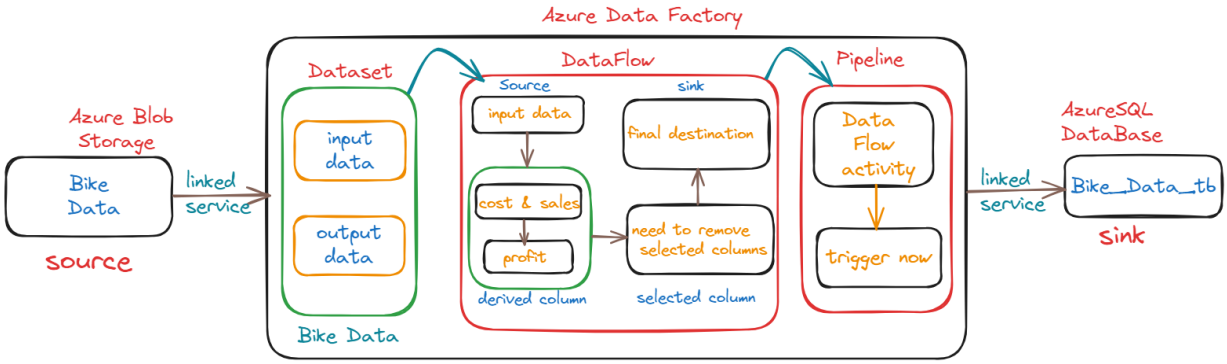
Azure Data Factory Pipeline – I

Architecture of the Pipeline -I



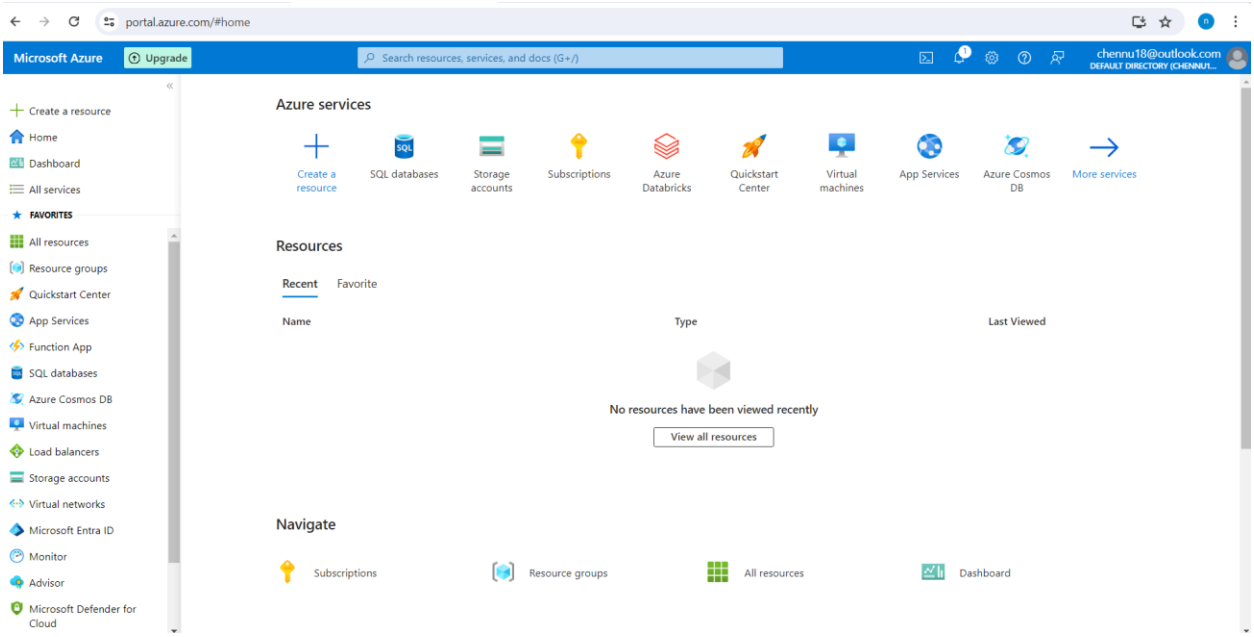
Azure services to be Used:

1. Azure Data Lake Storage
2. Azure SQL DataBase
3. Azure Data Factory

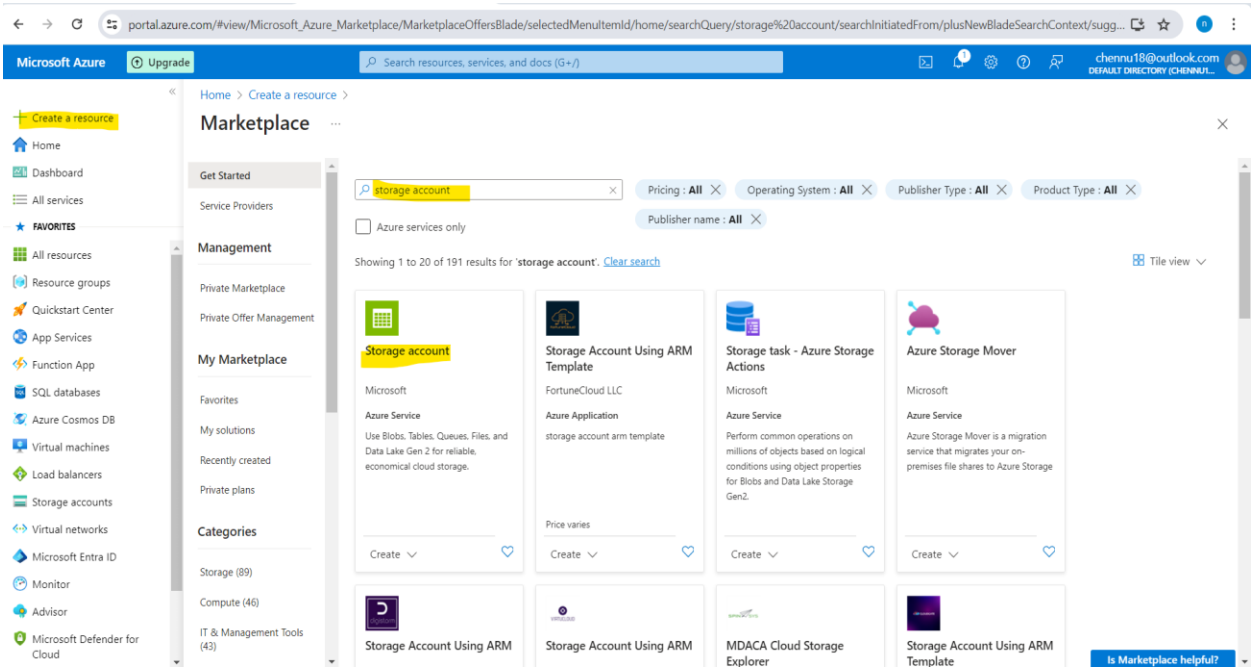
Creating Azure Data Lake Storage service from portal.azure.com

➔ Login to azure portal using this link https://portal.azure.com/

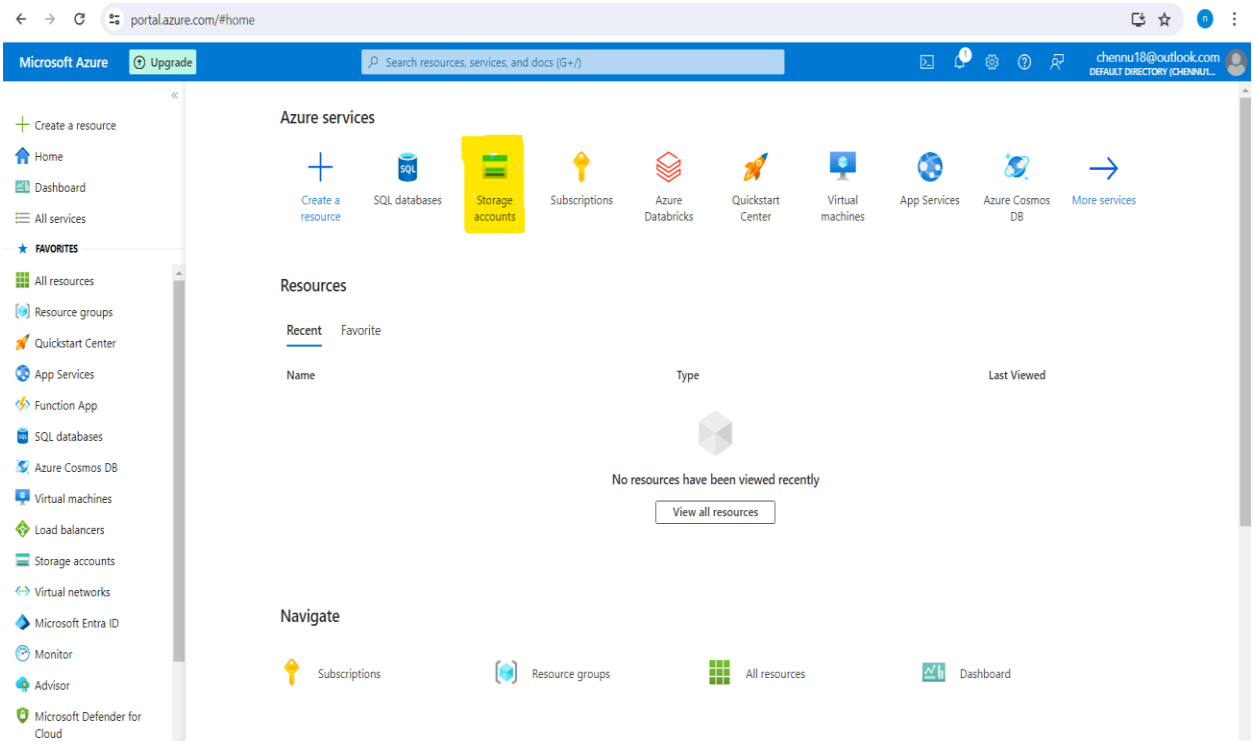
➔ After successful login you will able to see home page like below



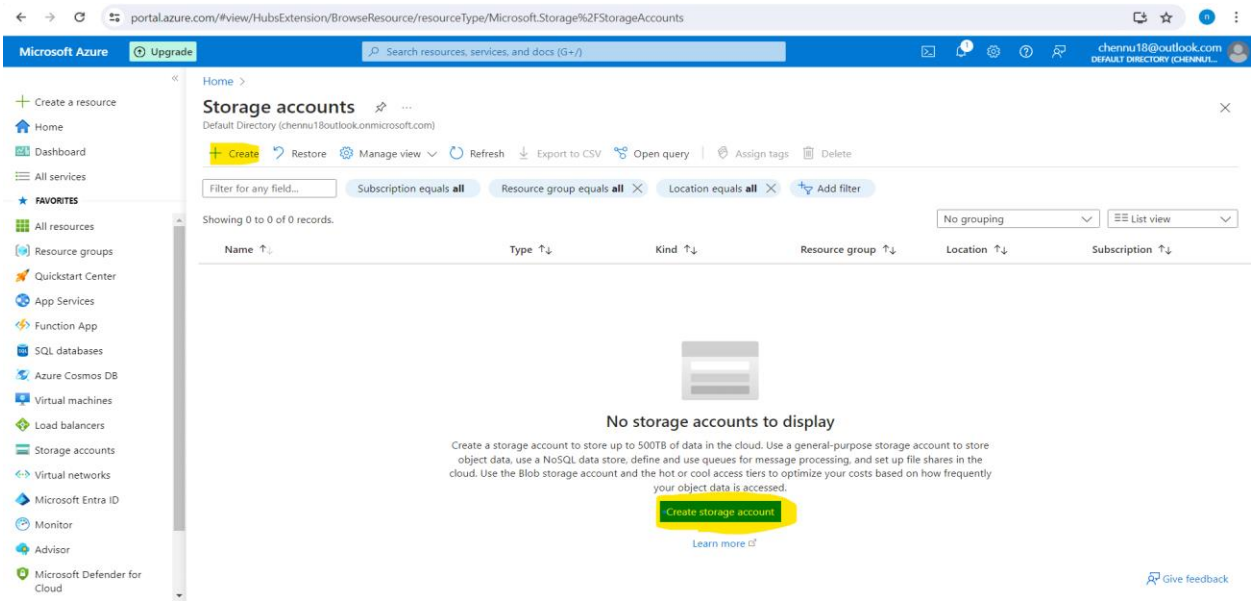
➔ To create azure storage service from Azure market place



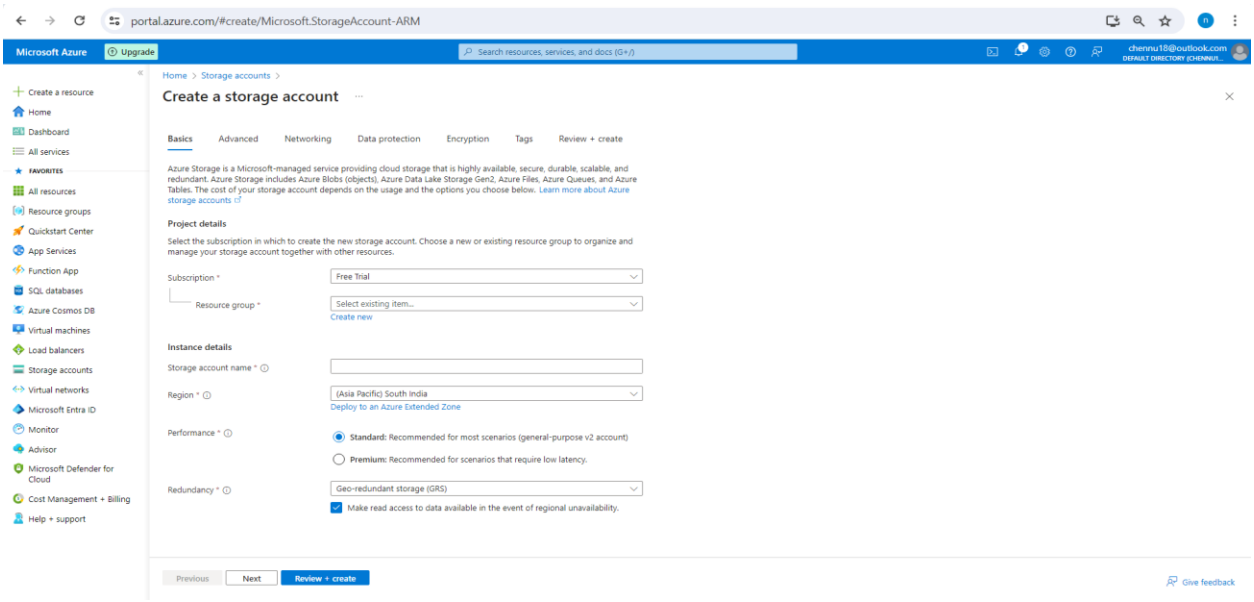
➔ To create azure storage service from Home Page directly click on storage account icon



➔ There are two ways to create storage account service see in below figure



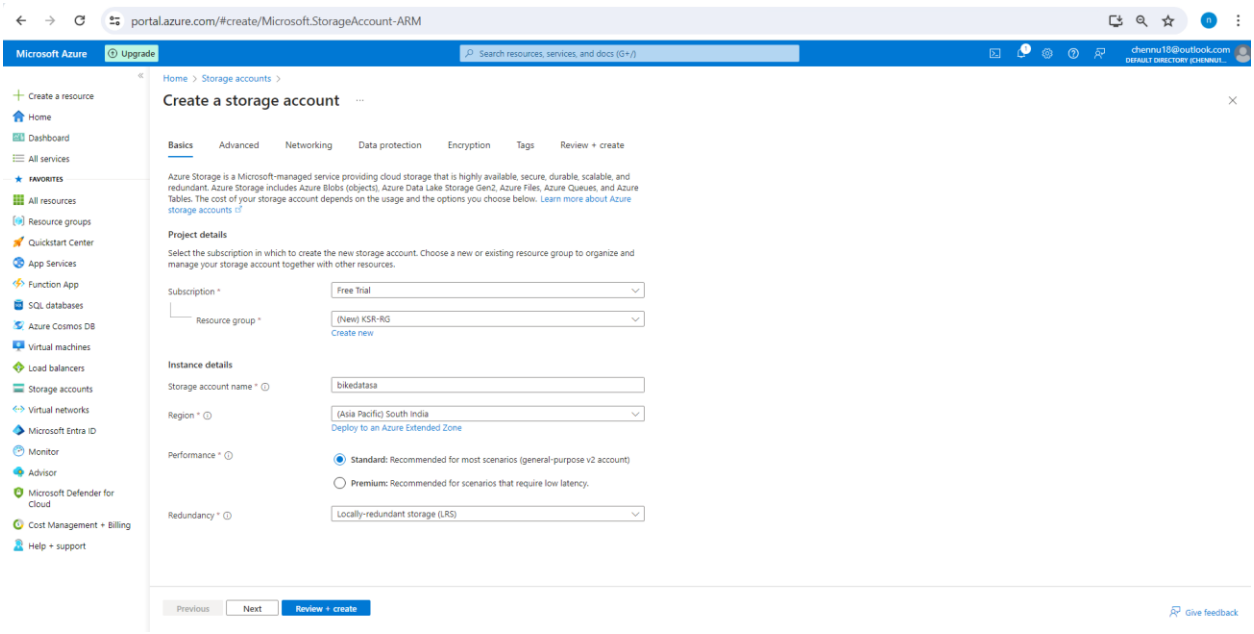
➔ After click on “create or create storage account new page will open like below



➔ **We need to fill the required Basic details**

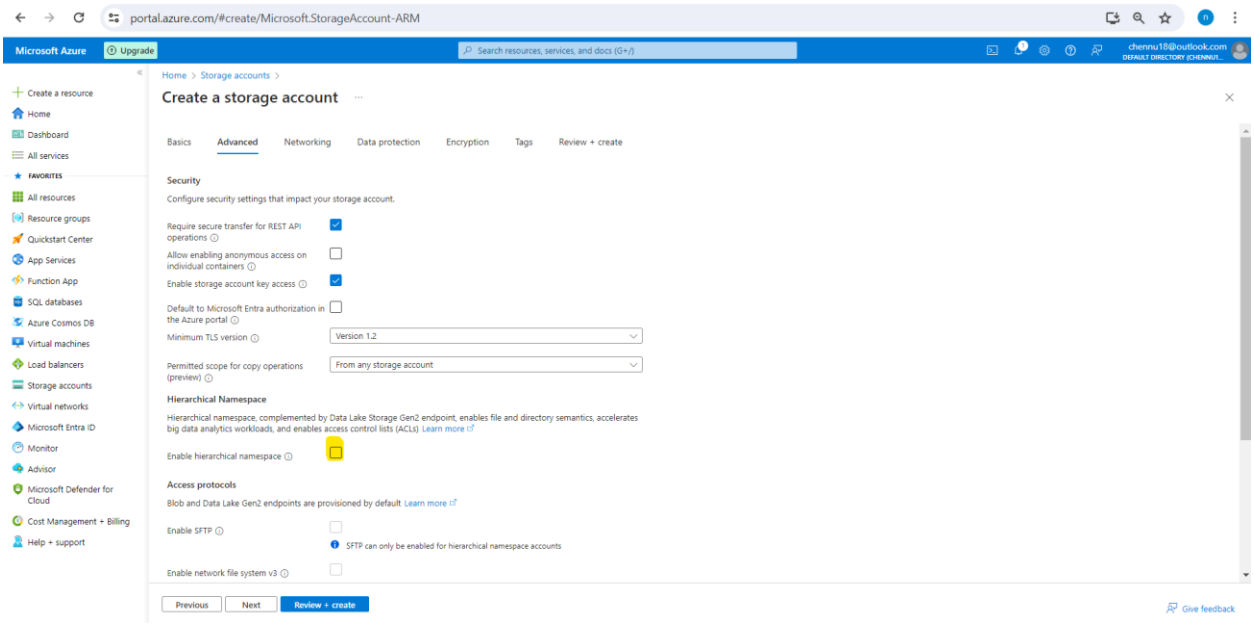
1. Subscription: Free or pay as you go (if you don’t have free subscription go with pay as you go)
2. Resource Group: Choose resource group if exists already if not create new resource group itself there

3. Storage Account Name: It must be unique across all Azure storage account names. It allows only lowercase letters and numbers, and its length must be between 3 to 24 characters. It doesn't allow uppercase letters or any special characters.
4. Region: Choose your preferred region
5. Performance: Choose "Standard" and "Premium" performance tiers based on your specific requirements.
6. Redundancy: Choose the redundancy level based on the criticality of your data and your business requirements.
    a. locally redundant storage (LRS),
    b. geo-redundant storage (GRS),
    c. zone-redundant storage (ZRS)



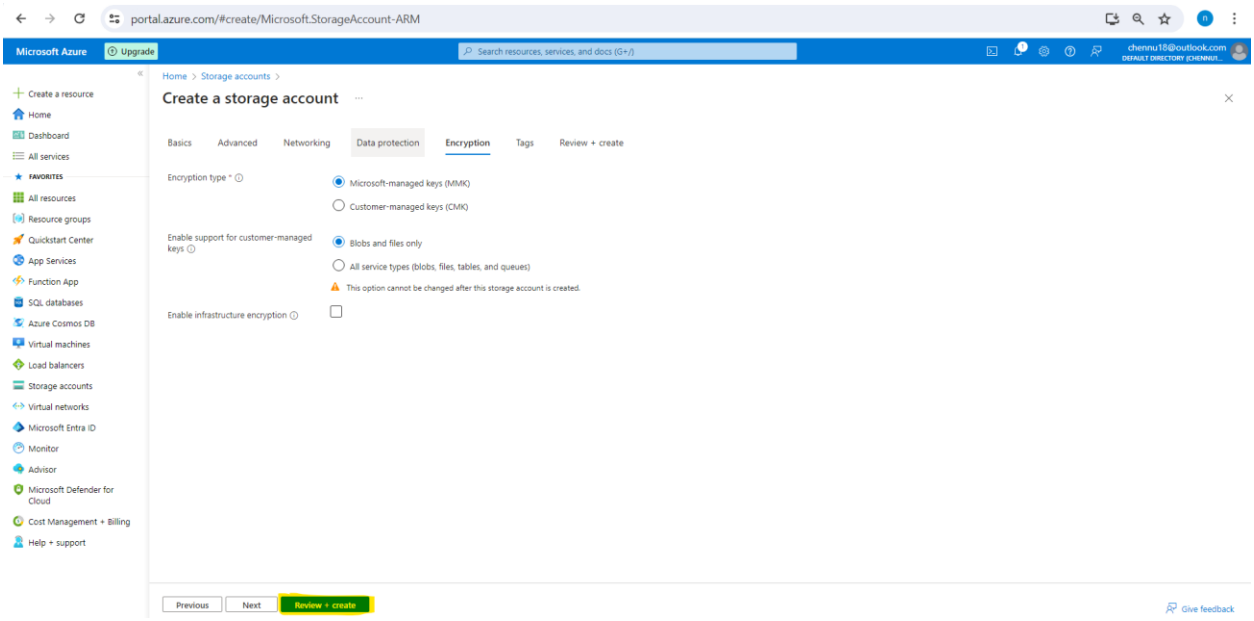
➔ Click on Next Button it will open advanced page like below

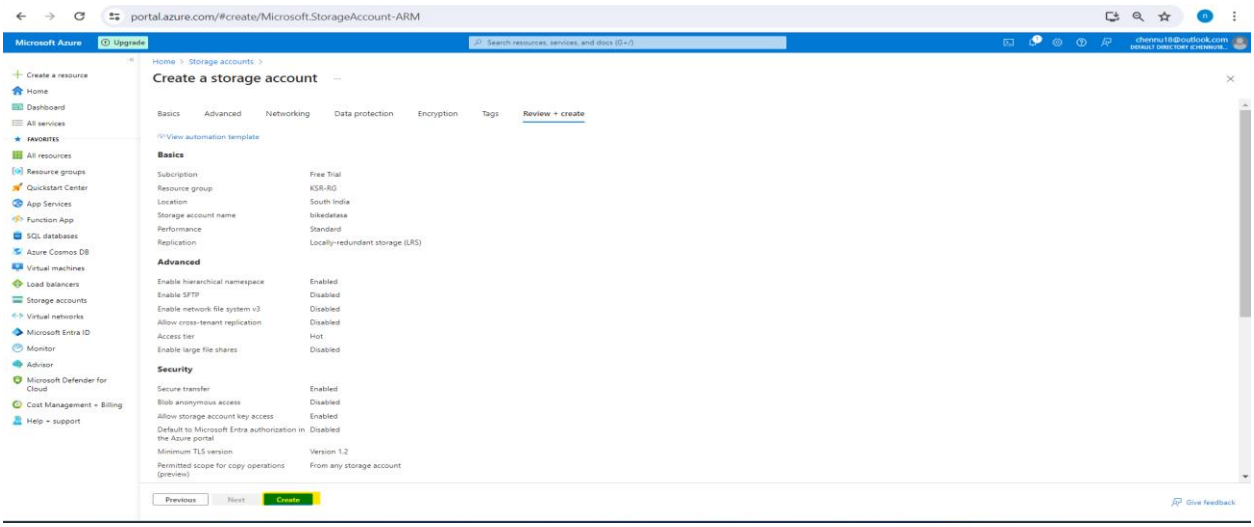
1. "In this, if the option 'Enable hierarchical namespace' is disabled, it should be a normal Azure Storage Account."
2. "In this, if the option 'Enable hierarchical namespace' is enabled (checkbox checked), it should be an Azure Data Lake Storage Account."
3. "The above two statements differentiate between an Azure Storage Account and an Azure Data Lake Storage Account based on the 'Enable hierarchical namespace' option."

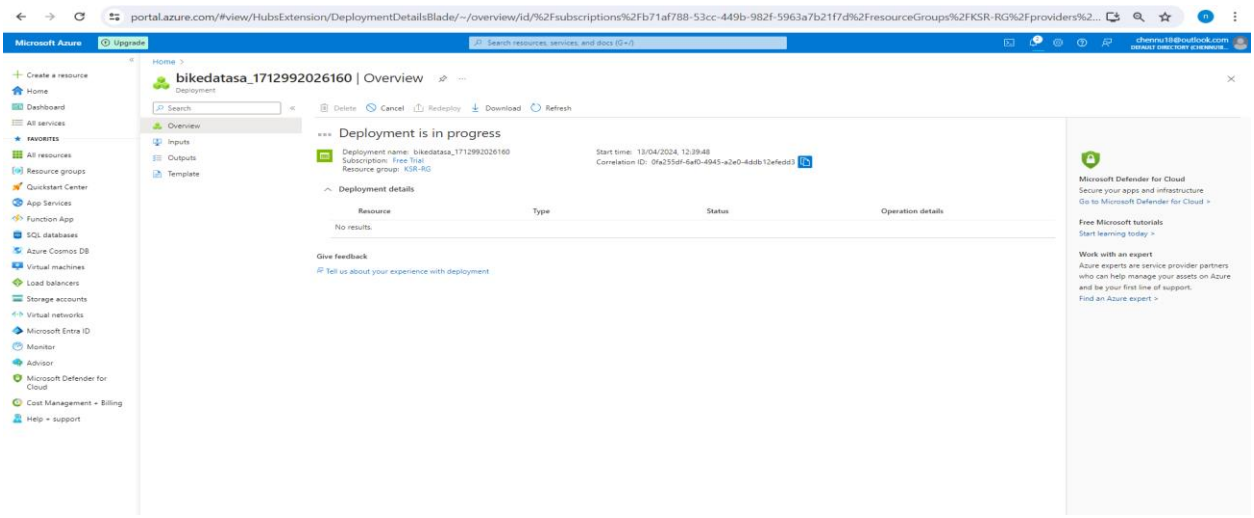


➔ Rest of the options we go with all default values.

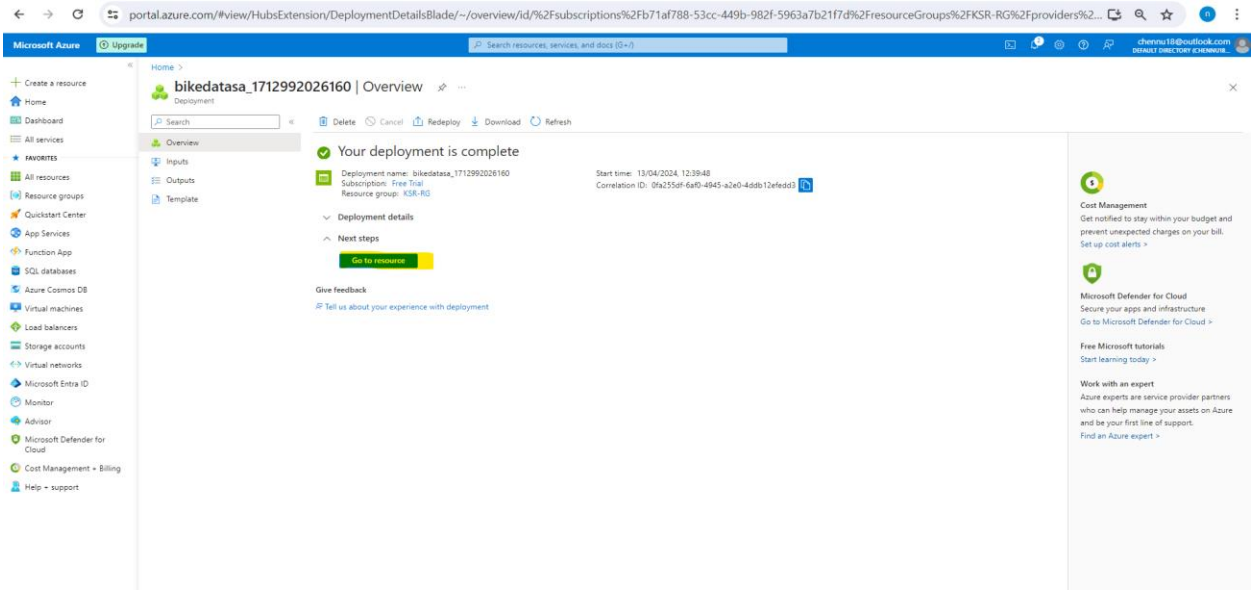
➔ Finally click on review and create button to validate the details and we can review then click on create button



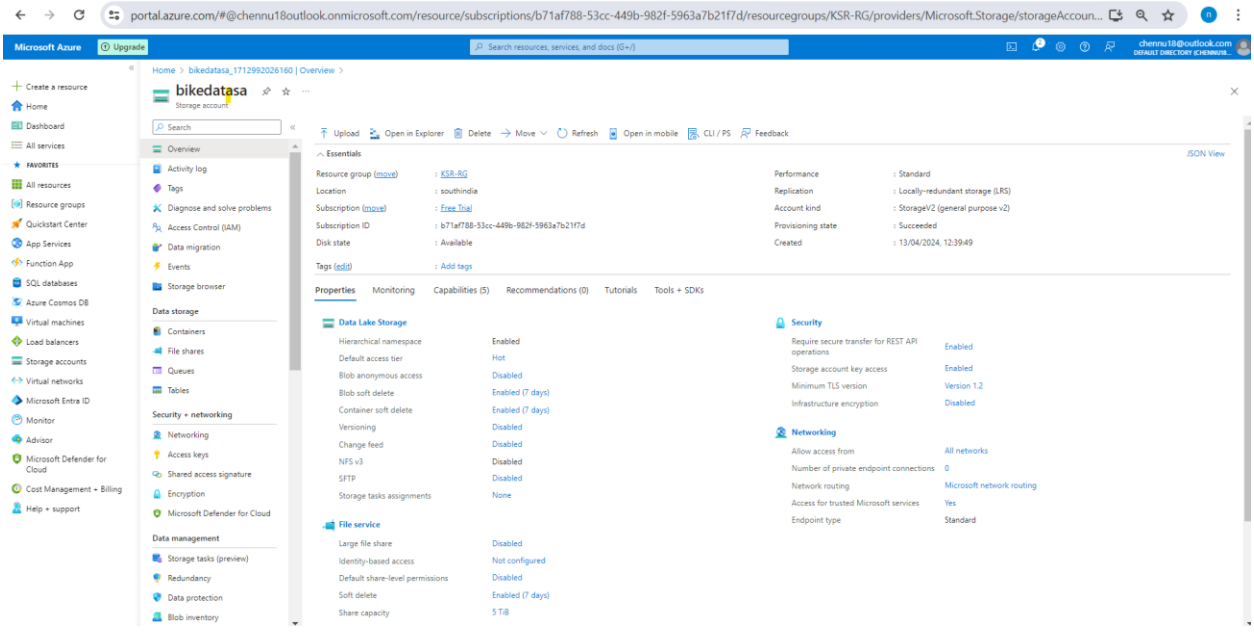




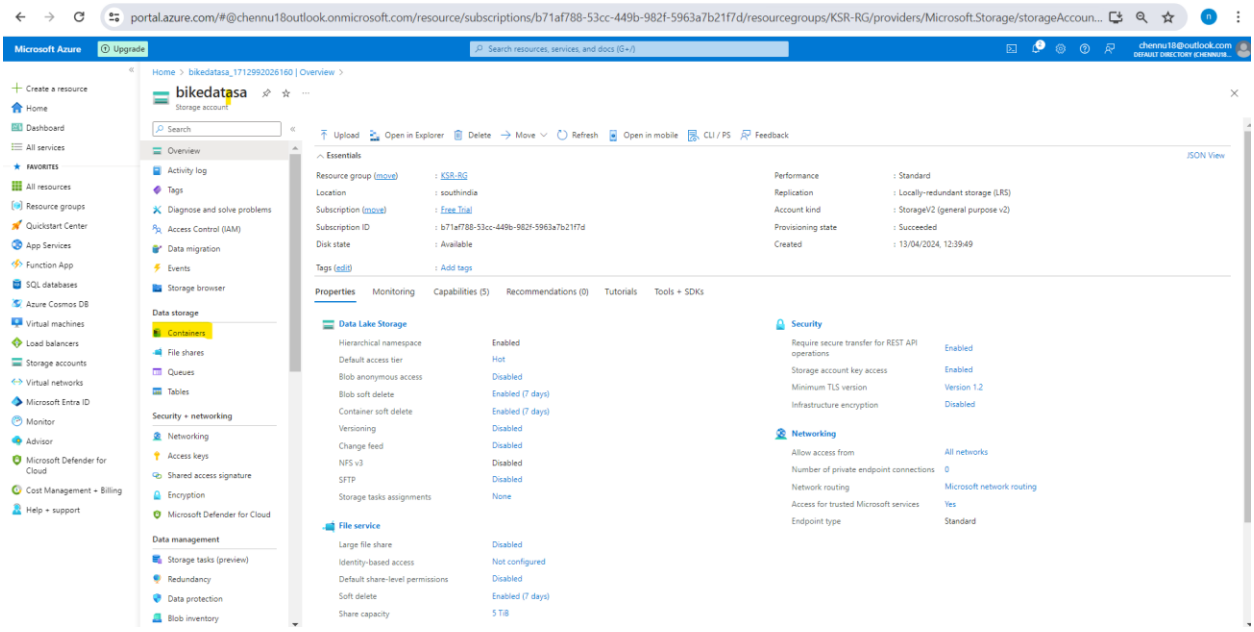
→After successfully create a service it will show like below fig

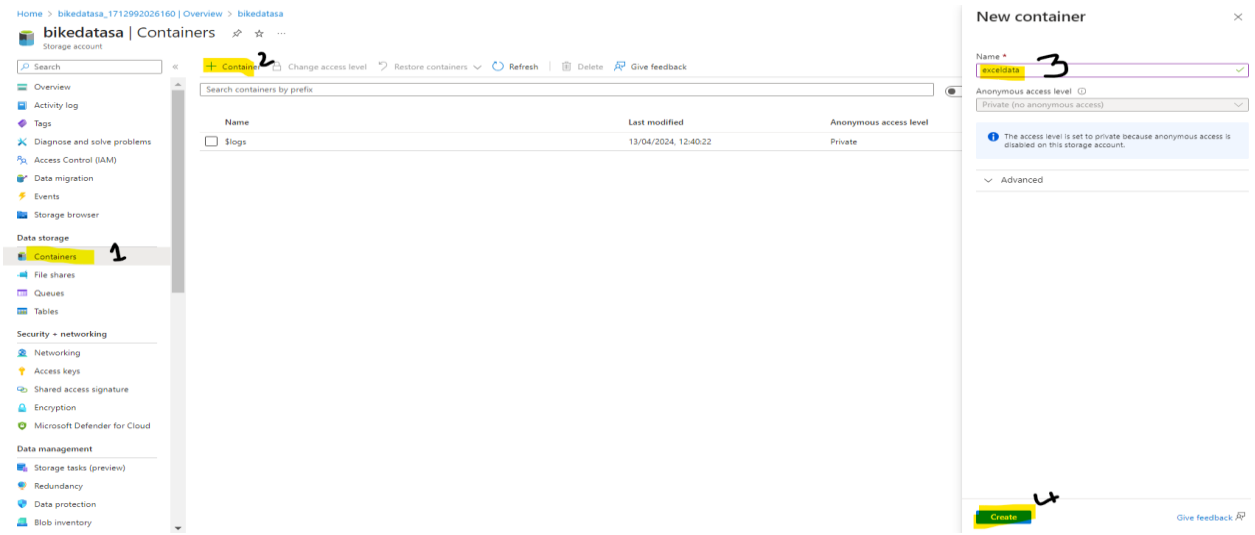


➔ Click on GotoResource page will open like below fig

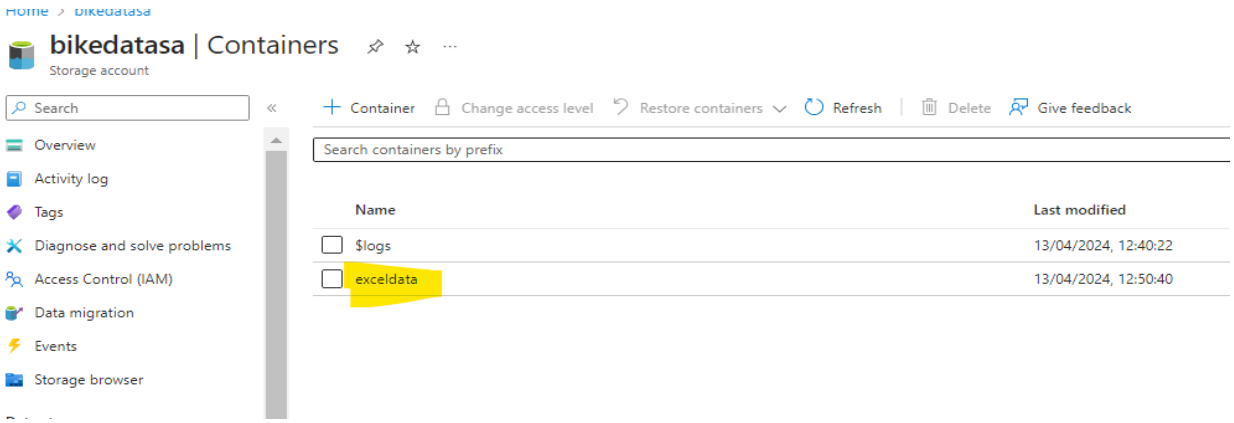


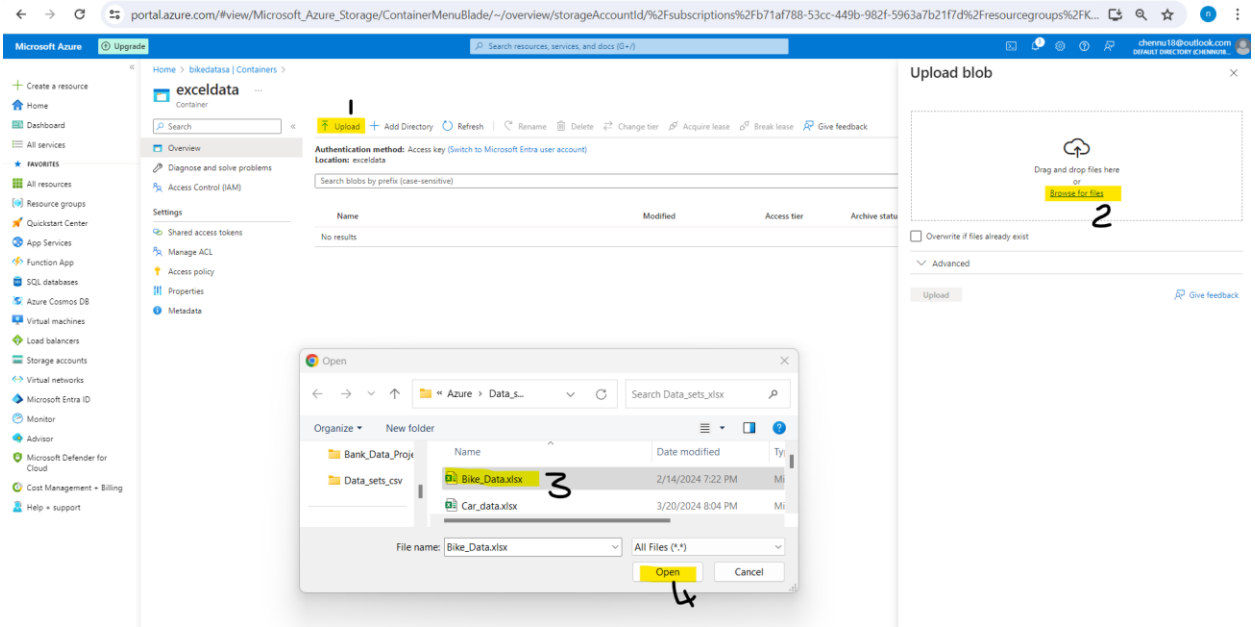
➔ We need to create a containers, just click on containers it will open a navigate new page



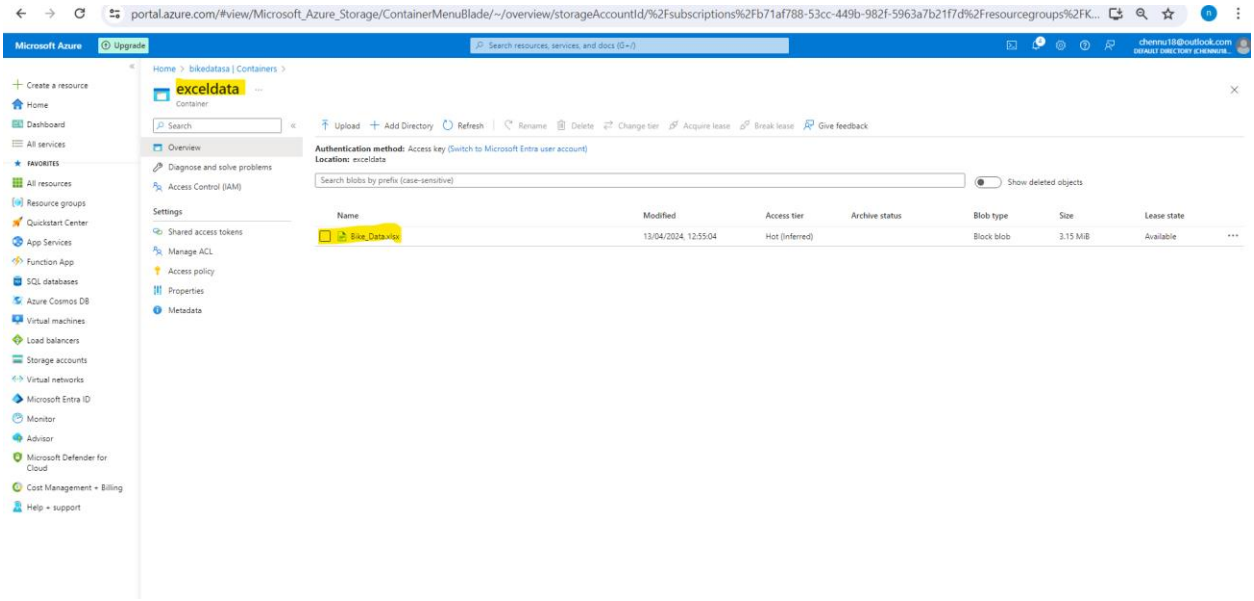


➔ Double click on excel-data it will navigate another page see below figures





➔ then click on open and click upload button you will get below screen if its uploaded successfully.

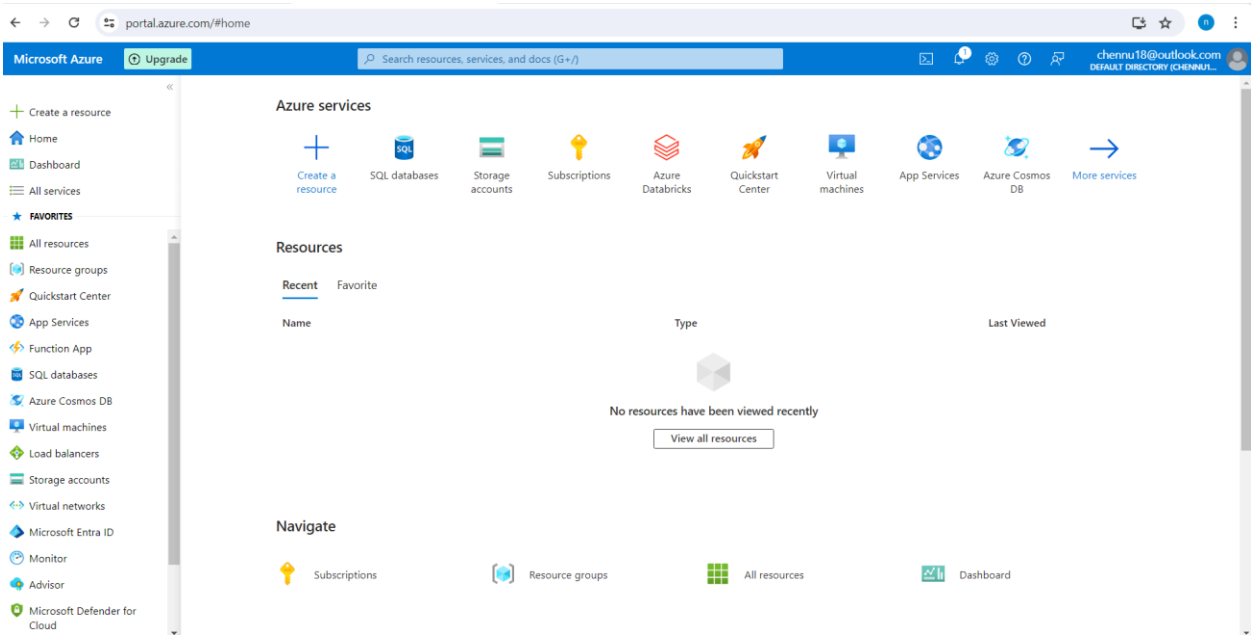


➔ Finally, Data Lake Storage Account Created successfully and uploaded bike data datset into that storage account

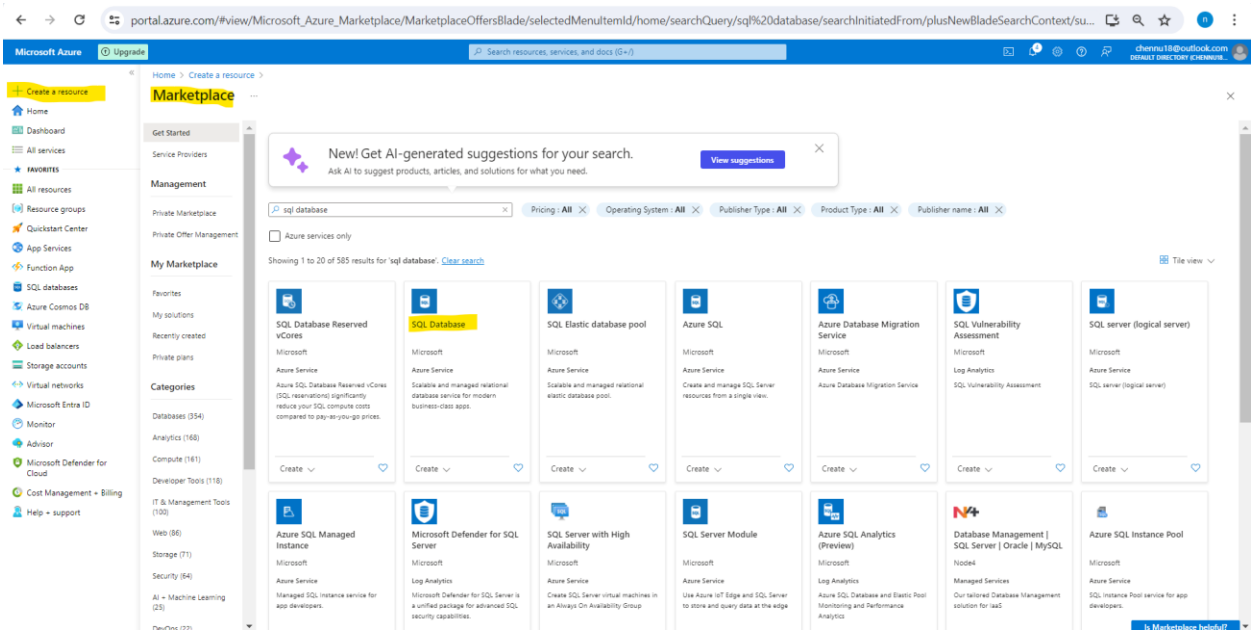
Creating Azure Data SQL DataBase service from portal.azure.com

➔ Login to azure portal using this link https://portal.azure.com/

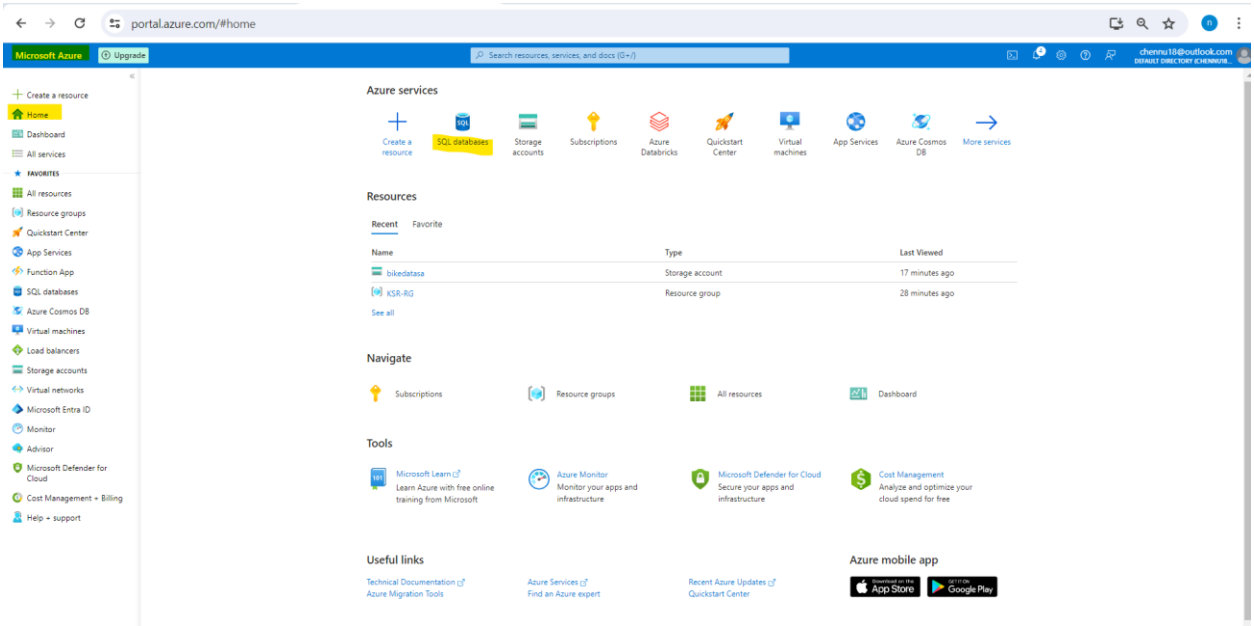
➔ After successful login you will able to see home page like below



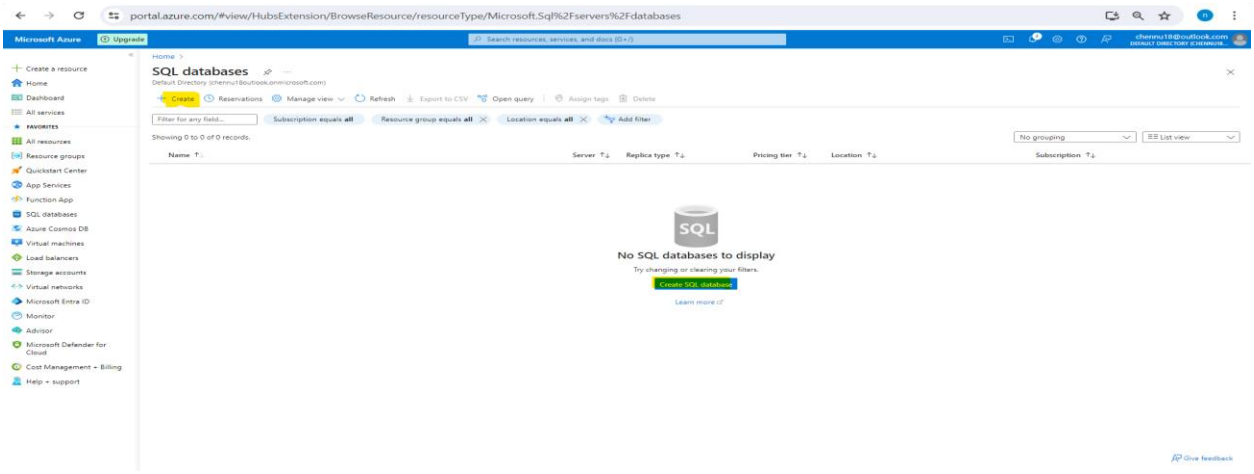
➔ To create Azure SQL Database service from Azure market place → create resource →search for SQL Database → hit enter → click on SQL Database



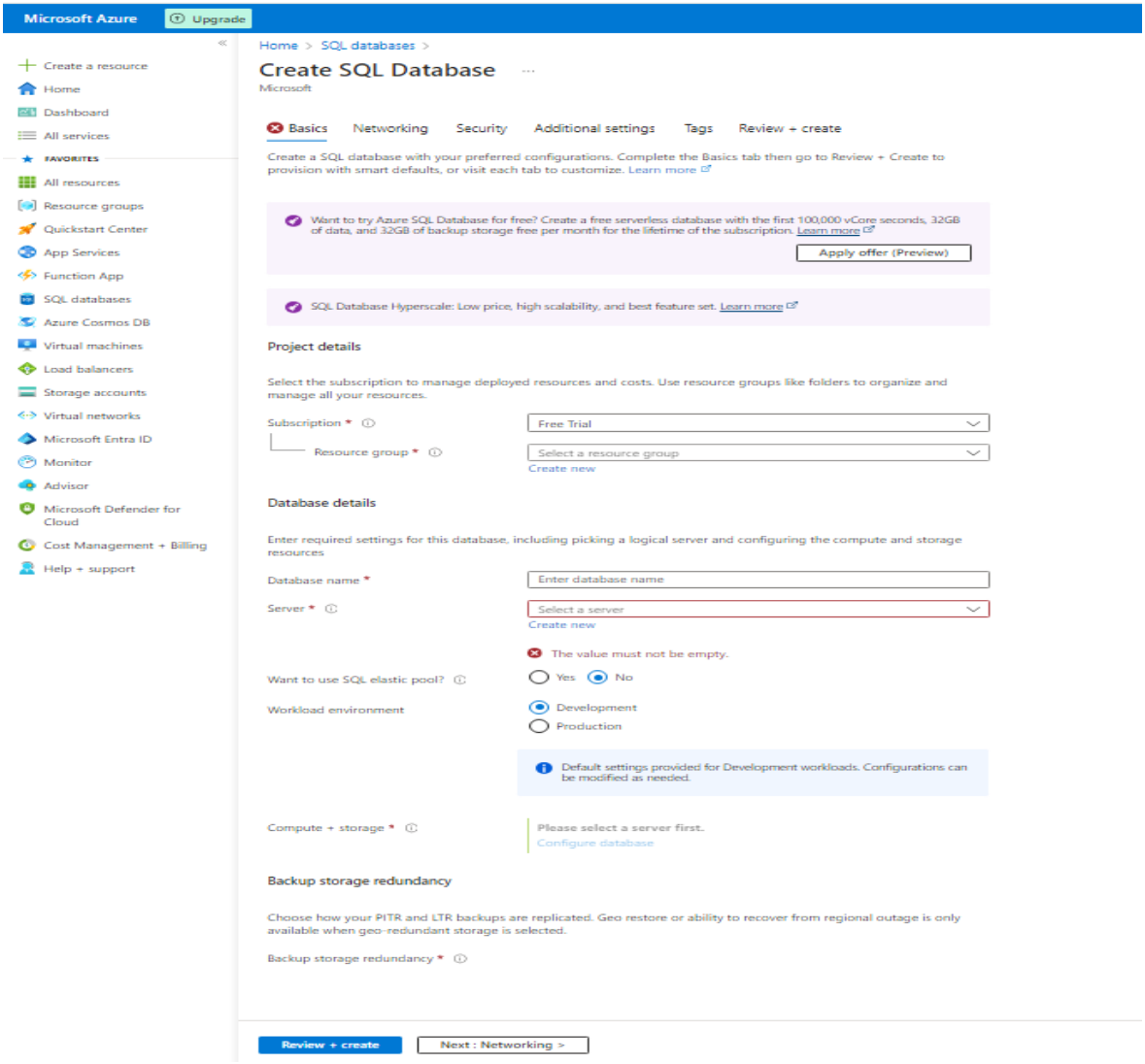
➔ To create azure storage service from Home Page directly click on storage account icon as you seen in below fig.



➔ Then click on create or create SQL Database button

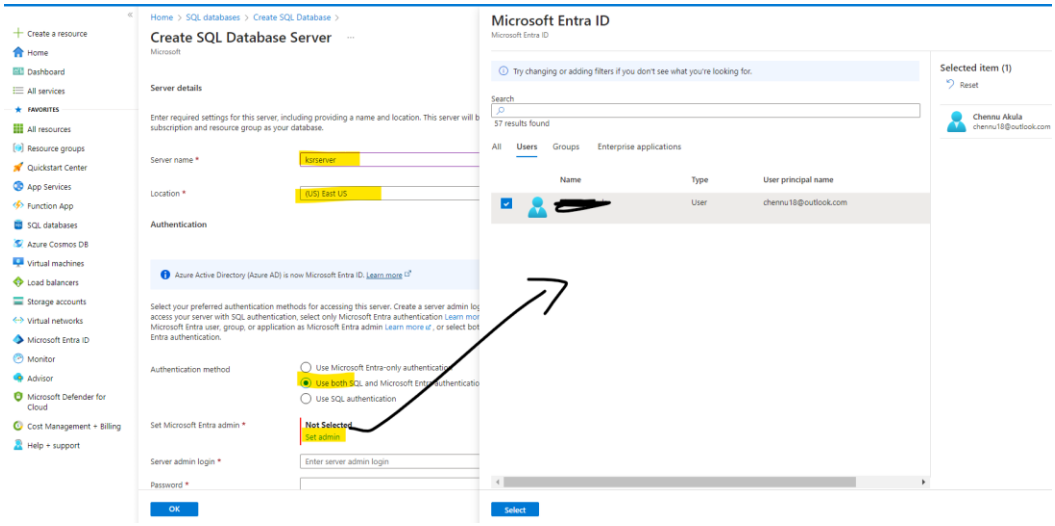


➔ After hitting create button it will navigate a new page like below

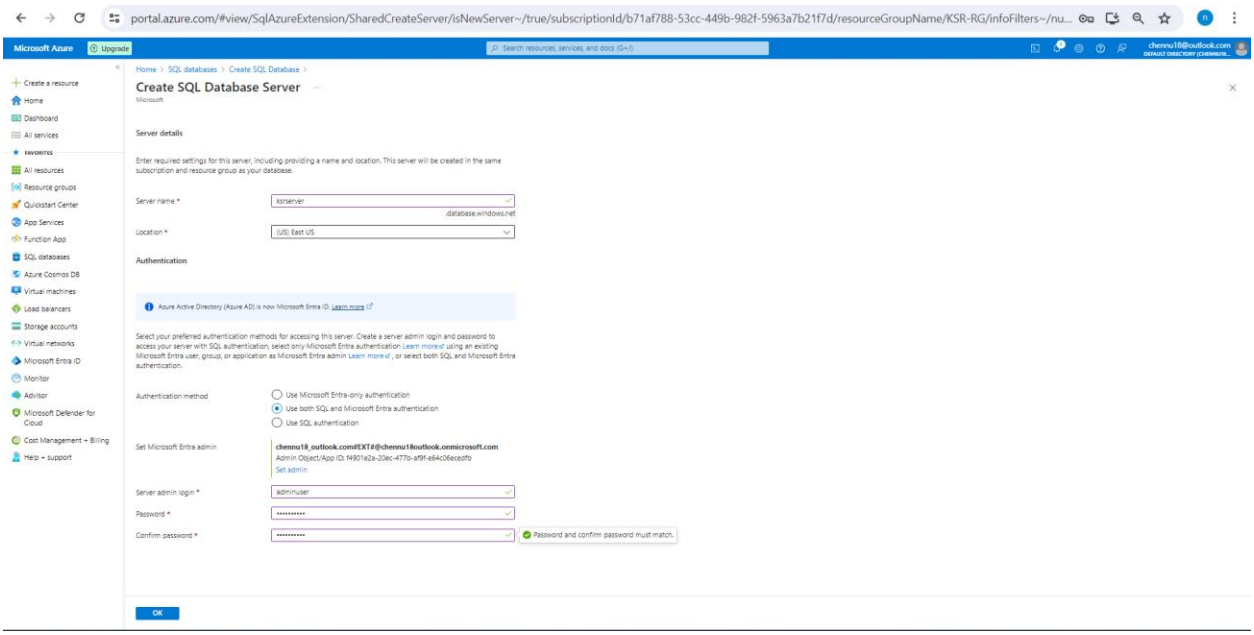


Need to fill basic details:

1. Subscription: Free or pay as you go (if you don't have free subscription go with pay as you go)
2. Resource Group: Choose resource group if exists already, if not create new resource group itself there
3. Database Name: Write your database name based on your own assumptions.
4. Server Name: if already server exists, we can use it, other wise create new server → click on server it will navigate another like below fig.--> after filling all the details click on ok button.
   a. Location : specify the preferred location
   b. Authentication: we can use both SQL and Microsoft Entra Authentication
   c. Set Microsoft Entra Admin: click on set → navigate users → choose active user as admin.



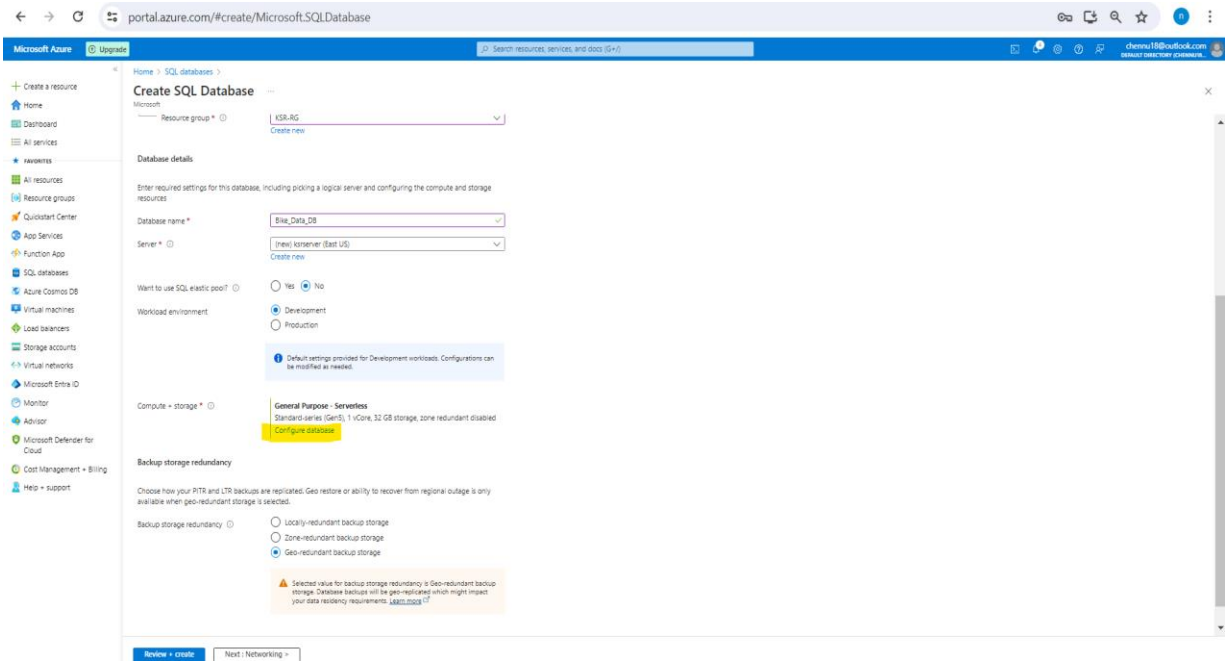
   d. Server Admin Login: provide server admin name
   e. Password: provide password, that must combination of uppercase, lowercase letters and special characters
   f. After providing all the details click on ok button same as below fig.

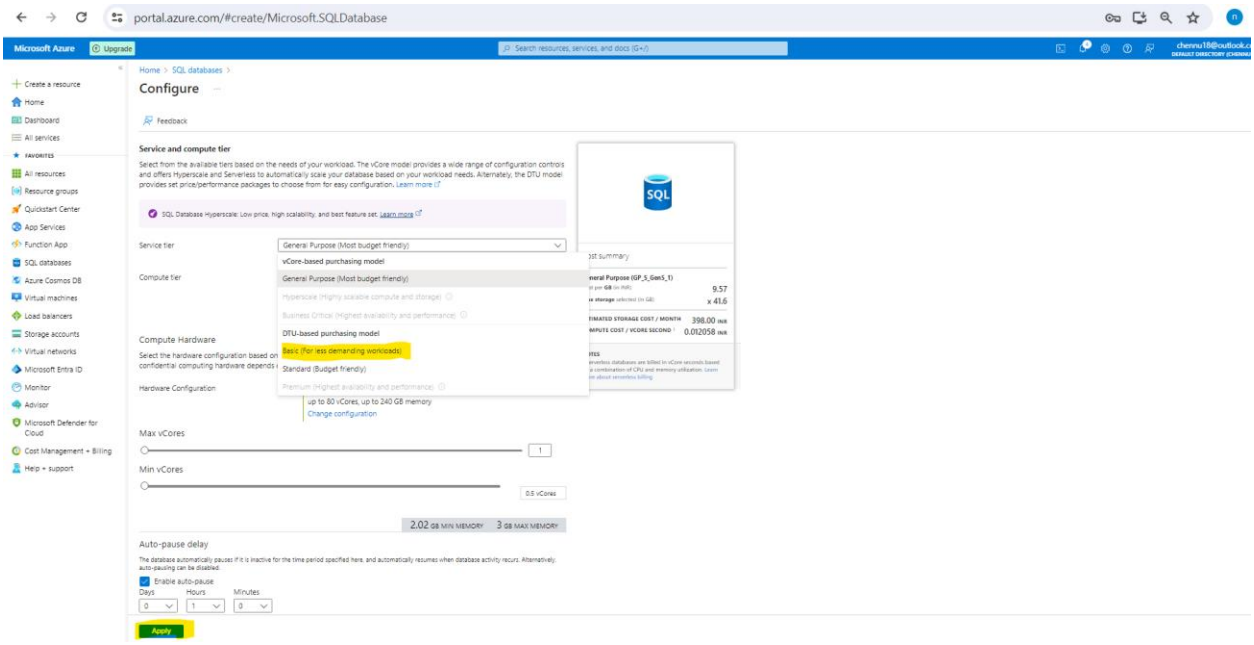


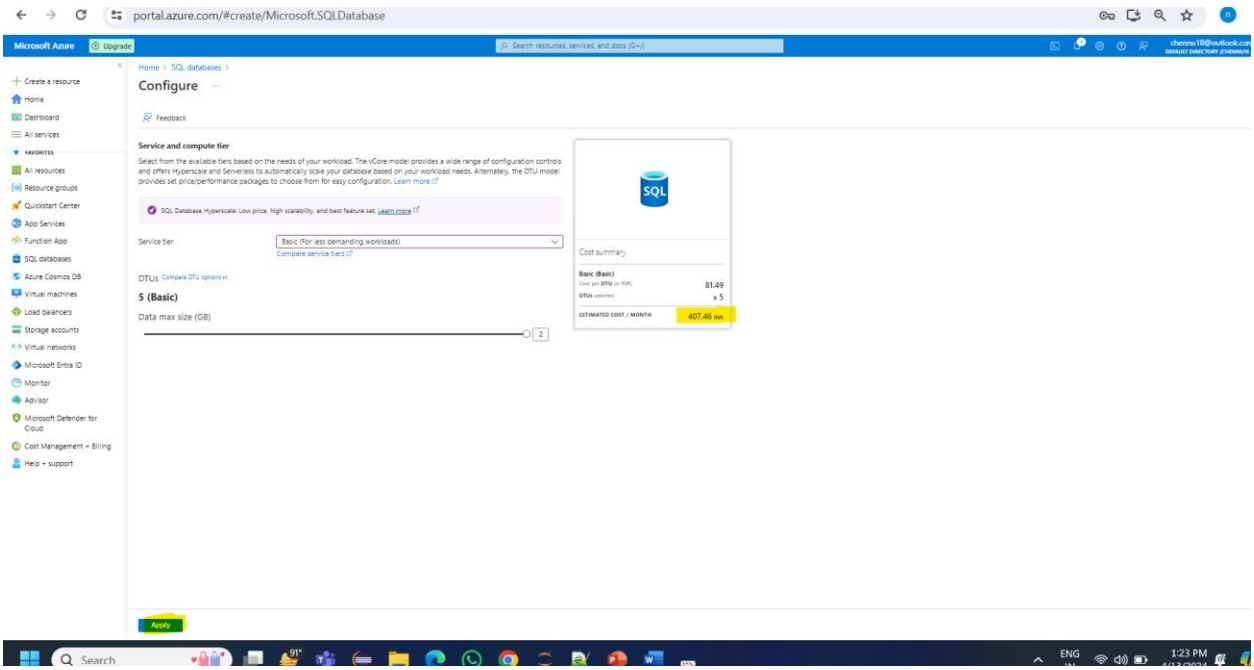
5.Want to use elastic pool: This option set to “No”

6.Workload Environment: we have two option here Development and Production, we need to choose “Development”

7. Compute Storage: click on configure database we need to select, select basic because we are using this service practice purpose, if we need to use business we have select appropriate compute storage as per our budget. Observe the below three figures

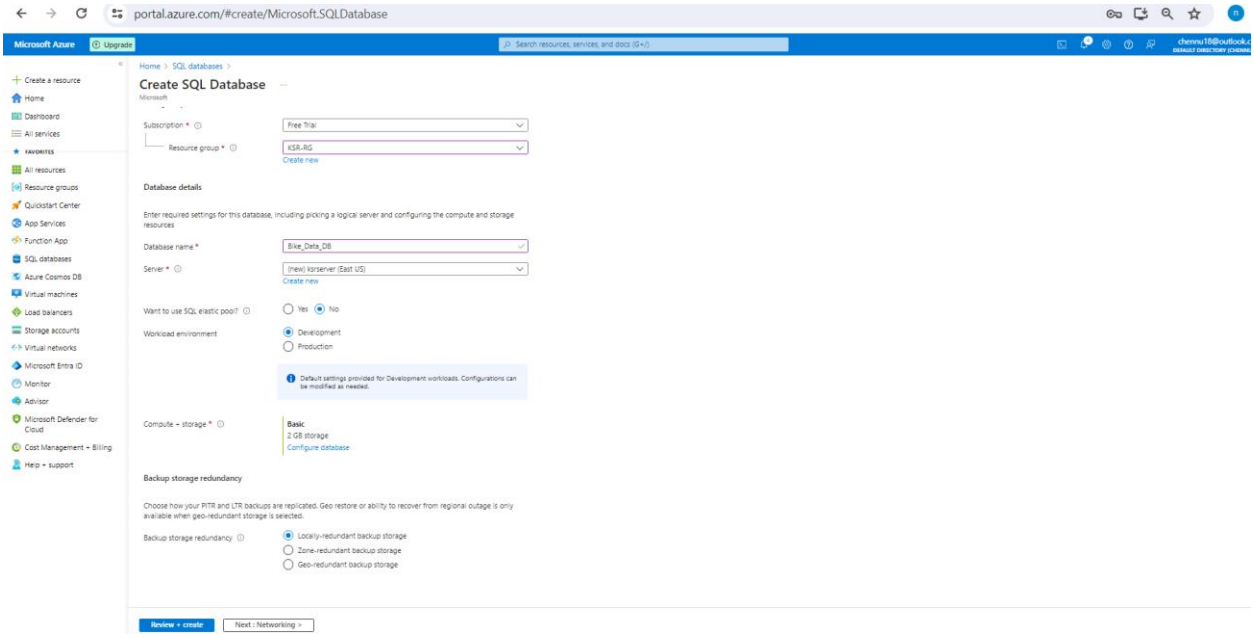






8. Backup Storage Redundancy: Choose the redundancy level based on the criticality of your data and your business requirements.
    d. locally redundant storage (LRS),
    e. geo-redundant storage (GRS),
    f. zone-redundant storage (ZRS)

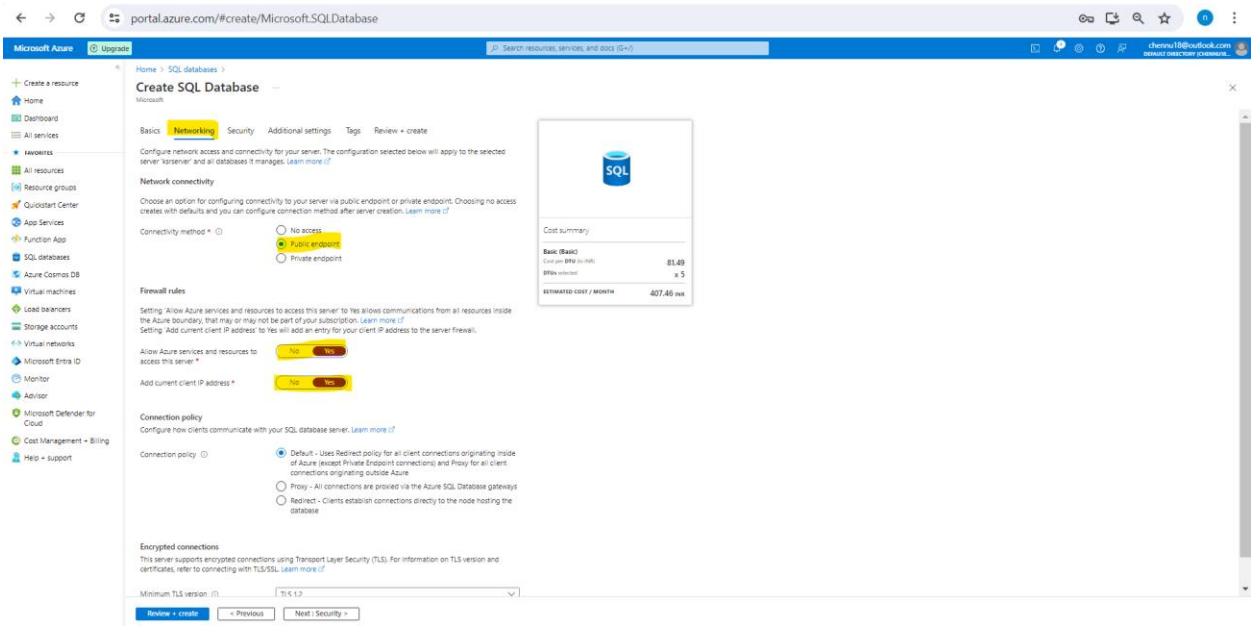
➔ After provide all details in basic tab click next button to set up networking



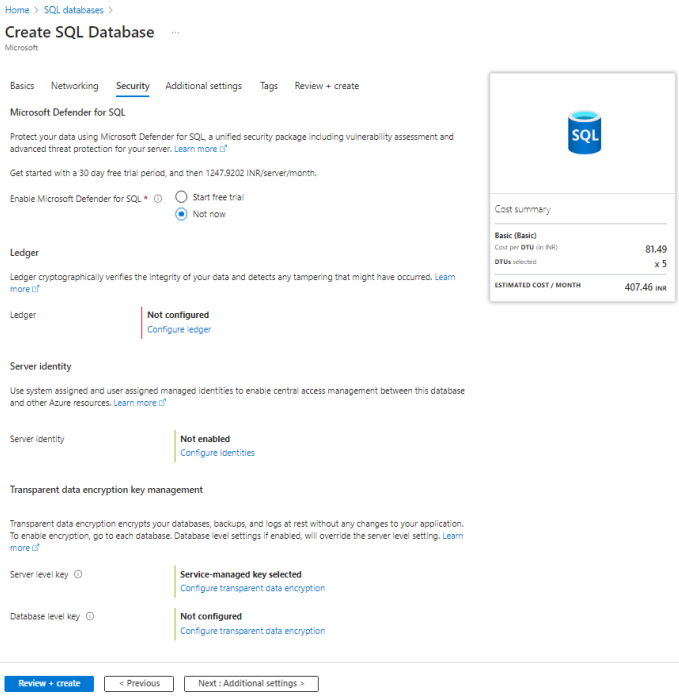
➔ In networking tab, we need to enable the public endpoint in connectivity method

➔ In firewall rules needs to enable,

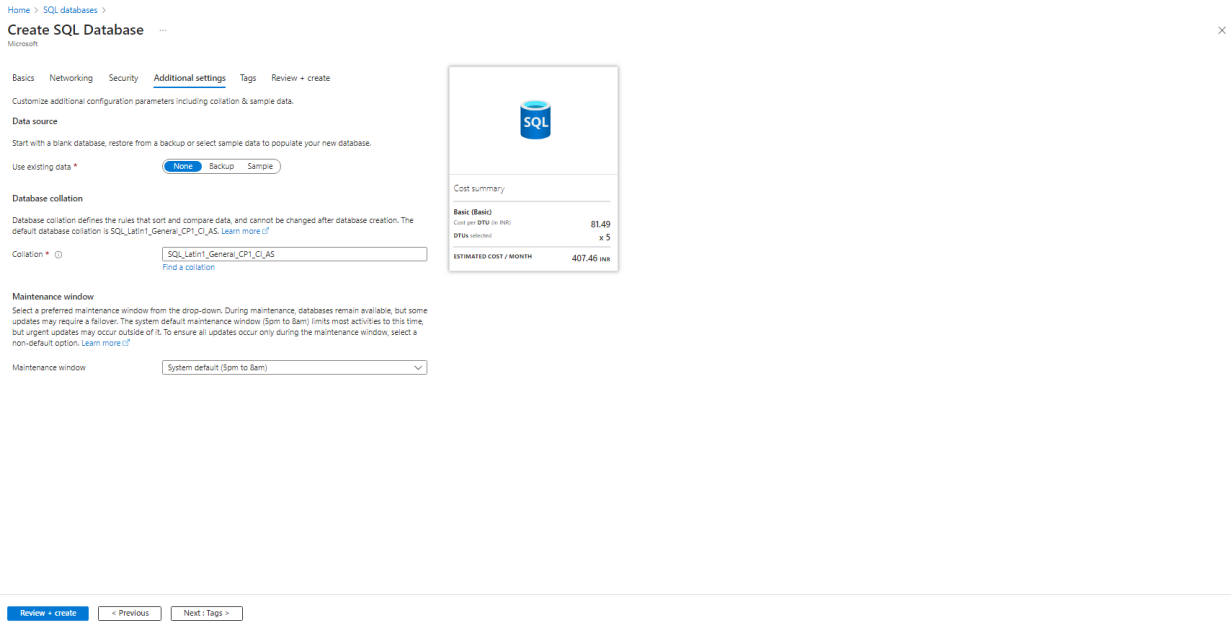
1. Allow Azure services and resources to access this server
2. Add current client IP address. as you seen like below fig



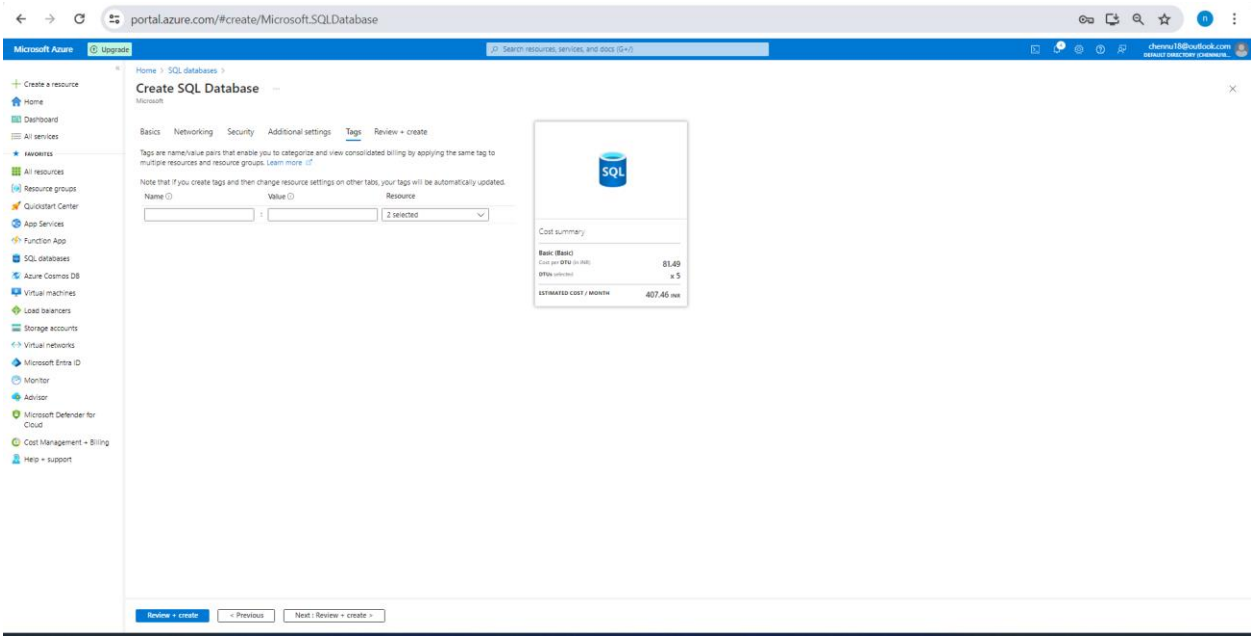
➔ In security tab go with all the defaults



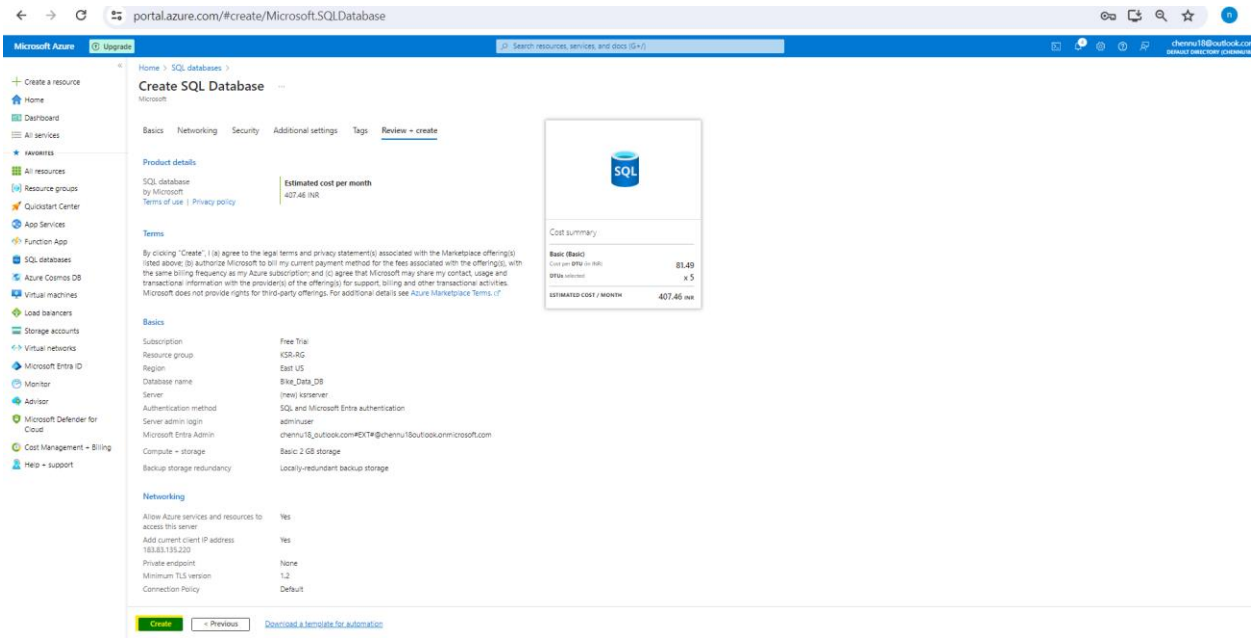
➔ In Additional settings tab also go with all the defaults

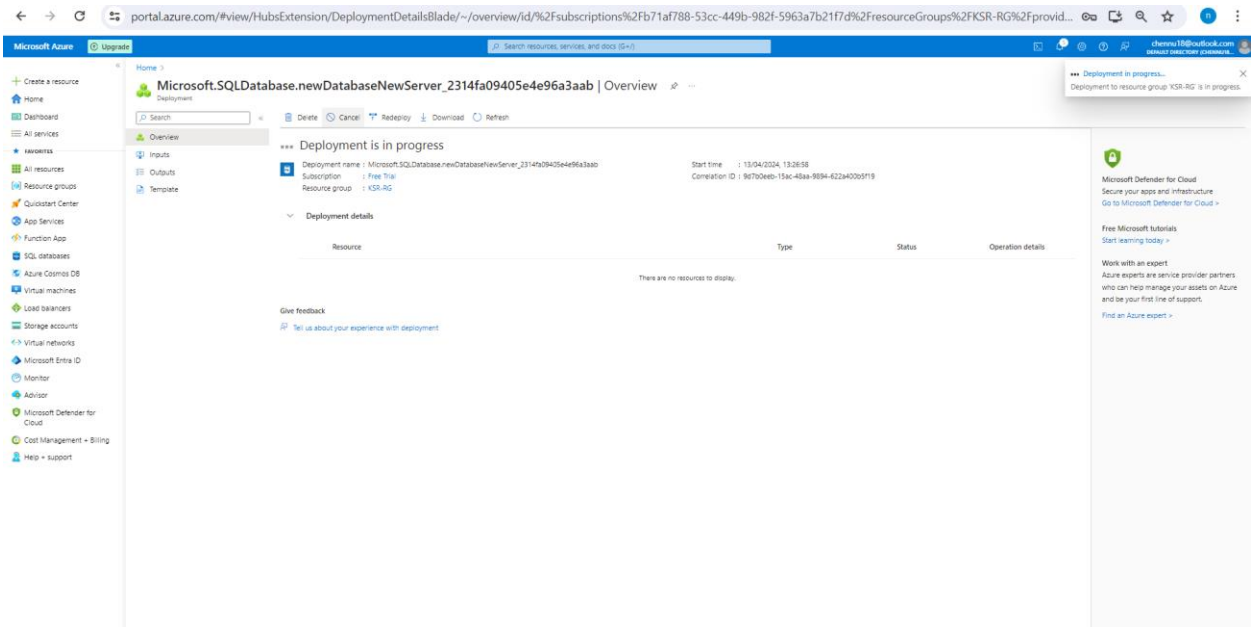


➔ In tags tab also go with all the defaults

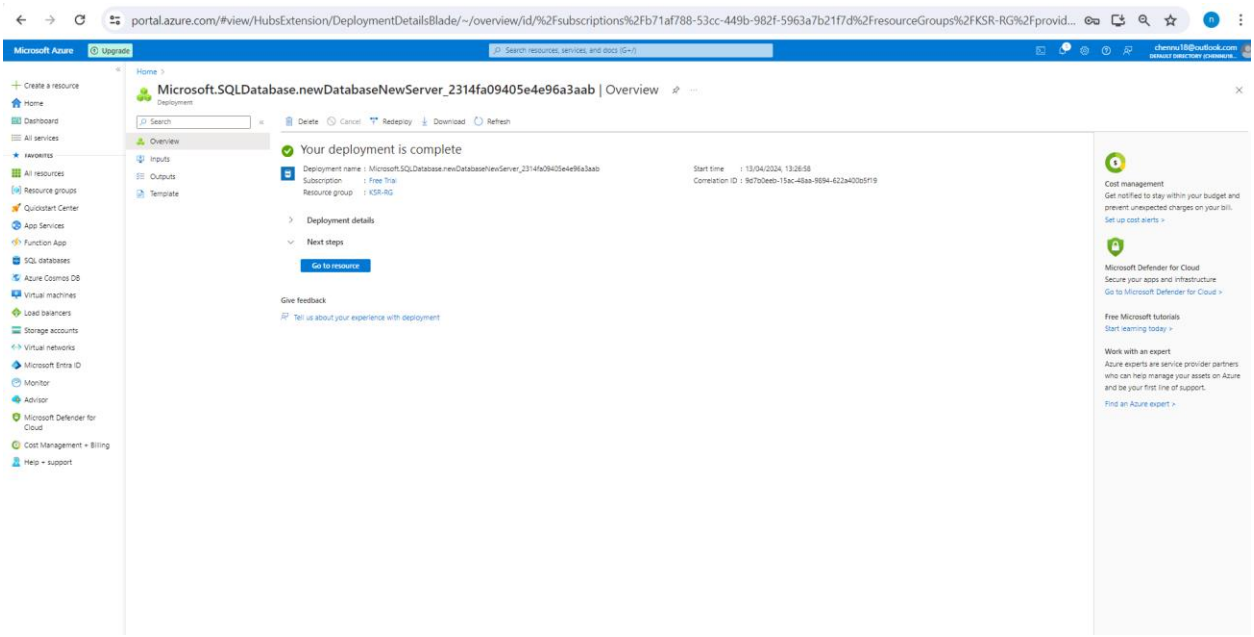


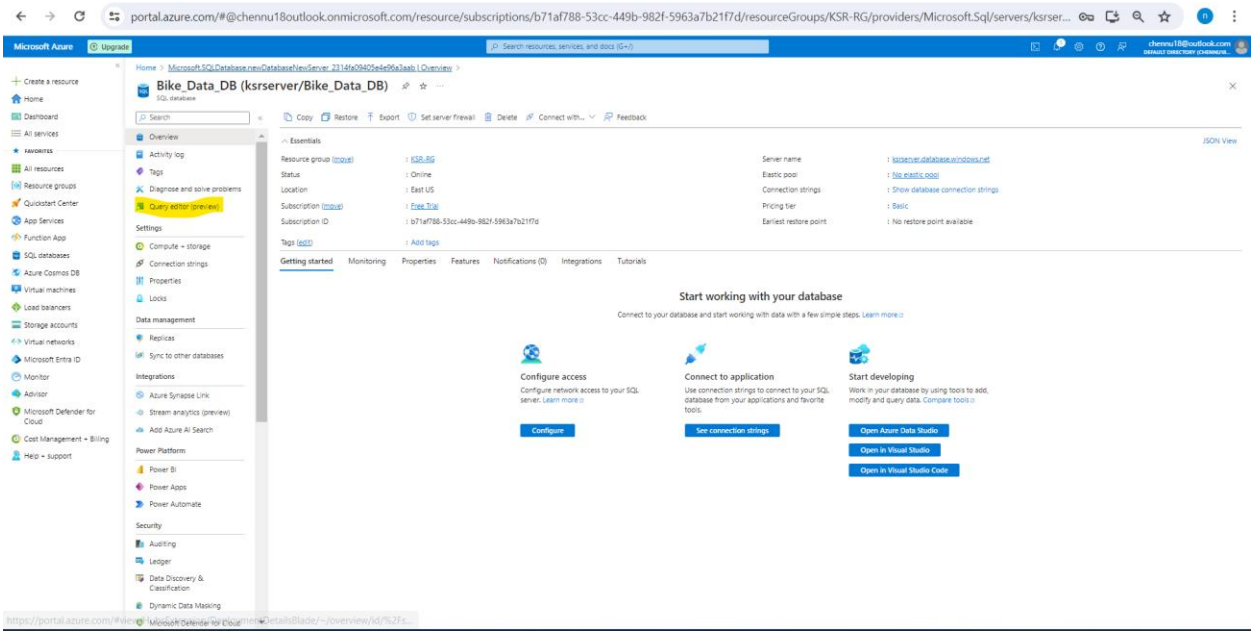
➔ Finally click on Review and create button to validating the details what you provided so far → then click on create button →it take less than 5 min to deploy the services shown below figures



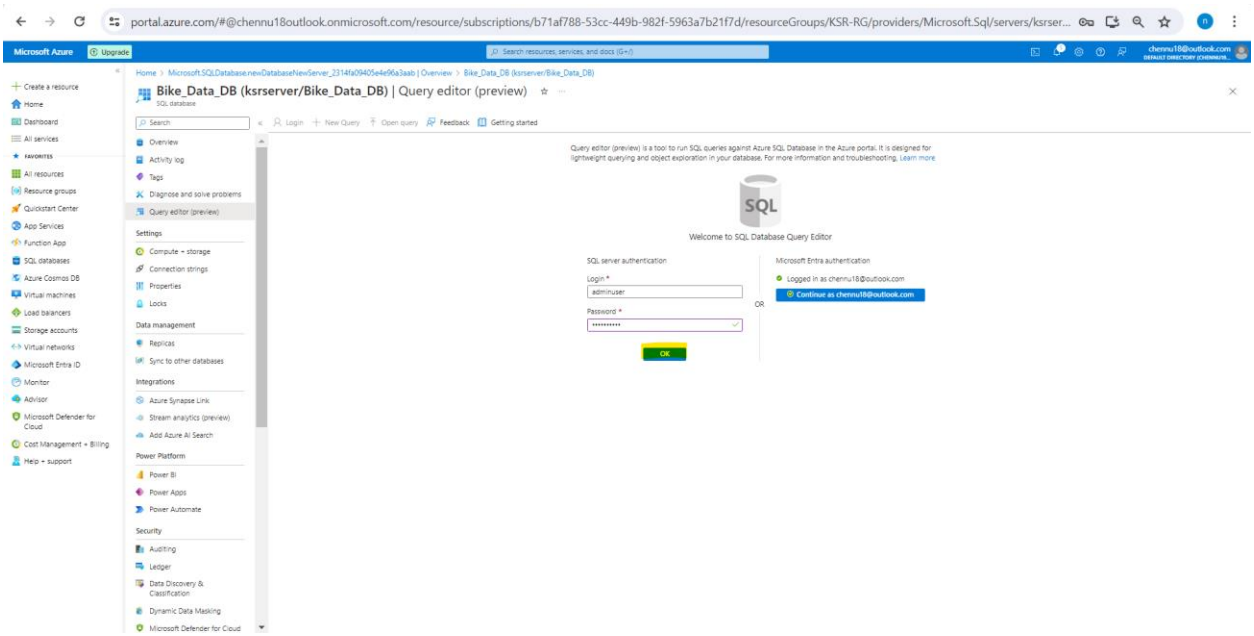


➔ After successfully completing deployment, you are able to see deployment is complete as you seen like below → click on Go to resource it will navigate to Azure SQL Database page.

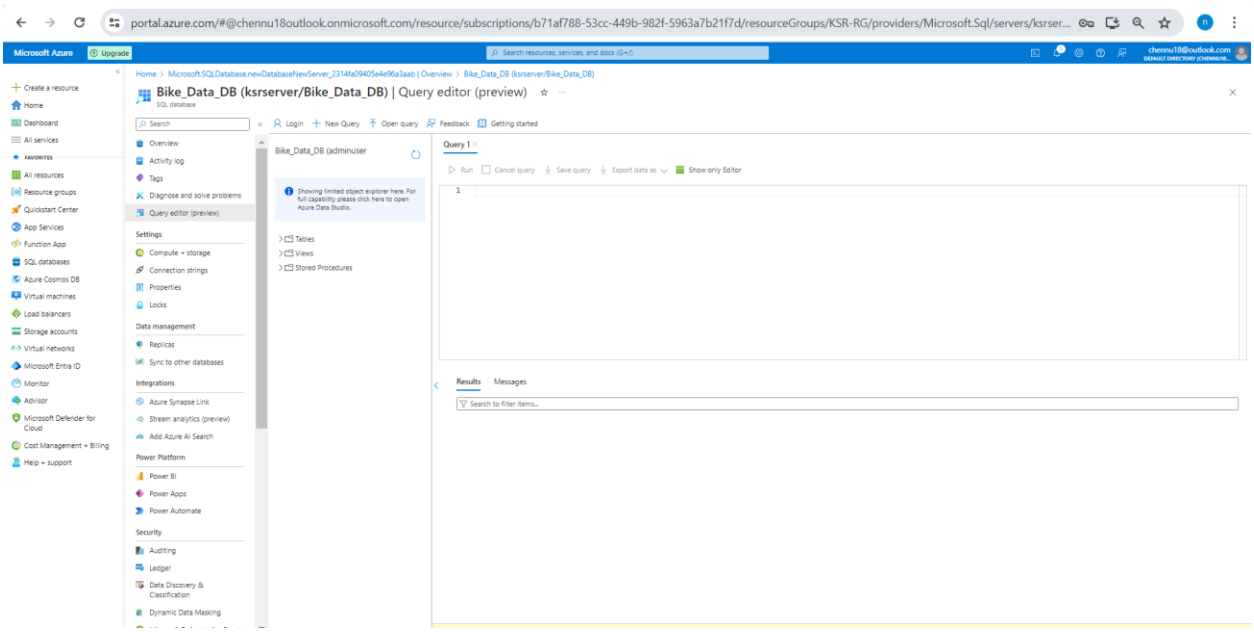


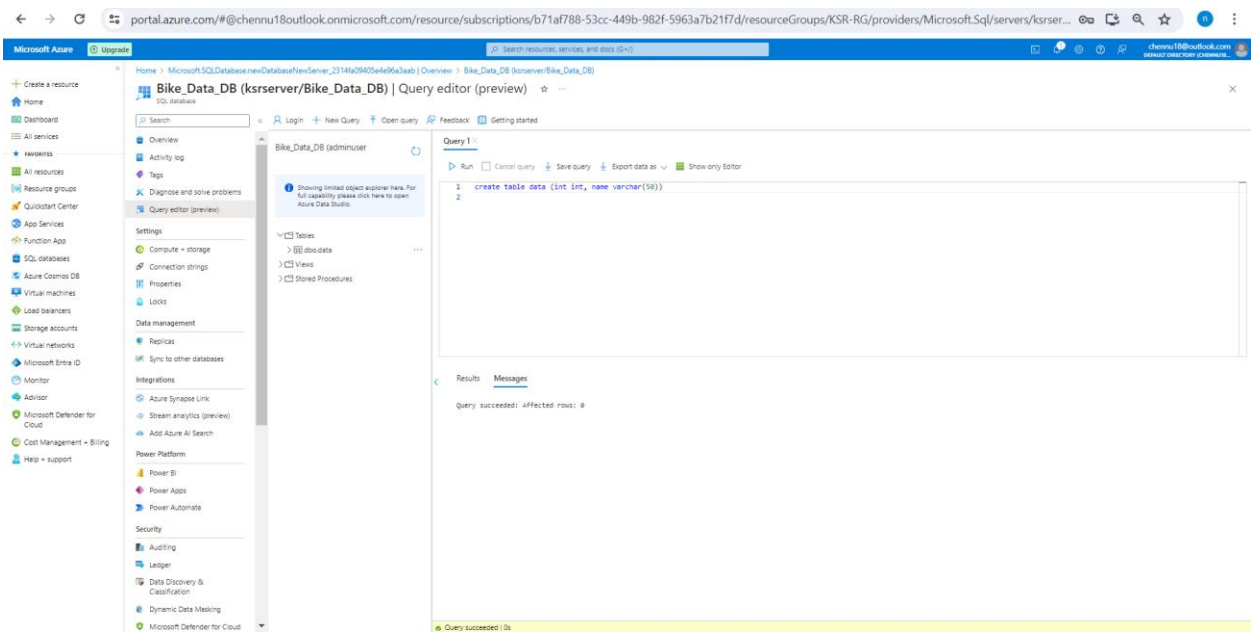


➔ In SQL Database overview page, in left navigation to find query editor → click on that → it will navigate to the SQL Server page → we need to provide username and password → then click ok.



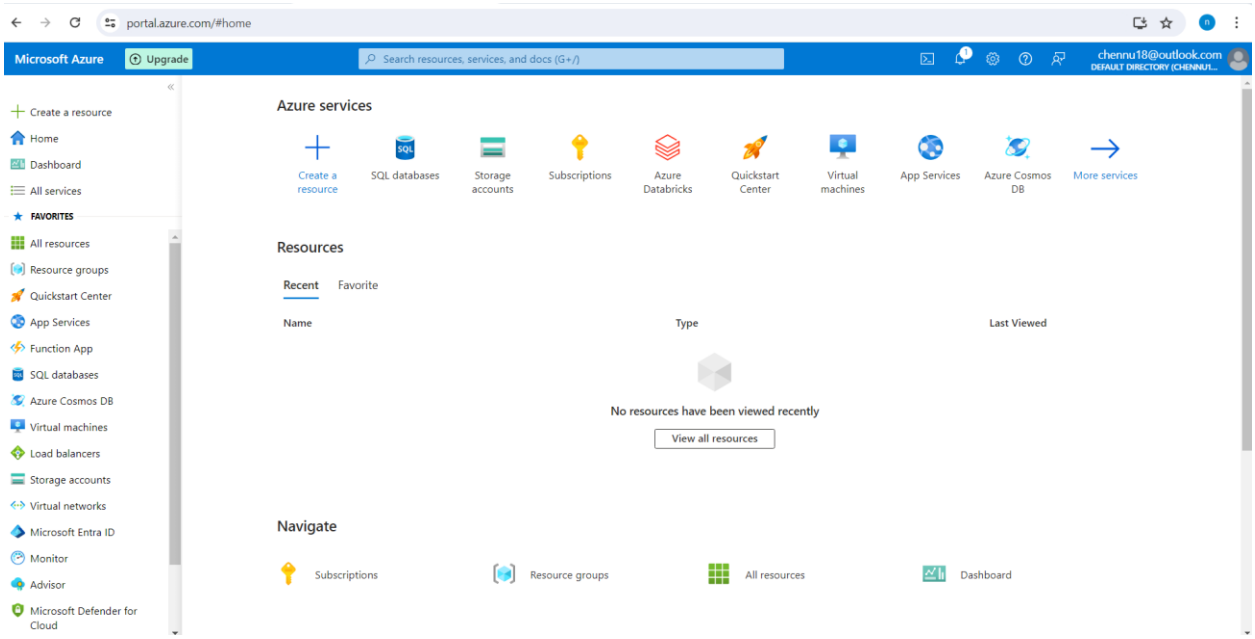
➔ After successfully login into the server → create table for checking purpose as shown in below figures



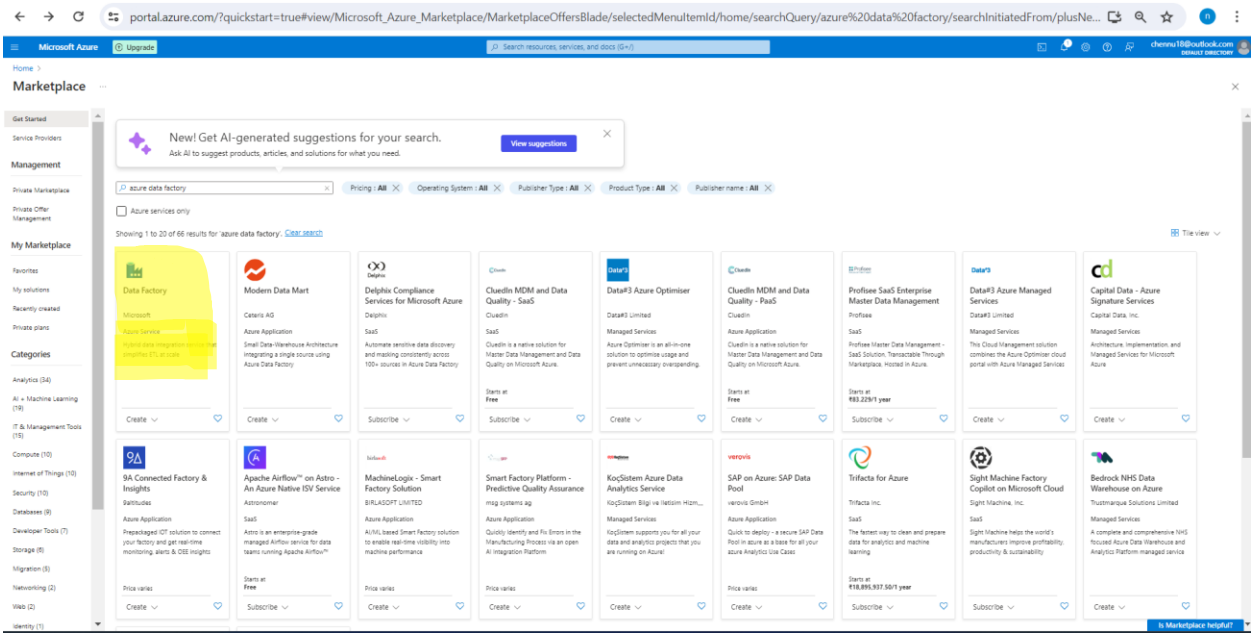


Creating Azure Data Factory service from portal.azure.com

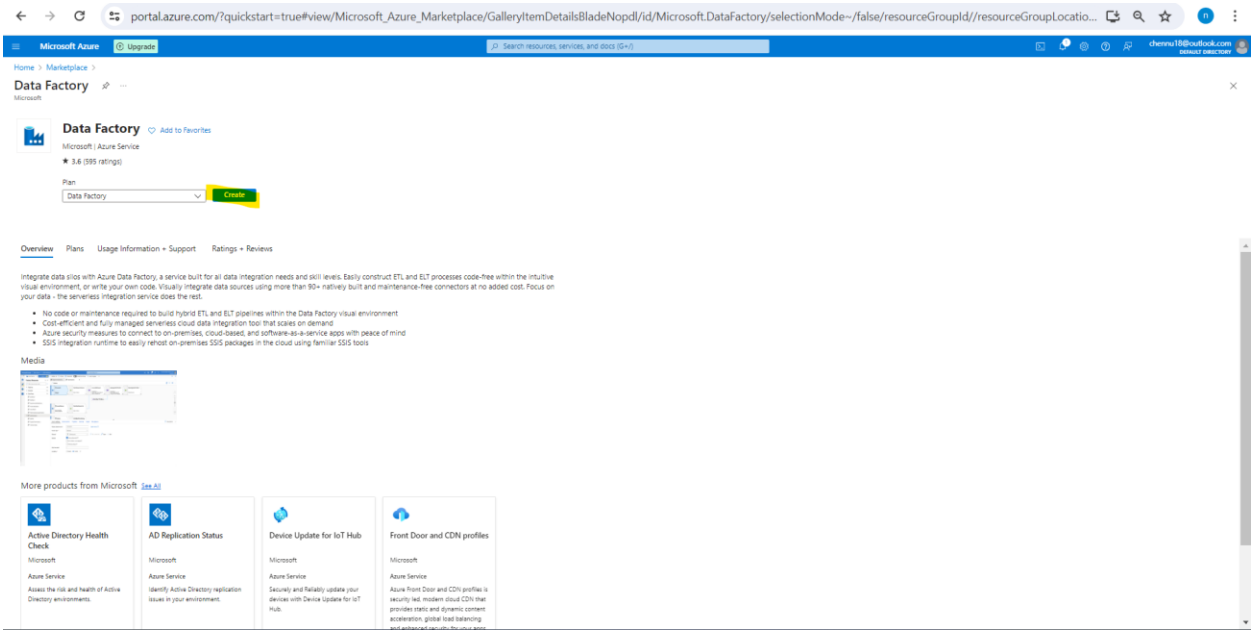
- ➔ Login to azure portal using this link https://portal.azure.com/
- ➔ After successful login you will able to see home page like below



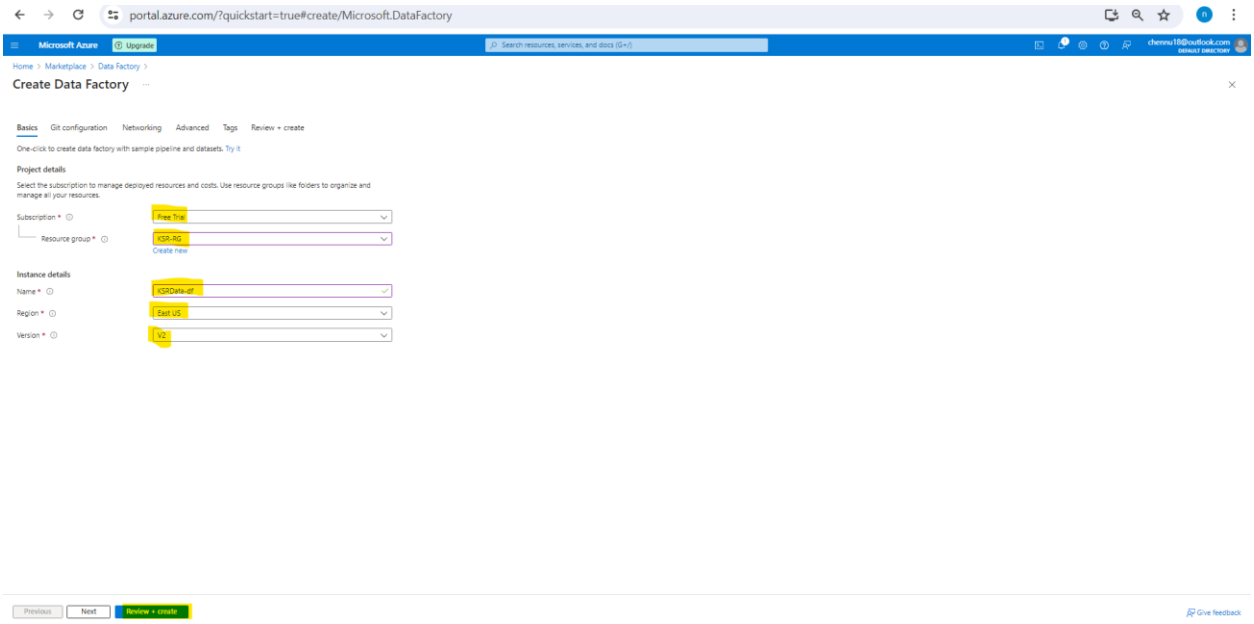
- ➔ Go to Azure Market place search for Azure Data Factory and hit enter, you will get page like below



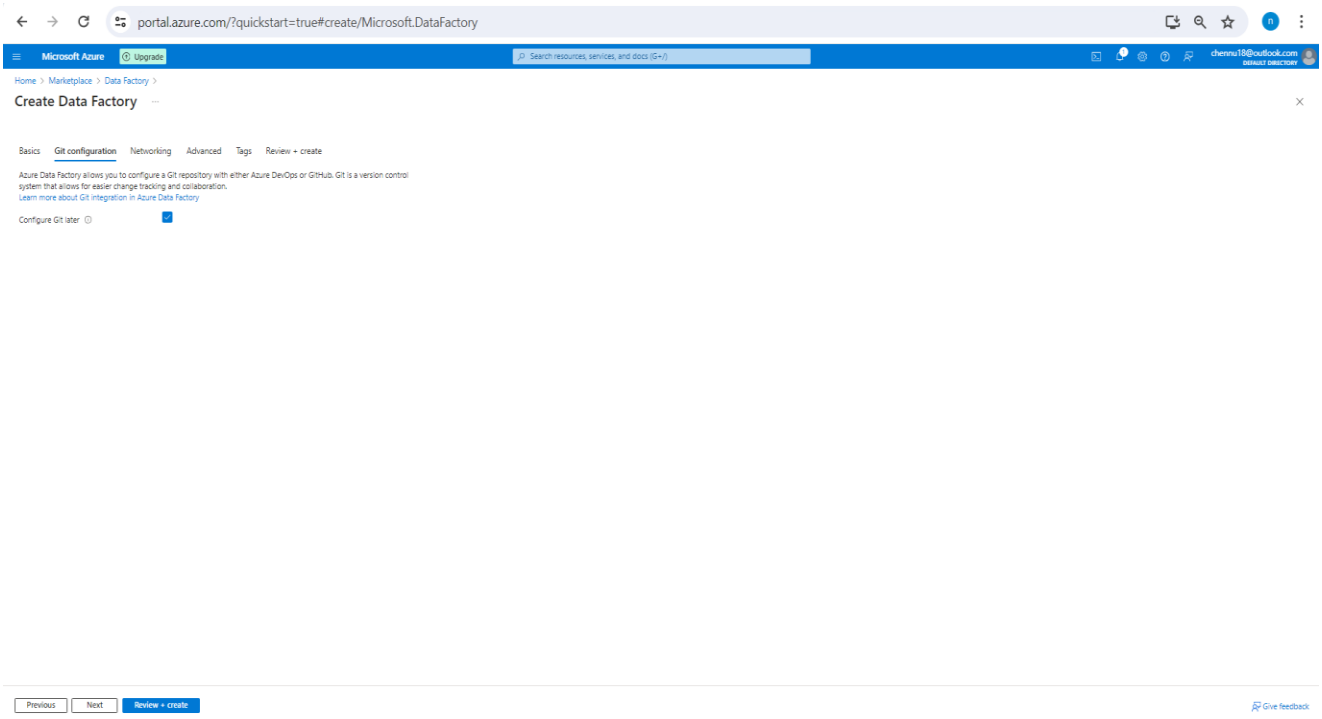
➔ To create Azure Data Factory, using Create button see in below fig



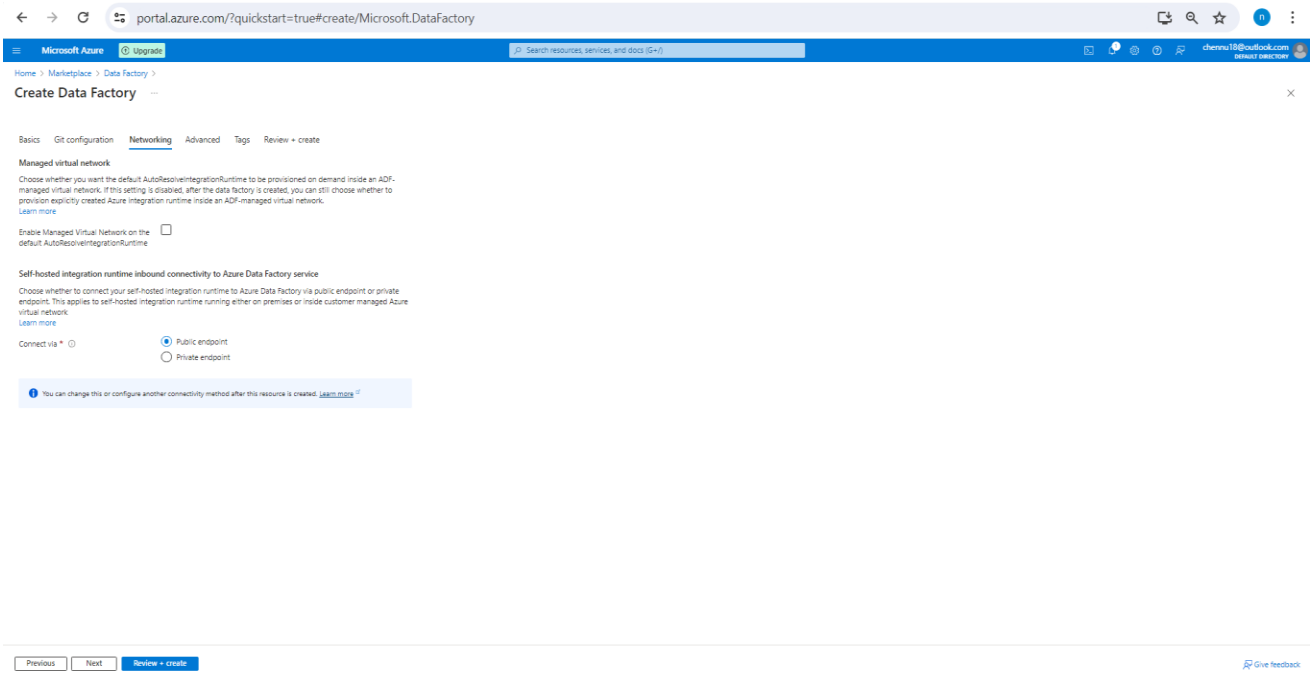
➔ We need to provide subscription details and resource group, data factory name, region and select version. All are there below figure take a look into that.



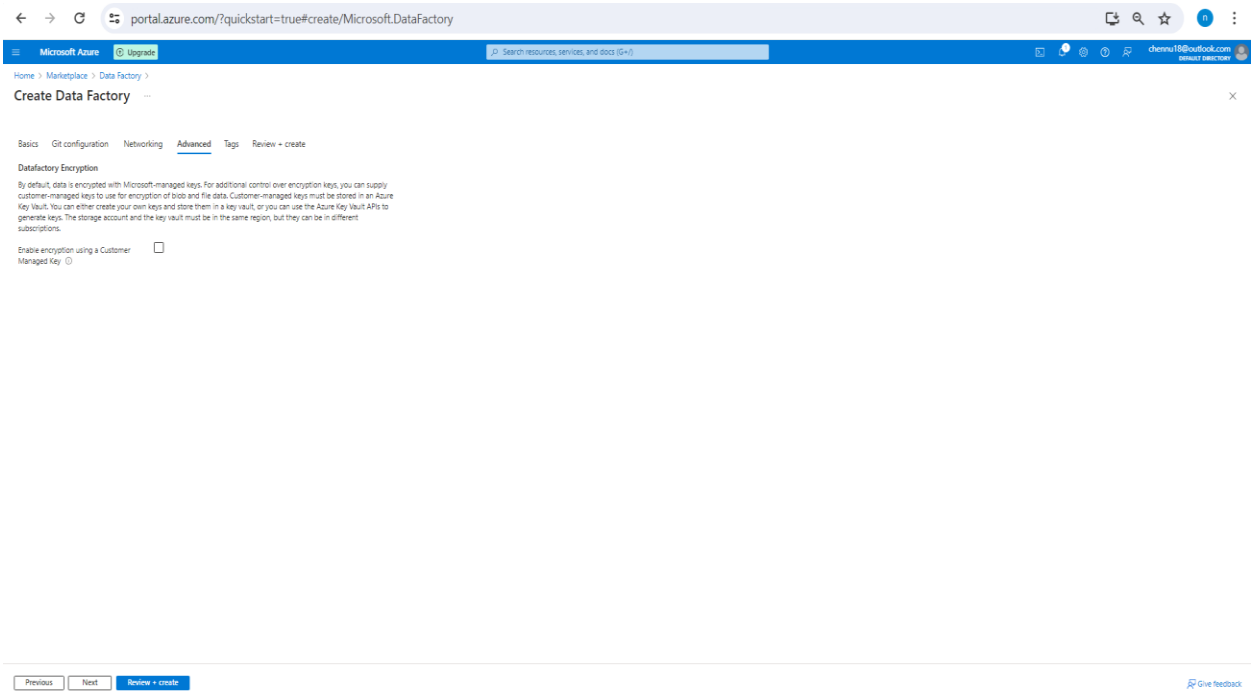
➔ For this time we don't have a git repository, we can configure later go with default.



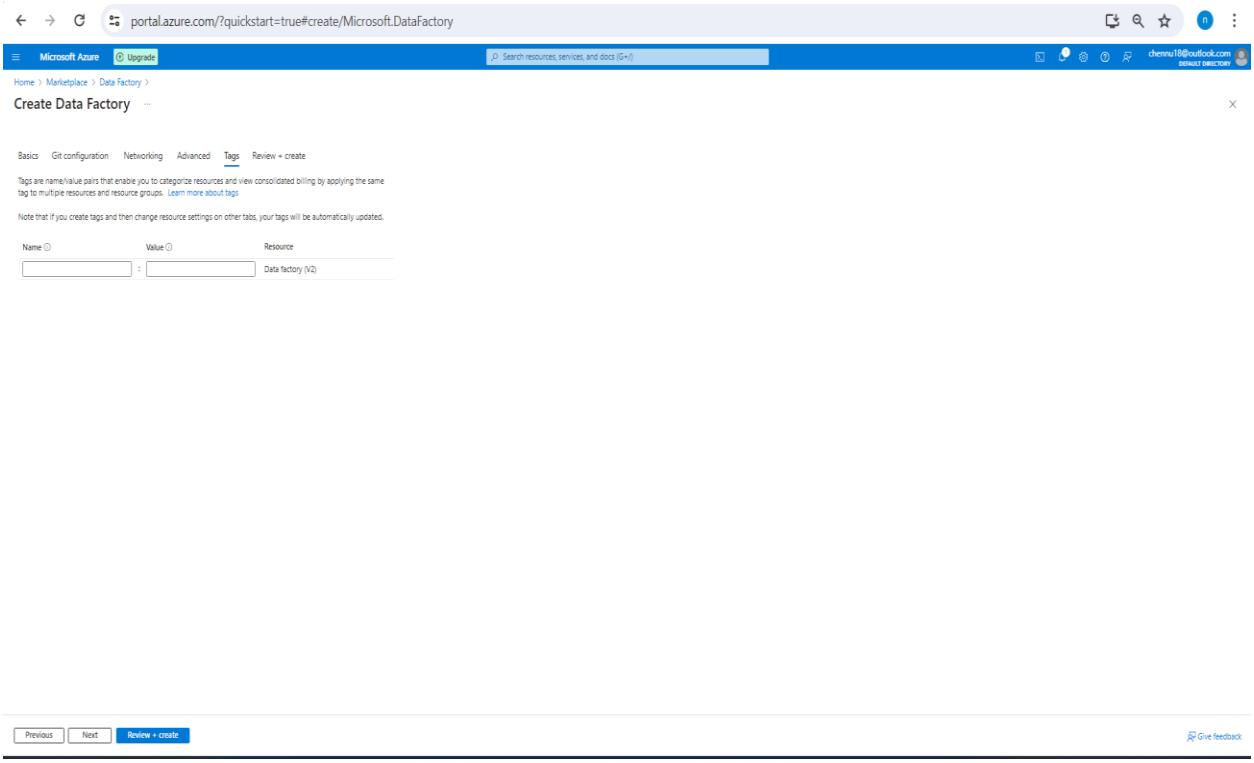
➔ Networking tab also go with default settings



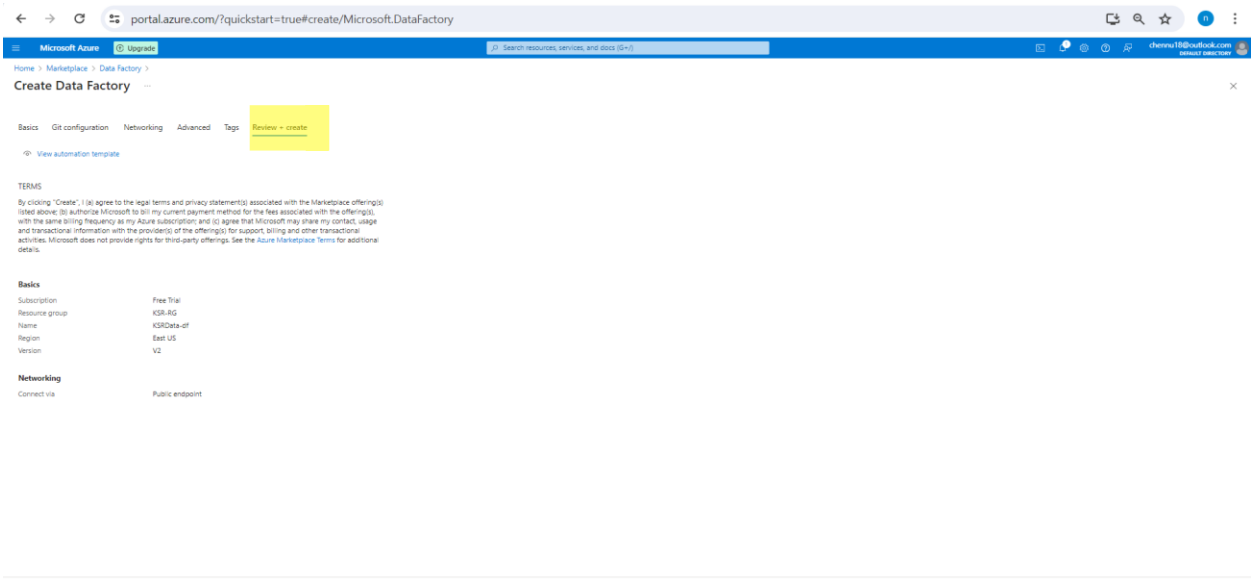
➔ Advanced tab also go with default settings



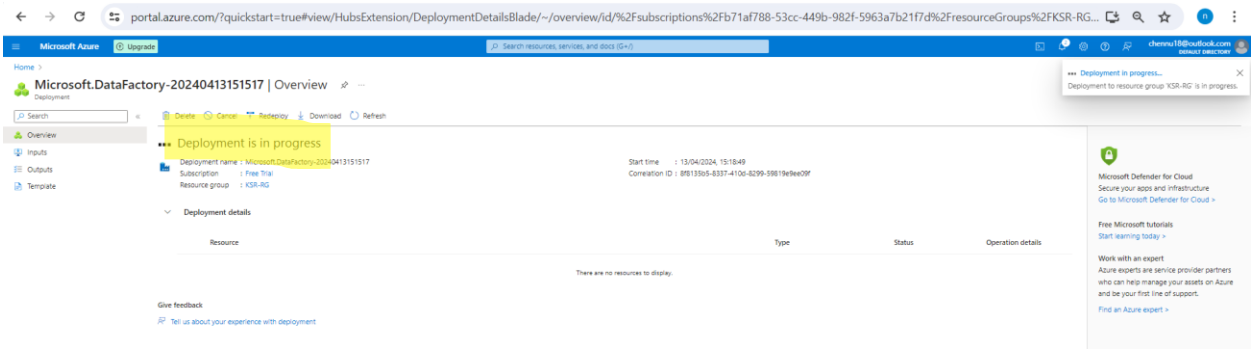
➔ Tags tab also go with default settings



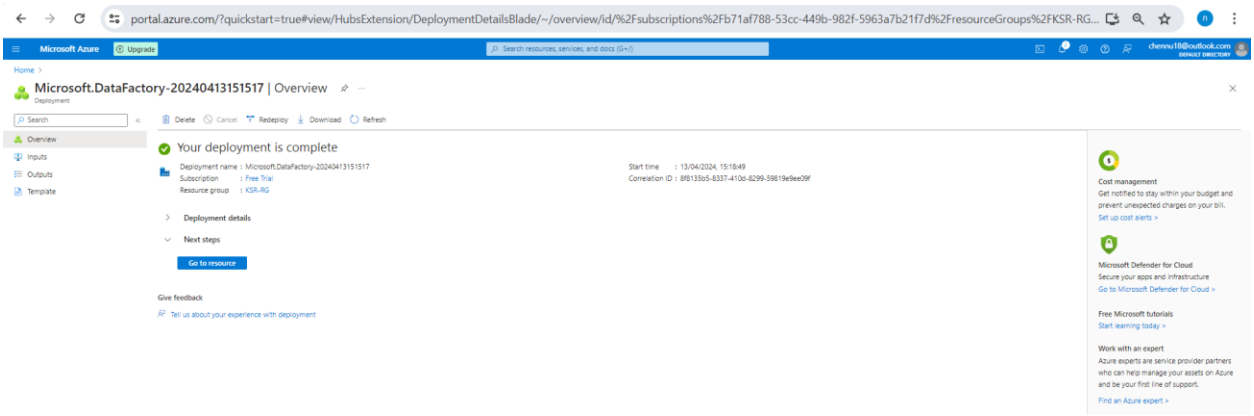
➔ Click on review and create button validate details what we provide , after validation pass then click on create



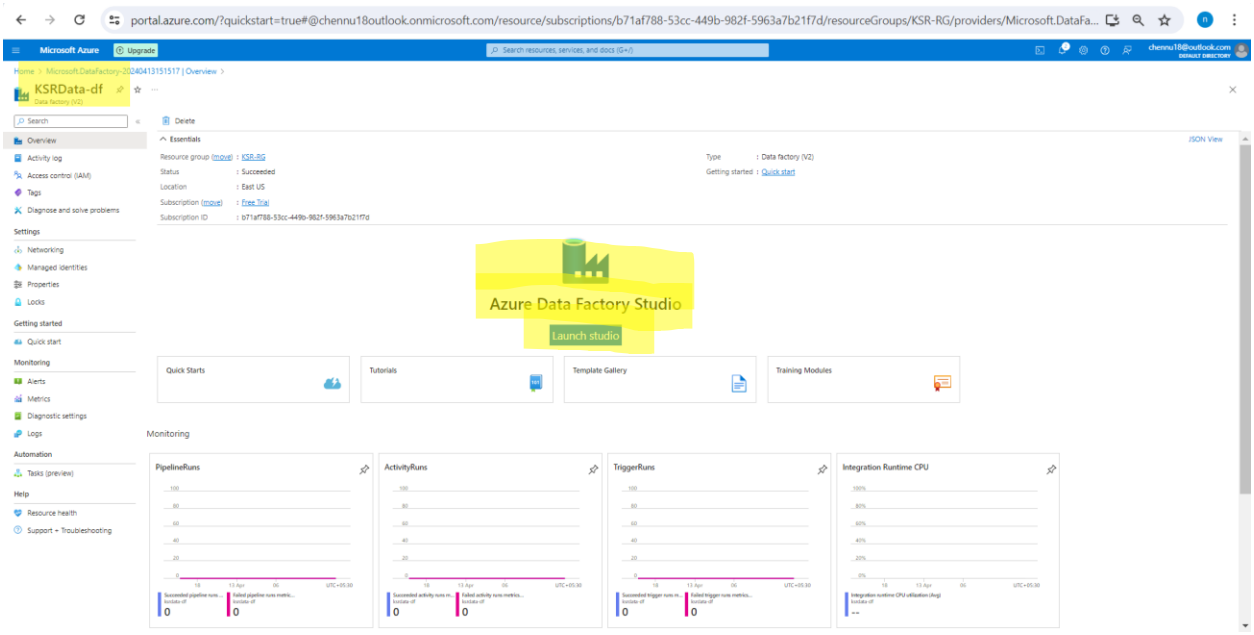
➔ Below figure shows the deployment in progress after hitting the create button



➔ Once deployment finish click on go to resource, we will navigate to azure data factory home page



➔ Switch from Azure Data factory home page to Azure data Factory Studio jus click on Azure data Factory Studio button seen below figure



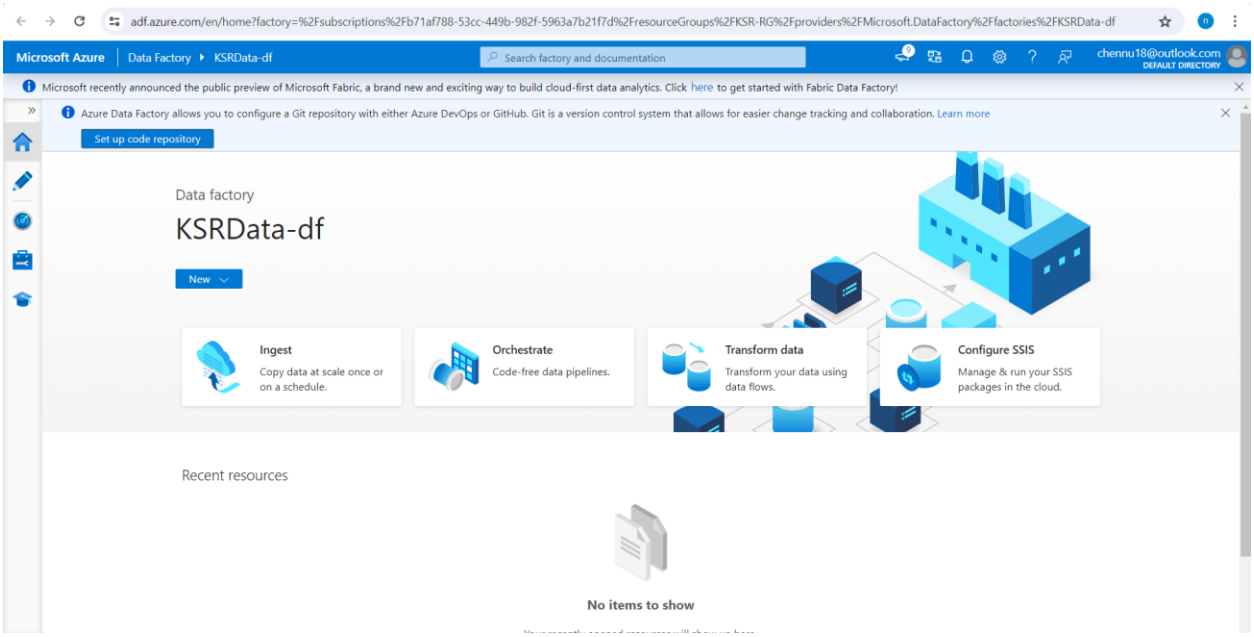
➔ Below figure represents the Azure data Factory Studio work environment. In left navigation we have

➔ 1. Home Tab

➔ 2. Author tab

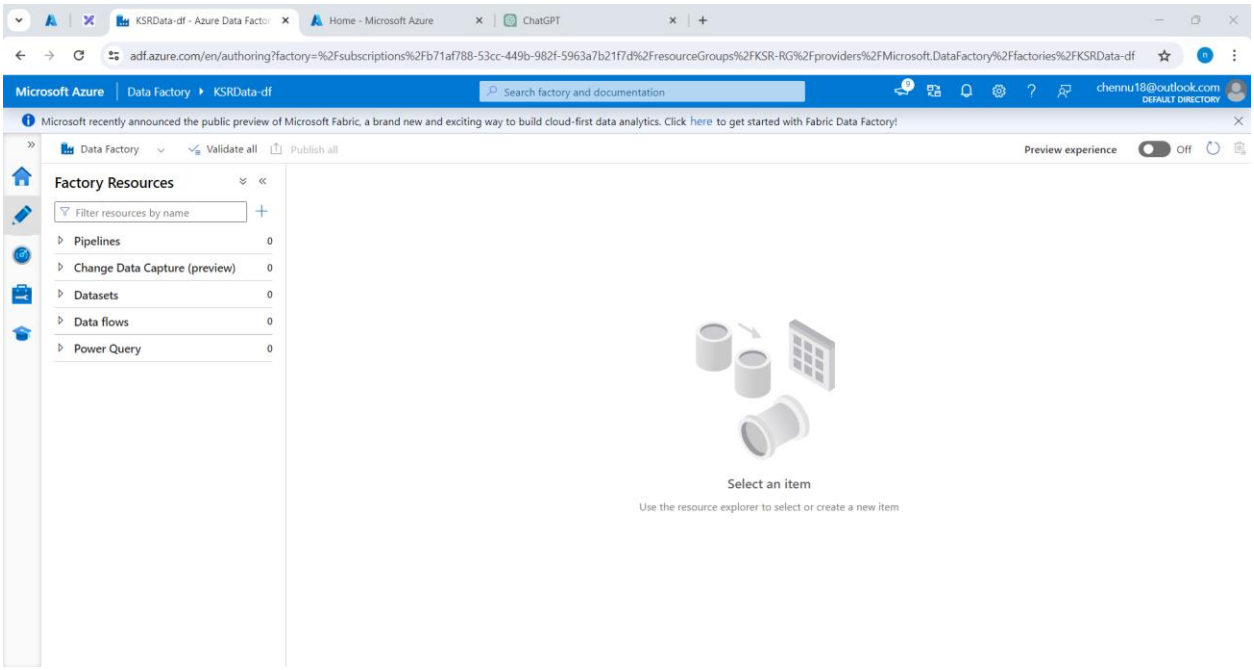
➔ 3. Monitor tab

➔ 4. Management Hub tab

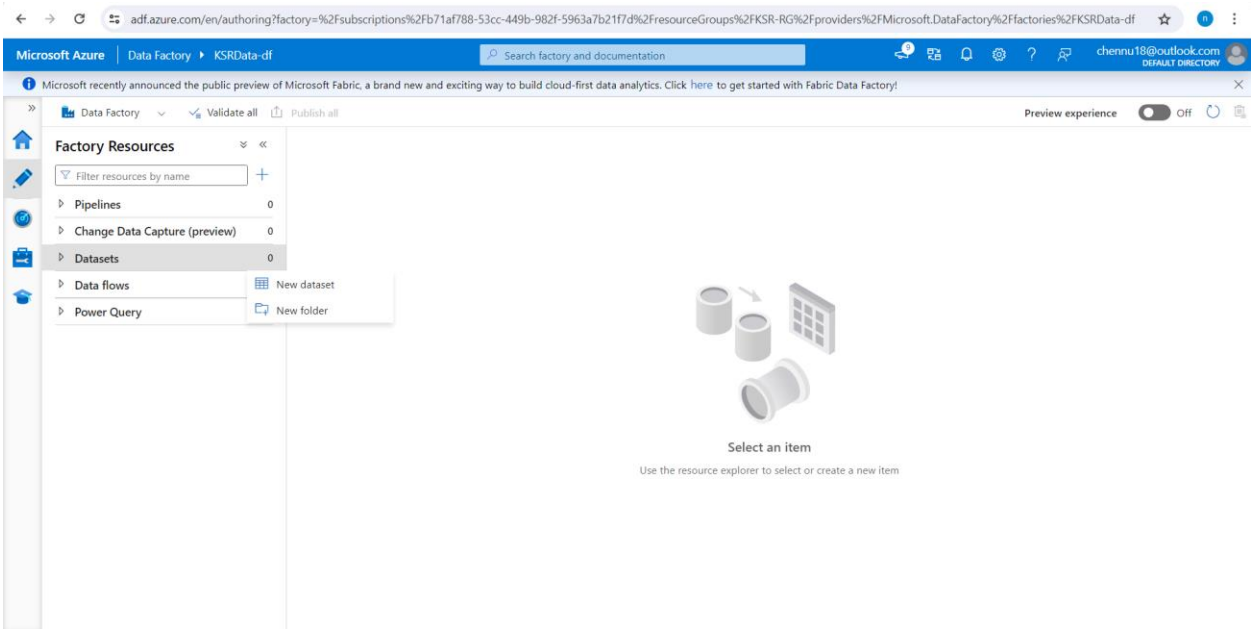


➔ Click on Author tab(pencil icon) to see the factory resources

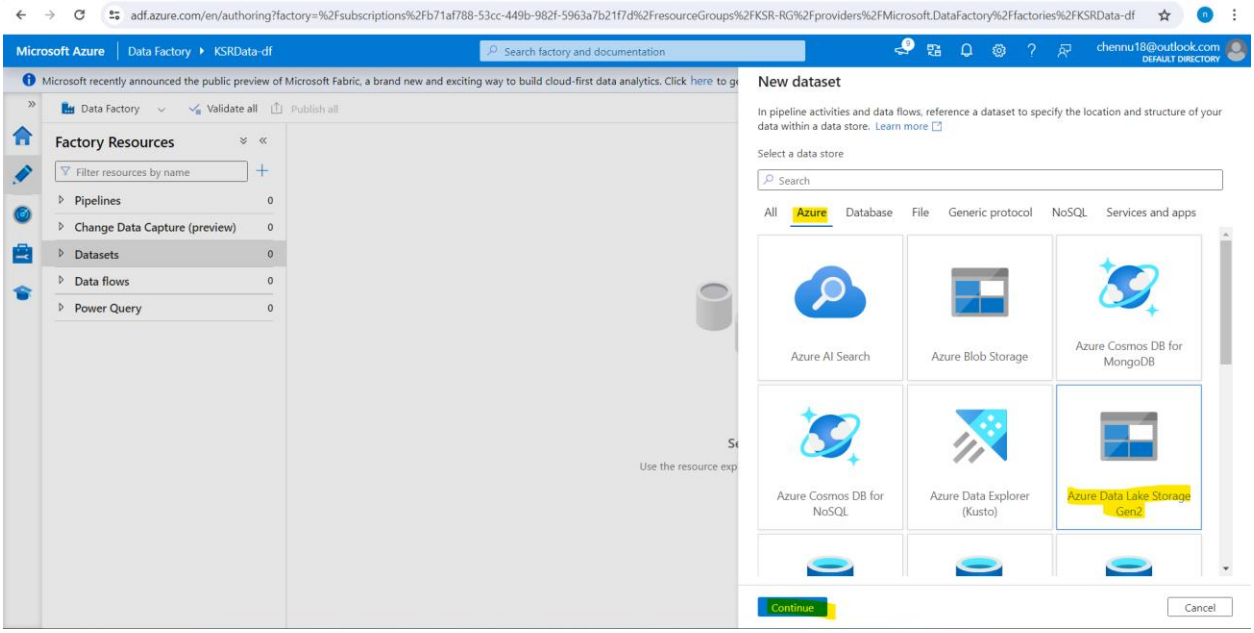
1. Pipelines
2. Change data capture(preview)
3. Datasets
4. Data Flows
5. Power Query



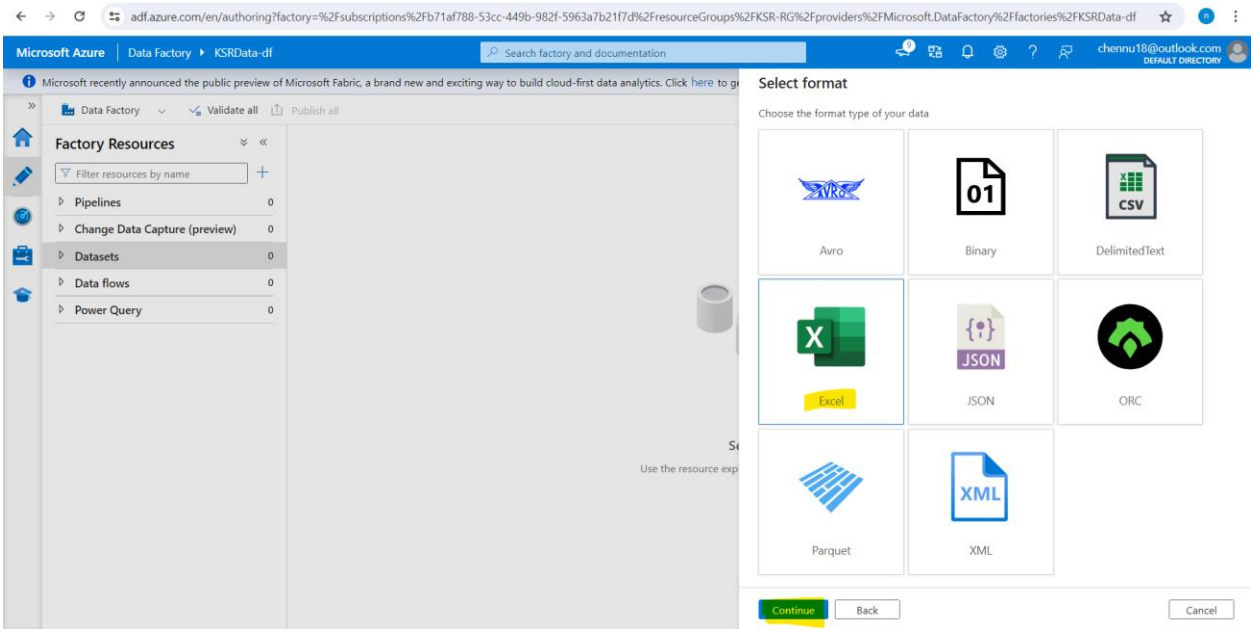
➔ Right click on datasets resources → click on new dataset → it will asking create a dataset



➔ Click on azure → click on azure data lake storage gen2→ click continue

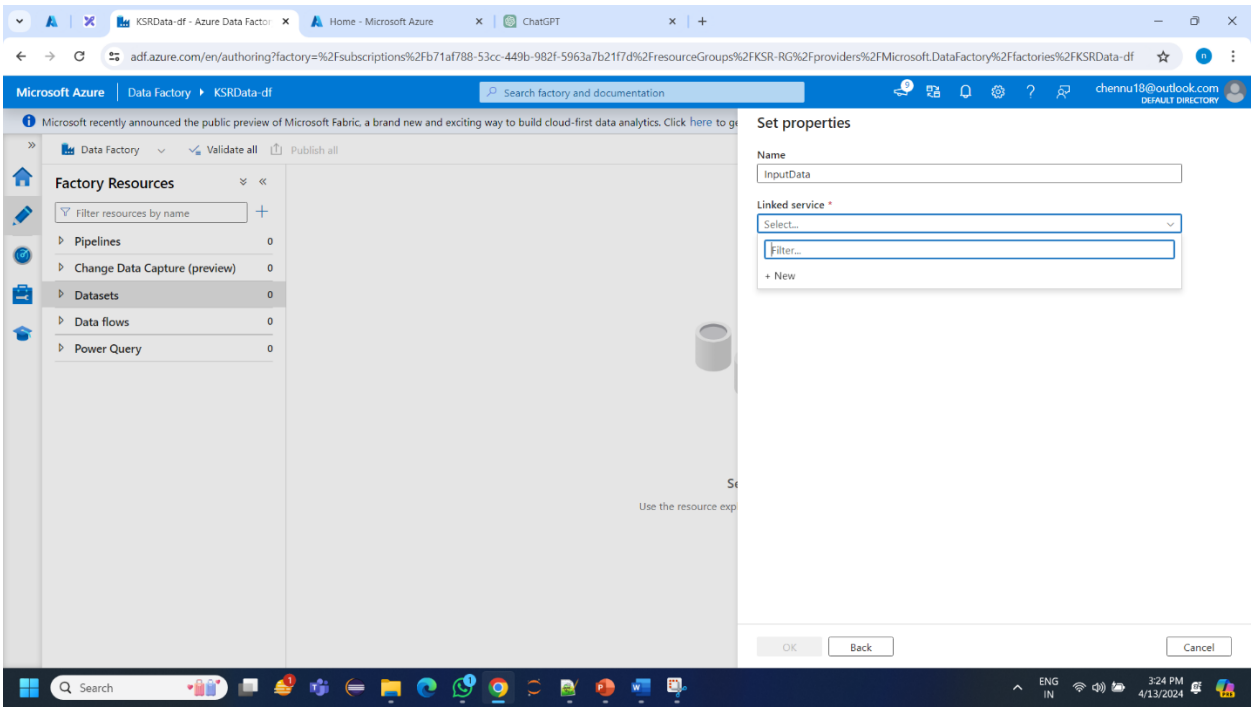


➔ After that choose source file type it is excel, csv, json, xml and etc

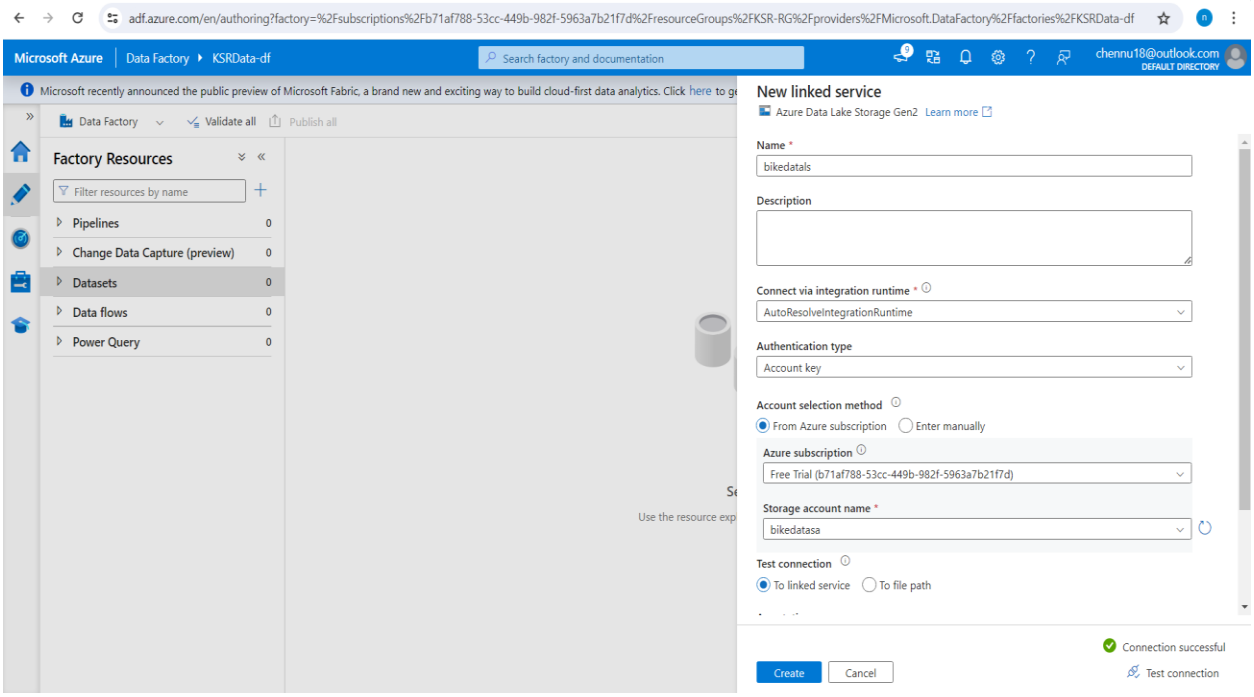


➔ Need to configure the linked services(which integration tool called connection strings.

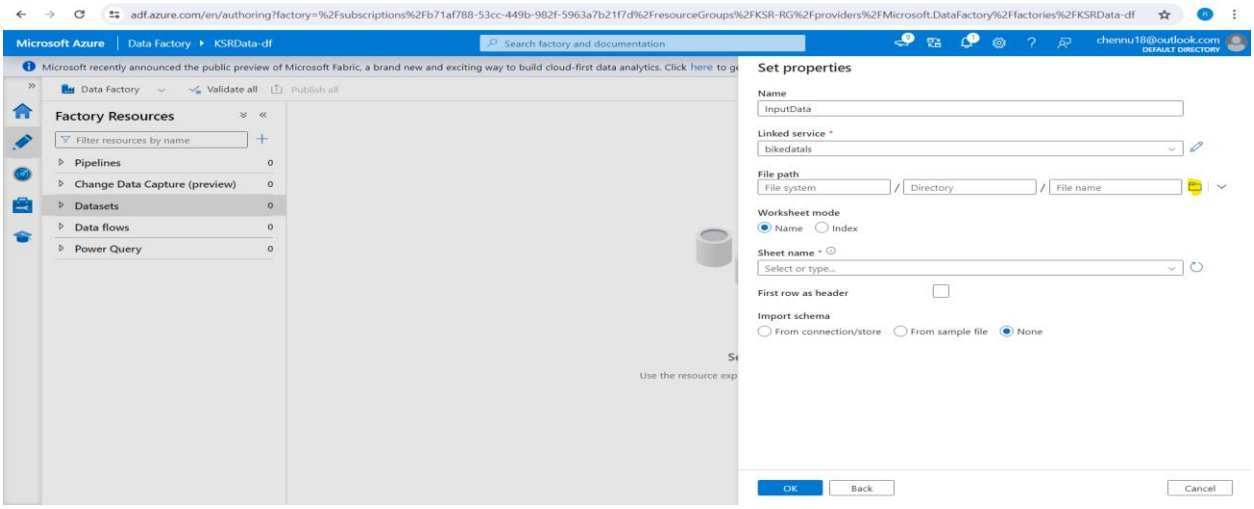
1. Provide linked service name
2. click new tab see in below figure

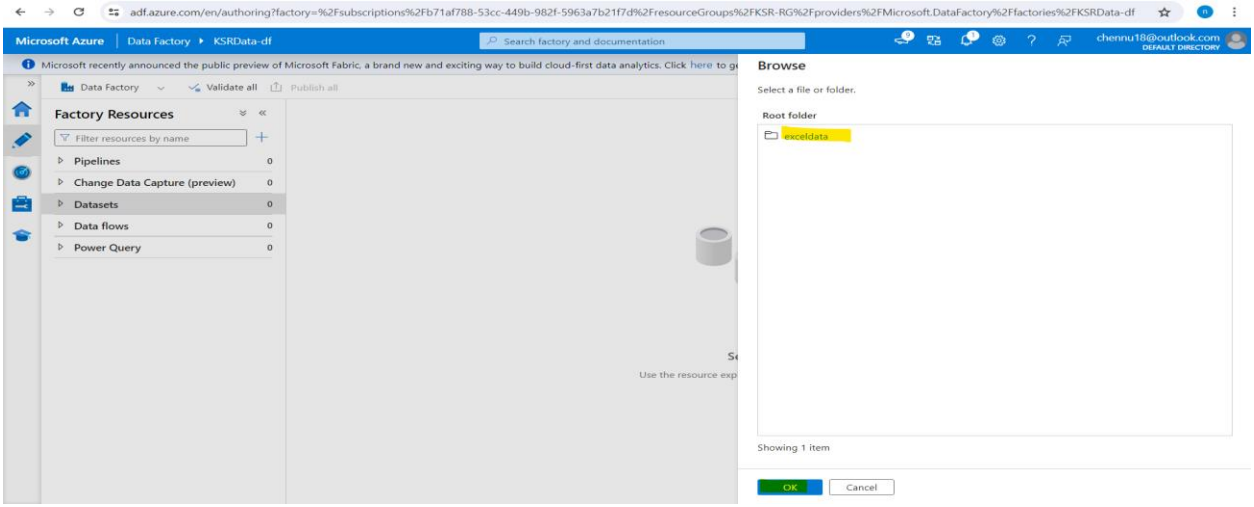


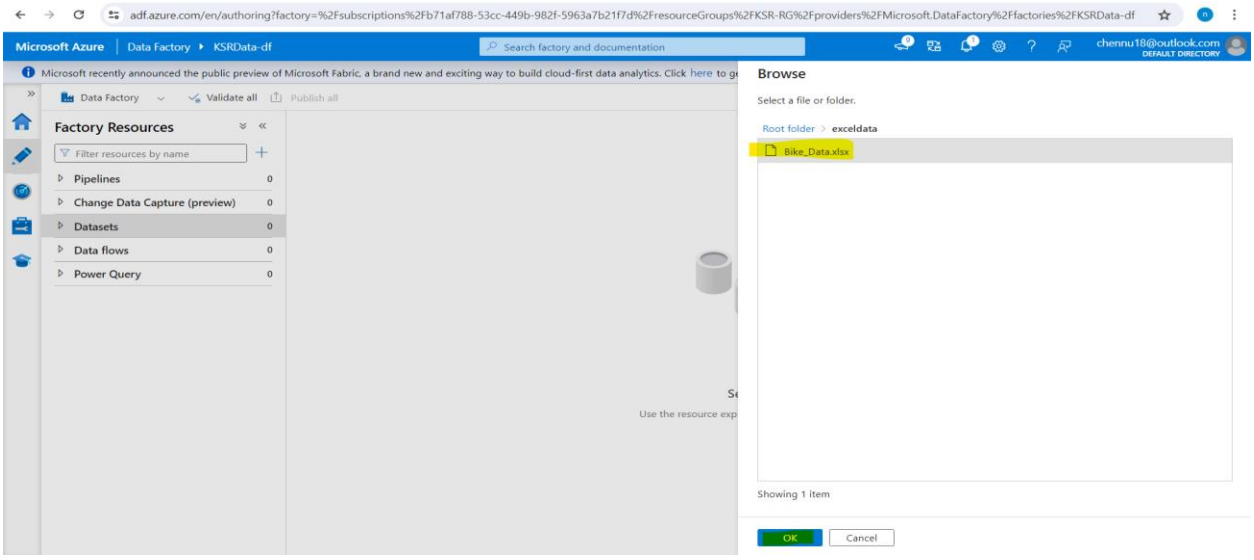
➔ After clicking new → need to provide required details like shown in below figure→ then click test the connection → to get message connection successful → click on create button

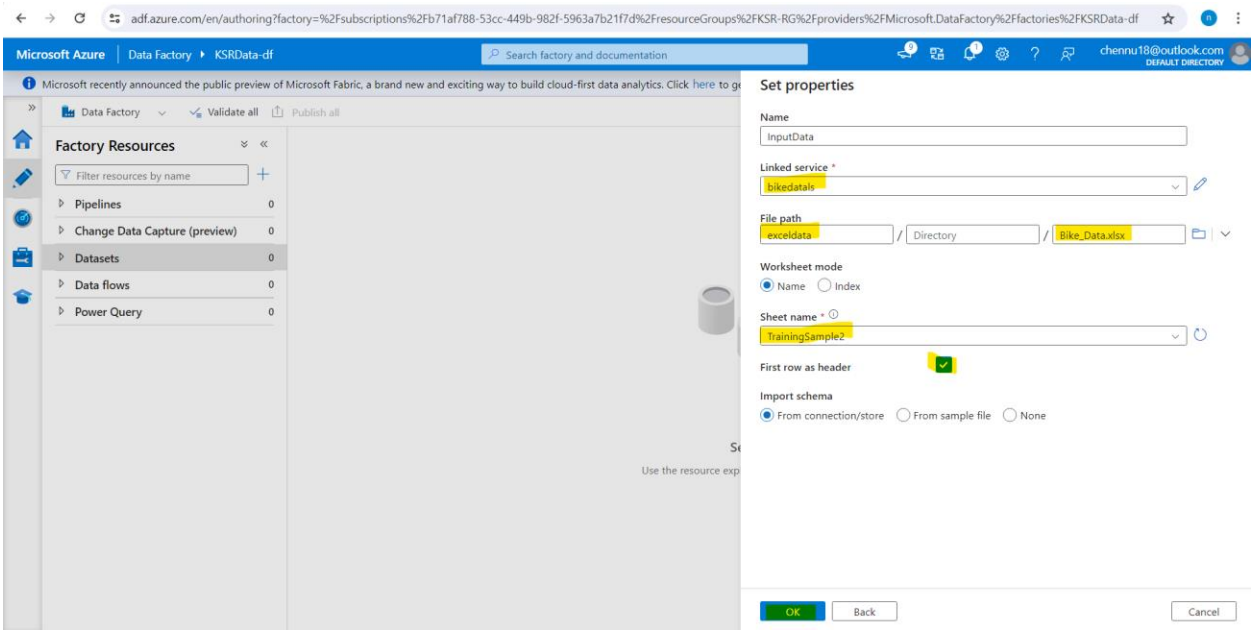


➔ We need to set properties of source file and browse source file location and select sheet name then click ok. See in following 4 figures

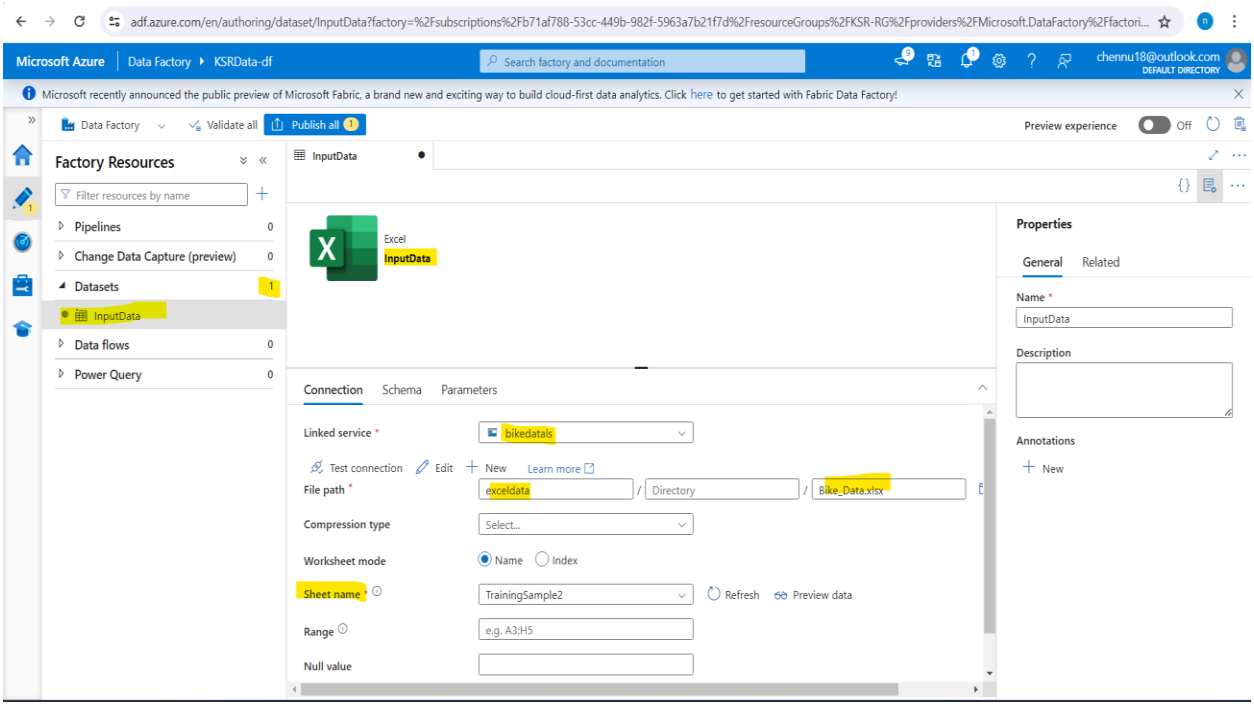




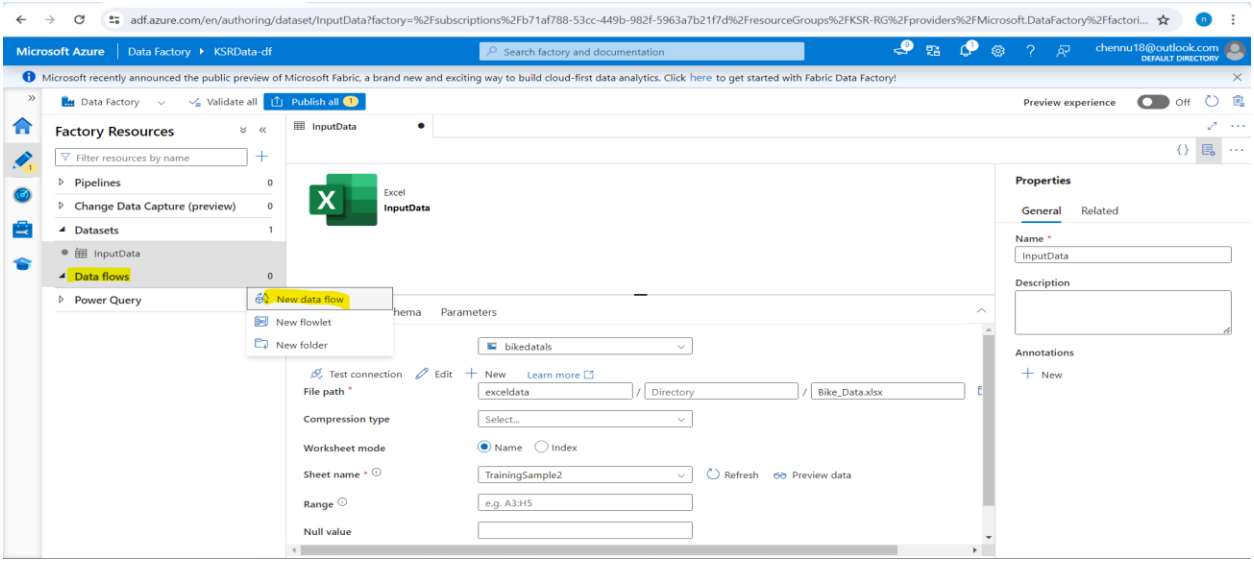




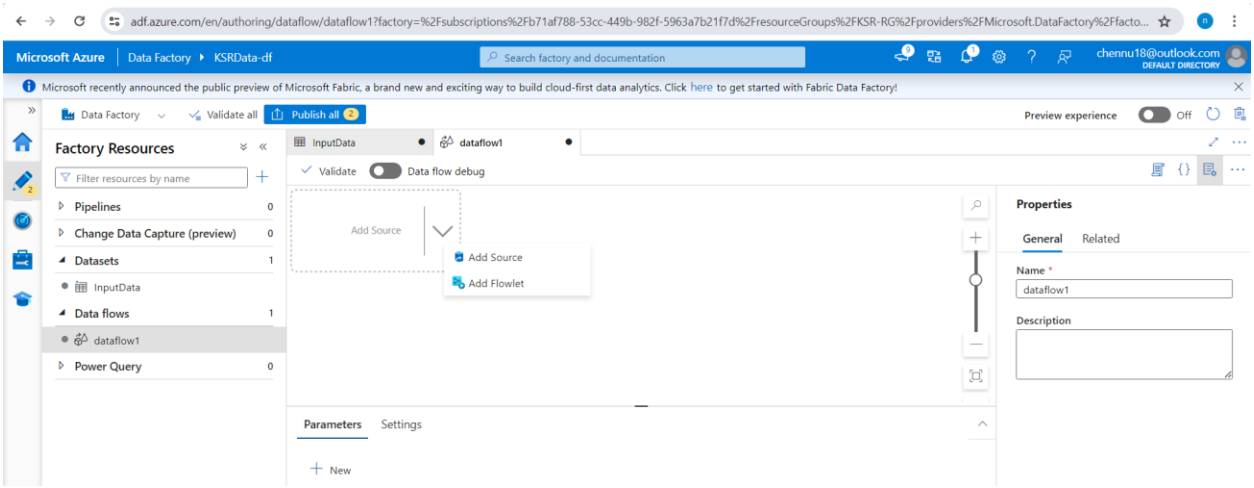
➔ Finally you see in the input data under dataset see in below figure

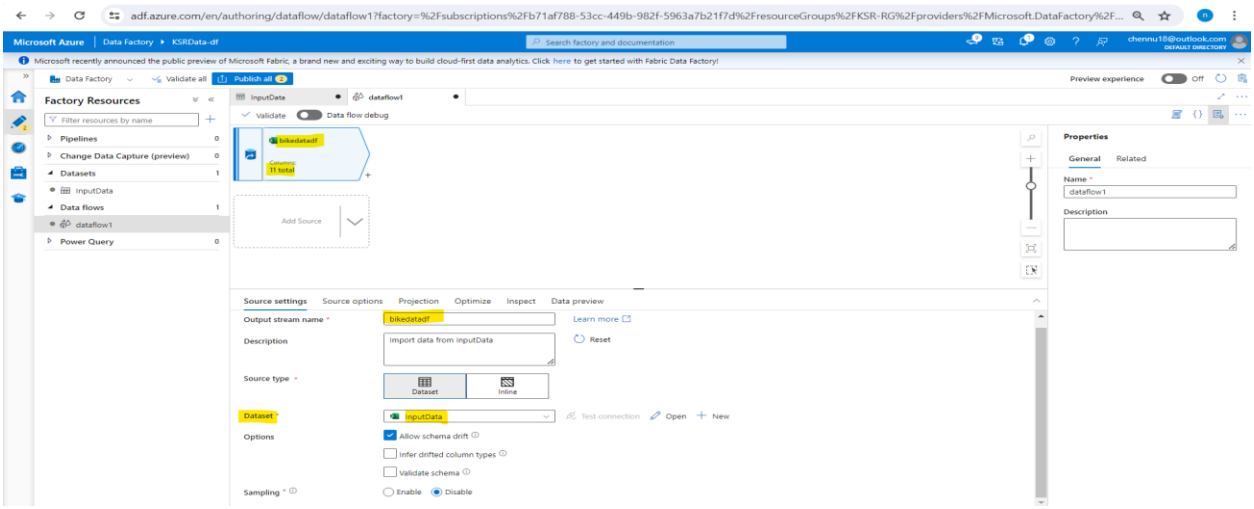


➔ Creating Data Flow under author tab → right click on Data Flow → click on New Data Flow →see in below figure.

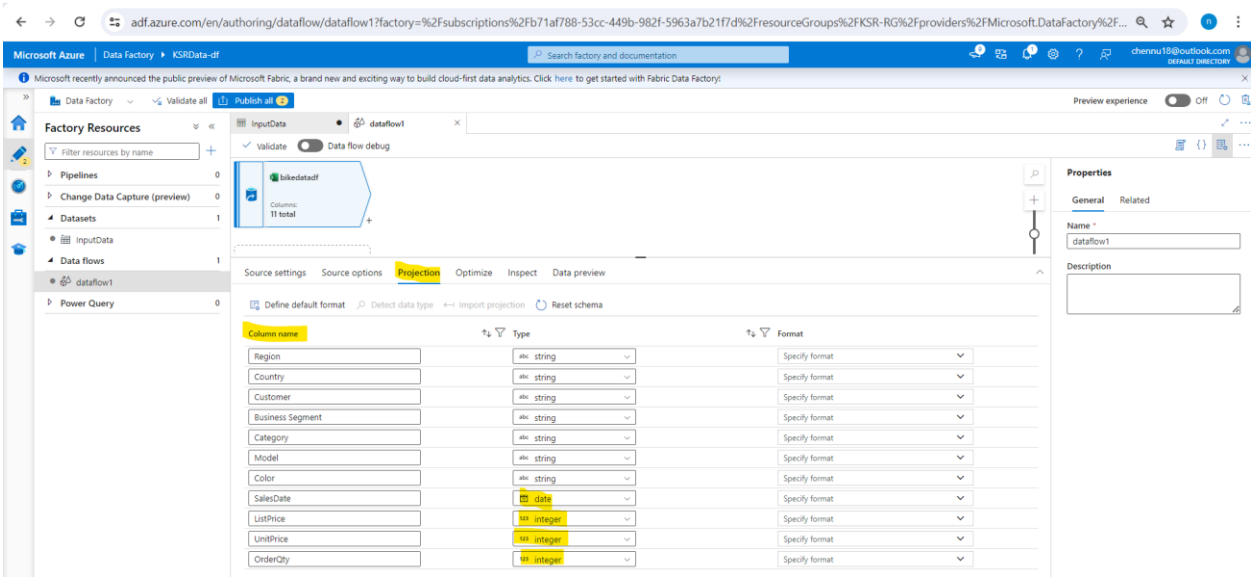


After creating Data flow, a page shown like below figure → click down arrow → click on add resource → fill the all required details follow the two diagrams.

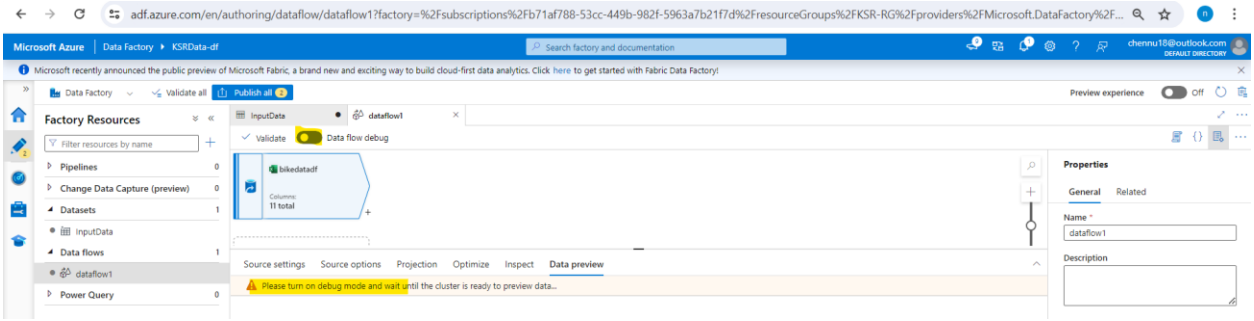


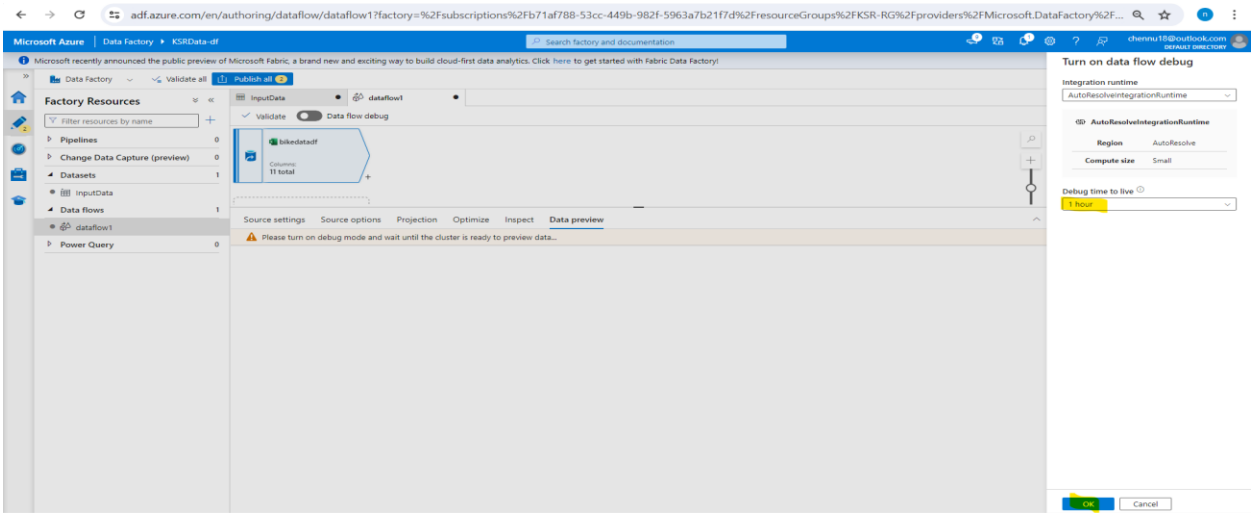


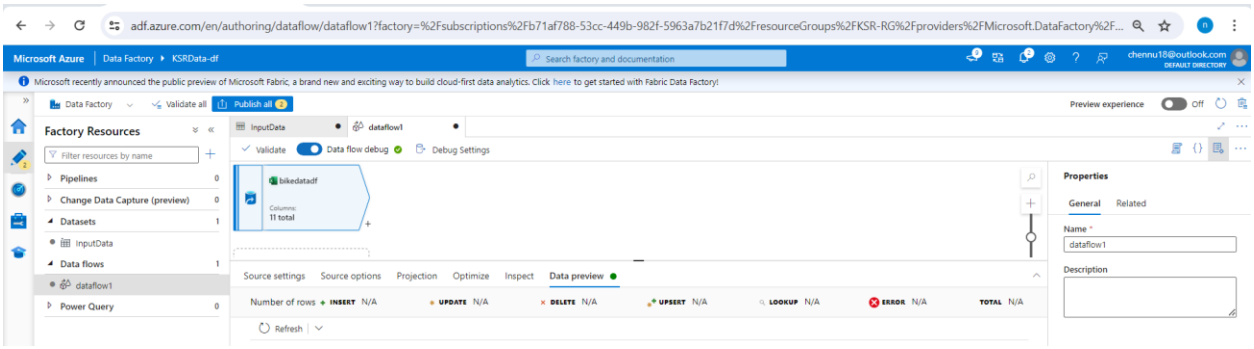
➔ In projection we need to change the data type as per the field values if it its numerical data, string data and data type.

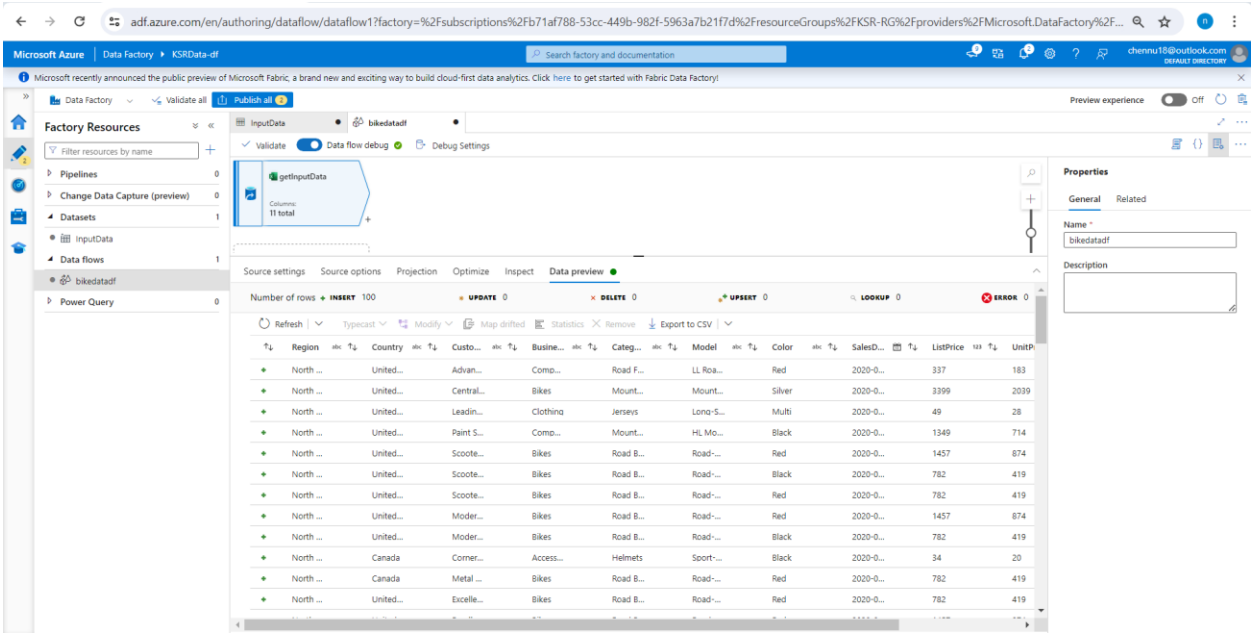


➔ After entering all the details if we want see the data → click on data preview tab, if you want see data preview, we need to enable data flow debug then only see the preview, while using data flow debug be careful because it will charge more amount.

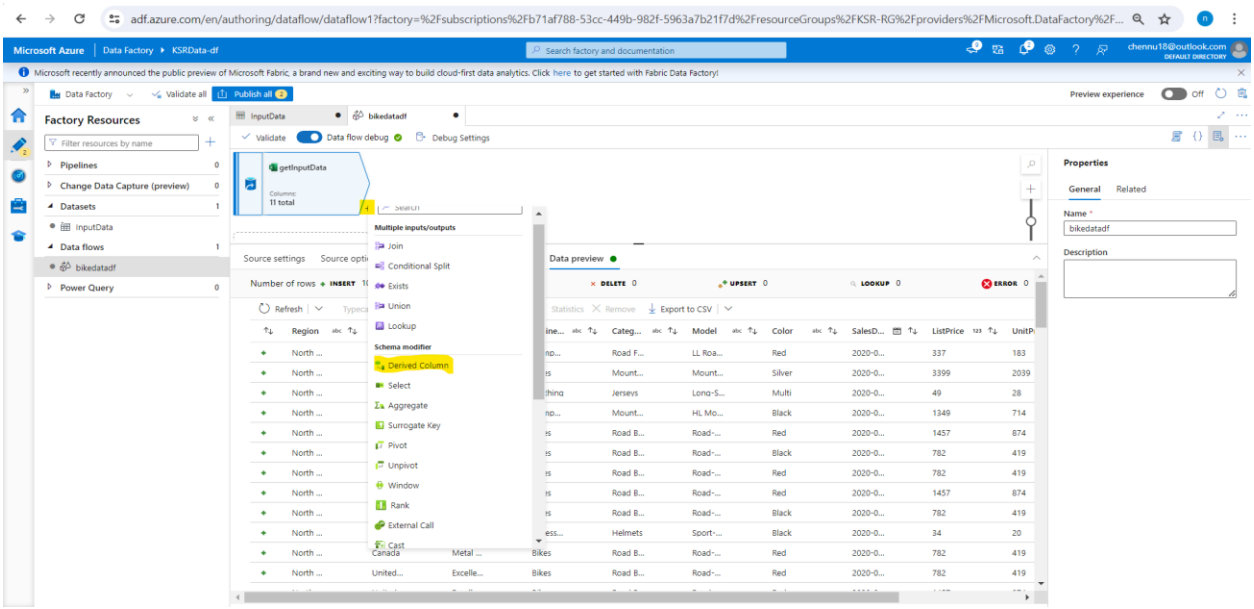




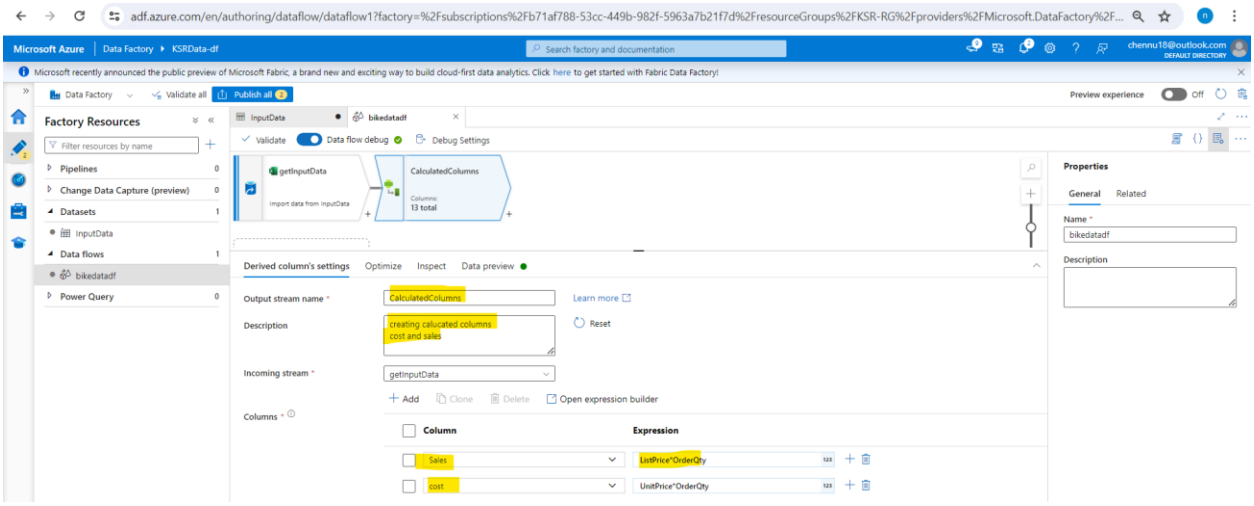




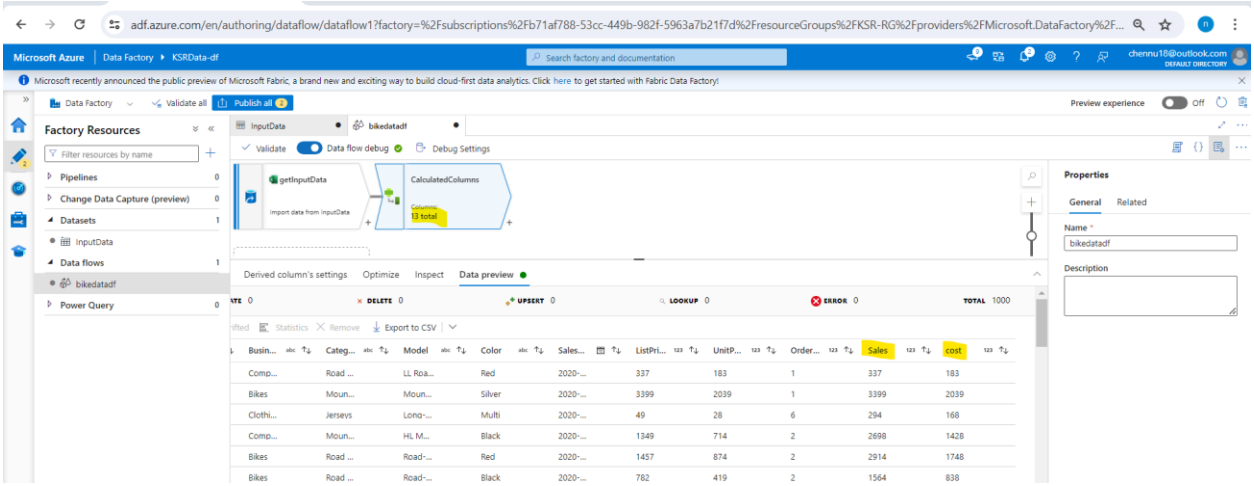
➔ We have create calculate column cost and sales using derived column activity, we can see below figure how to create derived column.



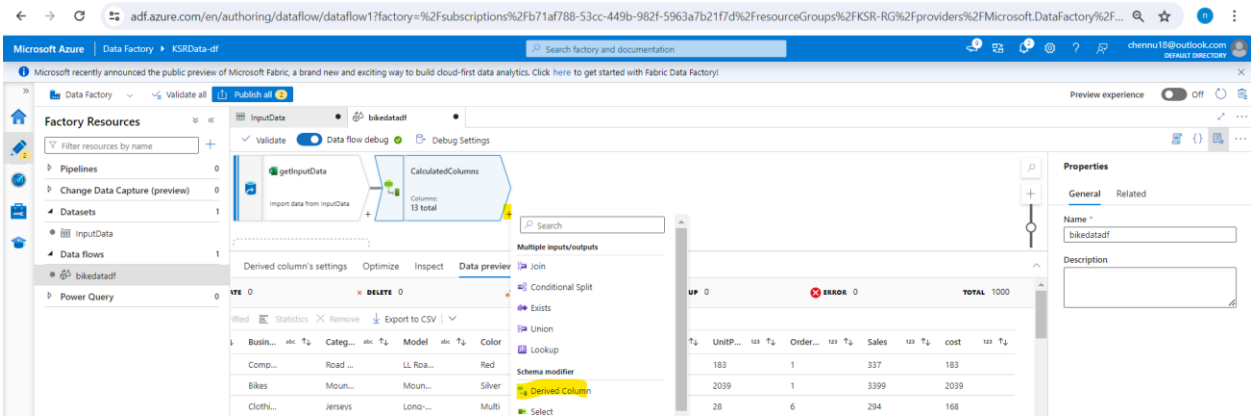
➔ Provide the required details for calculated columns → and you observe the column size increase 11 to 13



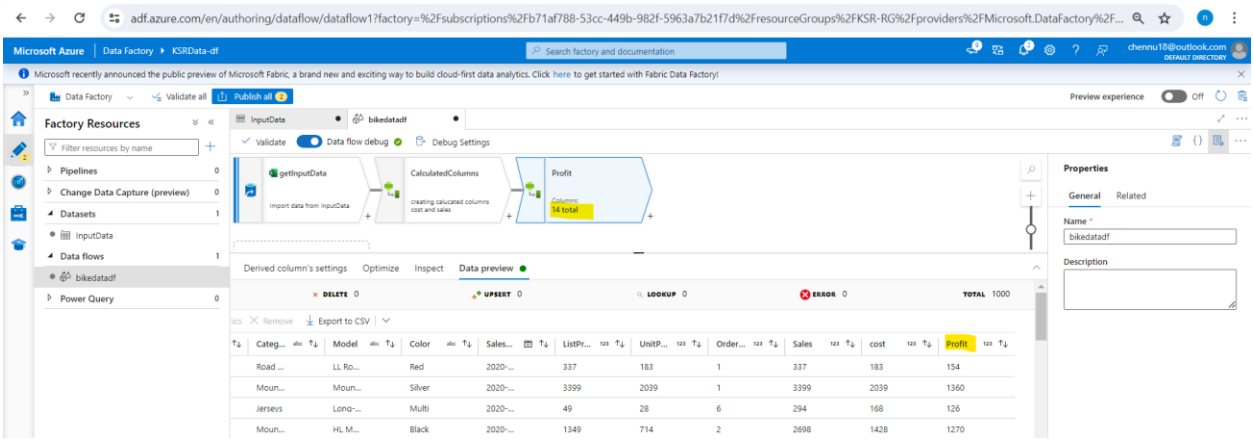
➔ Click on data preview see the calculated columns cost and sales in below figure.



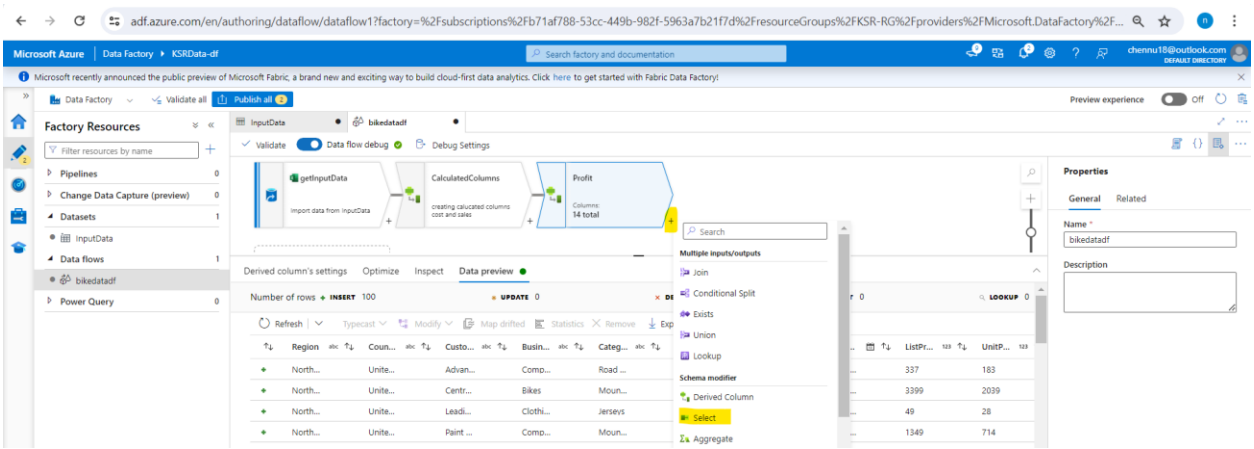
➔ Again Create derived column for calculating profit from the calculated column

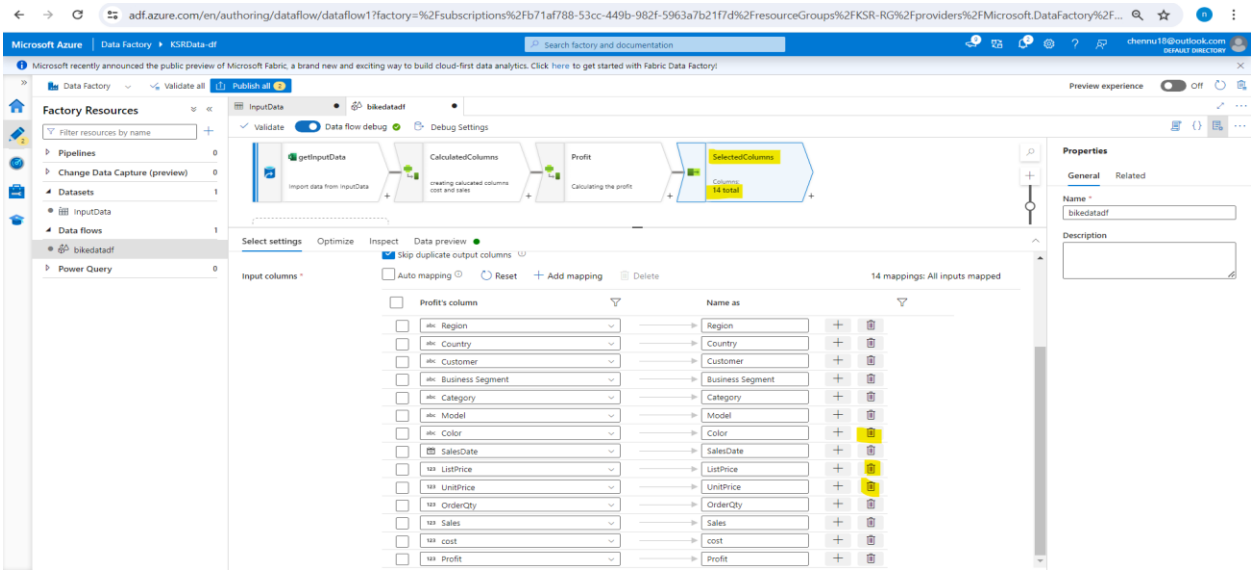


➔ When we click on data preview we can see profit column see in below diagram, the column size will increase 13 to 14

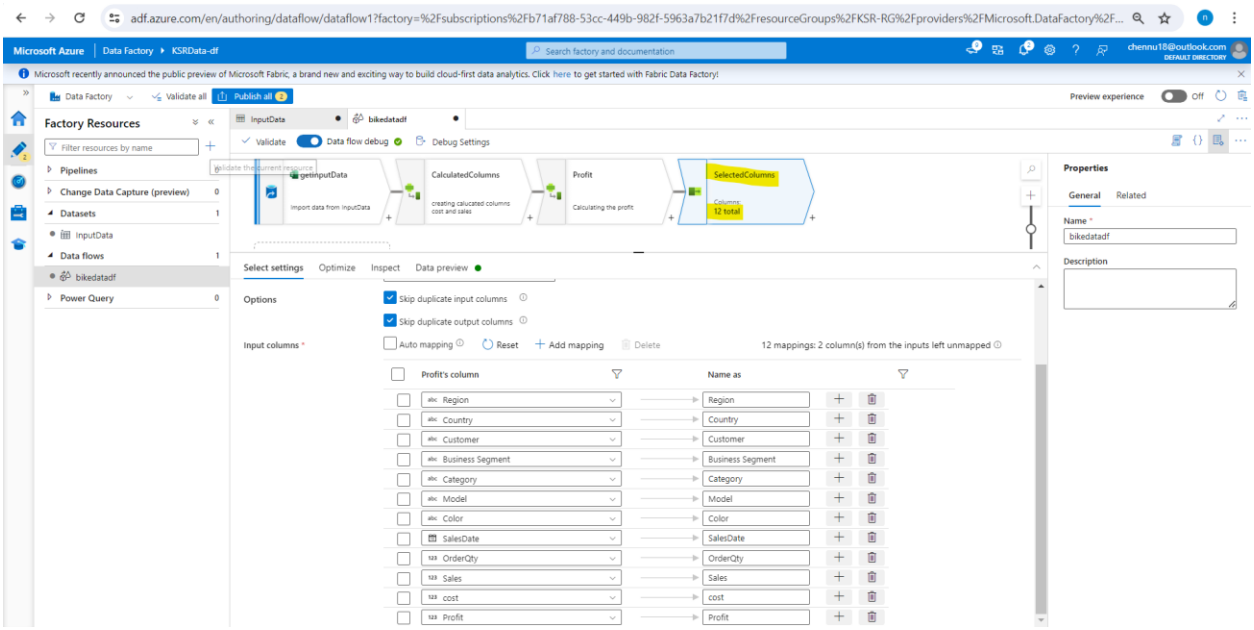


➔ We have to select required columns only, other columns we can delete list price and unit price, we can see following two diagrams

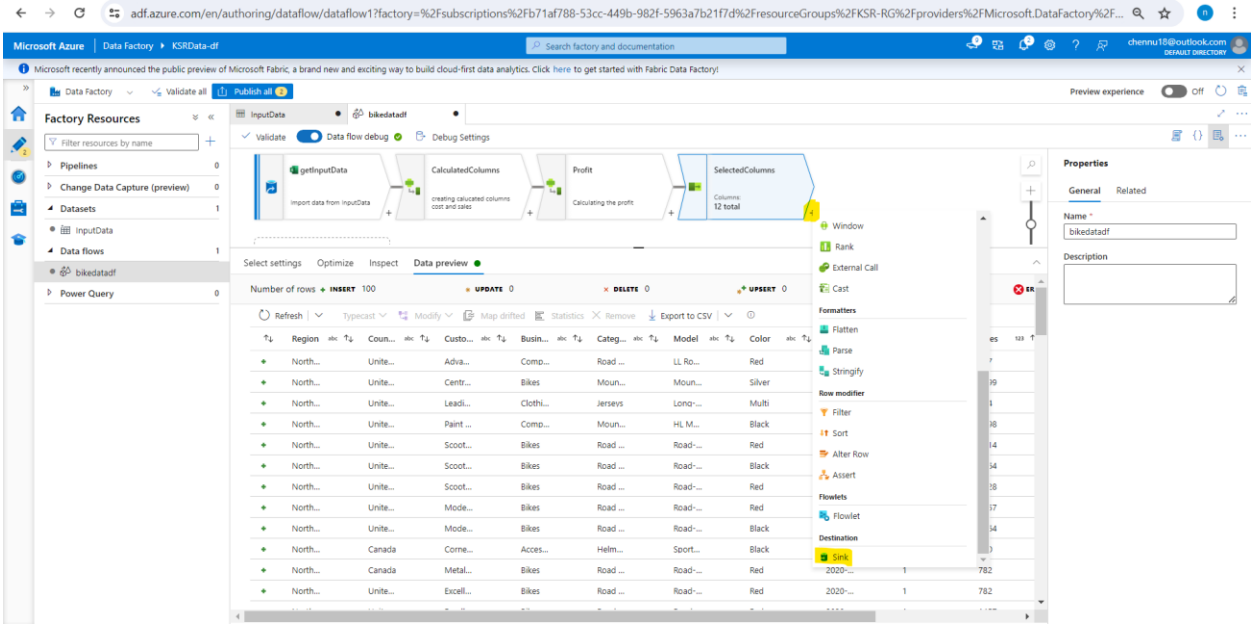




➔ Observe in below figure the column size reduces from 14 to 12



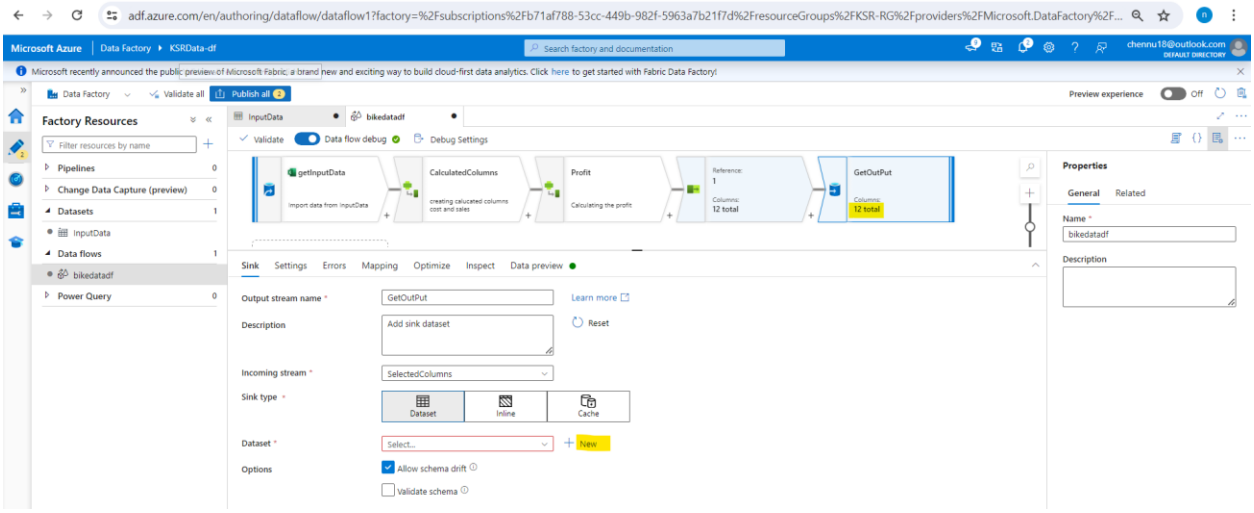
➔ Then Finally add destination here we called as sink activity how to create see below figure.

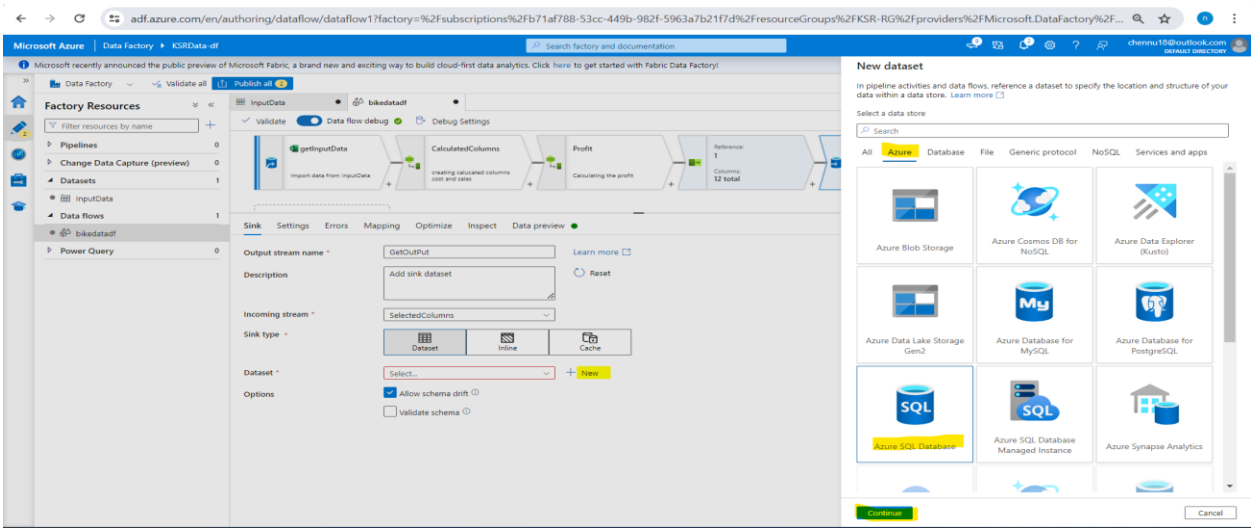


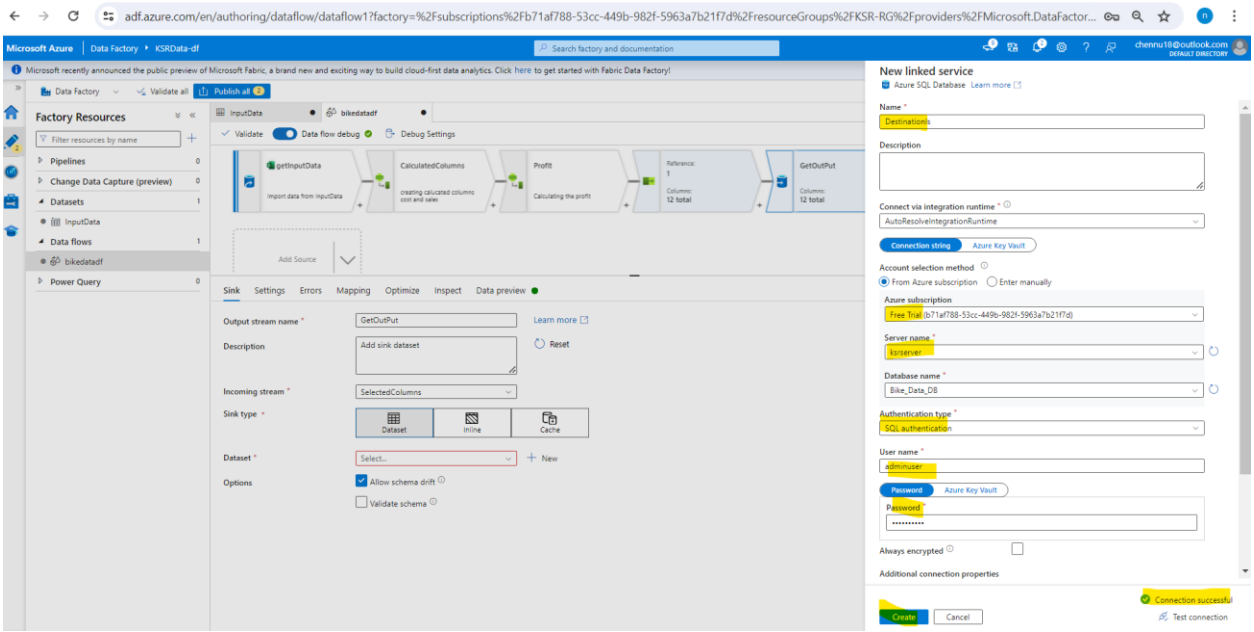
➔ Click on sink activity to provide required details and create new linked service for Azure SQL Database. You can see the process following figures how to do.

➔ Click new →Click azure → select Azure SQL Database → continue

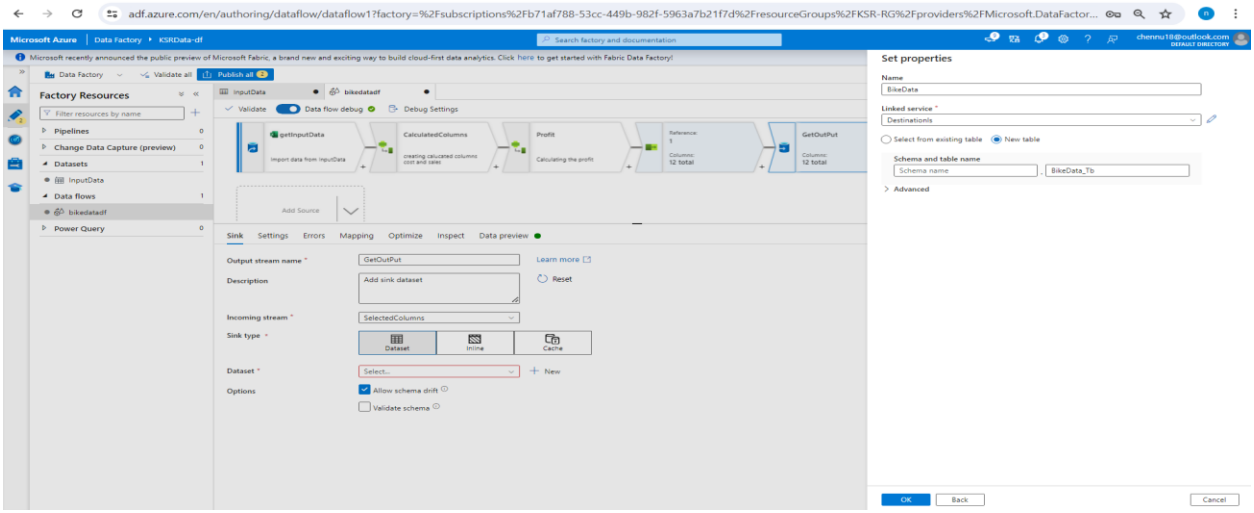
➔ To provide all the details for linked service and test the connection, once connection successfully → click ok.

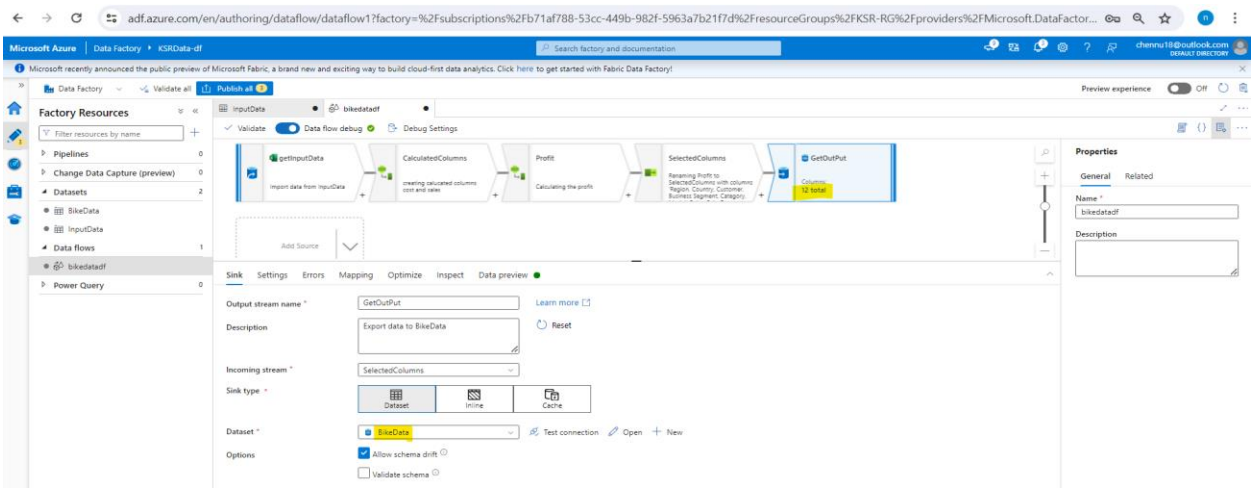

adf.azure.com/en/authoring/dataflow/dataflow1?factory=%2Fsubscriptions%2Fb71af788-53cc-449b-982f-5963a7b21f7d%2FresourceGroups%2FKSR-RG%2Fproviders%2FMicrosoft.DataFactory%2F...
Microsoft Azure
Data Factory ‣ KSRData-df
Search factory and documentation
chennu18@outlook.com
DEFAULT DIRECTORY
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all 2
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 0
Change Data Capture (preview) 0
Datasets 1
InputData
Data flows 1
bikedatadf
Power Query 0
InputData
bikedatadf
Validate
Data flow debug
Debug Settings
getInputData
Import data from InputData
CalculatedColumns
creating calucated columns cost and sales
Profit
Calculating the profit
Reference: 1
Columns: 12 total
GetOutPut
Columns: 12 total
Properties
General
Related
Name *
bikedatadf
Description
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Output stream name *
GetOutPut
Learn more
Description
Add sink dataset
Reset
Incoming stream *
SelectedColumns
Sink type *
Dataset
Inline
Cache
Dataset *
Select...
New
Options
Allow schema drift
Validate schema


adf.azure.com/en/authoring/dataflow/dataflow1?factory=%2Fsubscriptions%2Fb71af788-53cc-449b-982f-5963a7b21f7d%2FresourceGroups%2FKSR-RG%2Fproviders%2FMicrosoft.DataFactory%2F...
Microsoft Azure
Data Factory ‣ KSRData-df
Search factory and documentation
chennu18@outlook.com
DEFAULT DIRECTORY
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Factory Resources
Pipelines 0
Change Data Capture (preview) 0
Datasets 1
InputData
Data flows 1
bikedatadf
Power Query 0
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Output stream name *
GetOutPut
Learn more
Description
Add sink dataset
Reset
Incoming stream *
SelectedColumns
Sink type *
Dataset
Inline
Cache
Dataset *
Select...
New
Options
Allow schema drift
Validate schema
New dataset
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more
Select a data store
Search
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Blob Storage
Azure Cosmos DB for NoSQL
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen2
Azure Database for MySQL
Azure Database for PostgreSQL
Azure SQL Database
Azure SQL Database Managed Instance
Azure Synapse Analytics
Continue
Cancel

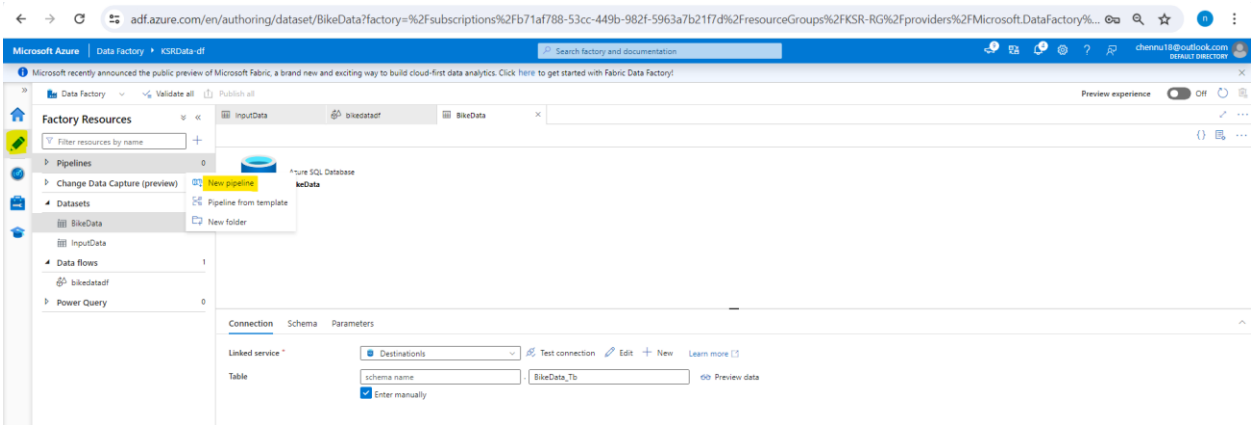

adf.azure.com/en/authoring/dataflow/dataflow1?factory=%2Fsubscriptions%2Fb71af788-53cc-449b-982f-5963a7b21f7d%2FresourceGroups%2FKSR-RG%2Fproviders%2FMicrosoft.DataFactor...
Microsoft Azure
Data Factory ‣ KSRData-df
Search factory and documentation
chennu18@outlook.com
DEFAULT DIRECTORY
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Factory Resources
Pipelines 0
Change Data Capture (preview) 0
Datasets 1
InputData
Data flows 1
bikedatadf
Power Query 0
Add Source
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Output stream name *
GetOutPut
Learn more
Description
Add sink dataset
Reset
Incoming stream *
SelectedColumns
Sink type *
Dataset
Inline
Cache
Dataset *
Select...
New
Options
Allow schema drift
Validate schema
New linked service
Azure SQL Database Learn more
Name *
Destinations
Description
Connect via integration runtime *
AutoResolveIntegrationRuntime
Connection string
Azure Key Vault
Account selection method
From Azure subscription
Enter manually
Azure subscription
Free Trial (b71af788-53cc-449b-982f-5963a7b21f7d)
Server name *
ksrserver
Database name *
Bike_Data_DB
Authentication type *
SQL authentication
User name *
adminuser
Password
Azure Key Vault
Password *
Always encrypted
Additional connection properties
Connection successful
Create
Cancel
Test connection

➔ Now select new table option → then we provide new table name → click ok

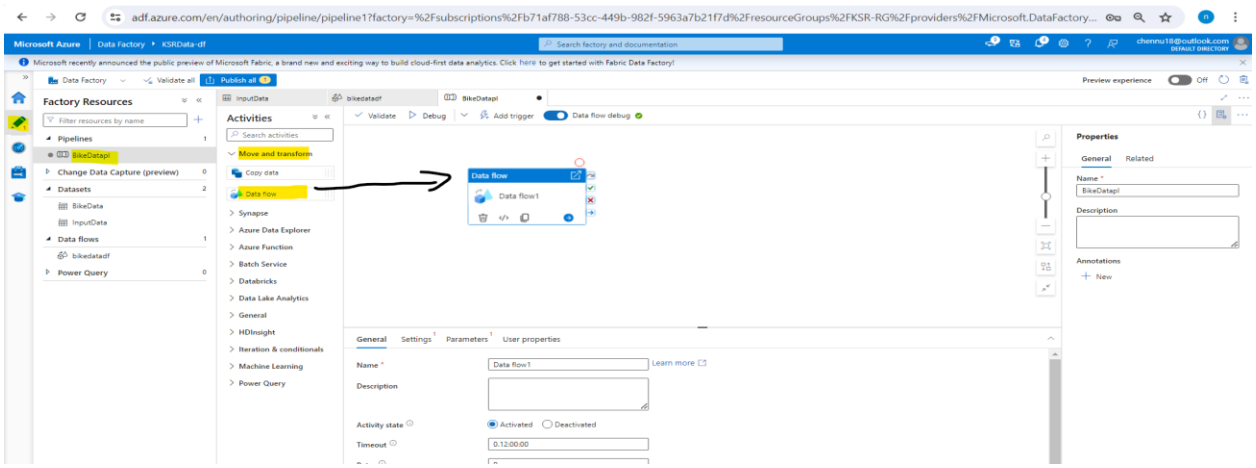




➔ Create a pipeline using Author Tab(Pencil Icon) → right click on that → click on new pipeline

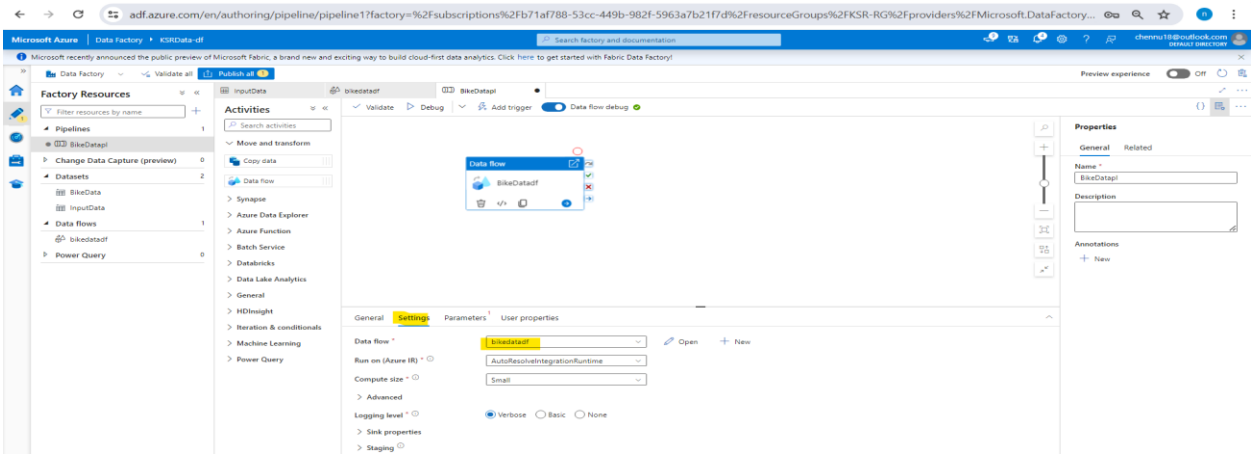


➔ Drop and drag the data flow activity into canvas page

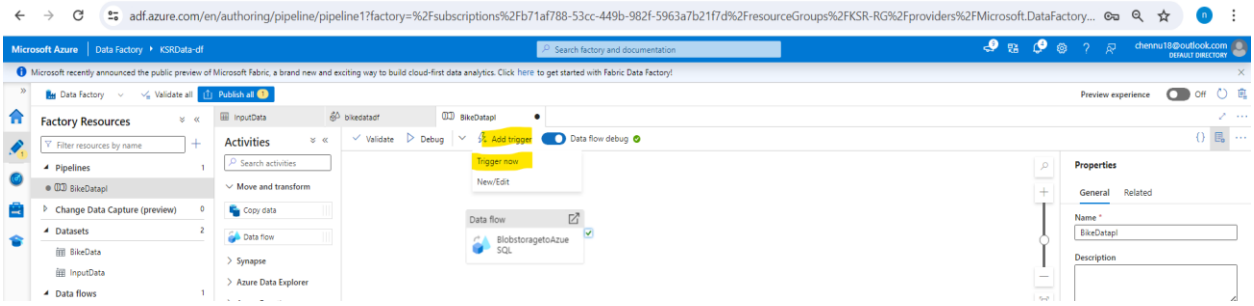


→ In settings tab we need select already created data flow under dataflow resource see in below figure

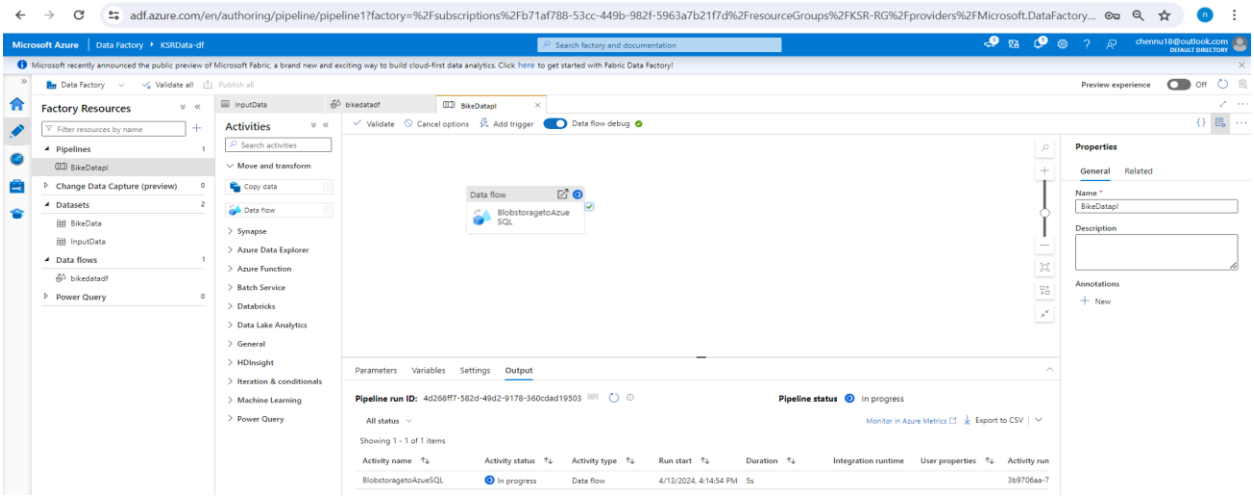
➔ Just click on publish all we can save the data, if you not click on publish all you lose all work whatever you done.



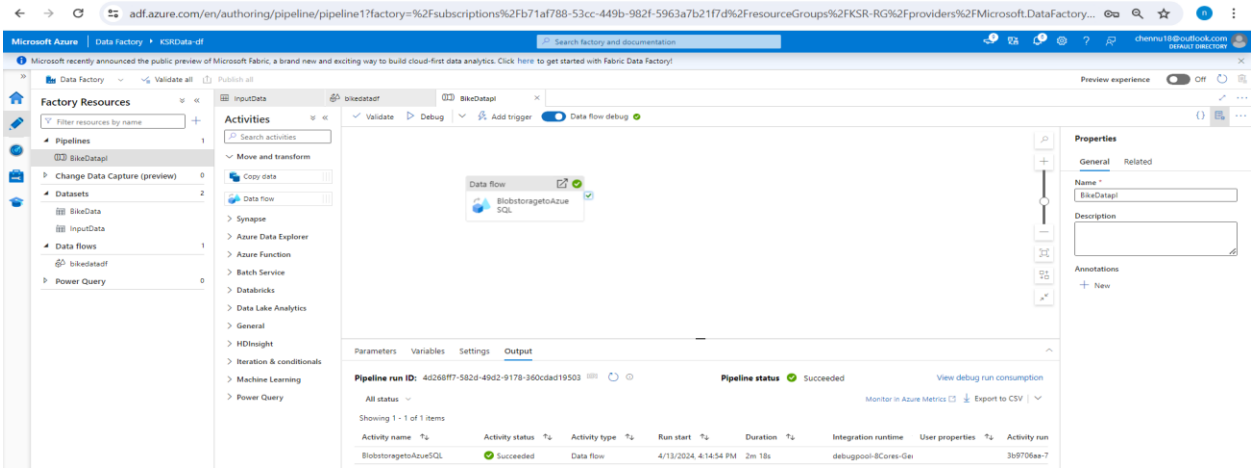
➔ After creating a pipeline, we can run that pipeline using trigger tab and use debug tab and also we can schedule it using triggers tab.



➔ After click on debug we will see the status like in below figure.



➔ It took less than 3 minutes to load data from source(Azure data lake) to destination(Azure SQL DataBase).



To validate, data loaded into destination → go to SQL Database → To SQL code see in figure