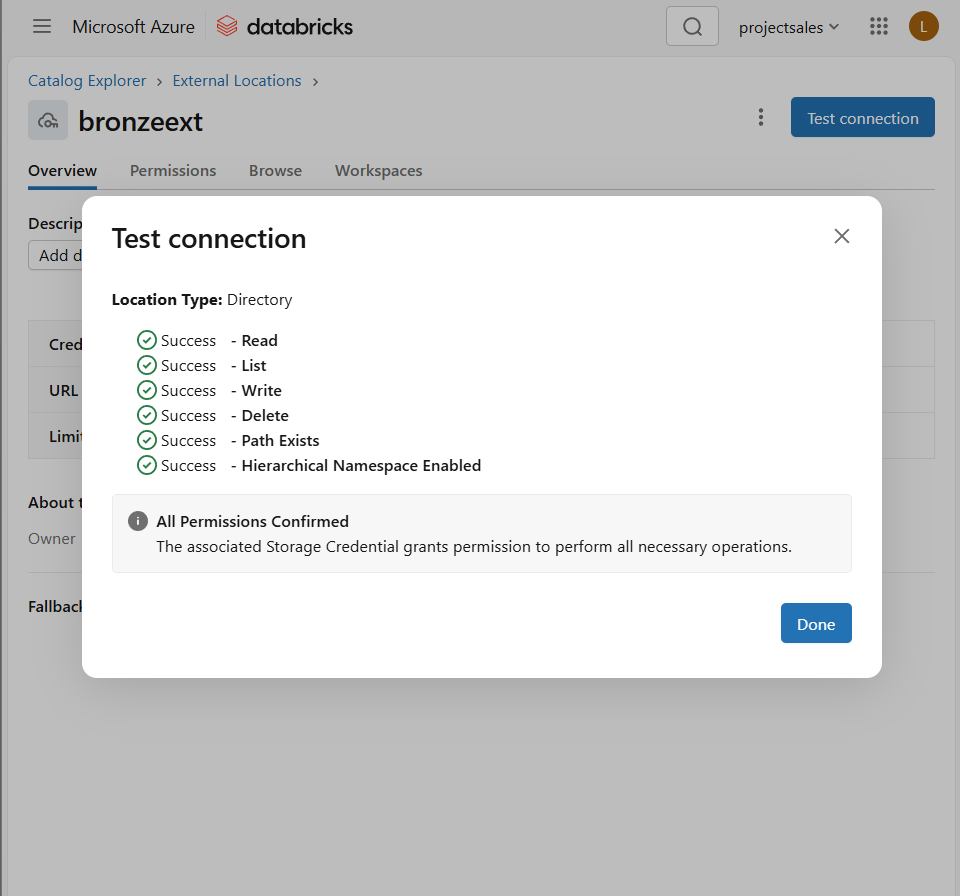
**Step 26: We create Storage Credential in Databricks to get the access for ADLS**

****

**Step 27: Create the external location in main Databricks**

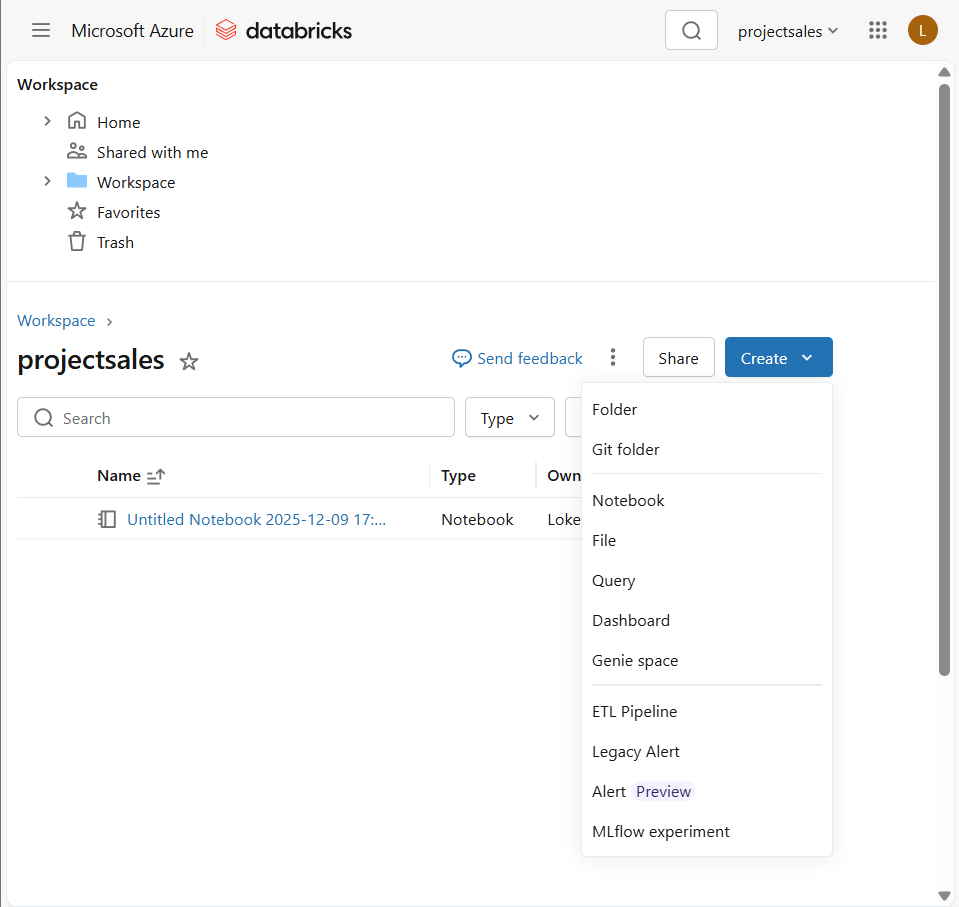
****

**If error occurs as create external location go to metastore and select the metastore we created and change it to admins 🡪 Once created click on Test connection and we can see as below**

****

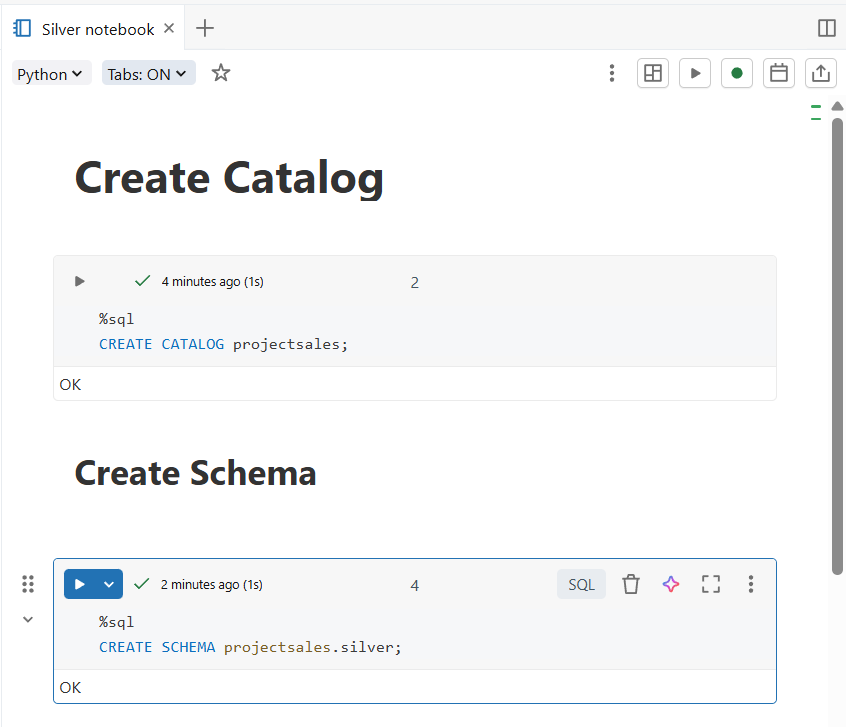
**Create similar external location for silver and gold**

**Step 28:Create the Folder in workspace folder and create the note book**

****

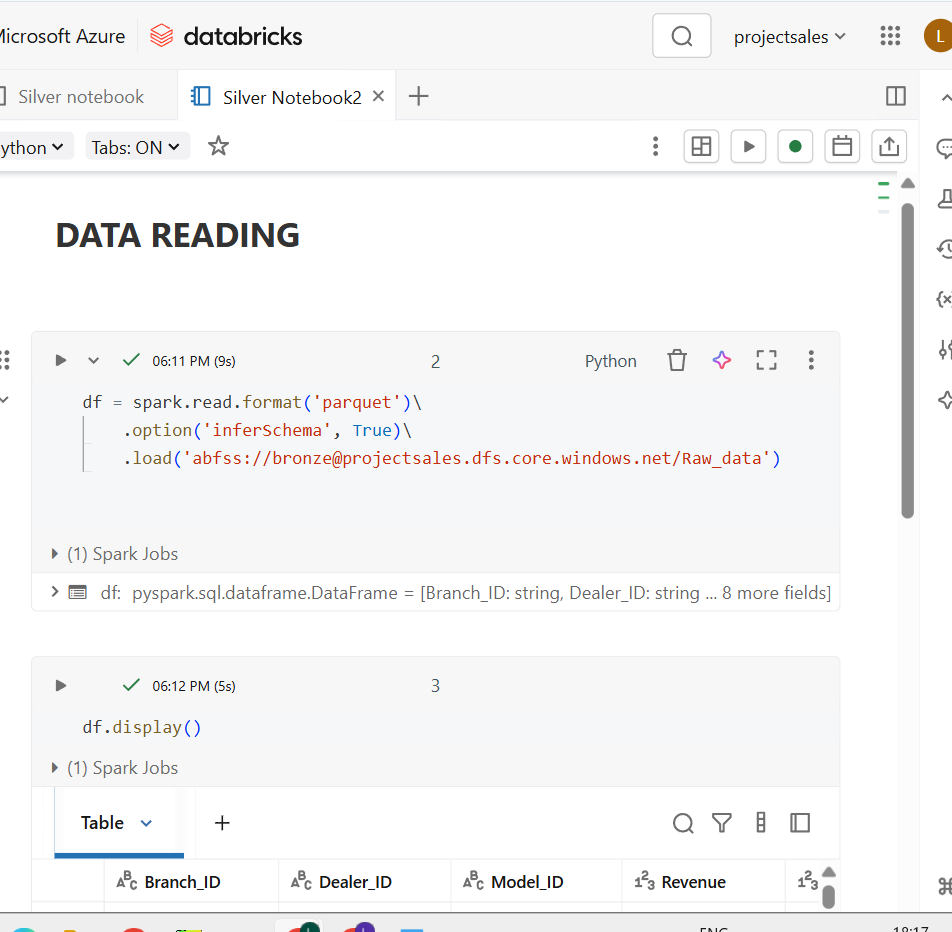
**Step 29: Initially follow the below notebook for the creating catalog,schema**



****

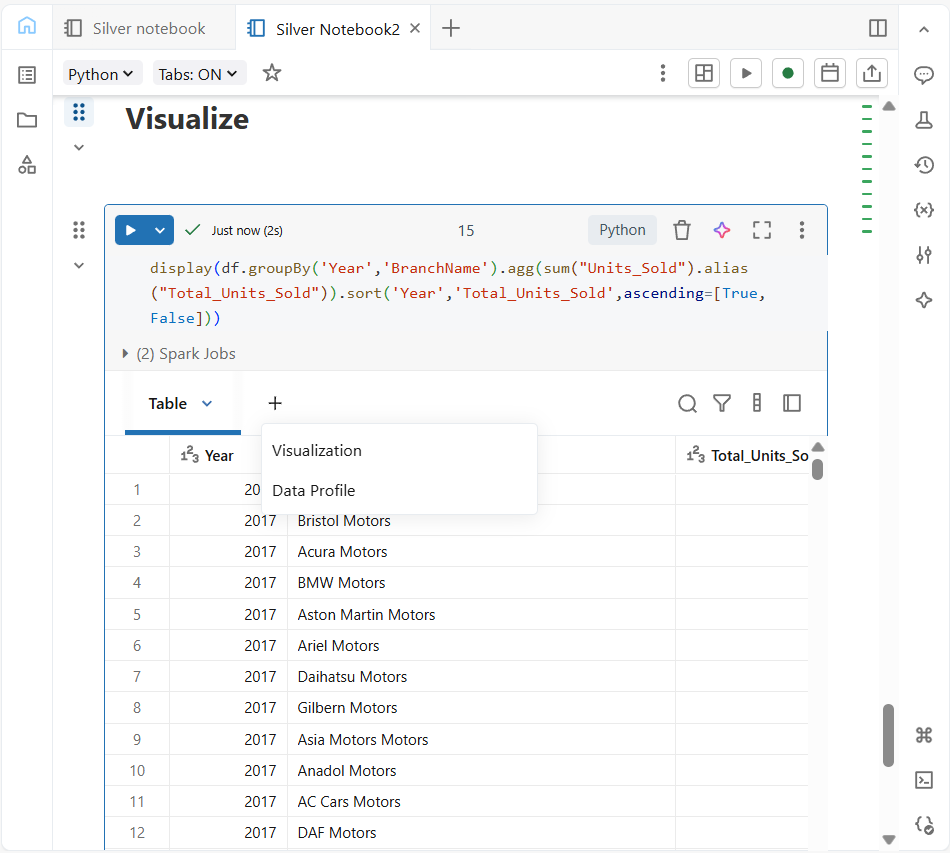
**Step 30: Create new notebook as Silver notebook for performing transmission**

**inferSchema – it reads the data from what are reading and suggests the best schema**

****



**Once all the transmissions are done, we can visualize the data**

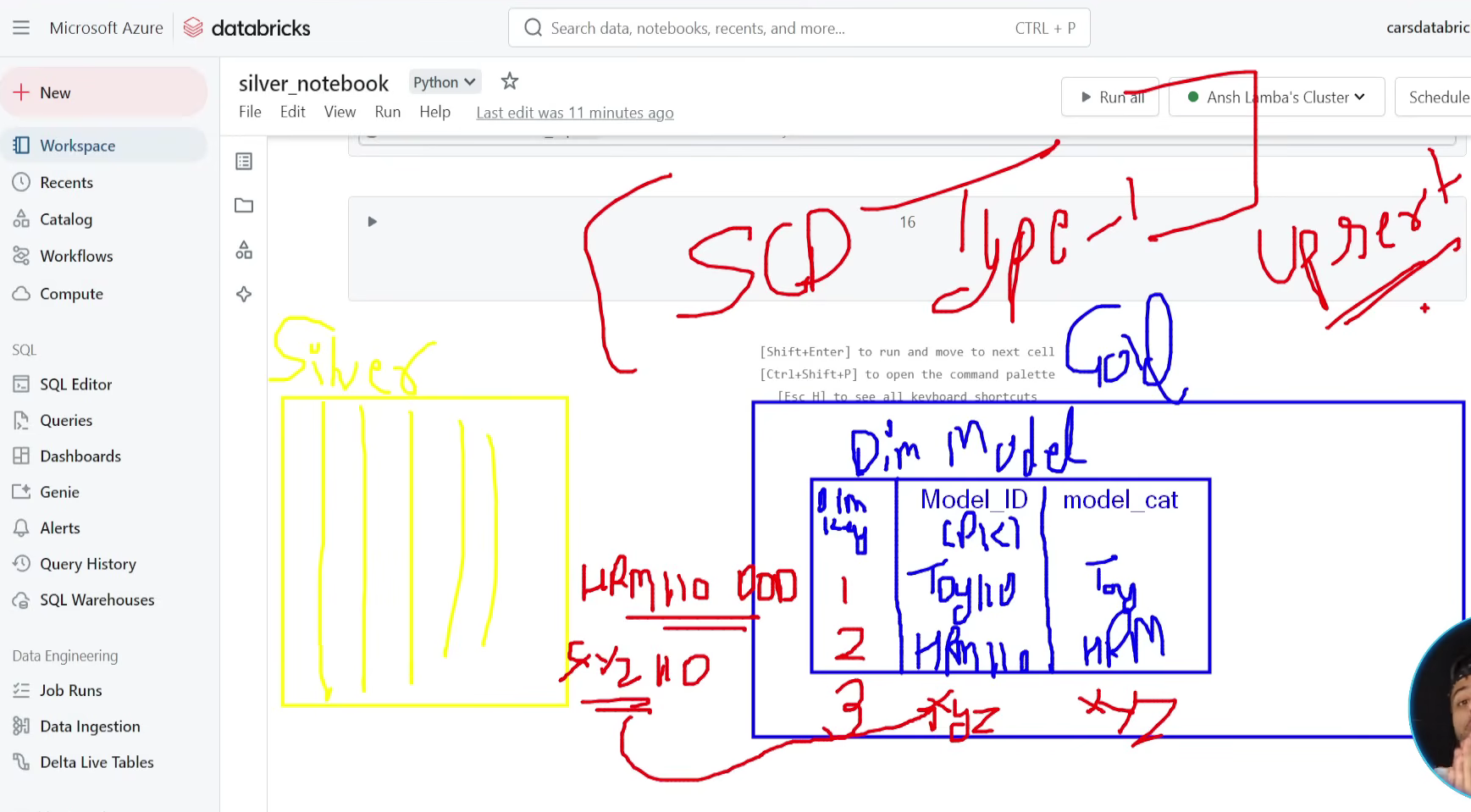
****

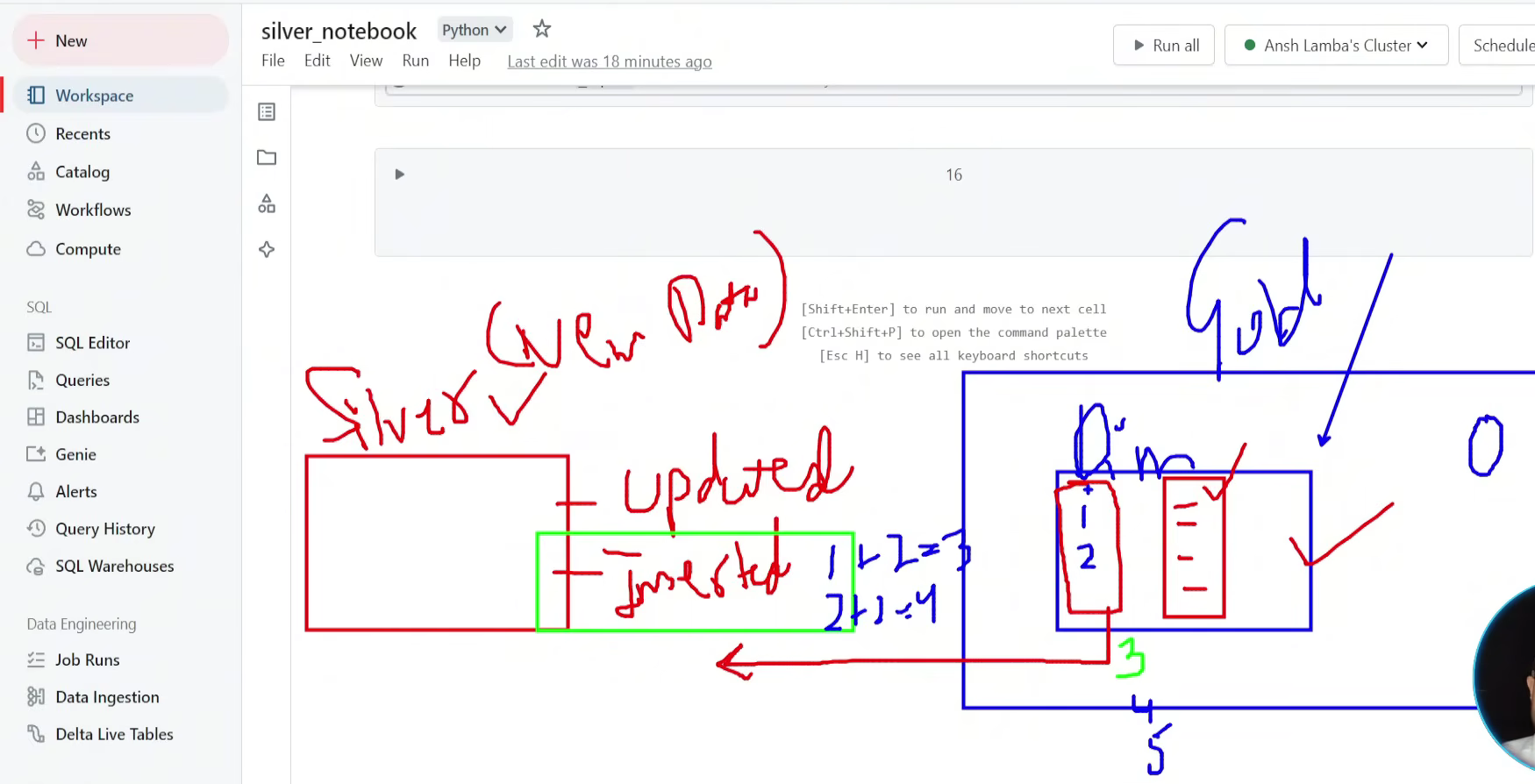
**Click on Visualize and select as type as per requirement**

****

**Step 31: Create the Gold note book in workspace**

**Here the flow we follow and it is SCD( Slowly Changing Dimension table ) Type-1 we as doing upsert( Update + Insert ). Here newly added rows will be inserted and update the rows which are already existed**

****

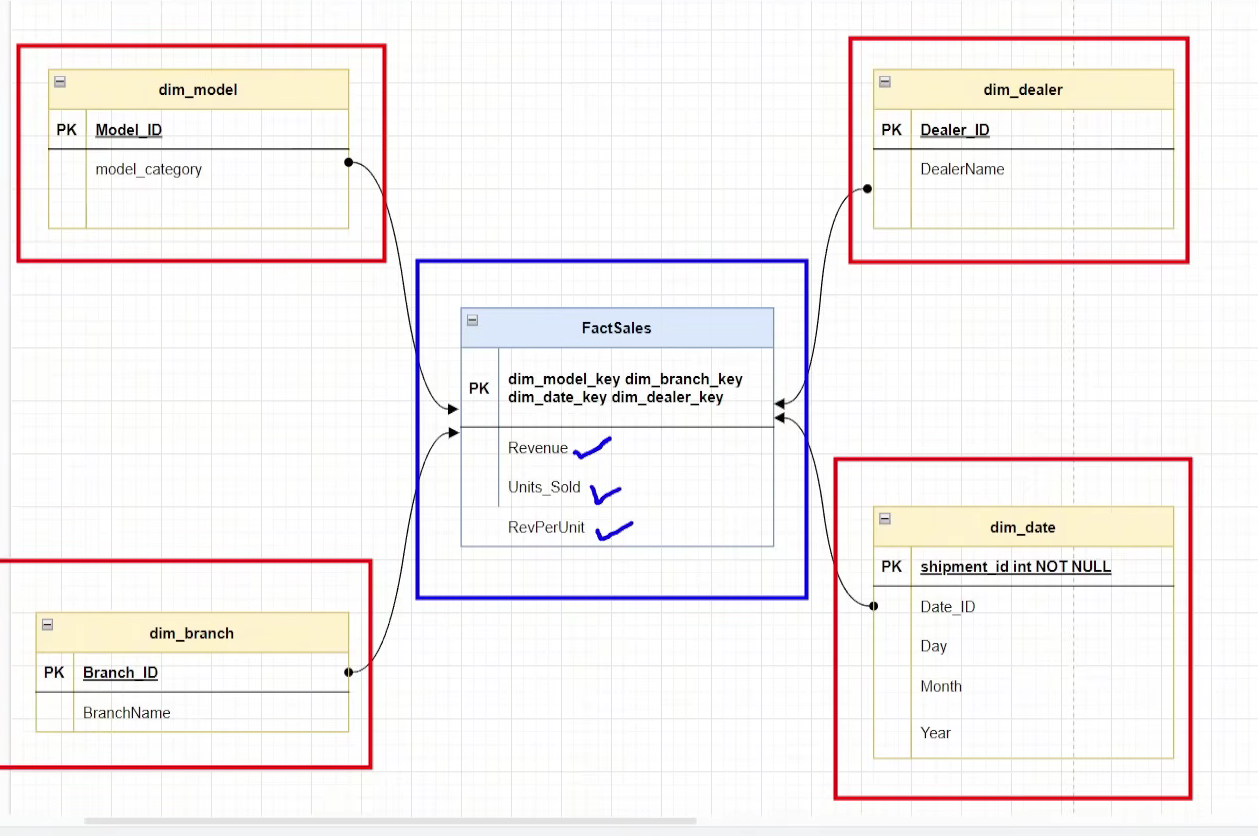
****

**Step 32: Create a Gold note book**

**Create the dimension tables using gold notebooks in the below**



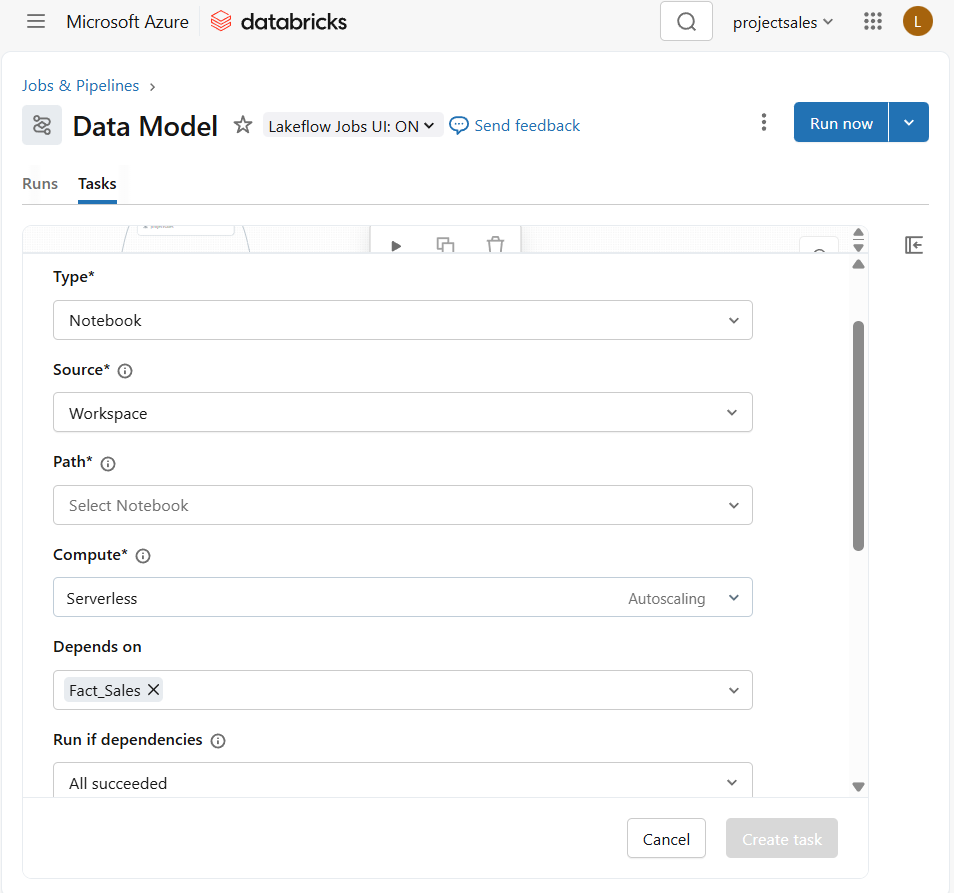
**Star Schema:**

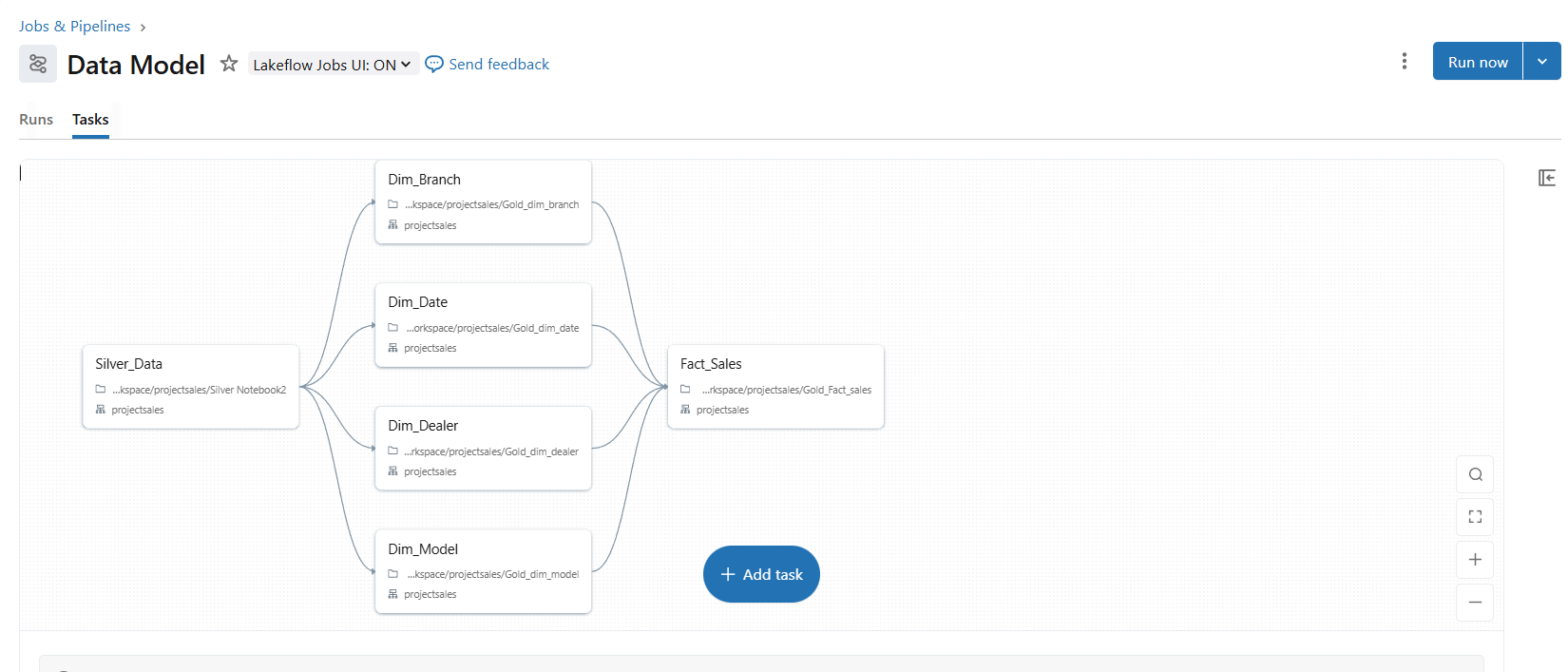
****

**Step 33: Once dimension table is ready create the fact table using the above star schema Fact table is ready. I have attached the document below to create the fact table**

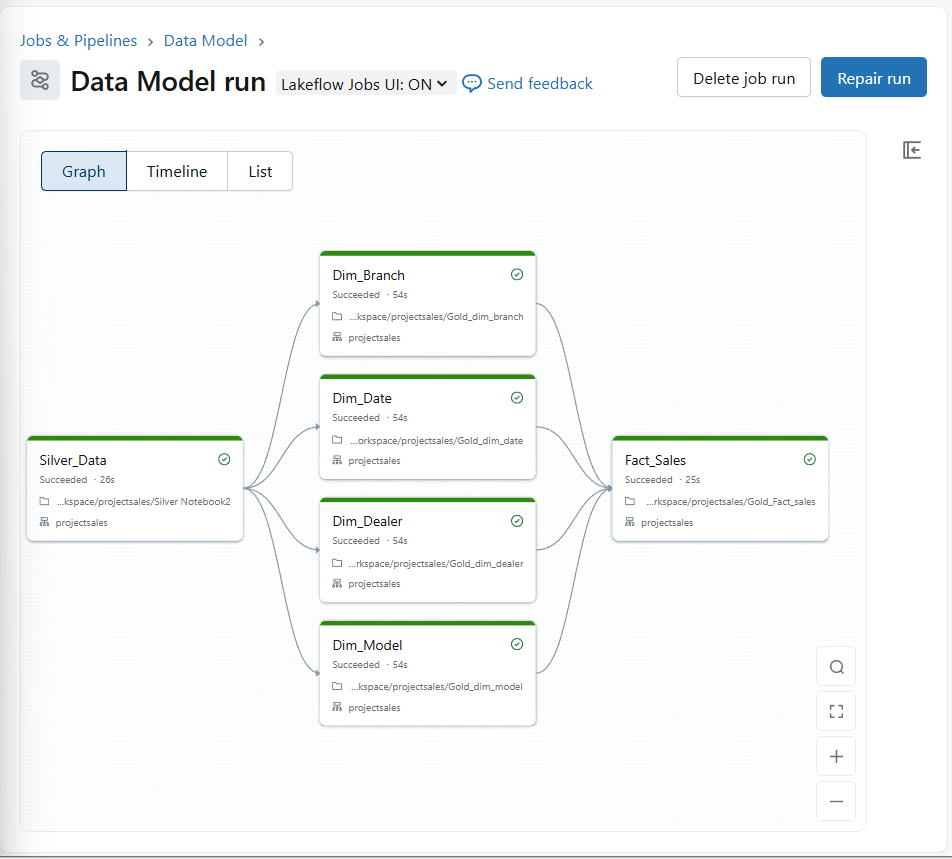


**Step 34: In Data Bricks 🡪 Go to Jobs and Pipelines and create the job🡪 provide the required details**

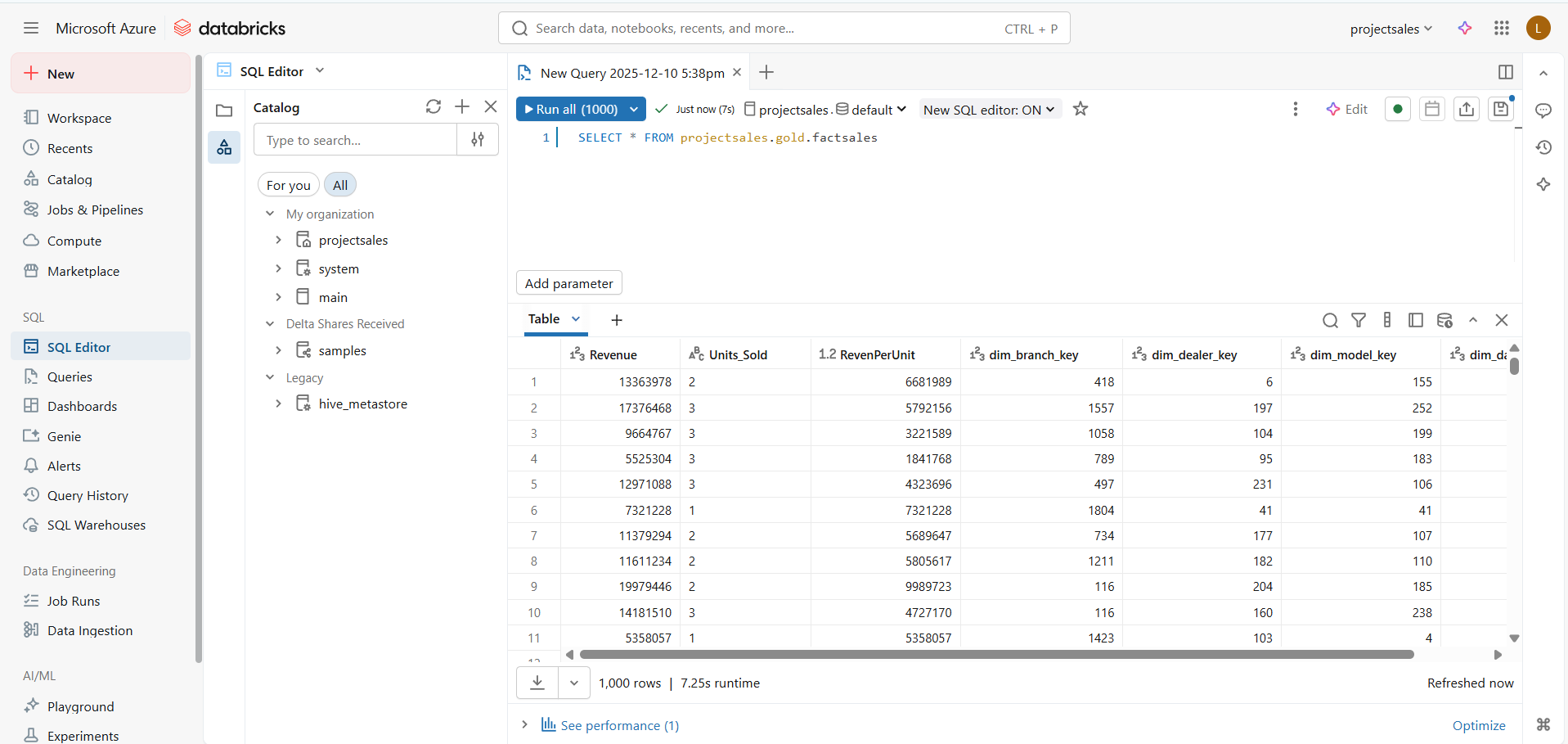
**Depends on will be selected based requiredment**

****

**Step 35: Once all the table are running fine it will show as below**

****

**Step 36: query the row is the sql editor**

****

**Step 37: In the src table change the file name in the source and debug then the newly added rows will be appended to src table**

**Step 38: Just click on debug in incremental pipeline and it will be added to ADLS and then click on debug in data bricks pipeline the newly added rows will be upserted in to the fact table**