Links

Friday, February 3, 2023 12:45 PM

Microsoft documentation	https://learn.microsoft.com/en-us/azure/data-factory/
Azure data Factory	#8. Azure Data Factory - Load data from On Premise SQL table to Azure Blob Storage All About BI channel on Youtube has 155 plus videos

https://azurelib.freshlearn.com/member/#/login

For examples of data factory
https://www.youtube.com/@cloudguru3178/videos

Pending
CI/CD pipeline build - adding paramter file
Connect adf to git

Lookup
-- can we lookup for a value in table like customer id from customers exits in Lookup_C table
-- Lookup at the target if table exists or create it and then insert the data

Course summary

Thursday, February 2, 2023 3:39 PM

1. Storage account - create blob storage and ADLS (gen2)
2. Within storage create a container / Root directory - name staging/ landing / raw .
3. Set the public access level and upload the file
4. For ADLS create a directory and sub directory and then upload a file
   ***NOTE - usually we do not come and create a storage, containers, upload the files . It is already setup for automatic load. Example facebook app is linked directly to store images in BLOB.
   Or we have some database from which we pull the data using pipeline /ADF , store the data in Blob/ ADLS , then we use databricks /sql to prepare a result

5. Not down the the access key

1. Note down the shared access signature SAS ****

5. Download storage account explorer and login via subscription so that you will see all underlying storages .
6. Create a SQL server and database
7. Create auzre data factory account and open azure data factory studio
8. Create a new Azure linked service to azure sql db
9. Create another azure linked service to connect to ADLS storage
10. Create a dataset from sql db employee table and link to DB service
11. Create a dataset from ADLS and link to ADLS storage - select target as a CSV
12. Create a pipeline with Copy data source task . Add source, Sink , target sink set row header

Video 4
Designing Dynamic flow

1. Select dynamic Source DB - using parameterized schema and db name
2. Select a Target ADLS , and add a parameter to select a path to create or select a dynamic folder … Paramter value should be DeptFolder Contact with date time , so everytime new file is added it will not overwrite existing one , it will create new folder

3. Create a pipeline and add above source and target. At source append a paramter value for file path as

```
@concat('DeptFolder/',formatDateTime(utcnow(),'yyyy'),
formatDateTime(utcnow(),'mm'),
formatDateTime(utcnow(),'dd'),
formatDateTime(utcnow(),'hh'),
formatDateTime(utcnow(),'mm'),
formatDateTime(utcnow(),'ss'))
```

Video 5
--Parameterized linked service - not so common
Used to make linked service common to multiple database. Here we will need 3 params - db name,uname,password
At dataset In ADF new link service property will be shown to enter dbname,uname,pwd - At dataset level - add a one more parameter to enter this db name dynamically at pipeline level

-Lookup Activity in ADF
Example - Create lookup activity to check if employee table count is greater than 1 , Create a If activity , True create a copy activit y from Source DB to ADLS file

--Nested Lookup - If elseif else

Check If empCount>100 then Copy emp data and within this task Check if Product count is >100 then copy product data
For this Inside first If True activity call another pipeline which will check => If pdorcut count >100 then copy product data . We do not have Nested if else , but we can call pipeline within tasks

--Video 6

For Loop -- SELECT SEQUNTIAL OPTION IF WE WANT TO RUN THE ETL ONE AFTER OTHER , IF YOU WANT TO RUN IT IN PARALLEL THEN LEAVE THE OPTION
With sequential and batch count set to 2 - it will perform the 2 parallel ETL at a time

Example 1 - **Lookup_department pipeline**

Apply lookup function on department - pass the department id to for loop- for each department fetch employee details and create department wise list

FAIL EXCEPTION activity
IF pipeline is failed then call this activity to give customize exception

--Video 7

Metadata activity - file is greater than 10mb, if file is modified one hour back / one day back / if file exists , list of child items -list of folder files and scan
VARIABLES
We can use it to record start and end time
SET VARIABLE FUNCTION

--**Video 8**

Link service - REST API |REST GET METHOD
USE BASE URL and then relative url to filter the data example www.data.com is base url and then www.data.com\country\India So \Country\parameter will become relative URL
Link service - REST DATA TYPE

Video 9

Incremental pipeline

Copy data after specific timestamp only

Video 10

Pipeline parameters - are defined at pipeline level and can be edited at run time .
When we defined inside at data source level or any other element,,you will be able to run it at debug level and define it in advance .

Global parameter
This is global across the data factory and can be consumed by any pipeline . Not at underlying data set levels (@pipeline.glo balname.databasename())

Debug - here we can click on red circle top on each activity and change it to debug mode , so that pipeline will run only till that p oint
Trigger- when u set a trigger , it means u want to run entire pipeline

Video 11

Create a key vault - secret key

Video 12

Scheduling a pipeline using triggers
4 types of triggers
Scheduled - regular pipeline for ETL
Tumbling window
Storage events - Run it when storage is available . Means file is placed in source | specify storage name and container name and path | Ignor e empty file |blob create or delete
For this under subscription , Microsoft eventGrid and Hub must be registered under Resource Providers
Custom events

One trigger can be attached to multiple pipelines
If file upload is failed then we need to manually load it after , by clicking on restart

Video 13
Tumbling window trigger - frequency based trigger ,we can set dependency , failover retry policy | attached to one pipeline only

Integration runtime - provides computation power . Information is in different format so we need some mechanism for converting the format. So this comes in picture
For any linked service default is - AutoResolveIntegrationRuntime. At run time it checks which region is efficient to provide IR , either at source region or target region .
For faster performance IR should be near target . It try to provide nearest region. We ca

3 types - Azure integration runtime ()
Self hosted Integration runtime (on premise - private network / office network)
Azure -SSIS IR - lift and shift SSIS package to azure

Video 14

Virtual machine creation - we can create windows machine . We can download software on this machins , visual studio virtual machine
Self hosted Integration runtime (on premise - private network / office network) - So this we need to install integration run time on local server . Or virtual machine where we may have our target DB

IF source is - self hosted IR and target is auto IR then by default self hosted IR will be used for operation

Video 15

Dataflows - Data flows allow data engineers to develop data transformation logic without writing code . No need to know more spark, Scala or language it uses spark cluster at the backend
Data flow debug creates a spark cluster at the back. Mainly used for datasets which is not a table where we can simply fire a query for transformation
Similar to Tableau Prep where we can work on data preparation and transformation
Dataflow is customize activity ,so to run it we need to create a data pipeline
So behind all, it runs on spark which is on big data , which is high performance engine . Also instead of creating data brick s we can use dataflow as we don't need separate maintenance.
But it can become lenghty as thing which we can do in one sql , we might have to add multiple steps

Video 16

Code management using - GIT (Internal) or Git Hub (Cloud -public or private repo)
- **Azure DevopsGit**
  - **Create a new orgs- add projects - new repo**
- **GitHub**

Create dev repo
Create UAT repo
Create a CI/CD deployment pipeline

Video 17 - PENDING
Logic apps -similar to data factory where we can create a workflow / orchestration - step by step workflow where we can integrate different services

Video 18,19,20
Python Intro

Video 21
Add annotation - to identify similar type of work / pipelines / team name . So in run history annotation may help
- to pipeline , trigger

User properties - we can add user properties to each step in data pipeline , data flow . We can use auto generate option which may generate source and target
we can mention query used in step or any other things which can help us to debug immediately at run time

Annotation is at pipeline level, user properties are at individual activity level

Reduce cost of pipeline
- No of DIU * duration * cost
- DIU - data integration unit (memory , machines ,processors)
  - Higher the DIU higher the cost . Keep DIU to 2 or try to reduce it
- Monitor Data movement activities at run window s
-
- Enable staging (valid for only few data source) - this improves the performance by staging a data in intermediate space from source and then copy to target

How will you manage the errors

- In data copy operation we may encounter bad records
- In setting we have **fault tolerance** option - skip incompatible rows, skip missing files , etc
- Enable logging - to catch the lines which are failing
- Logging level - warning for skipped file - specify the location for storing log files
- After running the pipeline we can check the run log how many rows read, how may rows skipped

Blob Storage

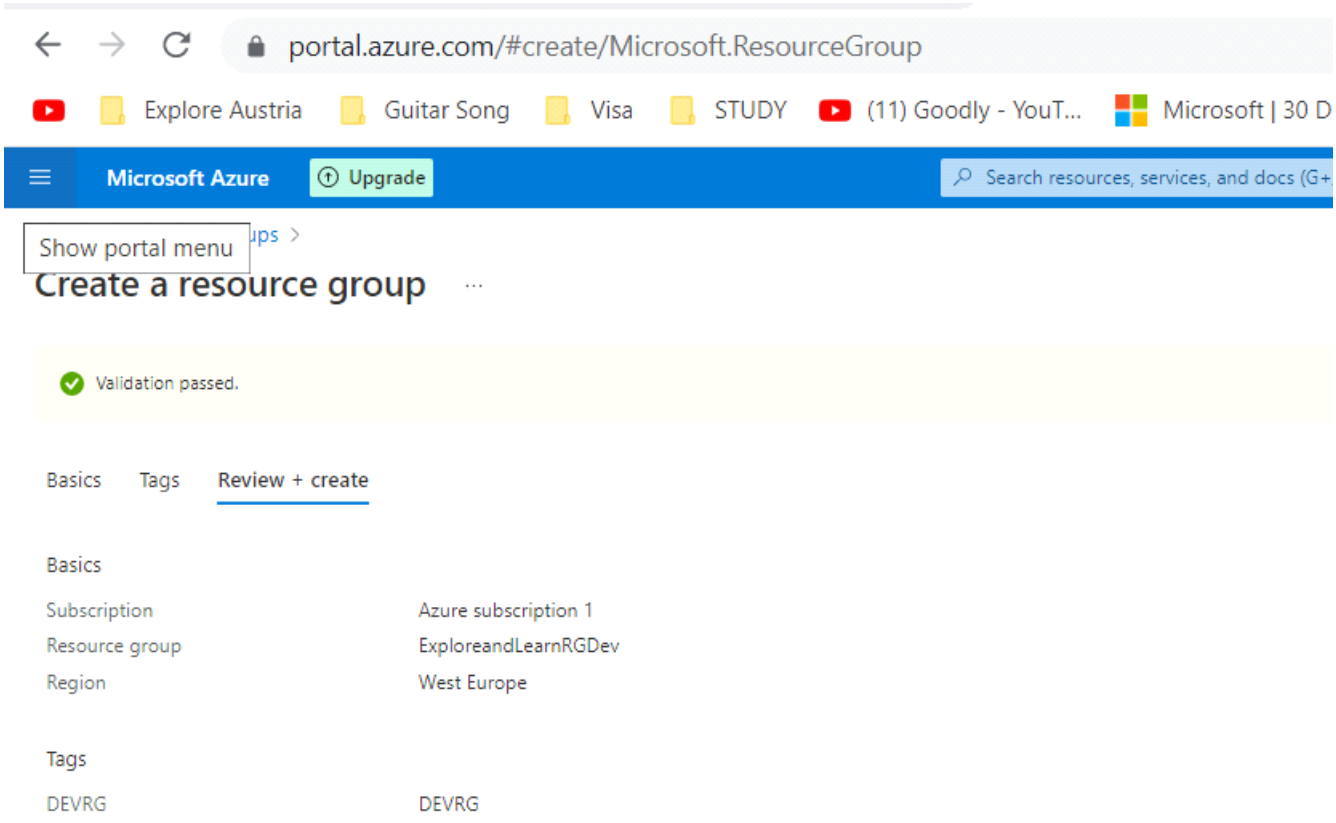
Wednesday, February 22, 2023 1:14 PM

Blob storage account -
ADLS Select data lake storage gen2 to create ADLS - data lake storage account
FileShare - is used by java dev/ cloud developer and used at application end. Data engineer uses only above two . You can map your local drive over netowrk to file share, so whatever file you add to local driver will get synced
Queue service- Applications which talk to each other | JMS
Azure tables - no sql database . Column based database - we use COSMOS DB instead which is more advance

Step1 Create Resource Group

Wednesday, February 22, 2023 2:32 PM

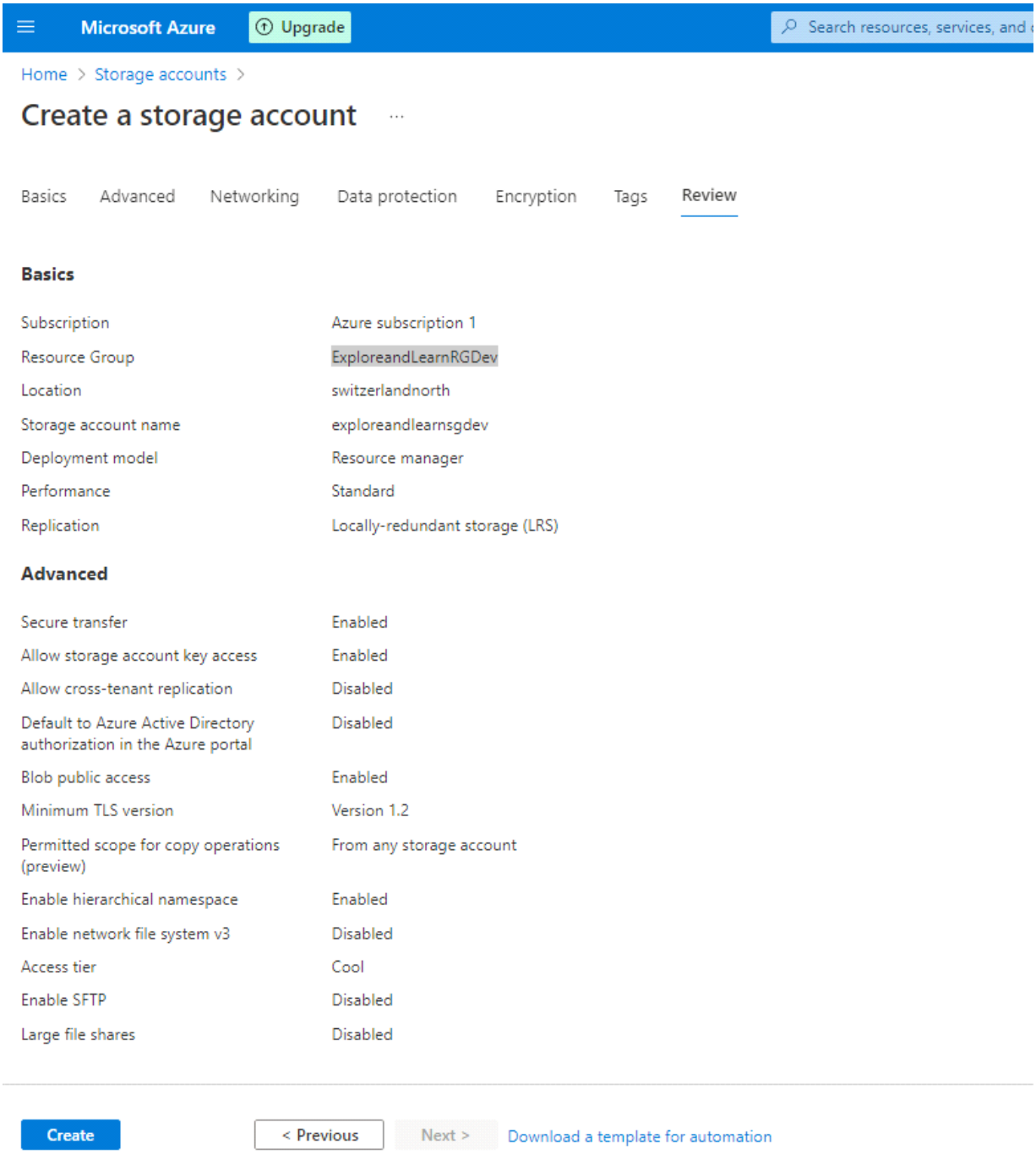
1 Enter Resource group name **ExploreandLearnRGDev**
2 Select the region closer to your physical location to get better performance
Skip Intermediate options
3 Add a tag for identification DEVRG



Step2 Create Storage Account

Wednesday, February 22, 2023 2:32 PM

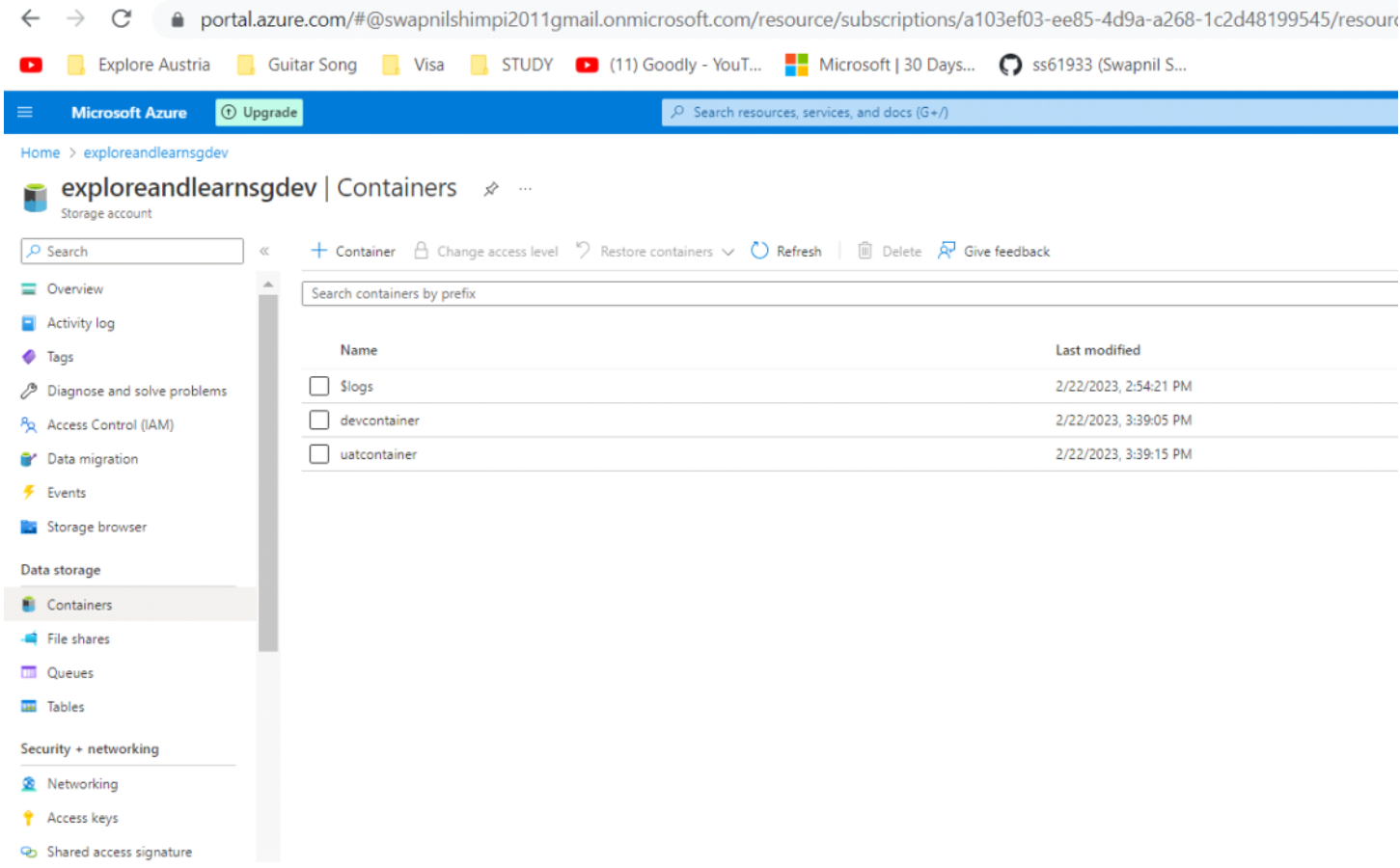
1. Click on create storage account and provide name exploreandlearnsgdev
2. Select Resource group name **ExploreandLearnRGDev**
3. Select the region closer to your physical location to get better performance
4. Redundancy - For demo purpose select local option , cost efficient
5. Select the check box for **Data Lake Storage Gen2.** It will provide access to ADLS
6. Access tier - select cool for demo purpose
7. Data protection - keep retention period of 20 days for all options
8. Skip intermediate options
9. Add a tag DEVSG



Step3 Create storage containers

Wednesday, February 22, 2023 3:36 PM

1. Navigate to Storage **exploreandlearnsgdev**
2. Click on container option on left side and add below 2 containers

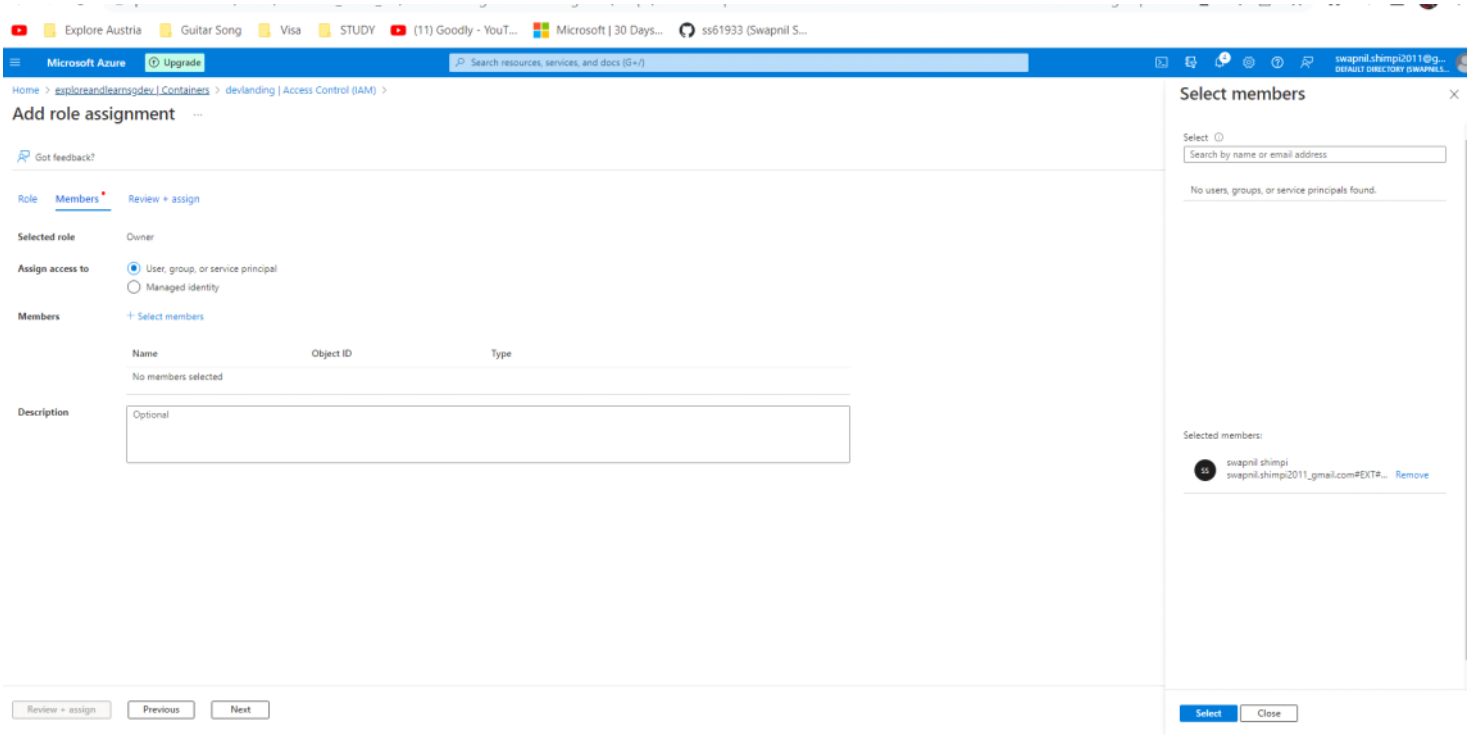


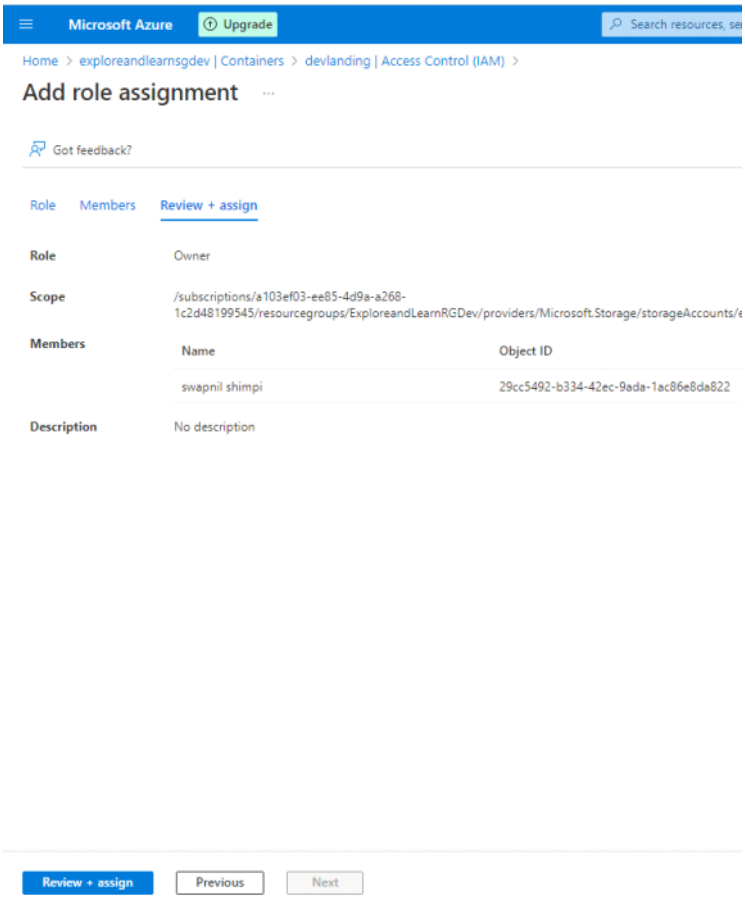
Theory - storage can be of Blob type or ADLS type
Blob will not have sub directory

Step4 Manage access to Container

Wednesday, February 22, 2023 3:45 PM

1. Select storage ->container->Access control (IAM)-> Role Owner ->
2. Select he member from left menu. (we can grant a role to groups also if present)

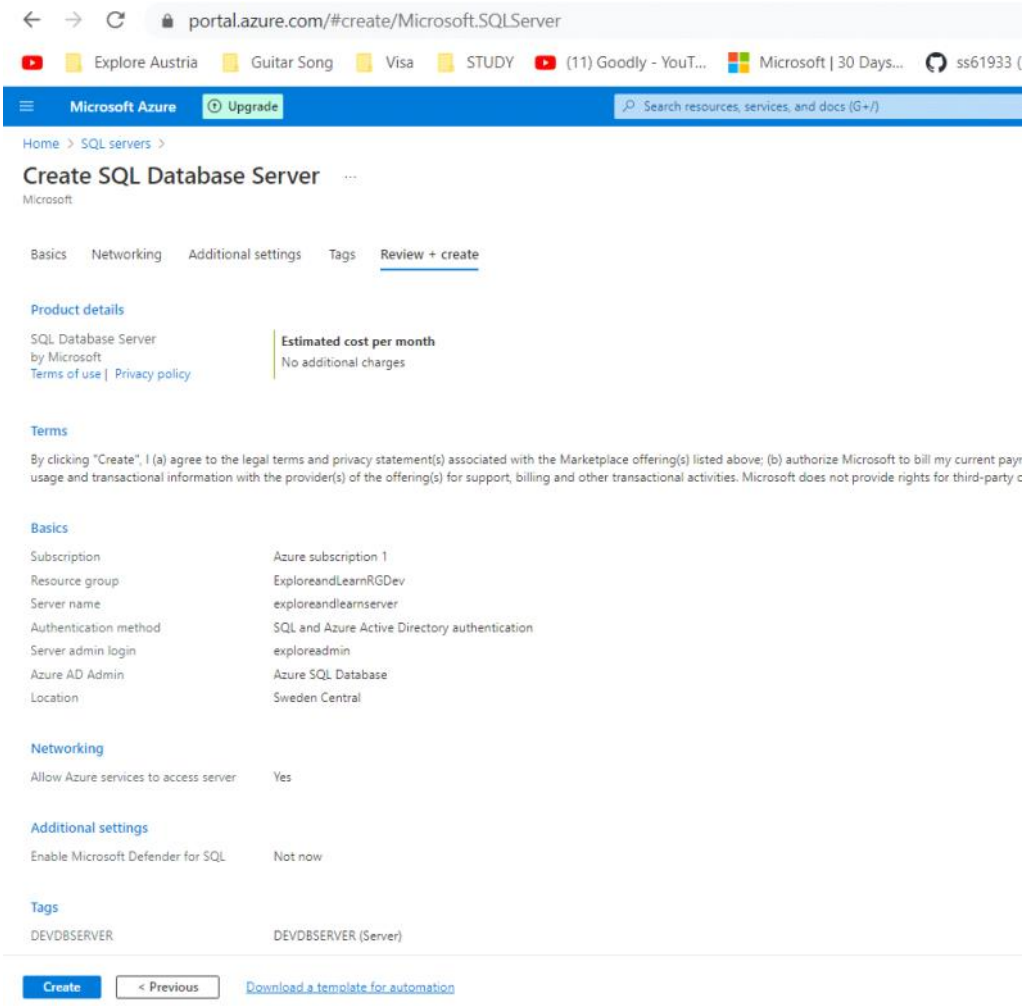




Step5 Create Sql server and database

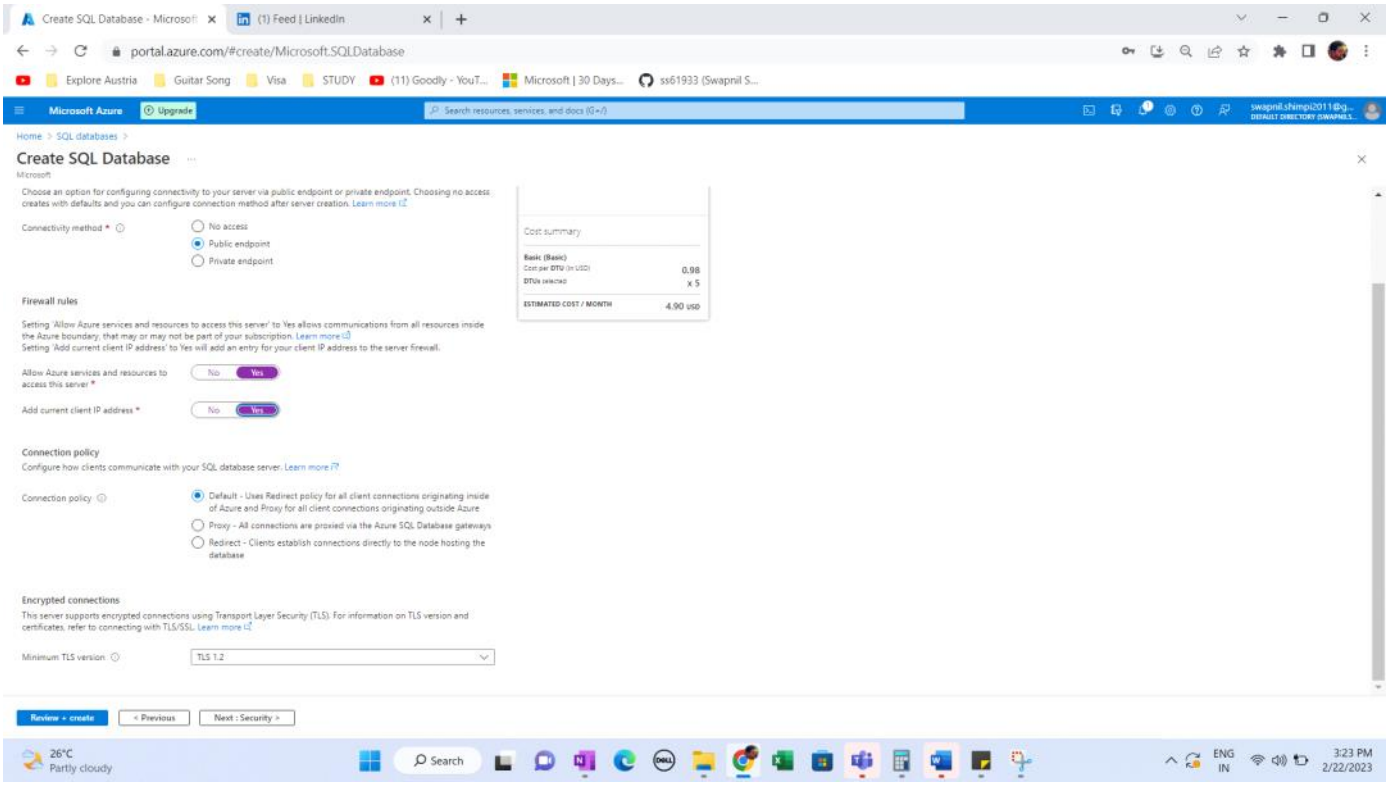
Wednesday, February 22, 2023 2:56 PM

1. Click on create Sql server , name the server **exploredbserver**
2. Select Resource group name **ExploreandLearnRGDev**
3. Select the region closer to your physical location to get better performance
4. Select " Use both SQL and Azure AD authentication " authentication method
5. Add server admin credentials- exploreadmin / Ss0061933!
6. Select Azure Active Directory as a "Azure SQL Database"
7. Set firewall rule "Allow Azure services and resources to access this serve" - YES
8. Add a tag DEVDBSERVER
9. Skip intermediate options
10. Add a tag DEVSG

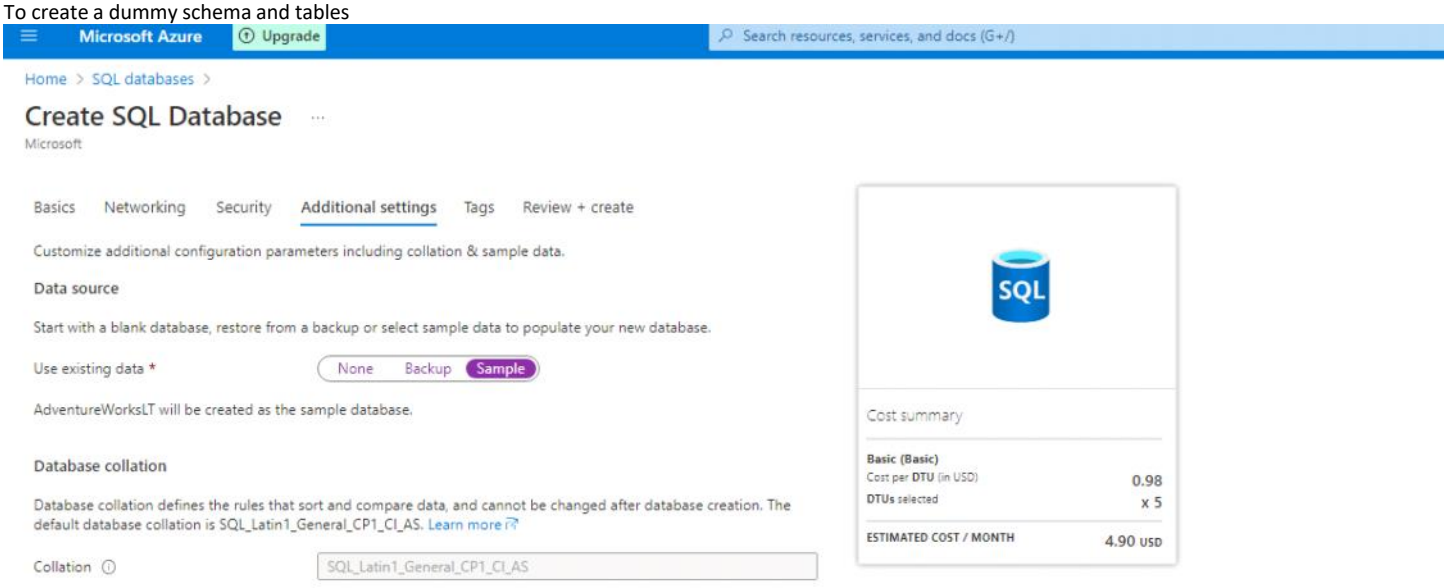


CREATING DATABASE

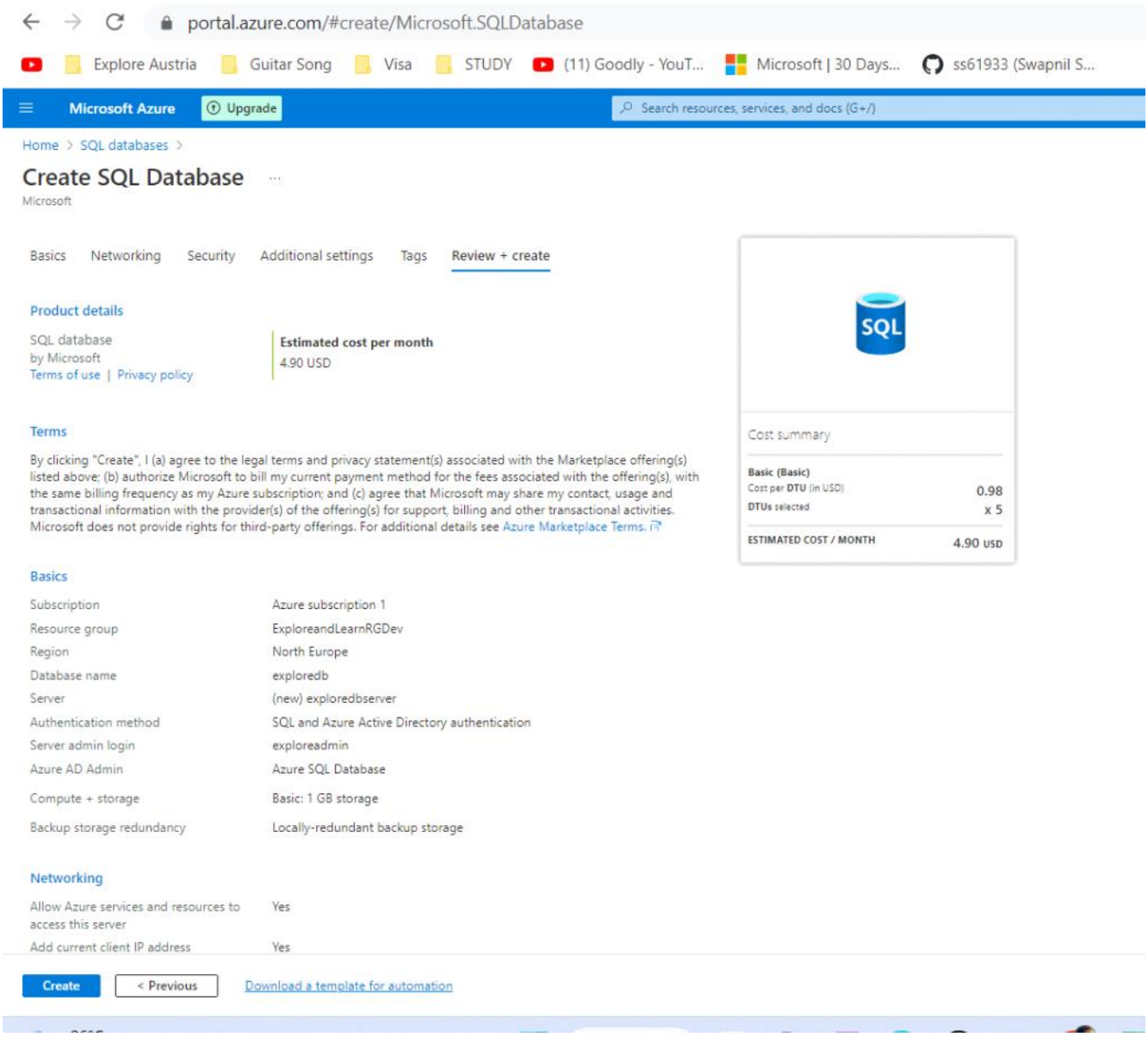
11. Click on create Sql database, name the server exploredbserver **under above server**
12. Service tier - Basic and Locally-redundant backup storage
13. Networking as below



14. To create a dummy schema and tables

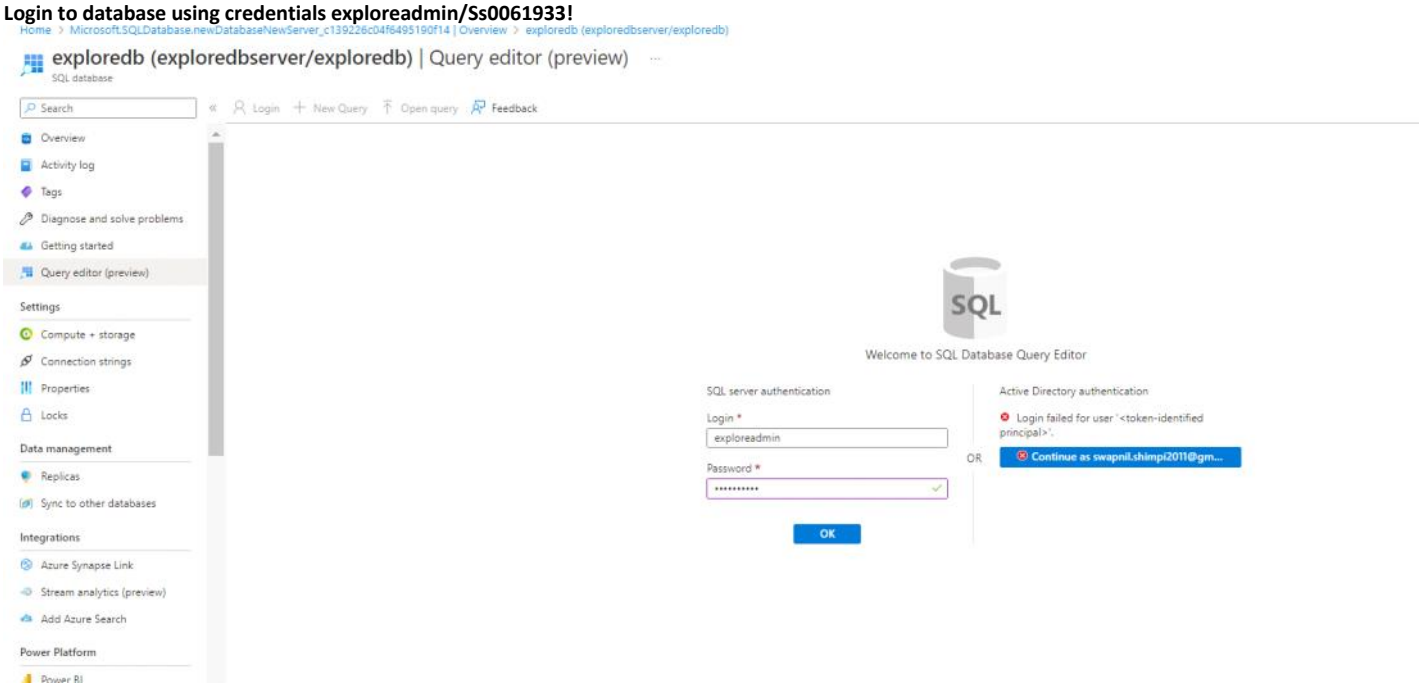


15. Add a tag DEVDB
16. Skip intermediate options
17. Add a tag DEVSG

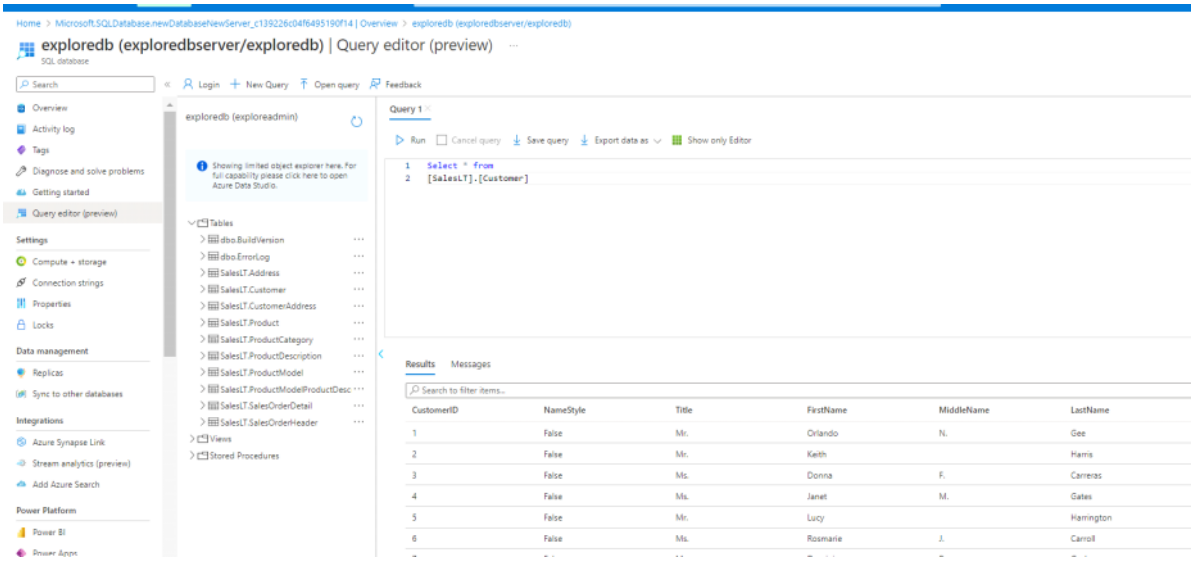


Validating database

1. **Login to database using credentials exploreadmin/Ss0061933!**



2. Query the database



Step6 Uploading files to ADLS account

Wednesday, February 22, 2023 4:16 PM

1. Select a storage account - > container - > sub container as shown below and uplaod a file