Varthya Akhil

Index

Agenda	1
Services Required	1
Pipeline Architecture	2
Azure Blob Storage Creation	3
Azure SQL Database Creation	9
Azure Data Factory Creation	13
Transformation of data using Data Flow	15

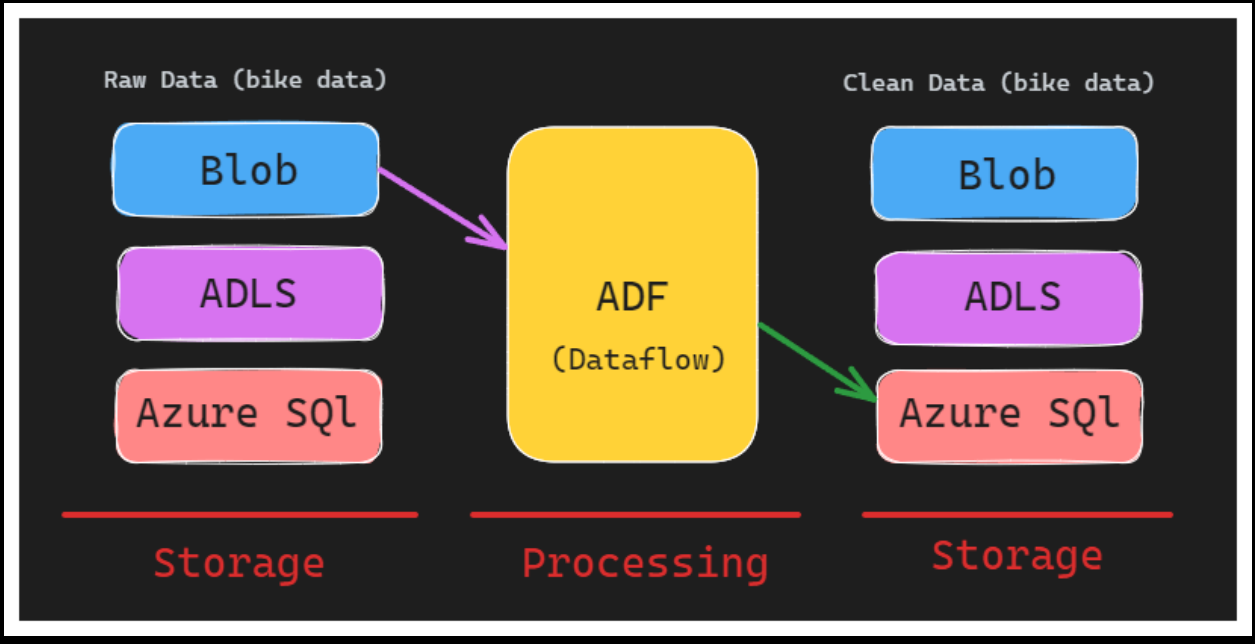
Agenda

Create a data pipeline to transform the bike data from azure blob storage using azure data factory and store the clean data into azure sql database.

Services Required :

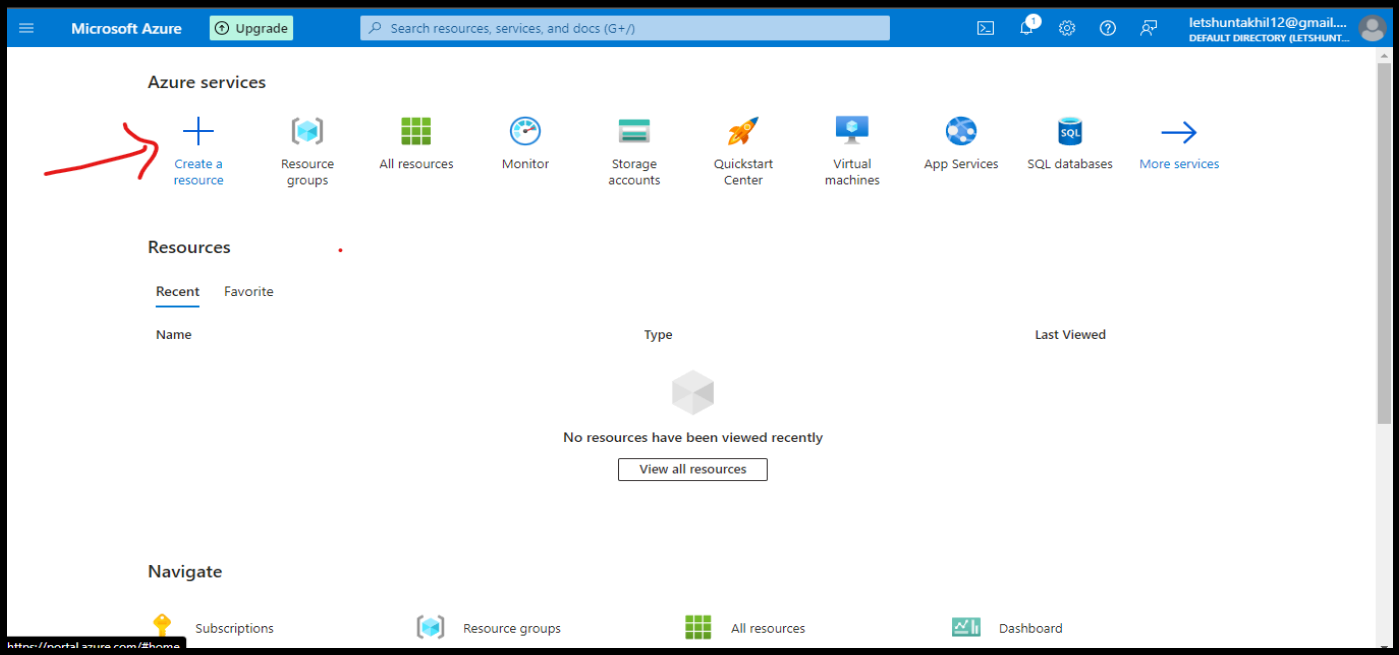
1. Azure Blob Storage
2. Azure Data Factory
3. Azure SQL Database

Pipeline Architecture :

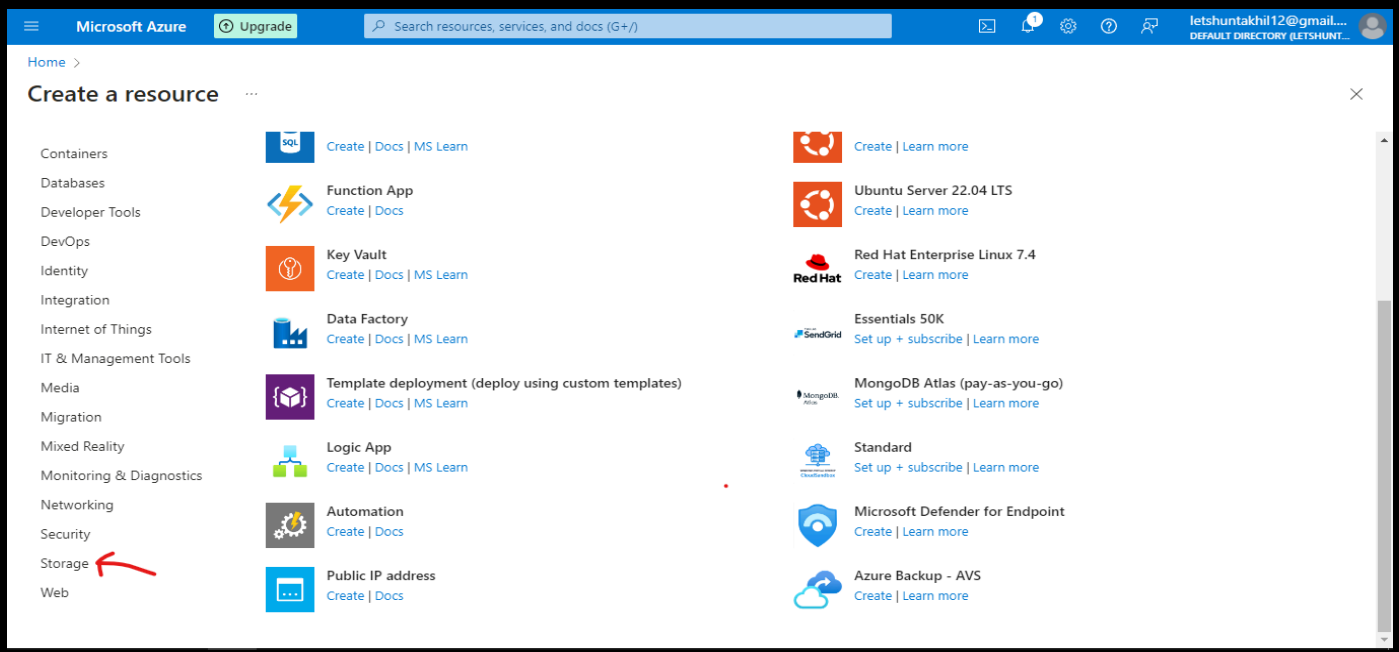


Azure blob storage Creation:

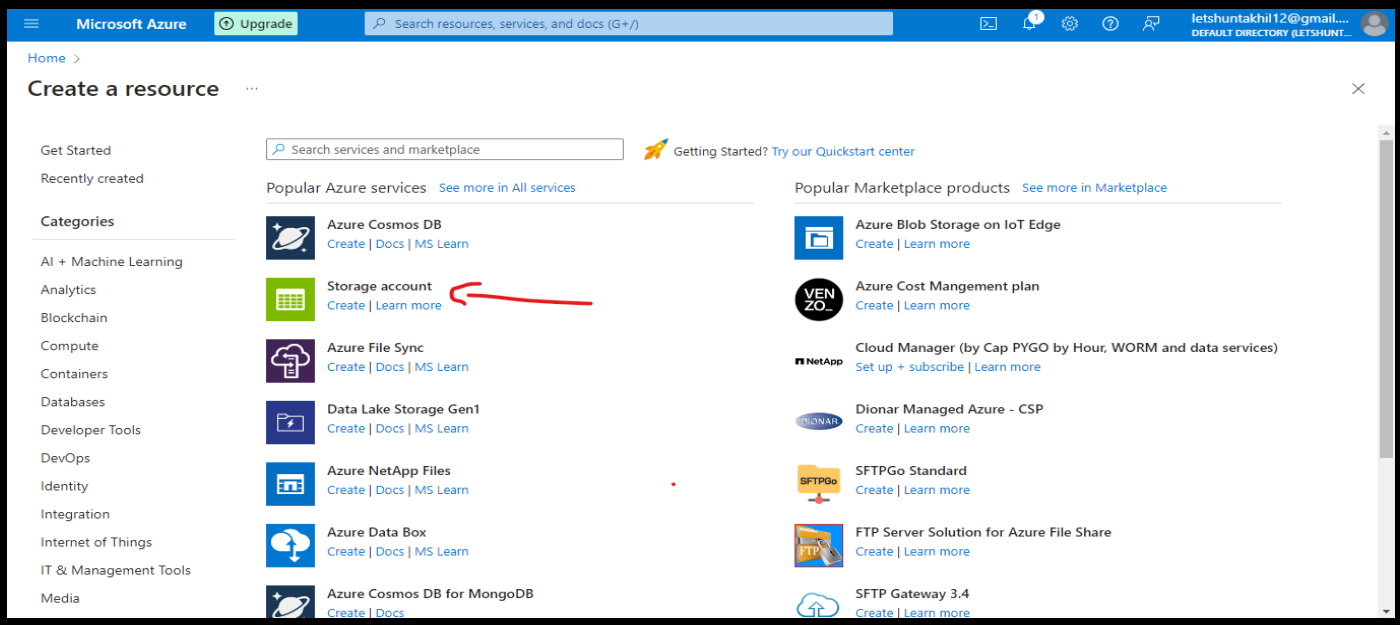
- Login the azure portal portal.azure.com by providing credentials.
- Home page will appear after successful login.
- Click on ' create a resource '.



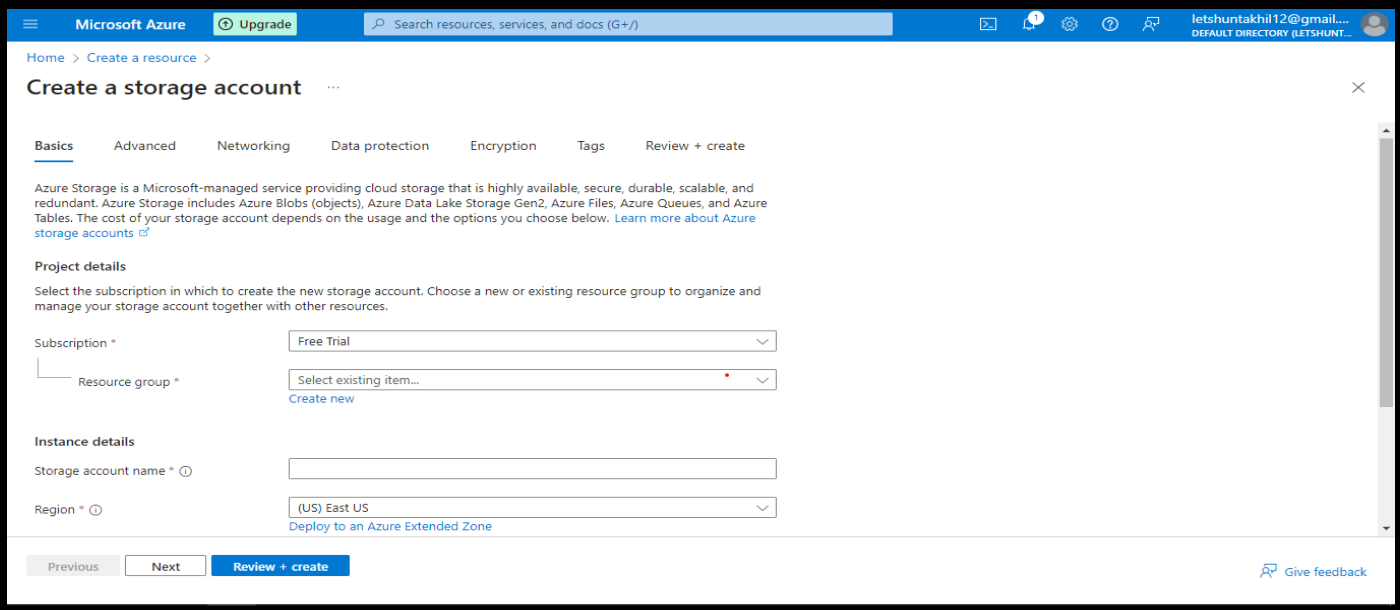
- Click on Storage & search for the storage account in the azure marketplace.



- Click on the storage account from the list.

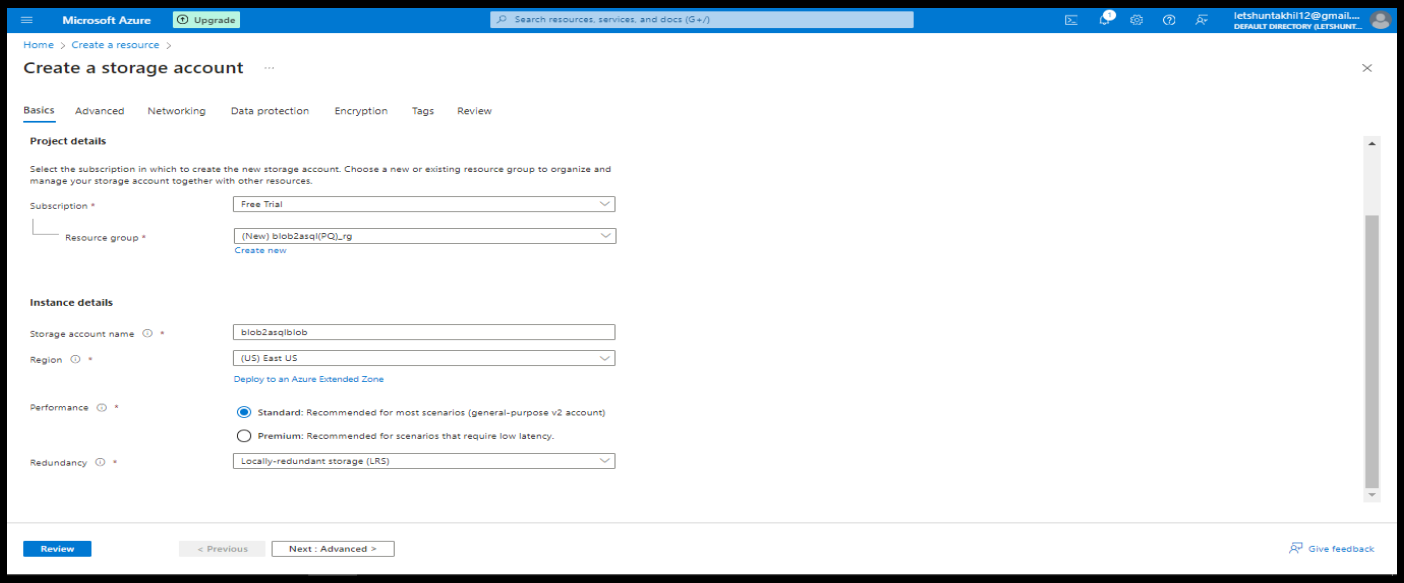


- You will get a form to fill up, so that we can make a request to create a storage account.

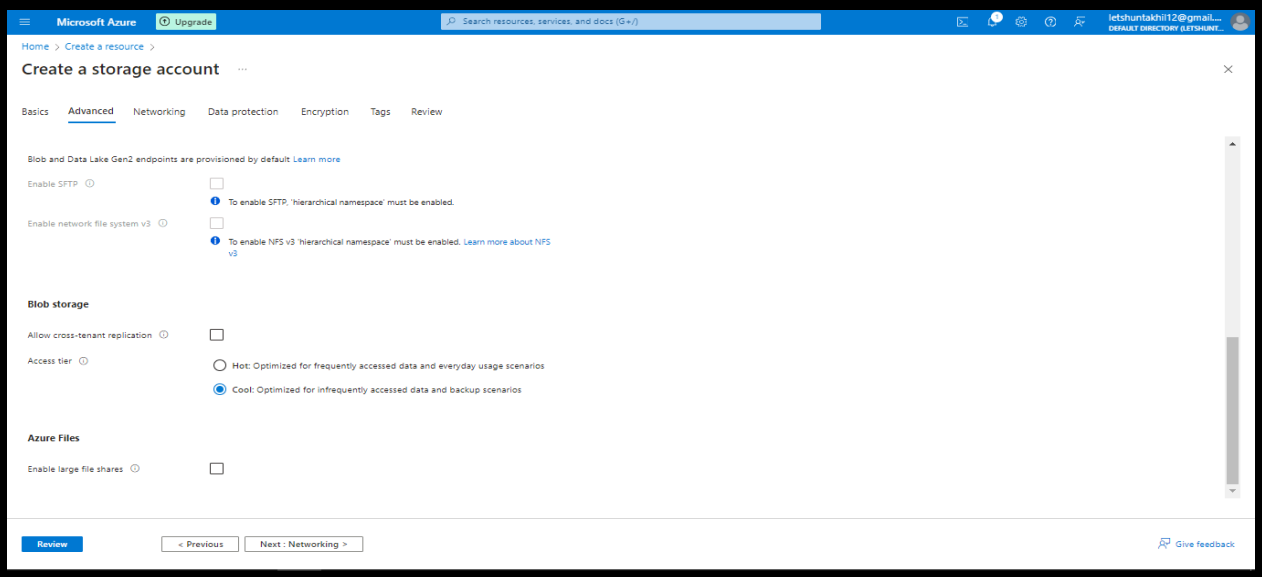


- In this form, we have six tabs to fill up the fields such as Basics, Advanced, Data Protection, Encryption, Tags and Review+Create.

- In the Basics Tab fill the details.
  - Project details
    - Subscription : Choose ‘ Free Trial ’ or ‘ Pay As You Go ’.
      - Choose Pay as You go if you don’t have the free trial
    - Resource Group : Create a new resource group if you don't have one.
  - Instance details
    - Storage account name : provide an account name such that **The field can contain only lowercase letters and numbers. Name must be between 3 and 24 characters and it should be unique.**
    - Region : choose a region such that this particular service I.e. azure blob storage can be created. Suppose if we choose East (US) , this service is created in this region.
    - Performance :
      - Standard : Recommended for most scenarios (general purpose v2 account).
      - Premium : Recommended for scenarios that require low latency.
    - Redundancy (Backup) : select anyone based on requirement.
      - Locally redundant Storage (LRS).
      - Geo Redundant Storage (GRS).
      - Zone Redundant Storage (ZRS).
      - Geo-Zone Redundant Storage (GZRS).

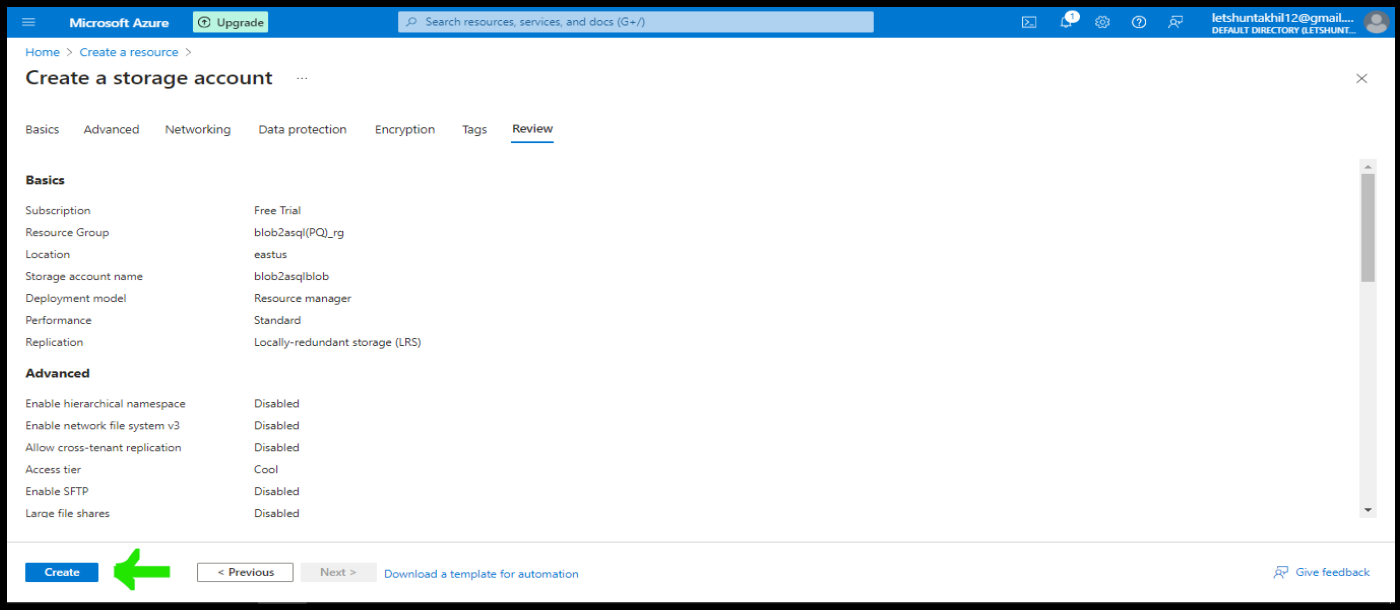


- Click next (Advanced) .
- In the advanced tab :
  - Leave the fields as it is but change the Access tier to Cool.
  - Access Tier :
    - Hot : Optimized for frequently accessed data and everyday usage scenarios.
    - Cool : Optimized for infrequently accessed data and backup scenarios.

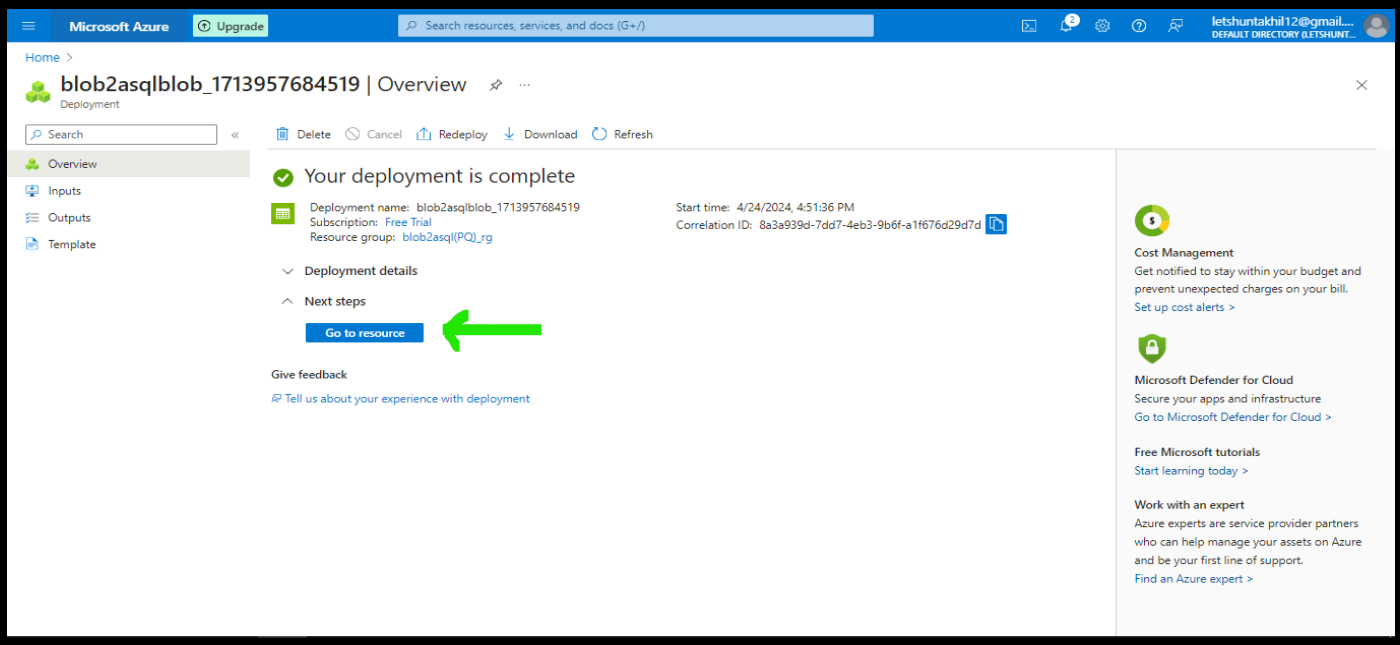


- Click next (Networking) .

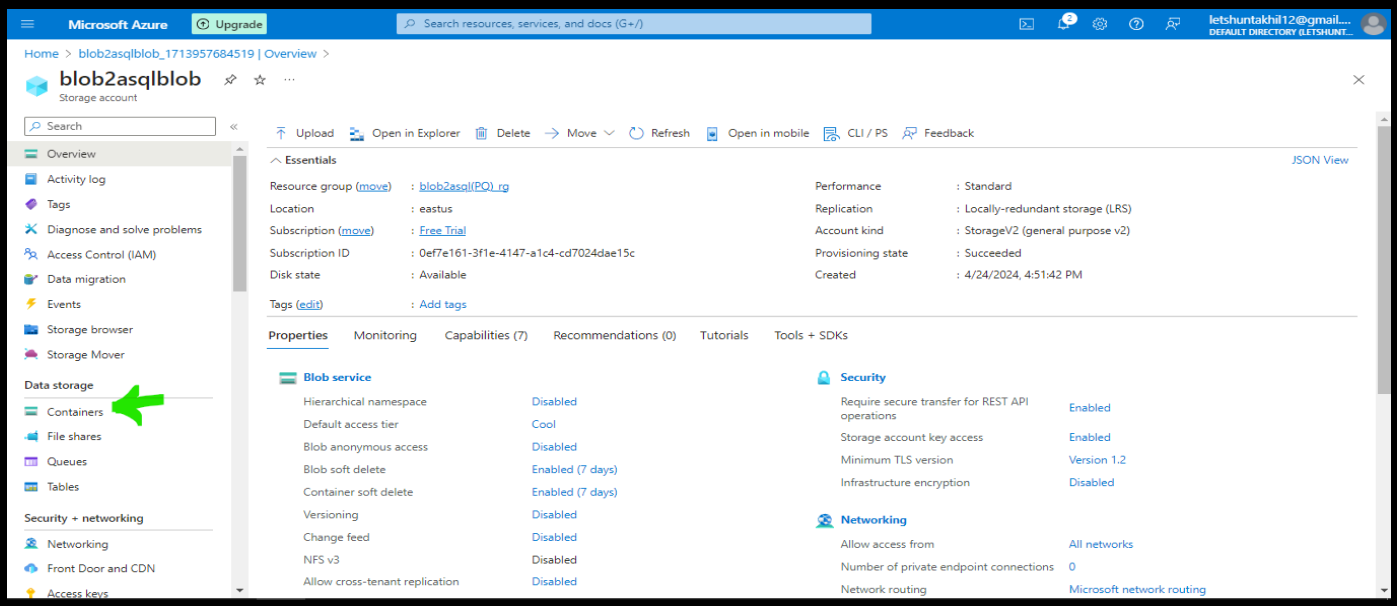
- Networking : Keep it default -> click next (Data protection) .
- Data protection : Keep it default -> click next (Encryption) .
- Encryption : Keep it default -> click next (Tags) .
- Tags : Keep it default -> click next (Review + Create) .
- In Review + Create verify the details and click create.



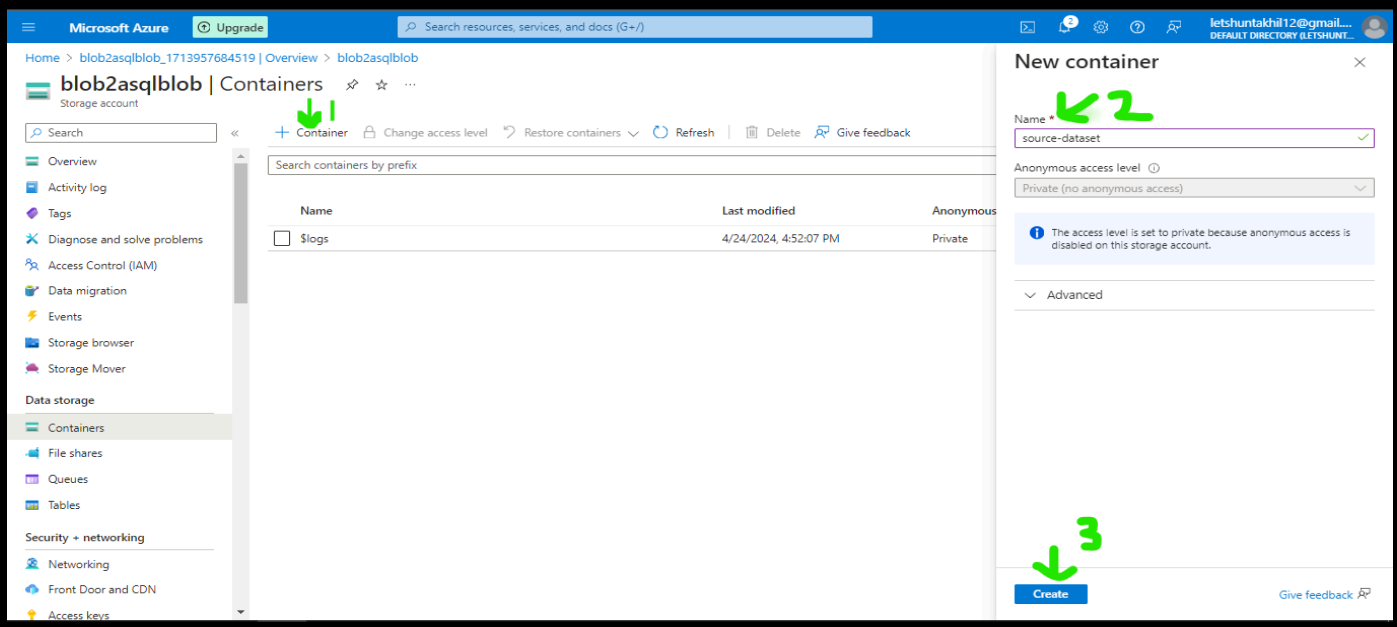
- Azure blob storage has been created and click on ' Go to Resource ' .



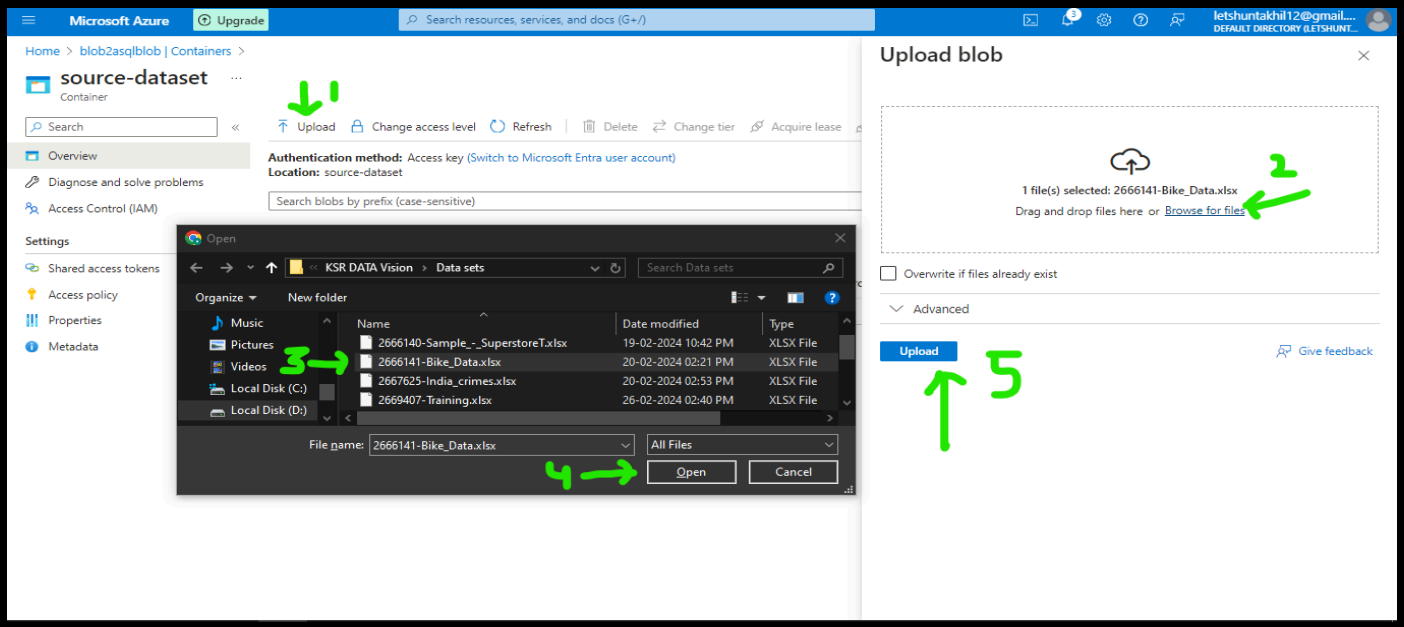
- Click on the ' containers ' in the menu list.



- Create a container by clicking ' + Container '.
- Fill the form after clicking ' + Container ' and then click ' create '.



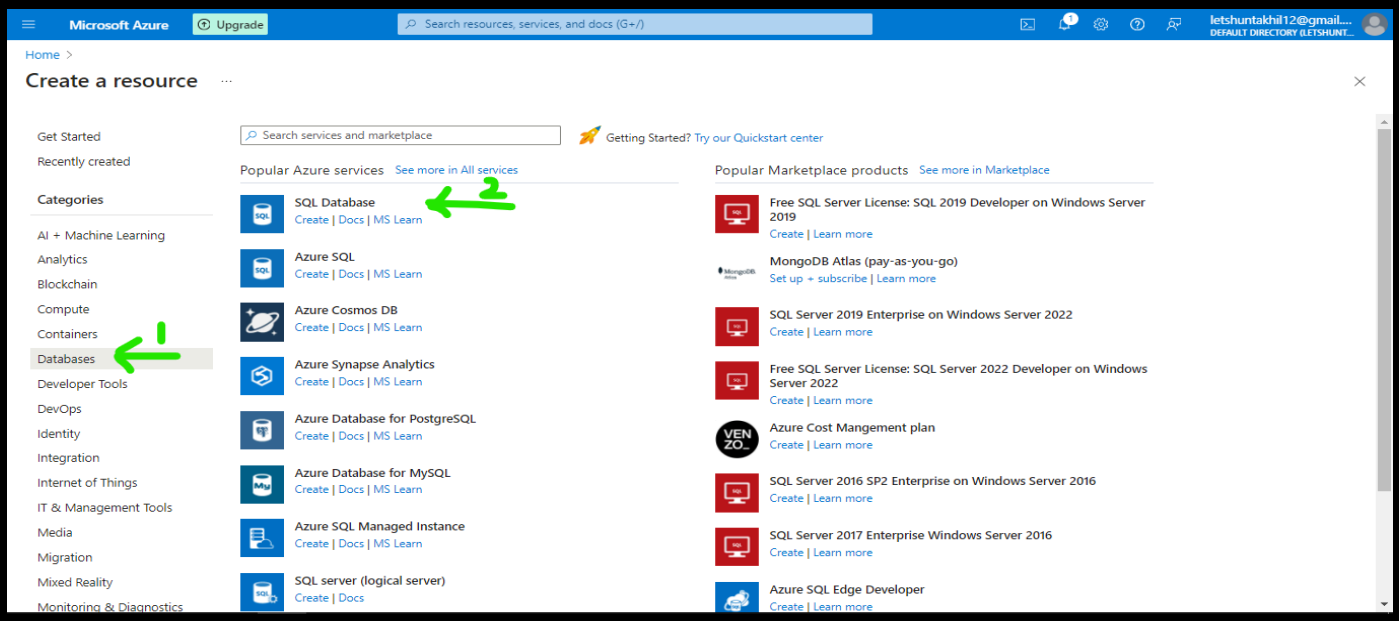
- Click on the container ' source-dataset ' after creating.
- Click Upload –> Browse for files –> select the file –> open –> upload .
- The bike data is uploaded.



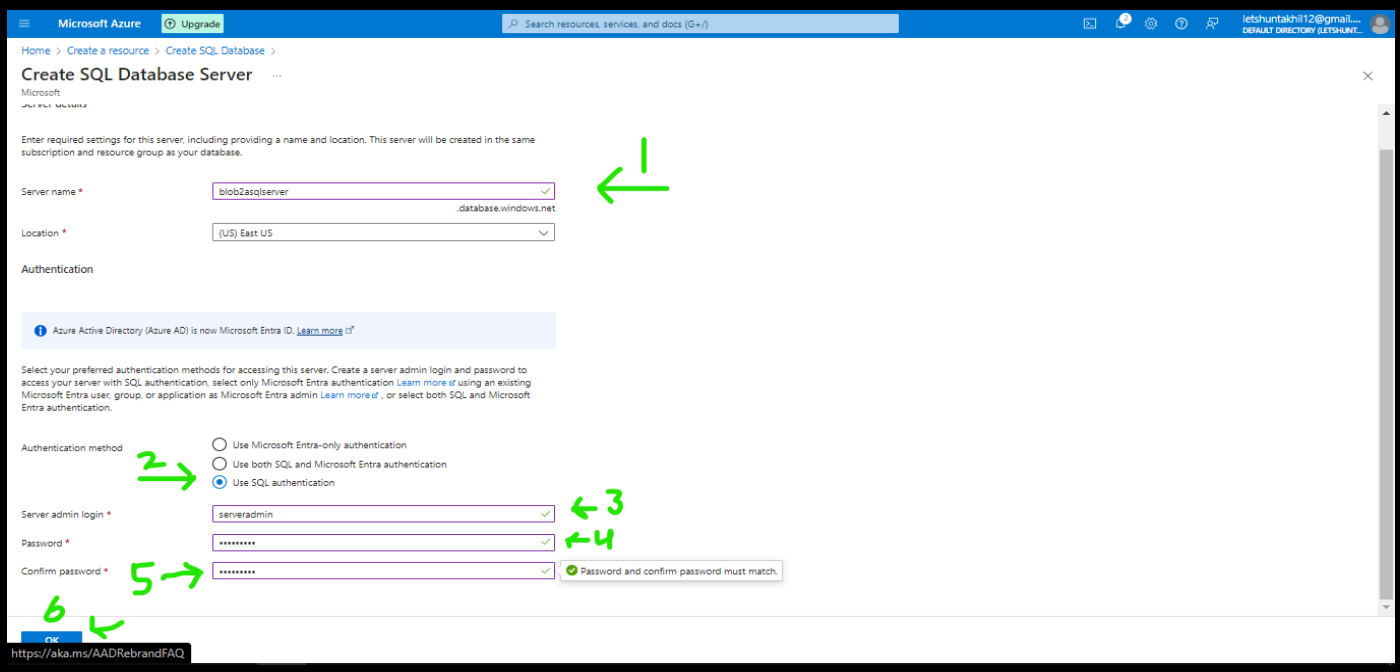
- Successfully we have created a storage account (blob) , container and uploaded the file.

Azure SQL Database Creation :

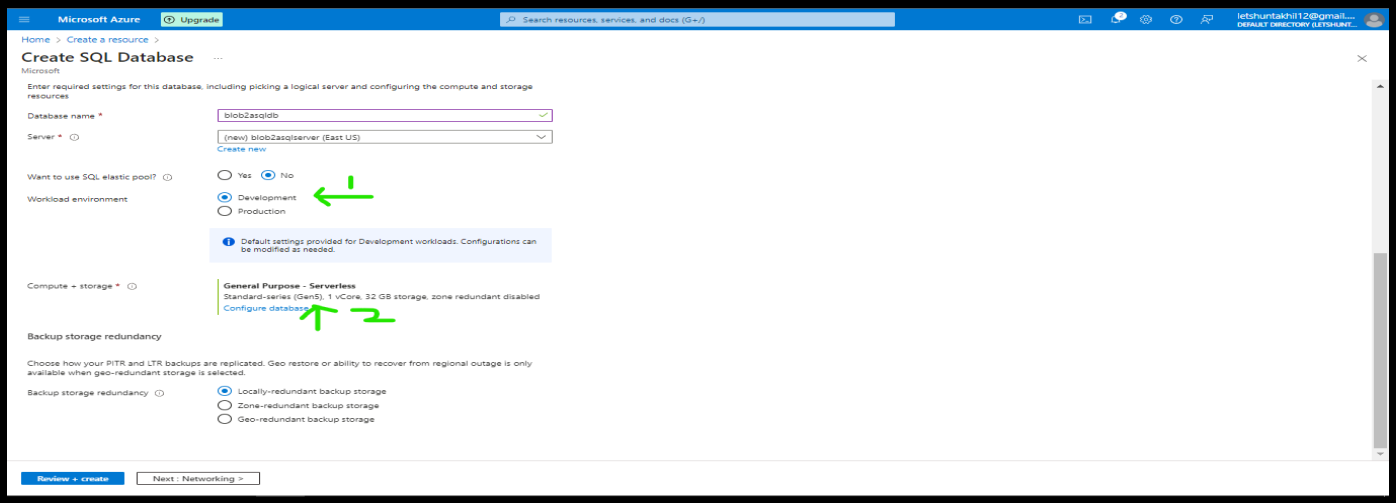
- Click on ' Go To Resource ' by opening the homepage in the new tab.
- In the menu list , search for ' Databases 'and click on it.
- Look for ' SQL Database ' in the azure marketplace — > click.

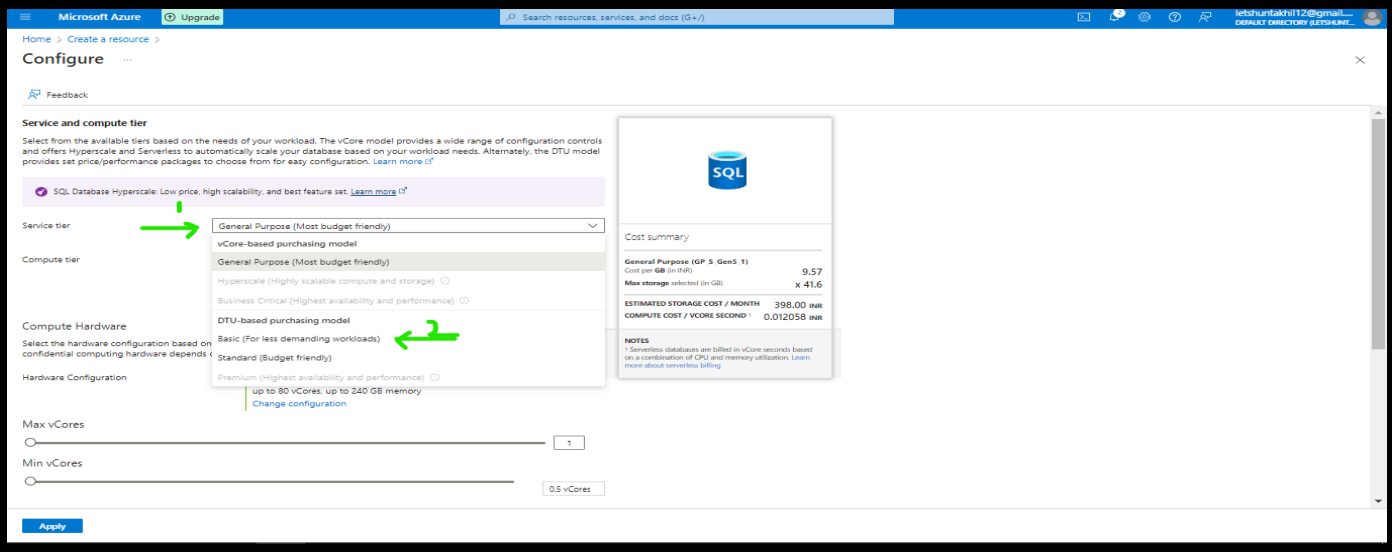


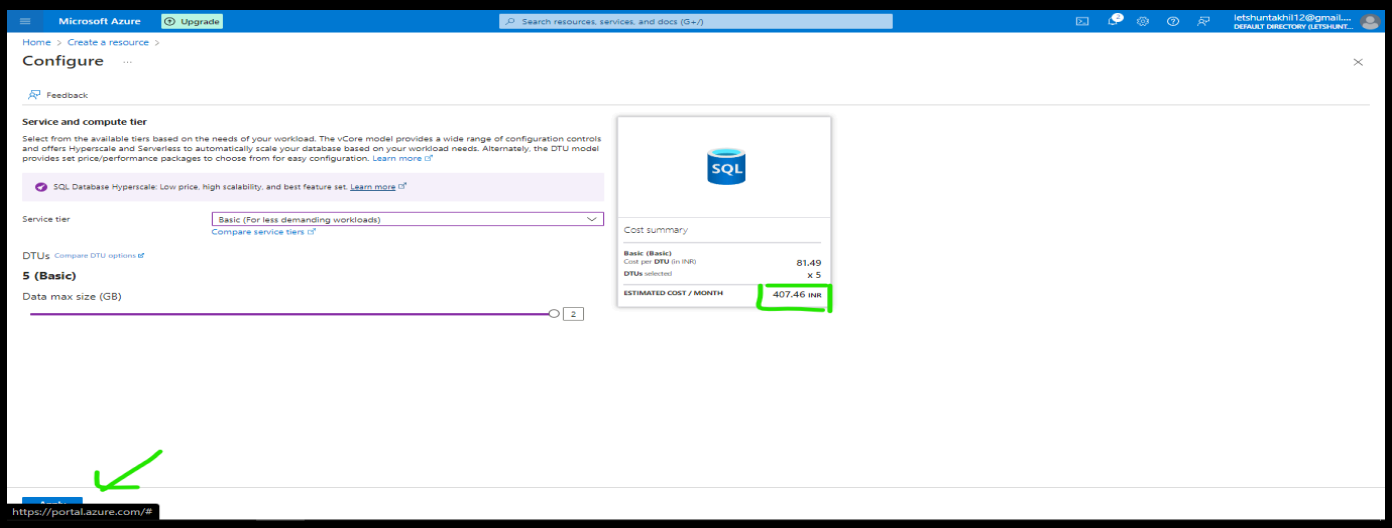
- Fillup the form to request creation of SQL Database .
- In the form we have 6 tabs → Basics, Networking, Security, Additional settings, Tags and ' Review + Create '.
- Basics form :
  - Project Details
    - Subscription : Free Trial
    - Resource Group : Select from the list
  - Database details
    - Database name : Provide the database name'
    - Server : Create a new server if you didn't create it. You need to fill the form when you create a new server.



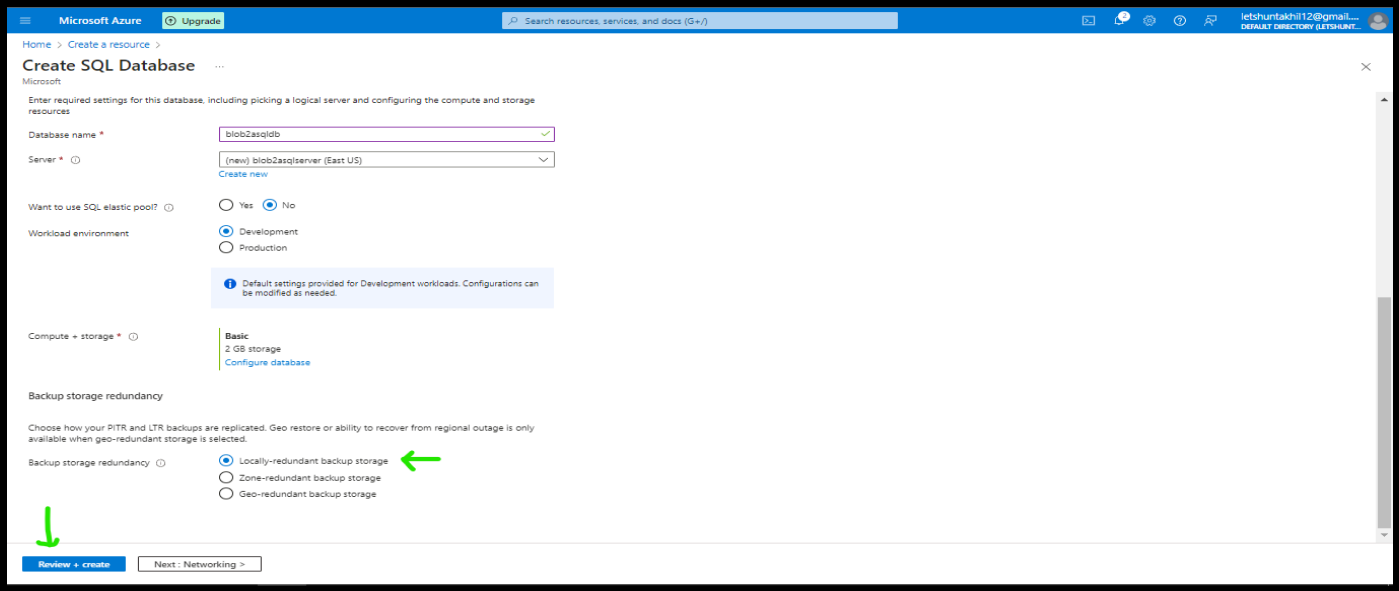
  - Select the Workload environment as development and configure the compute and storage as per the requirement.
  - Workload environment : Development
  - Compute + Storage : DTU → Basic (for less demanding workloads).


Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Home > Create a resource >
Create SQL Database
Microsoft
Enter required settings for this database, including picking a logical server and configuring the compute and storage resources
Database name *
blob2asqldb
Server *
(new) blob2asqlserver (East US)
Create new
Want to use SQL elastic pool?
Yes
No
Workload environment
Development
Production
Default settings provided for Development workloads. Configurations can be modified as needed.
Compute + storage *
General Purpose - Serverless
Standard-series (Gen5), 1 vCore, 32 GB storage, zone redundant disabled
Configure database
Backup storage redundancy
Choose how your PITR and LTR backups are replicated. Geo restore or ability to recover from regional outage is only available when geo-redundant storage is selected.
Backup storage redundancy
Locally-redundant backup storage
Zone-redundant backup storage
Geo-redundant backup storage
Review + create
Next : Networking >

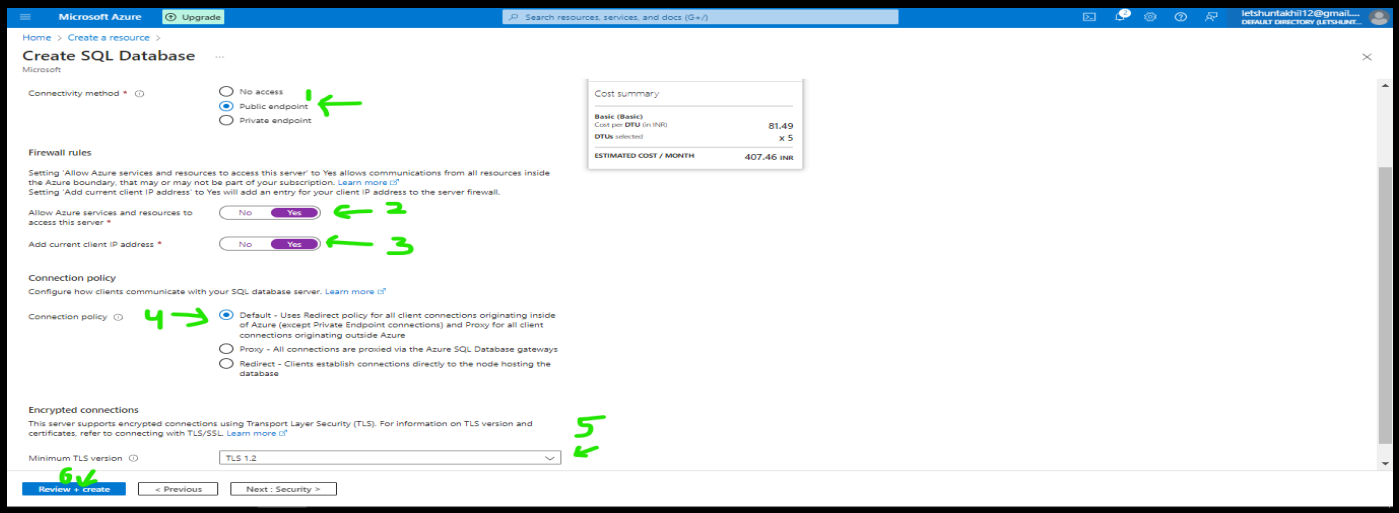

Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Home > Create a resource >
Configure
Feedback
Service and compute tier
Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. Learn more
SQL Database Hyperscale: Low price, high scalability, and best feature set. Learn more
Service tier
General Purpose (Most budget friendly)
vCore-based purchasing model
General Purpose (Most budget friendly)
Hyperscale (Highly scalable compute and storage)
Business Critical (Highest availability and performance)
DTU-based purchasing model
Basic (For less demanding workloads)
Standard (Budget friendly)
Premium (Highest availability and performance)
Compute tier
Compute Hardware
Hardware Configuration
up to 80 vCores, up to 240 GB memory
Change configuration
Max vCores
1
Min vCores
0.5 vCores
Cost summary
General Purpose (GP_S_Gen5_1)
Cost per GB (in INR)
9.57
Max storage selected (in GB)
x 41.6
ESTIMATED STORAGE COST / MONTH
398.00 INR
COMPUTE COST / VCORE SECOND
0.012058 INR
NOTES
Serverless databases are billed in vCore seconds based on a combination of CPU and memory utilization. Learn more about serverless billing
Apply


Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Home > Create a resource >
Configure
Feedback
Service and compute tier
Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. Learn more
SQL Database Hyperscale: Low price, high scalability, and best feature set. Learn more
Service tier
Basic (For less demanding workloads)
Compare service tiers
DTUs
Compare DTU options
5 (Basic)
Data max size (GB)
2
Cost summary
Basic (Basic)
Cost per DTU (in INR)
81.49
DTUs selected
x 5
ESTIMATED COST / MONTH
407.46 INR
https://portal.azure.com/#

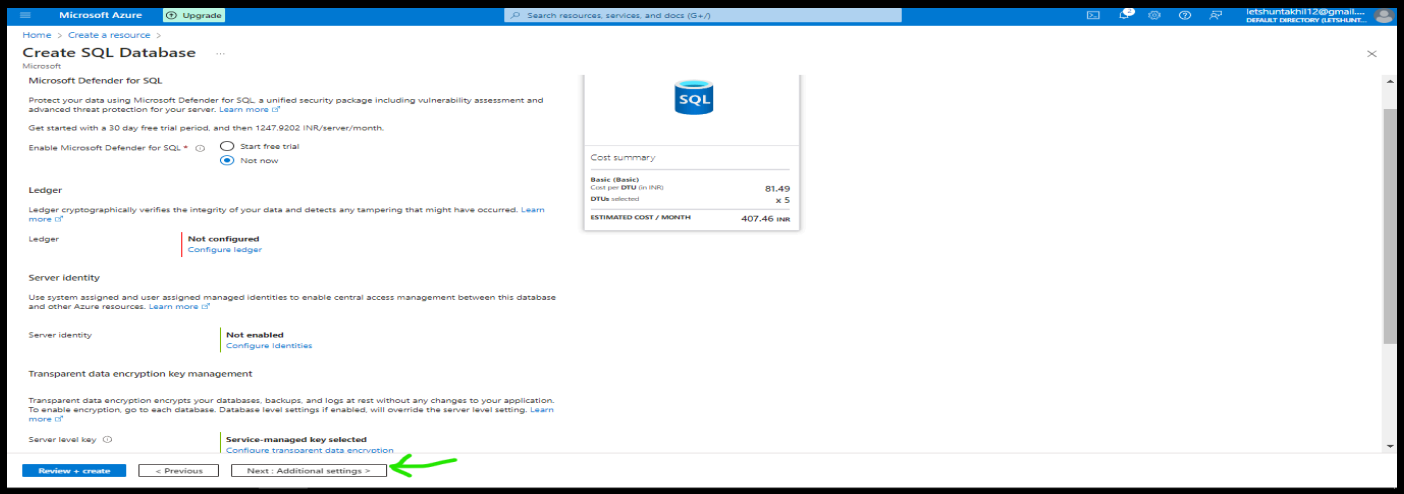
- Backup storage redundancy
    - Backup storage redundancy : Locally redundant backup storage
- Click next (Networking).



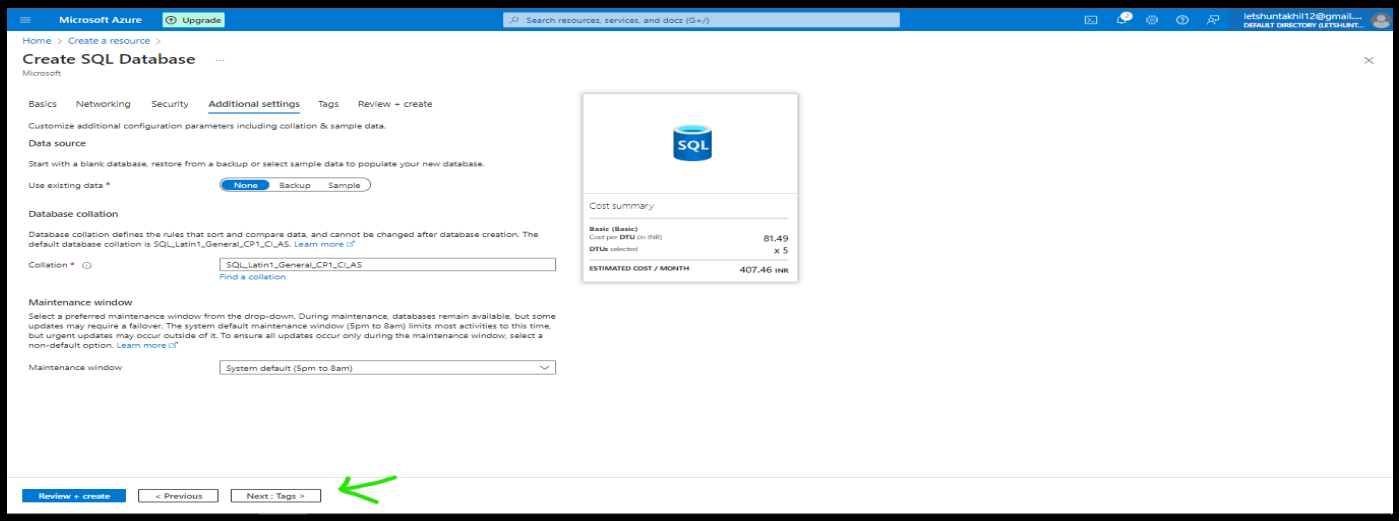
- Networking form :
    - Network connectivity
        - Connectivity method : public endpoint
            - Firewall Rules
                - Allow azure services and resources to access this server : Yes
                - Add current client ip address : Yes
            - Connection Policy
                - Connection policy : default
            - Encrypted Connections
                - Minimum TLS version : TLS 1.2
- Click on next (Security)



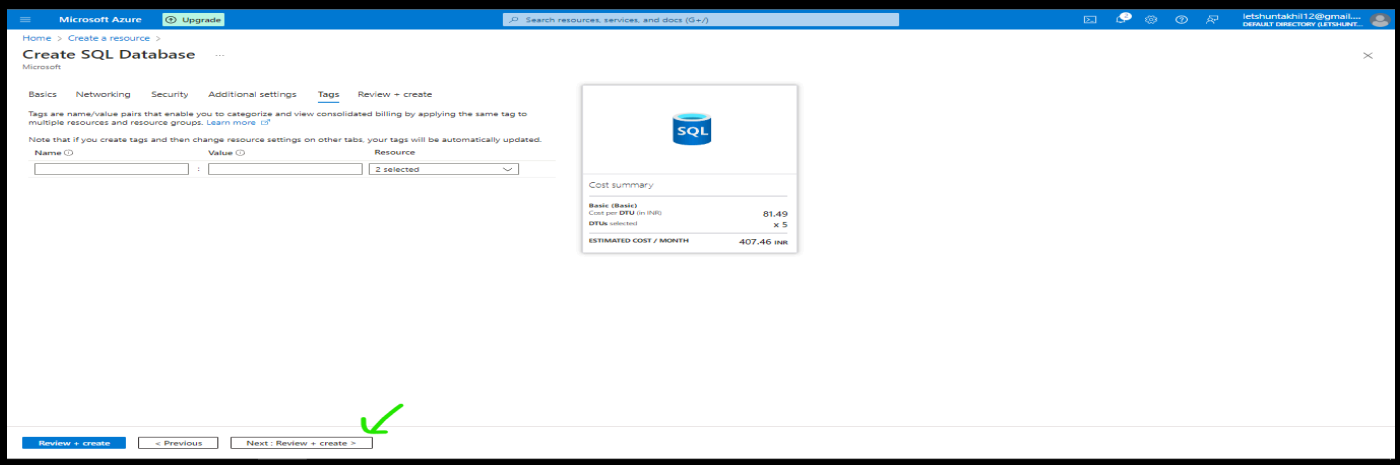
- Security form : Keep it default —> click next (Additional settings) .



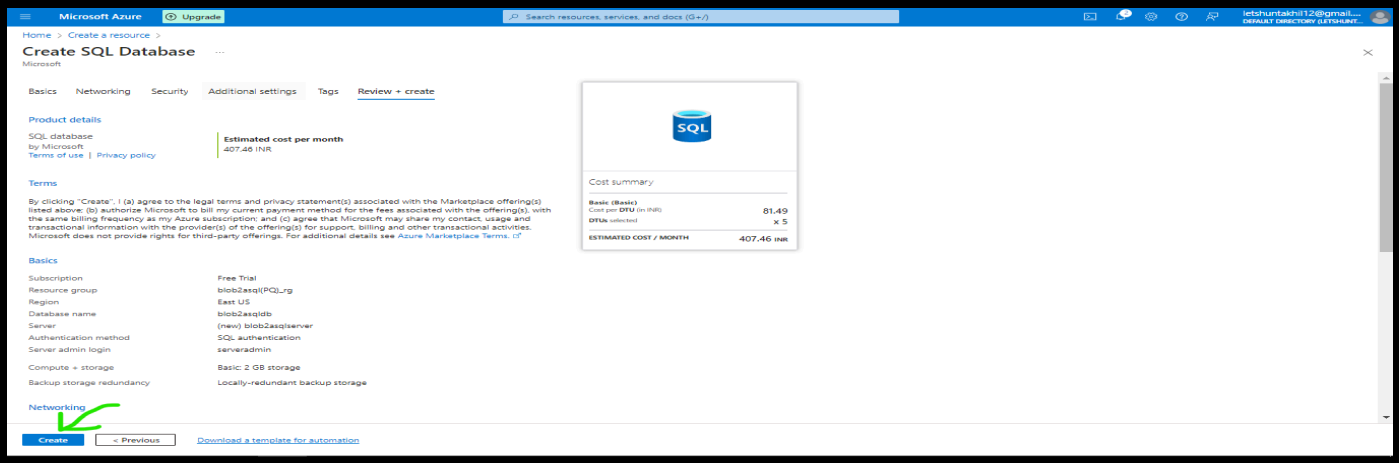
- Additional Settings form : Keep it default —> click next (Tags) .



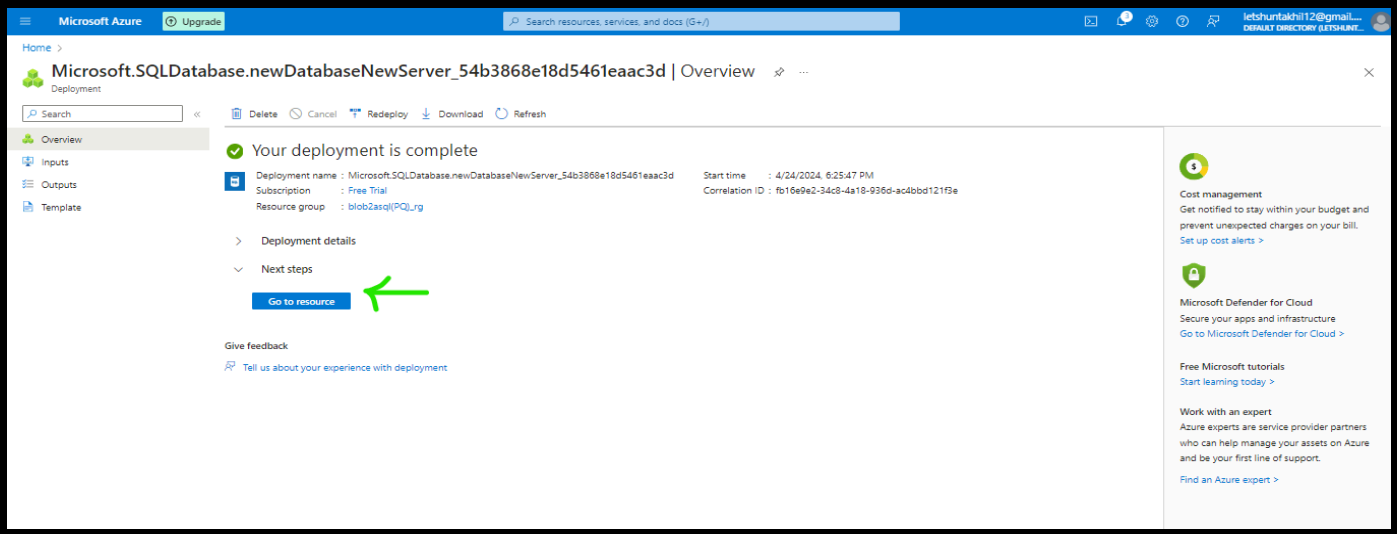
- Tags form : Keep it default —> click next (Review + Create) .



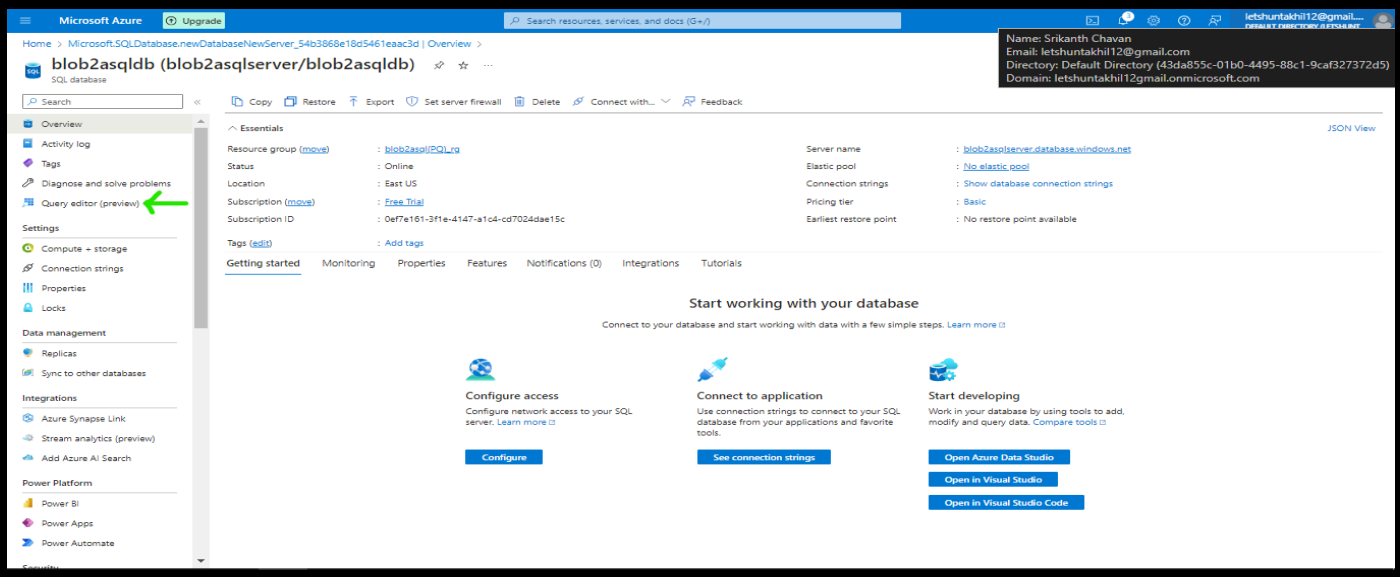
- Review the form and click ' create '.



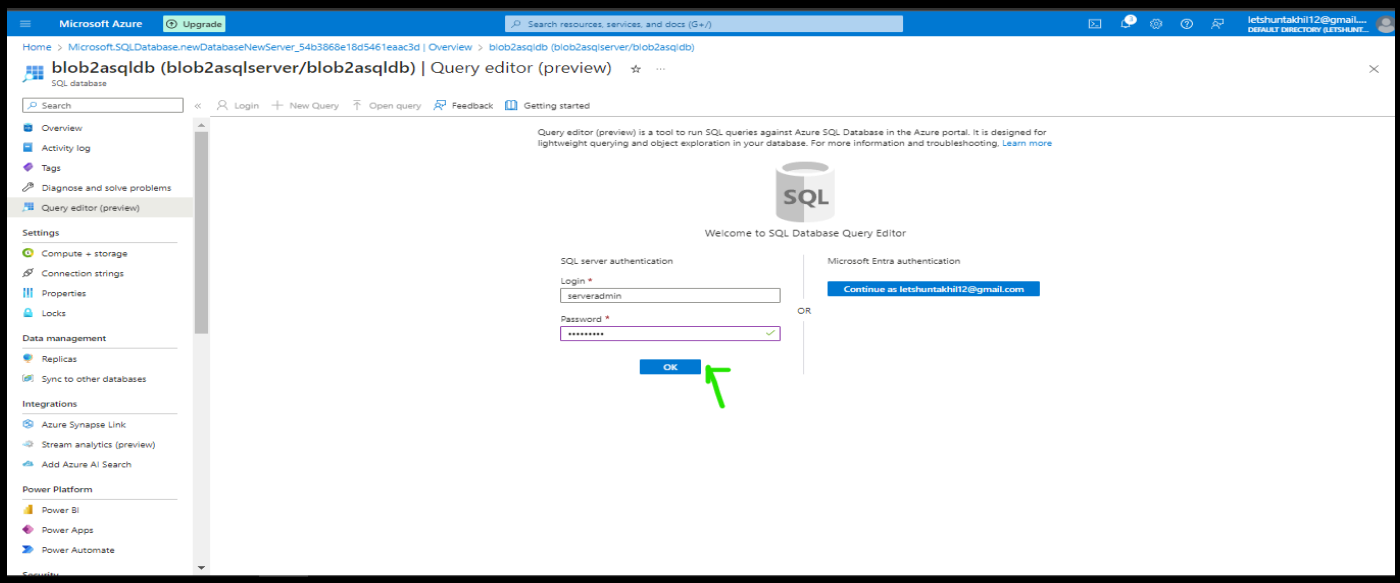
- After creating, wait for sometime to deploy and then click ' Go To Resource '.



- Click the Query editor in the menu to look for the tables and query the table .



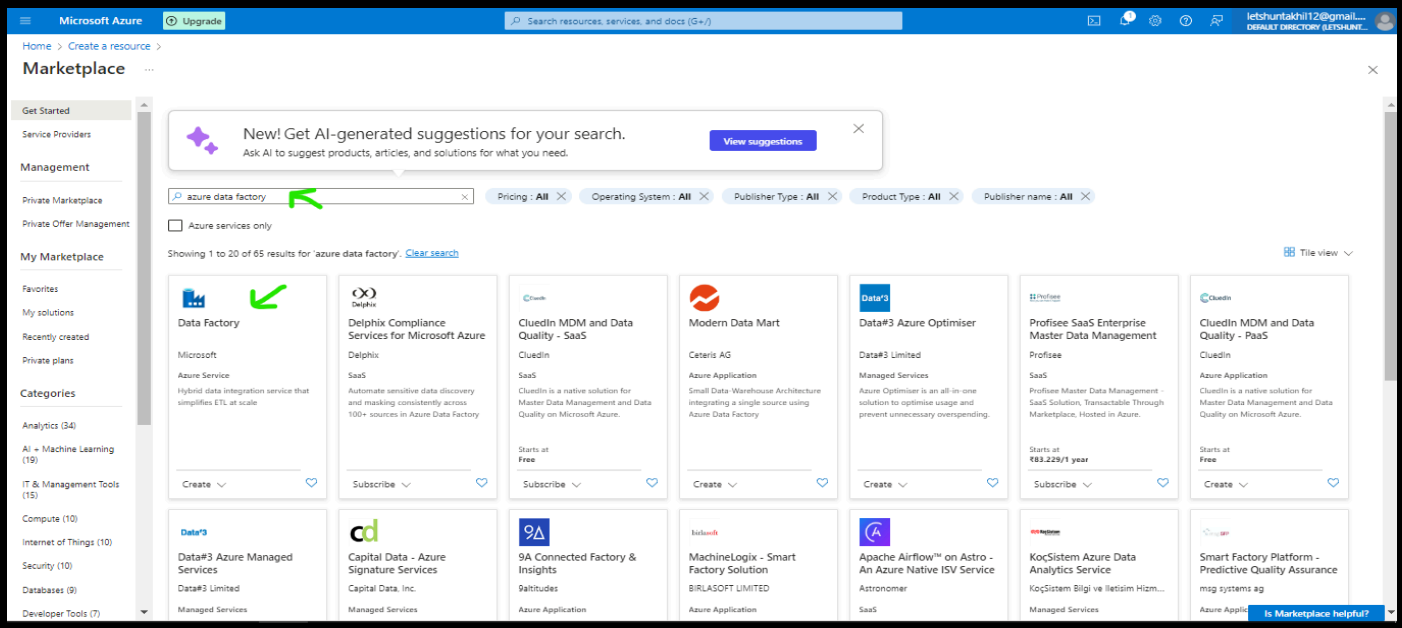
- Login to the database by providing credentials (details provided in the form) .



- Successfully we have created the azure sql database.

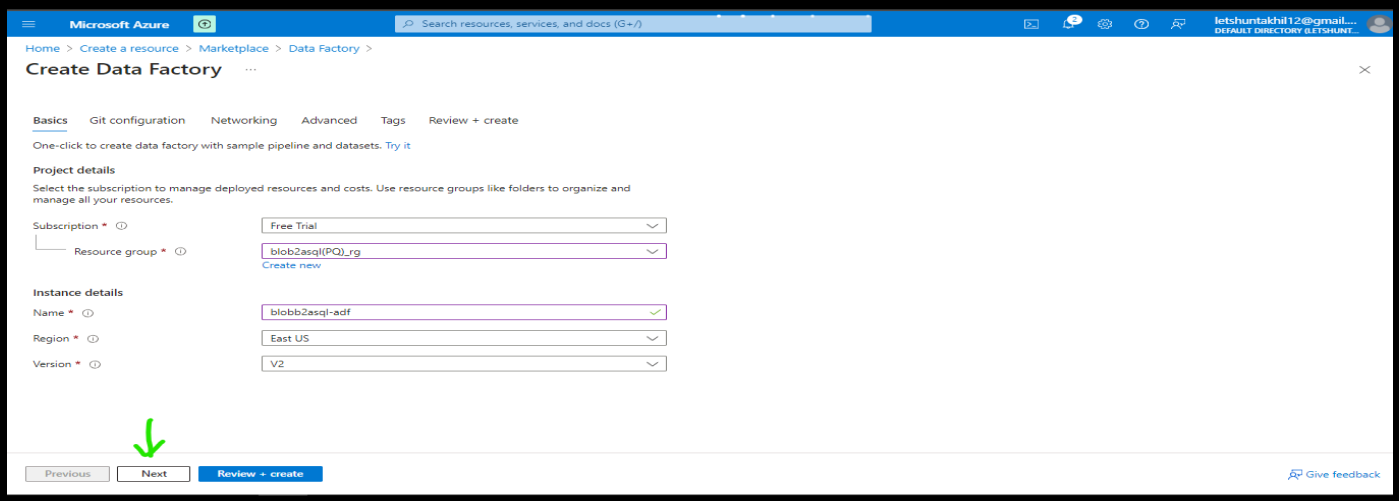
Azure Data Factory Creation :

- Click on ' Go To Resource ' by opening the homepage in the new tab.
- In the menu list, search for Azure data factory and click enter.
- Look for Data factory in the azure marketplace and click on it.

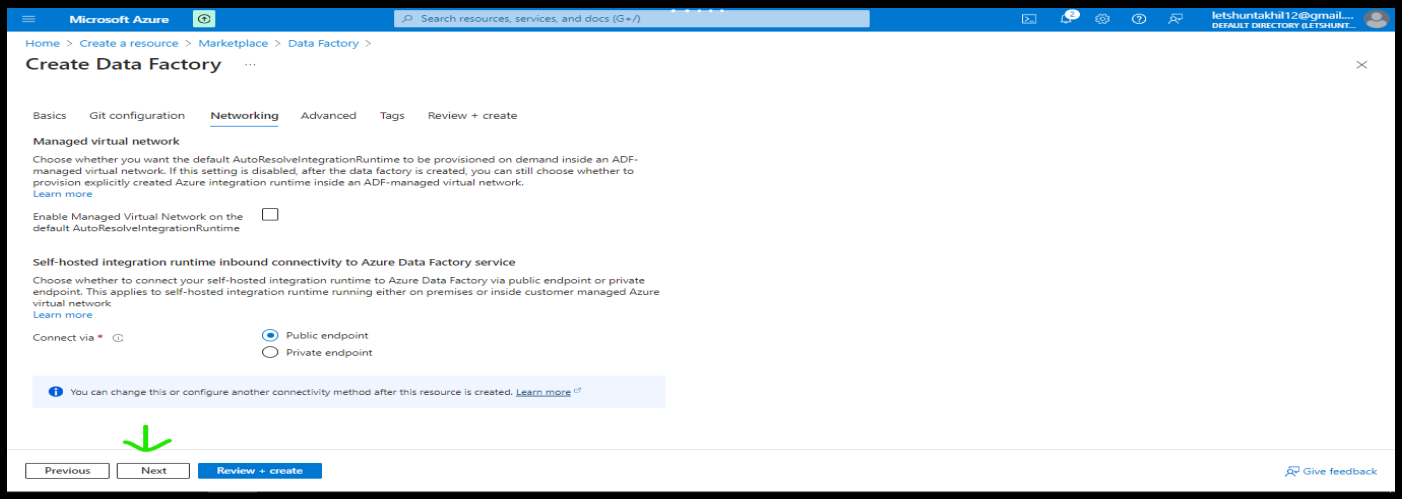


- Fill the form to create an azure data factory.
- In the form we have 6 tabs Basics, Git configuration, Networking, Advanced, Tags and Review + Create.
- Basics form :
  - Project details :
    - Subscription : Select from the list .
    - Resource group : Select from the list .
  - Instance details :
    - Name: Select from the list.
    - Region : East US.
    - Version: V2.

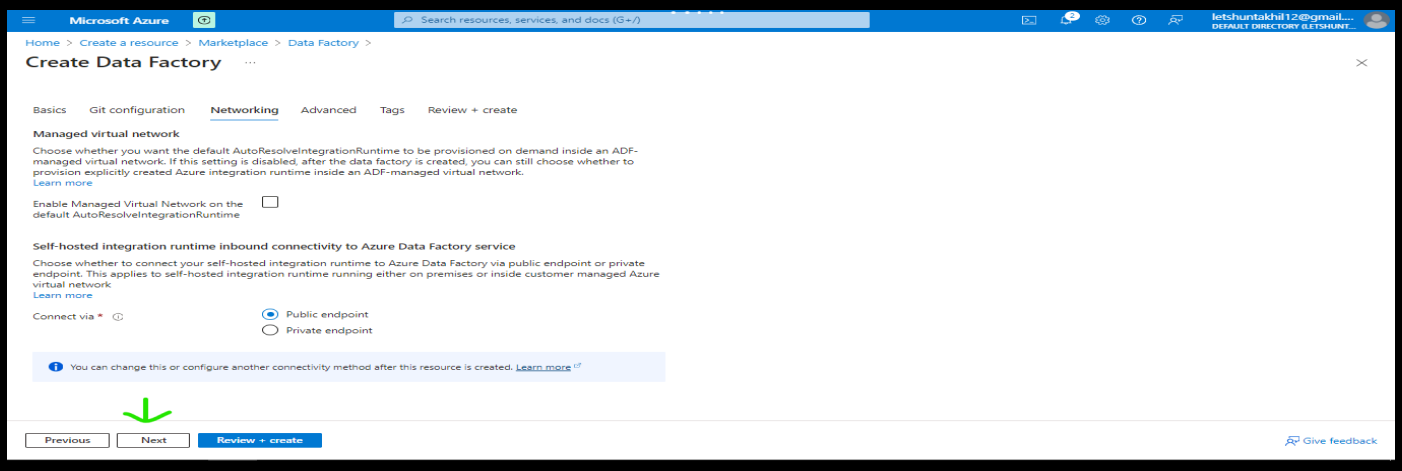
- Click next (Git Configuration) .



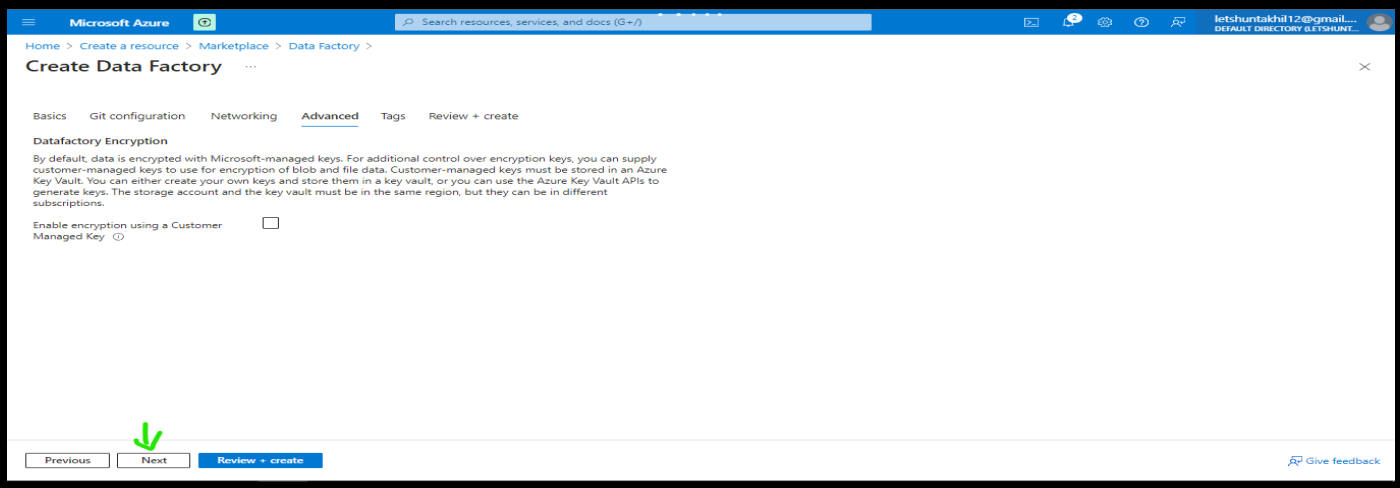
- Click next (Networking) .



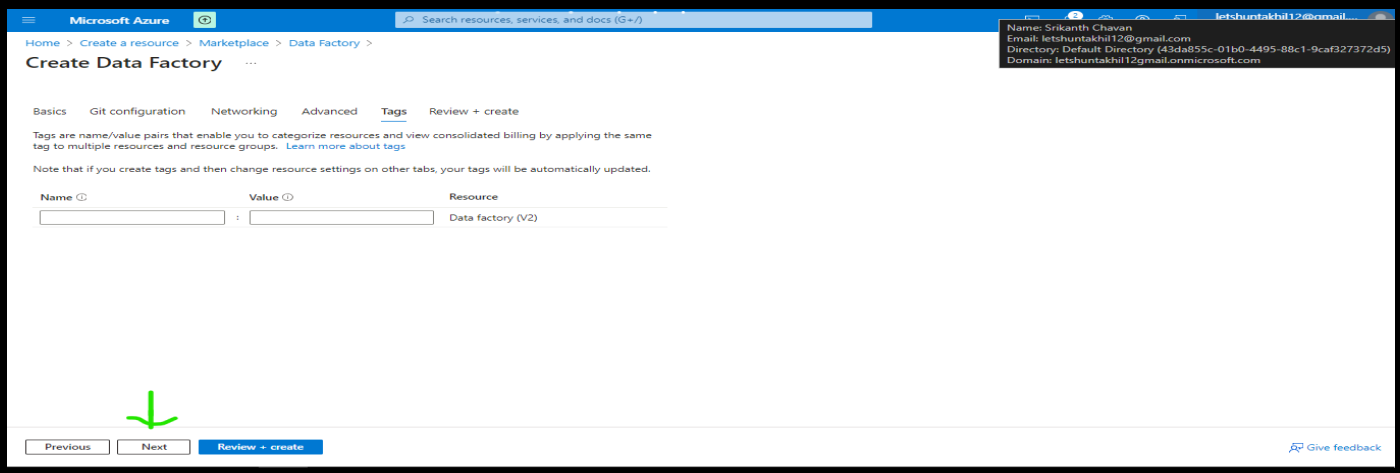
- Click next (Advanced) .



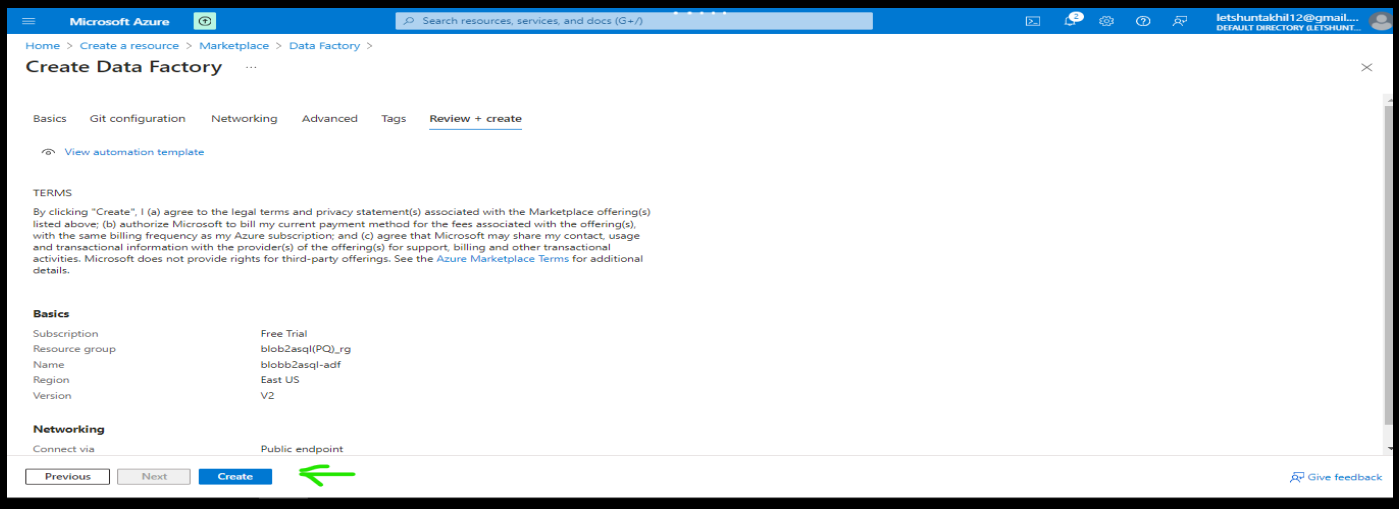
- Click next (Tags) .



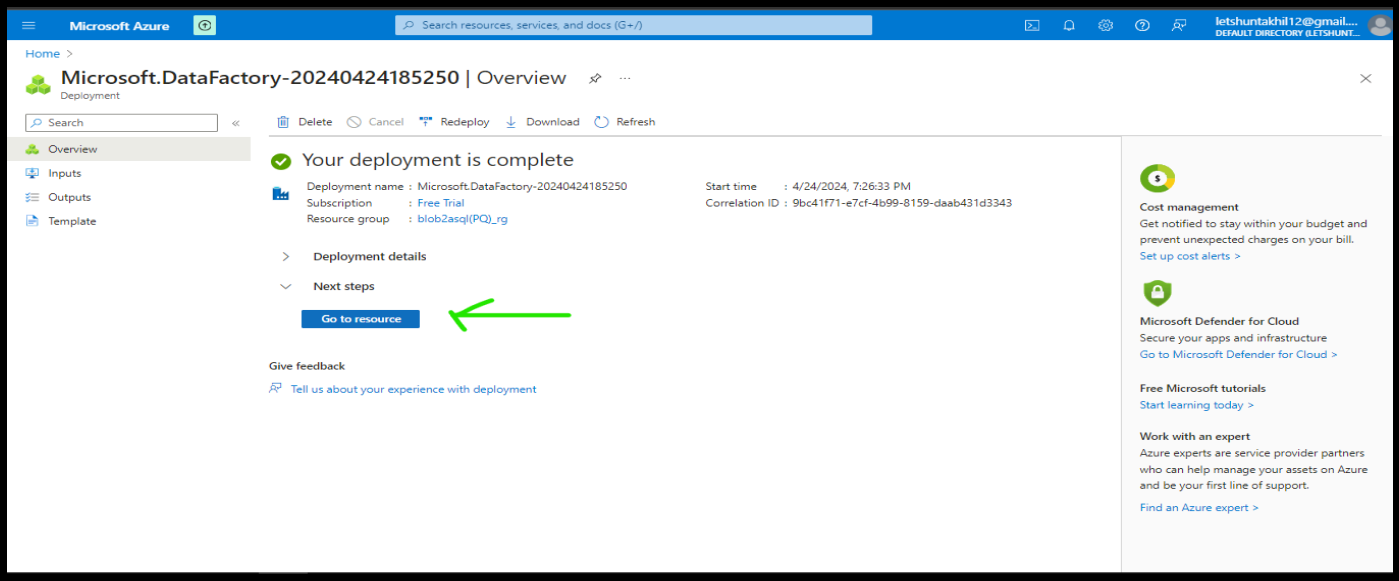
- Click next (Review + Create) .



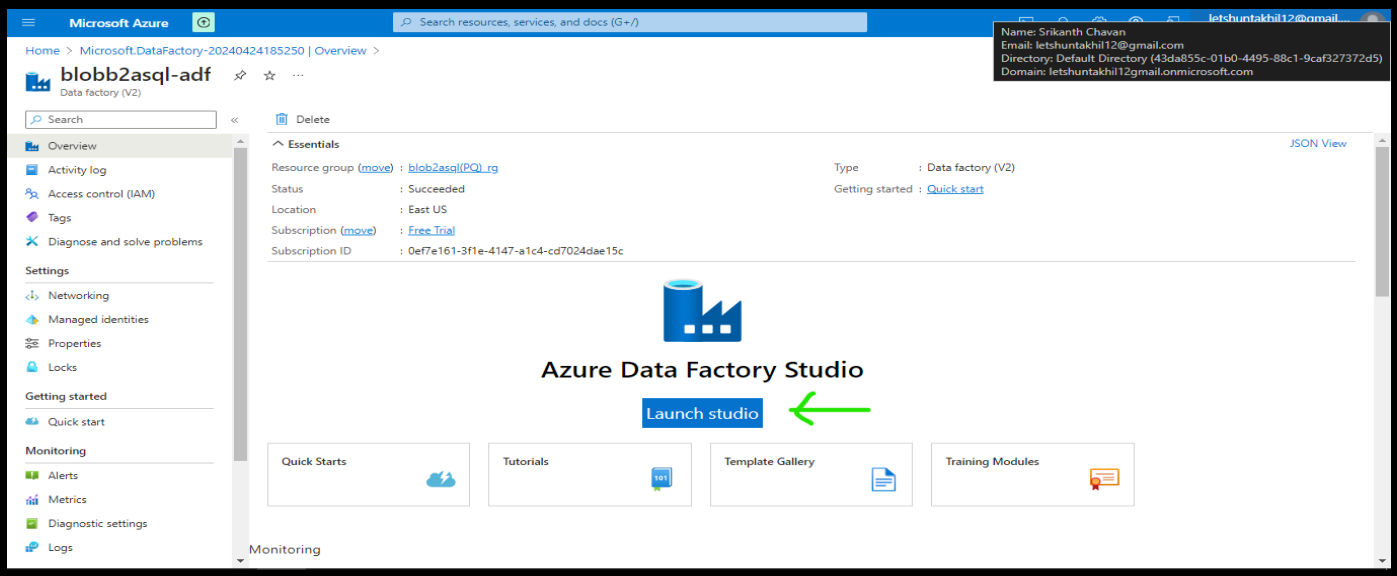
- Review the form and click create.



- After successful creation and deployment of azure data factory, click Go to Resource.

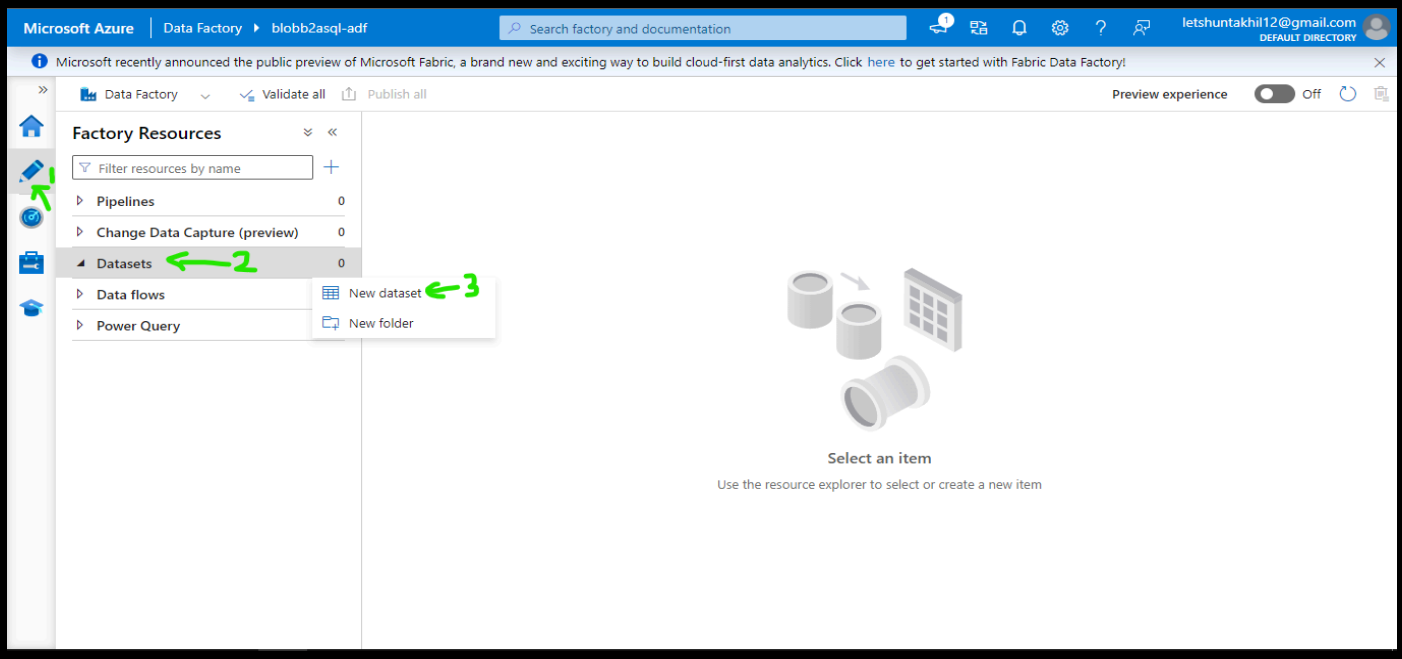


- Let's transform the bike data by clicking the launch studio.

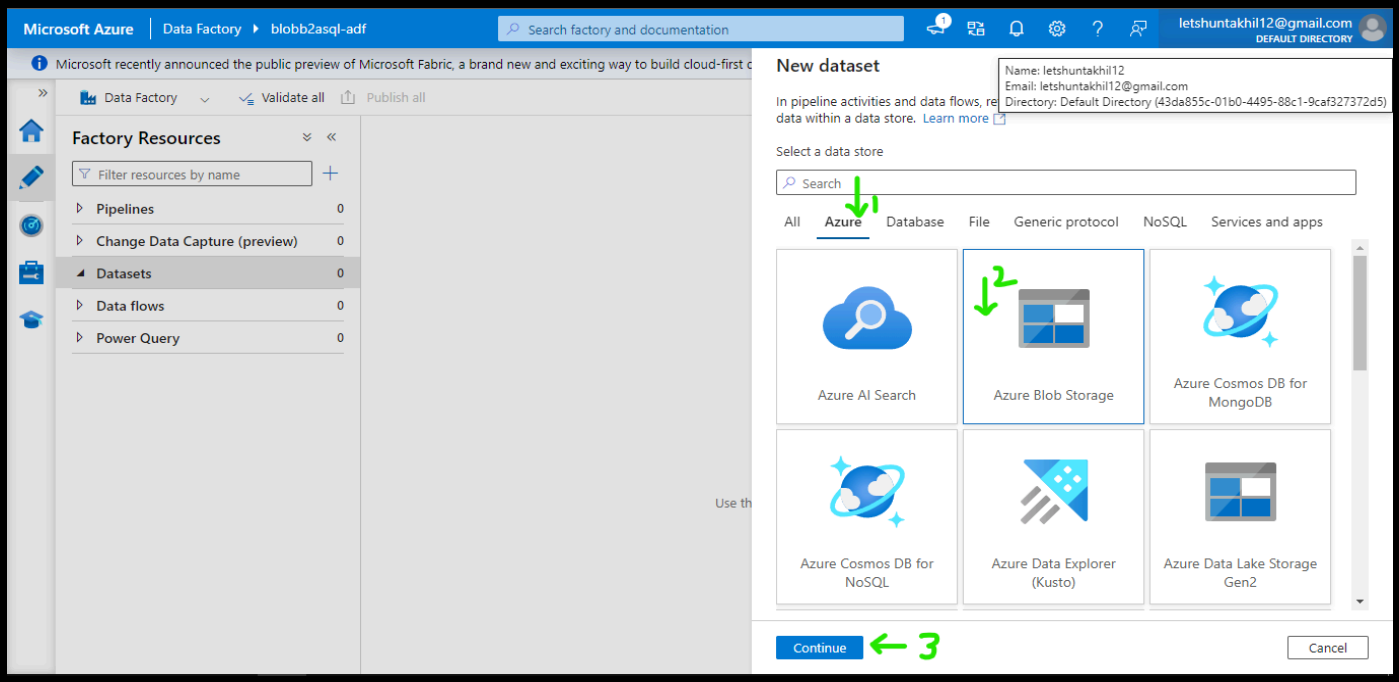


Transformation of data using Data Flow:

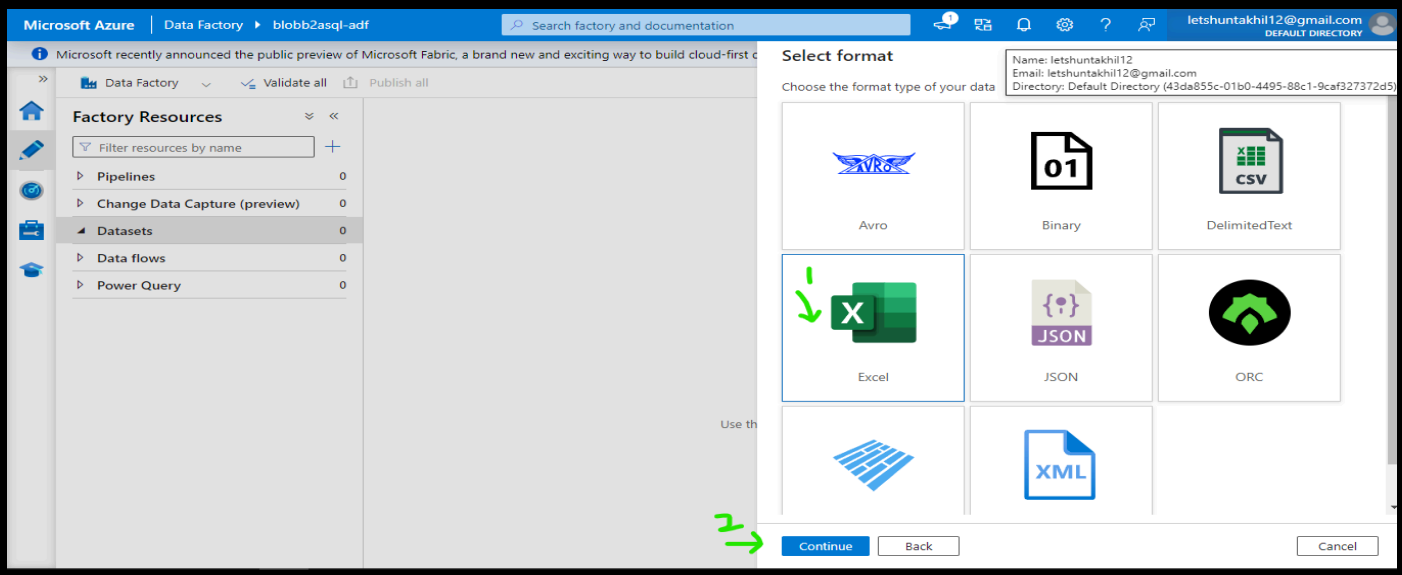
- Azure data factory studio will open in a new tab.
- Create a dataset (path) for the azure blob storage (source data) by clicking the pencil icon (✏️) → datasets → new datasets.



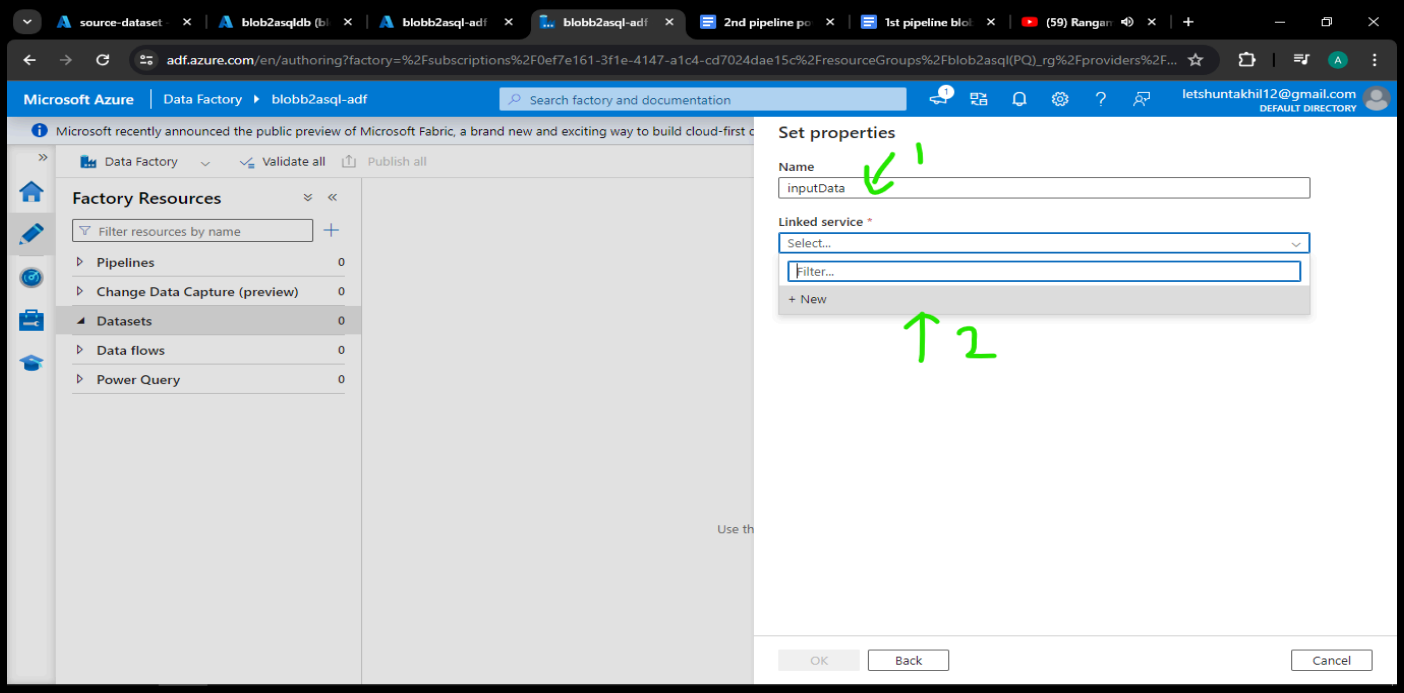
- Select the azure blob storage in the azure tab and then click continue.



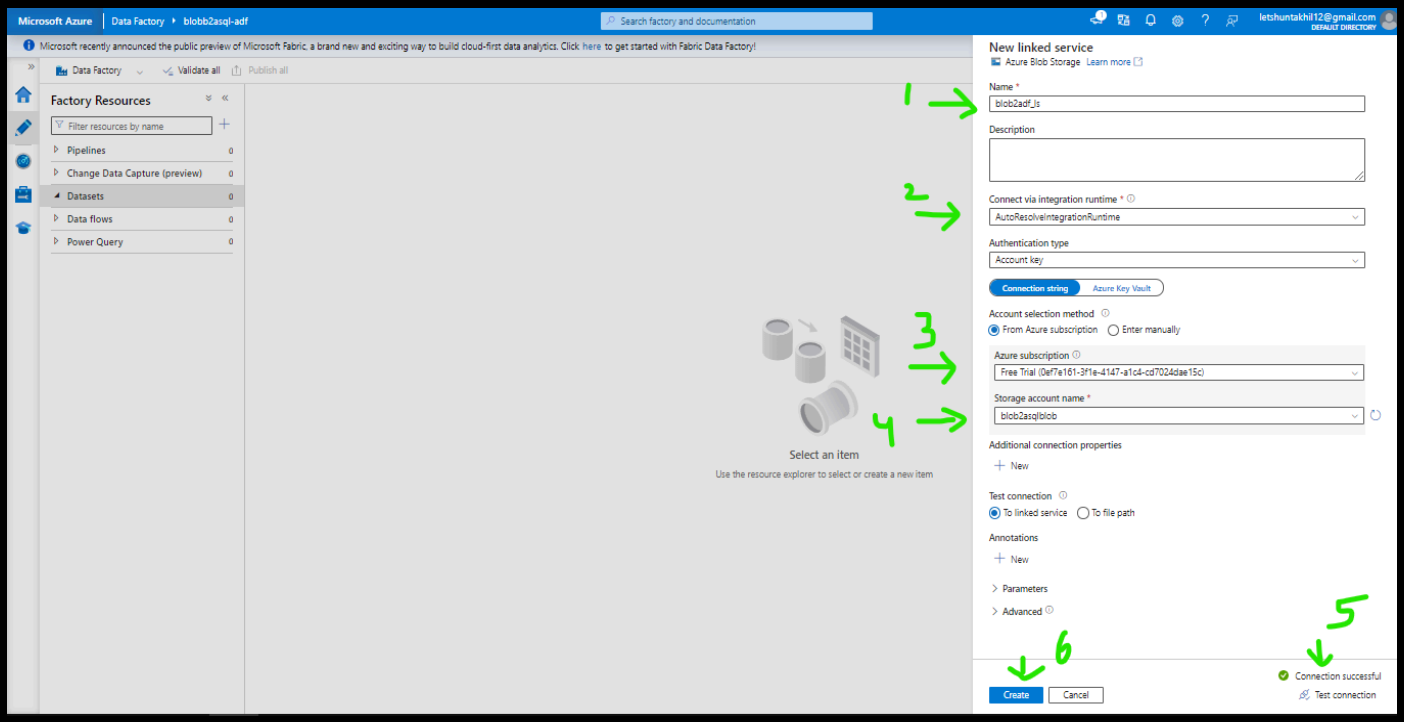
- Select the format (Excel) of the data to transform which is in the source dataset (azure blob storage) and then click continue.



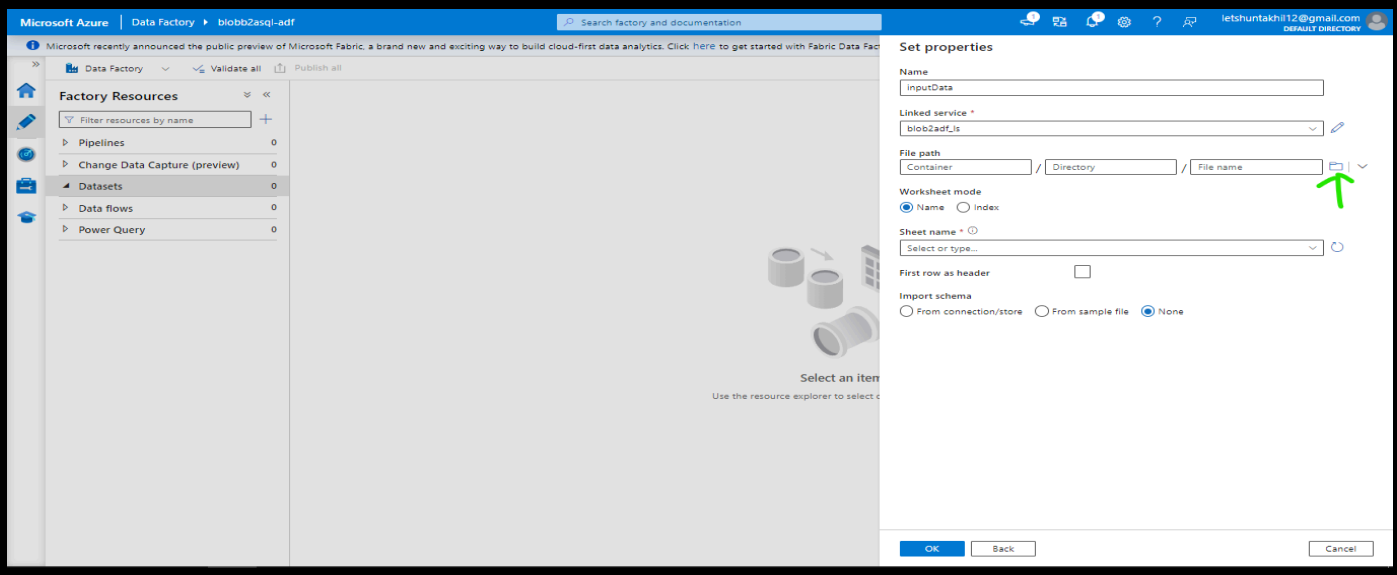
- To make a connection between blob storage and azure data factory we need a linked service to connect or make a bridge between them. So that communication can happen.
- The next step is to choose or create a linked service .

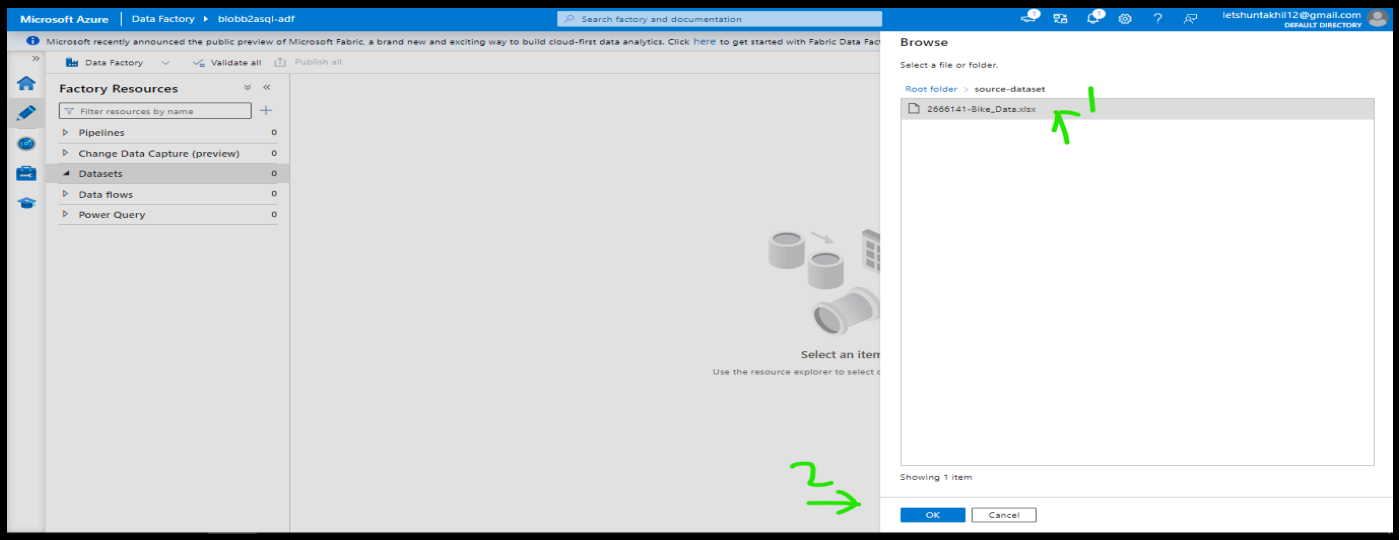


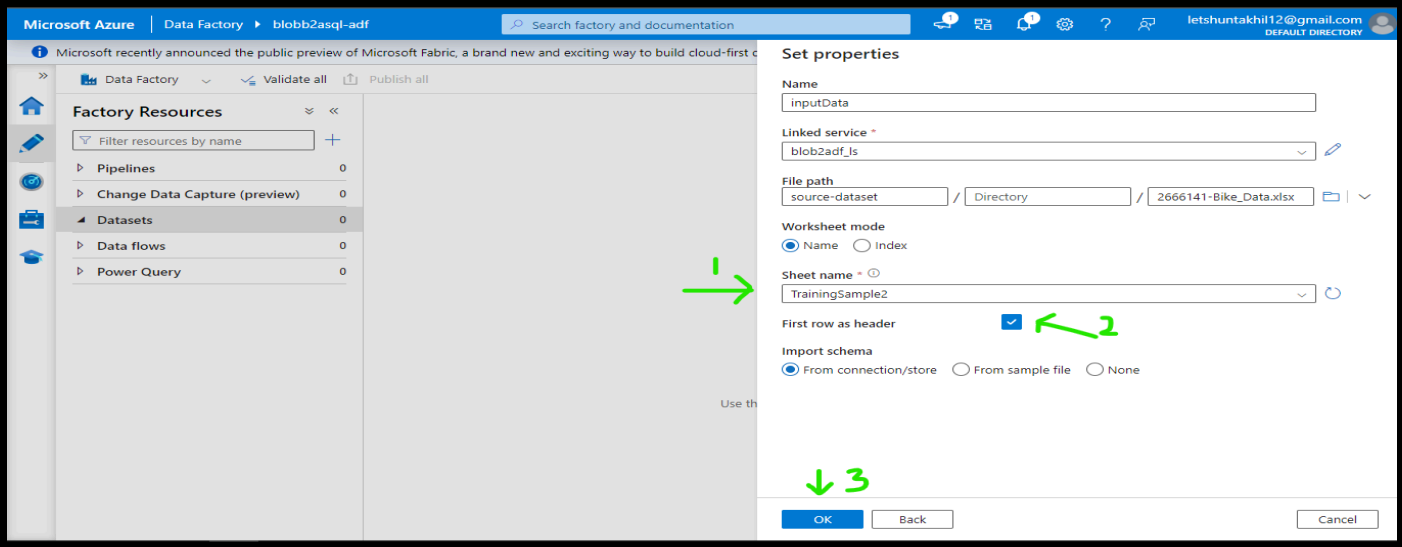
- Fill the form to create a linked service.
- In the form we have properties to fill in name, description, connect via integration runtime, authentication type, subscription, storage account name and some other properties.
- Connect via integration runtime → integration runtime is responsible for creating a connection between different environments.
  - Self Integration runtime : It is for the on premises environment.
  - AutoResolveIntegrationRuntime : It is for within the environment (azure to azure) .
- Azure subscription : select the subscriptions where the file is located.
- Storage account name : select the storage account where the file is located.
- Test connection → successful → Create.



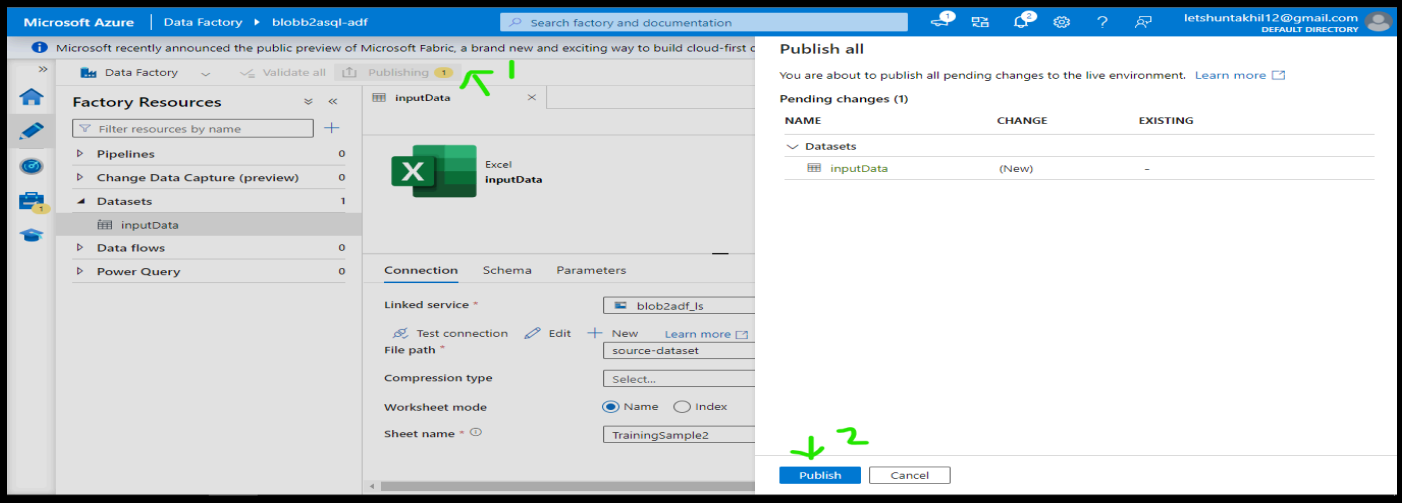
- Choose the file path by clicking the file icon (📁)→ select the first row as header → select the sheet name → ok.



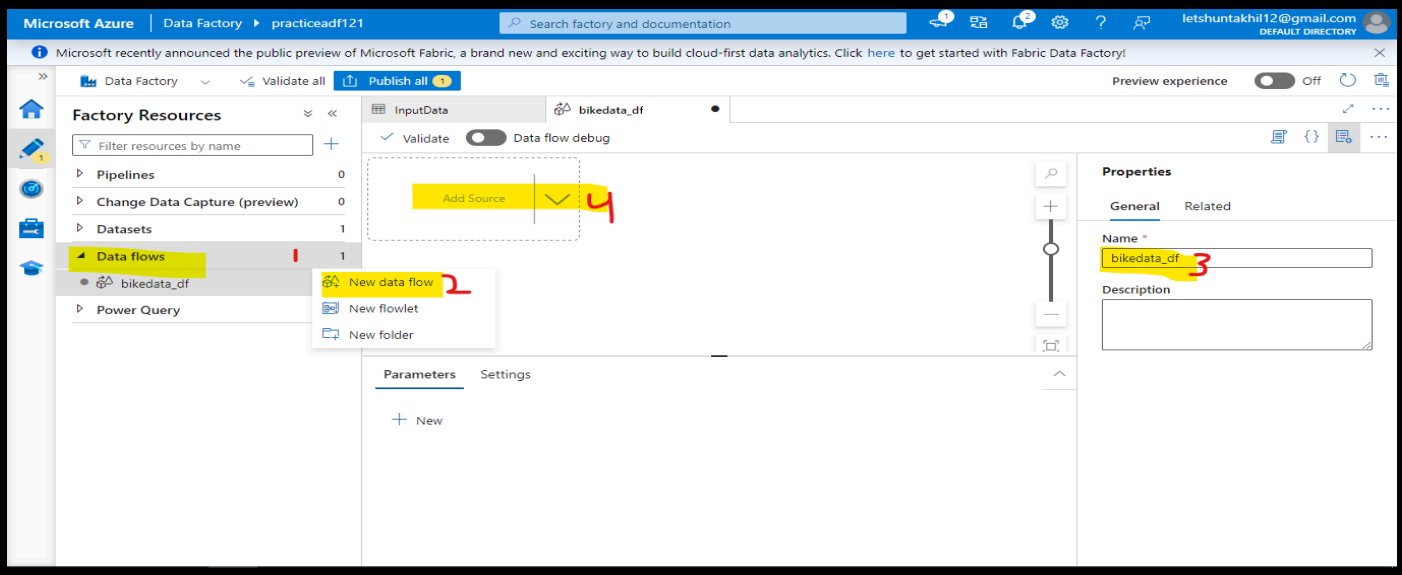




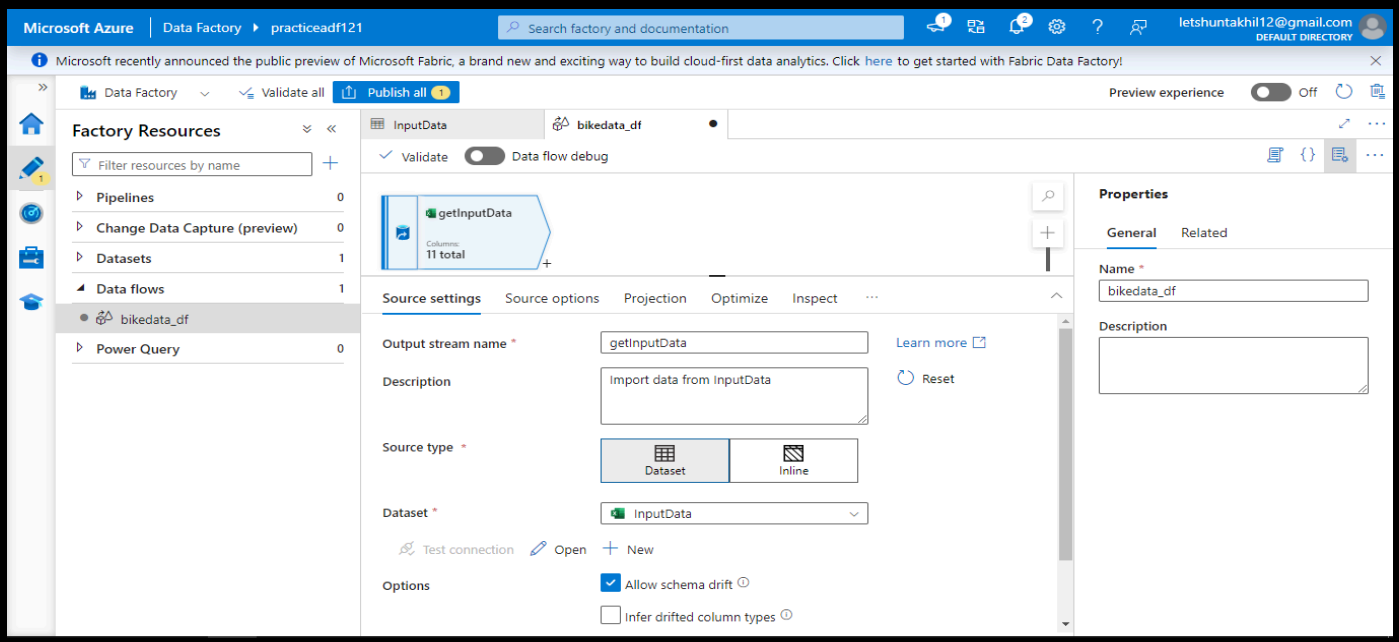
- Save the dataset by clicking publish all → publish.



- Successfully brought bike data to the data factory to transform.
- Create a Dataflow → new Data Flow (bikeData_df) which is in the pencil icon (✏️) menu.
- Click add source in the workspace.

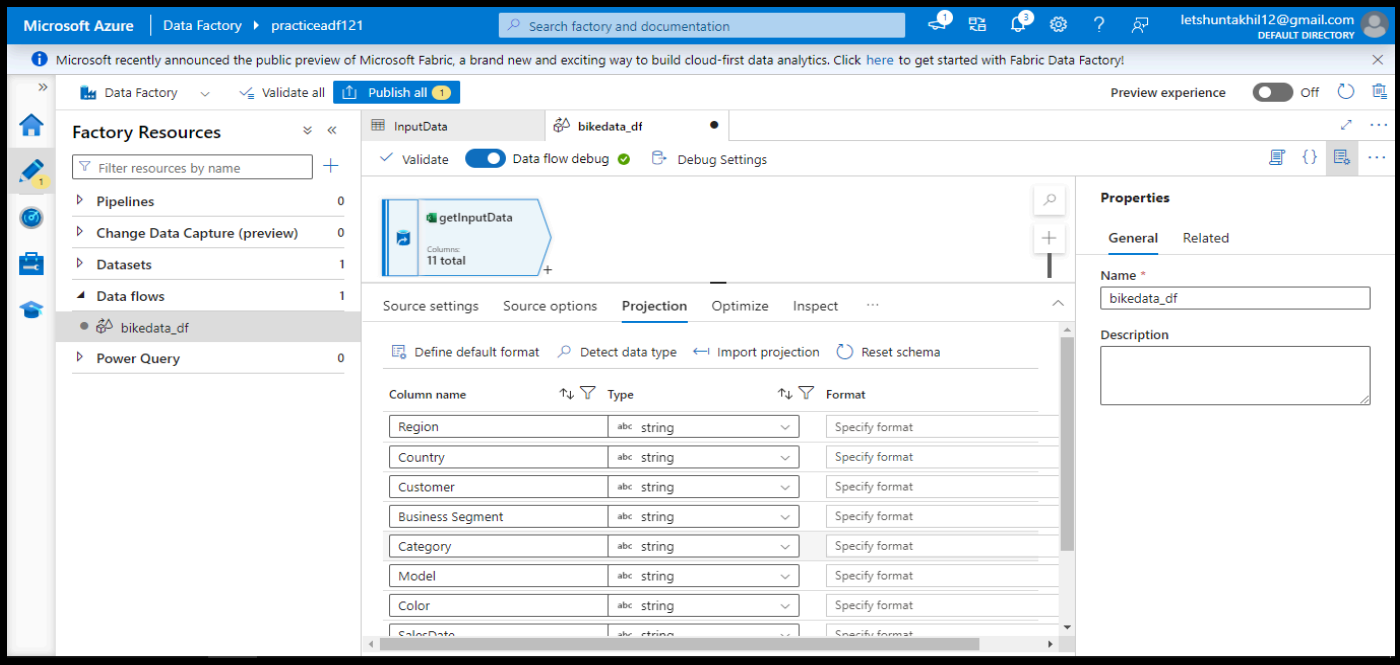


- You will get a form to provide the details of source data.
- bikeData_df → Settings → Dataset → select the file (inputData) .

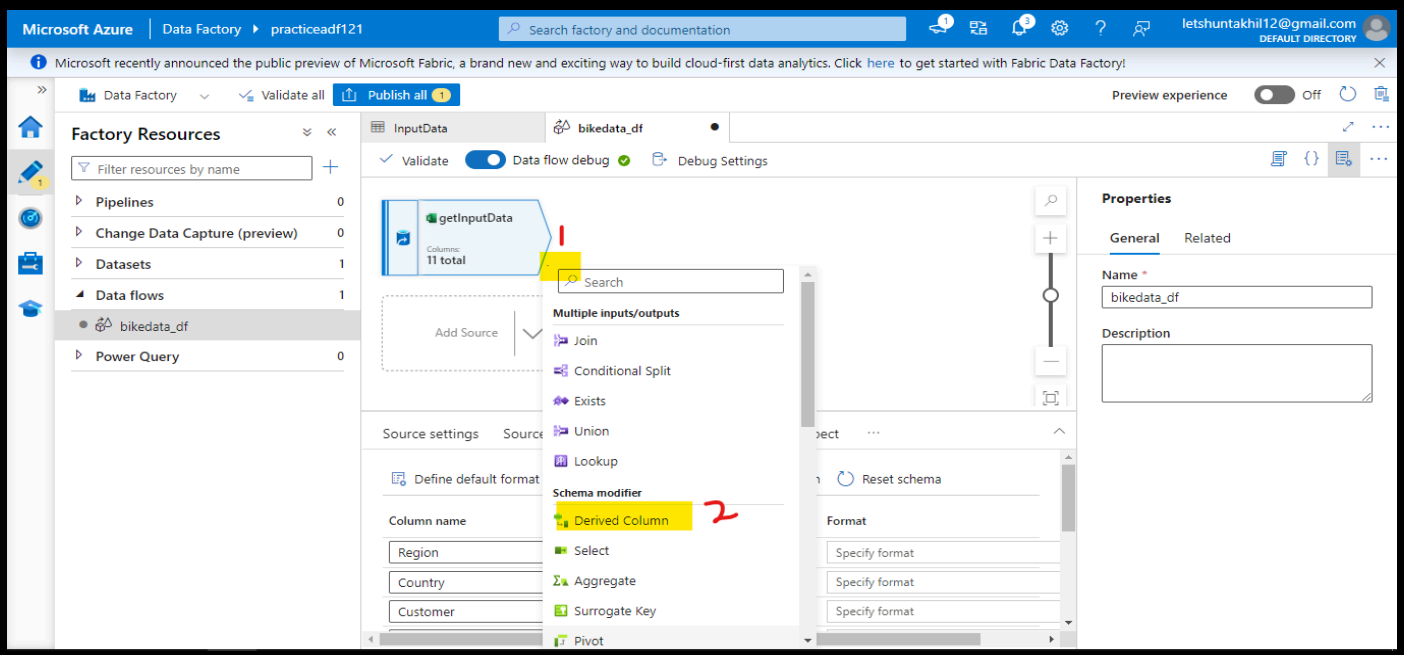


- Click on the data preview .

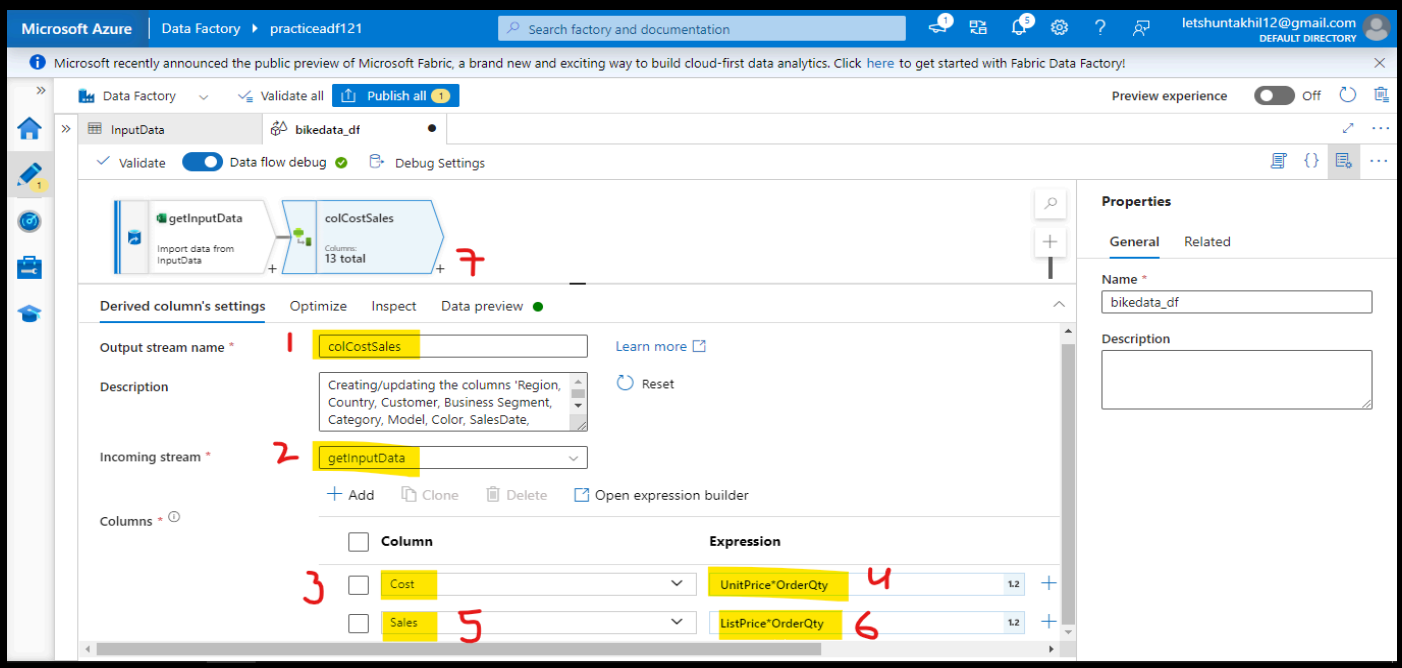
- It will ask you to turn on the Data flow debug.
- Click ok.
- Projection —> import projection



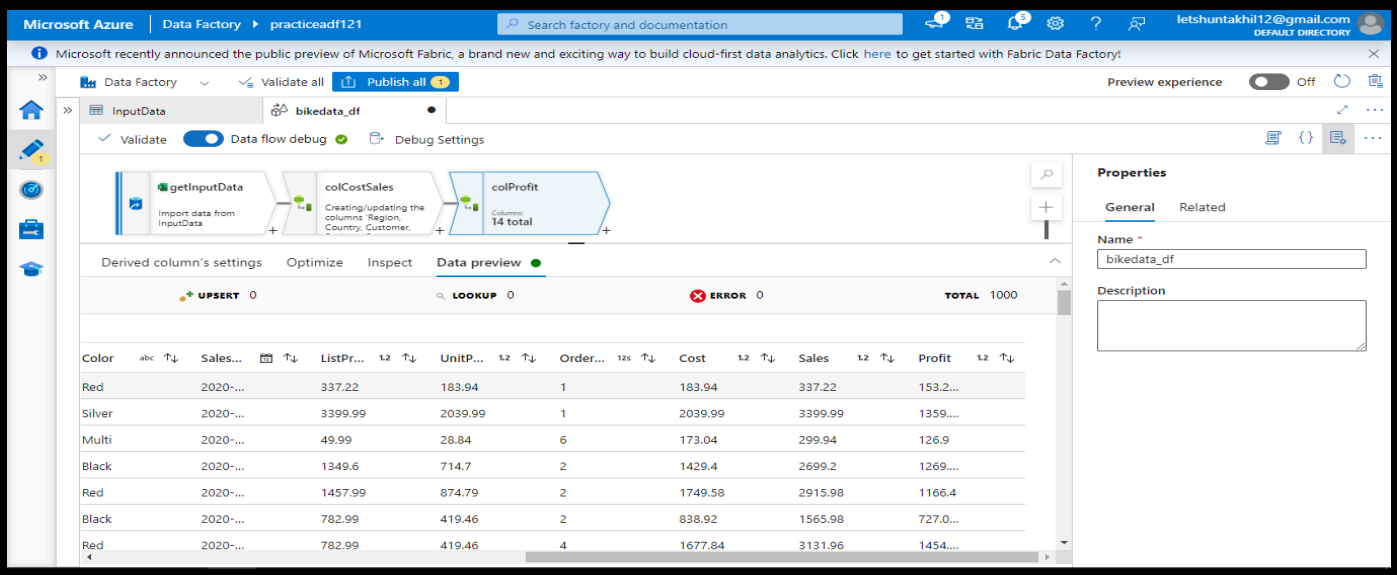
- Data preview by clicking refresh



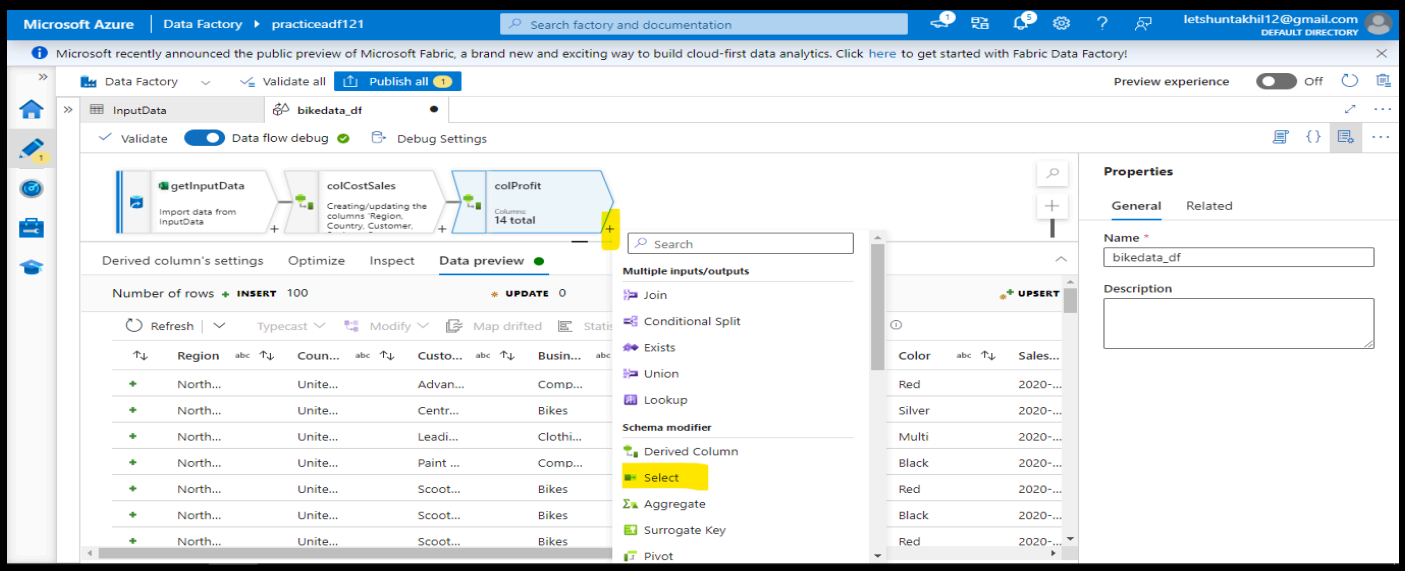
- Click on the + button to create a Derived column (new column)
- Create a new column cost and sales by multiplying UnitPrice*OrdreQty and ListPrice*OrderQty respectively.



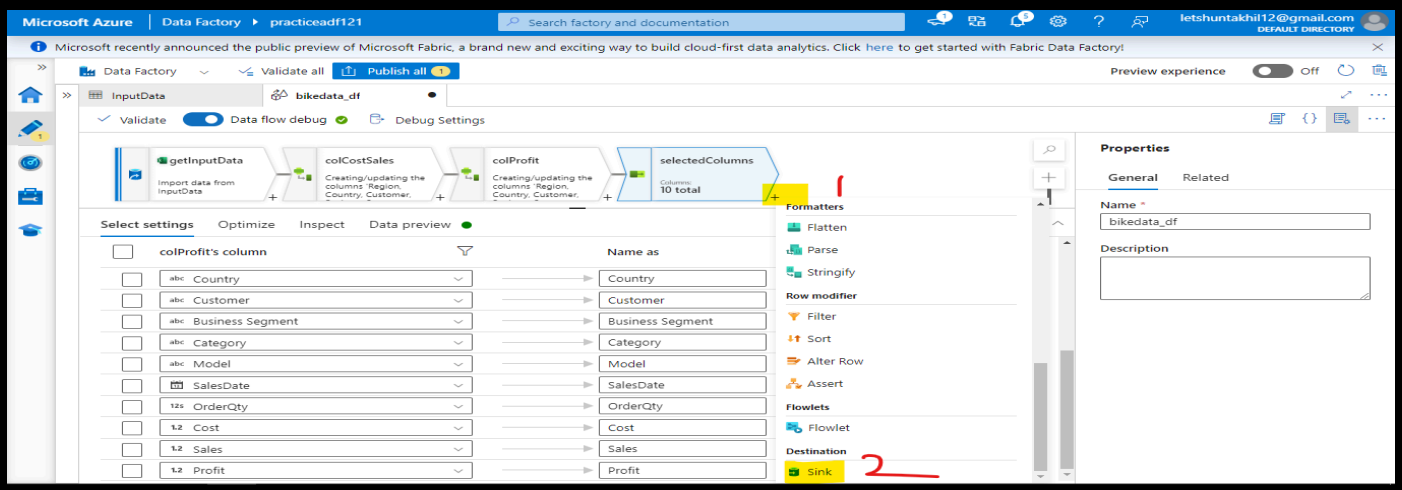
- Create a profit column → click on + after the colCostSales → derived column.

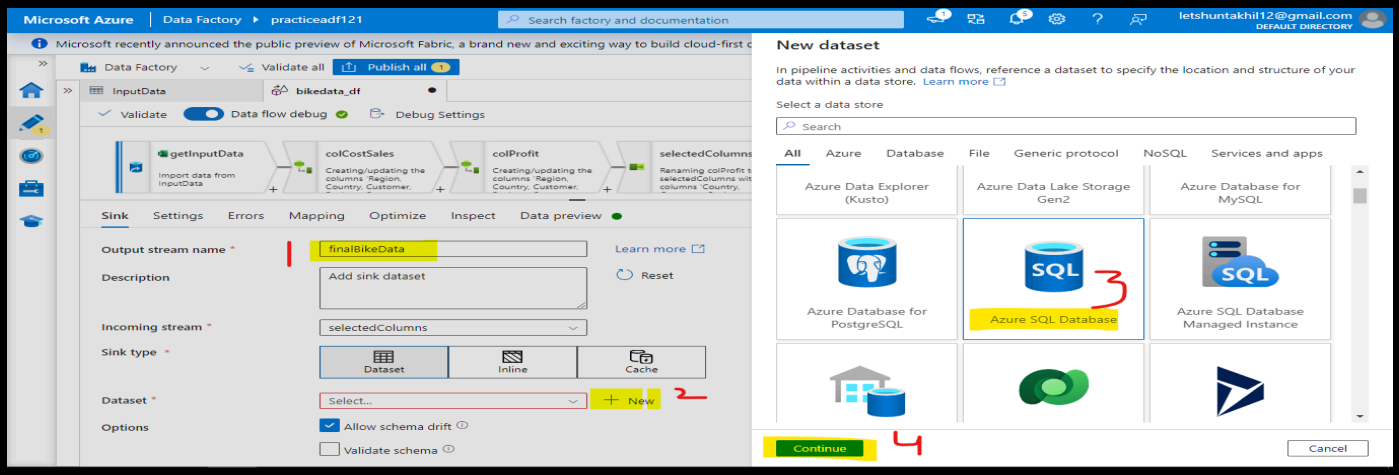


- Drop the columns which are not frequently used like UnitPrice , ListPrice,Region and Color.
- We can do it by clicking the + button and search for the ' select ' option.

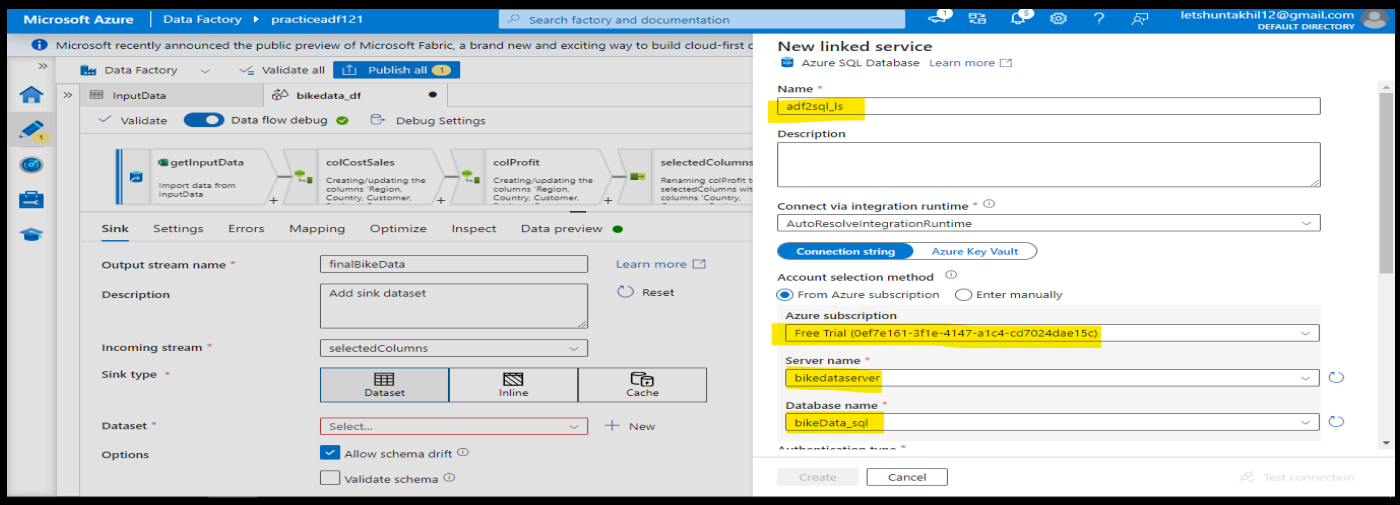


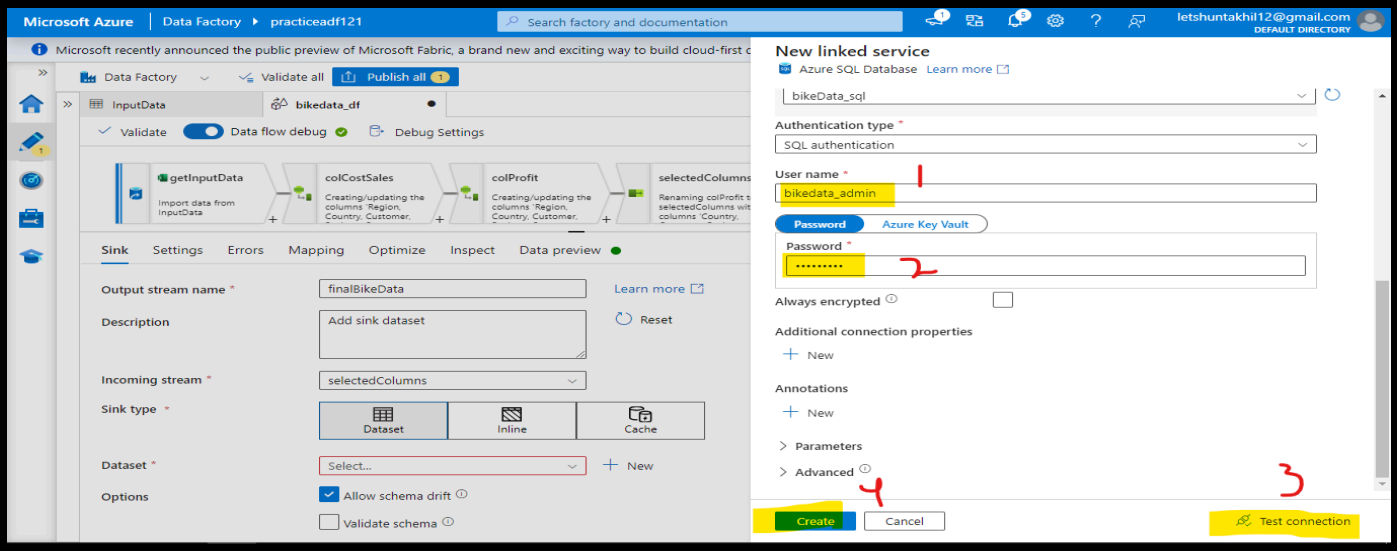
- Drop the selected columns by clicking the delete icon.
- Store the clean data in azure sql database by choosing the sink in dataflow after the selectedColumns.



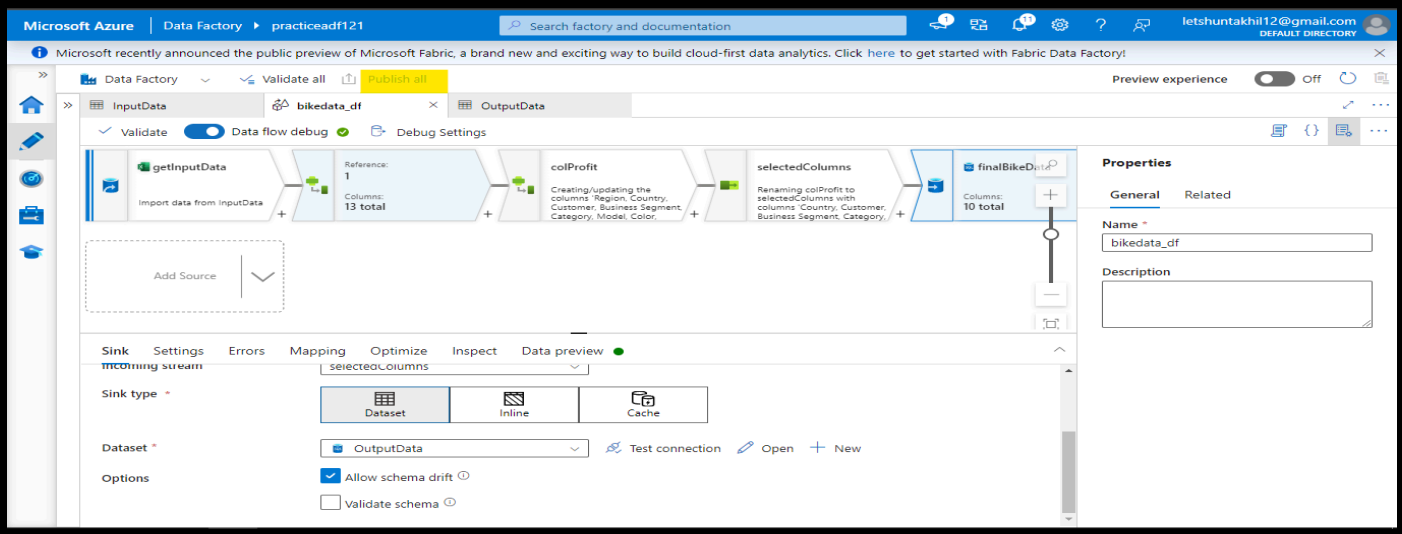


- Click on the new linked service.
- Create a new linked service by entering the details.Linked service is to connect or make a bridge between adf and azure sql database.

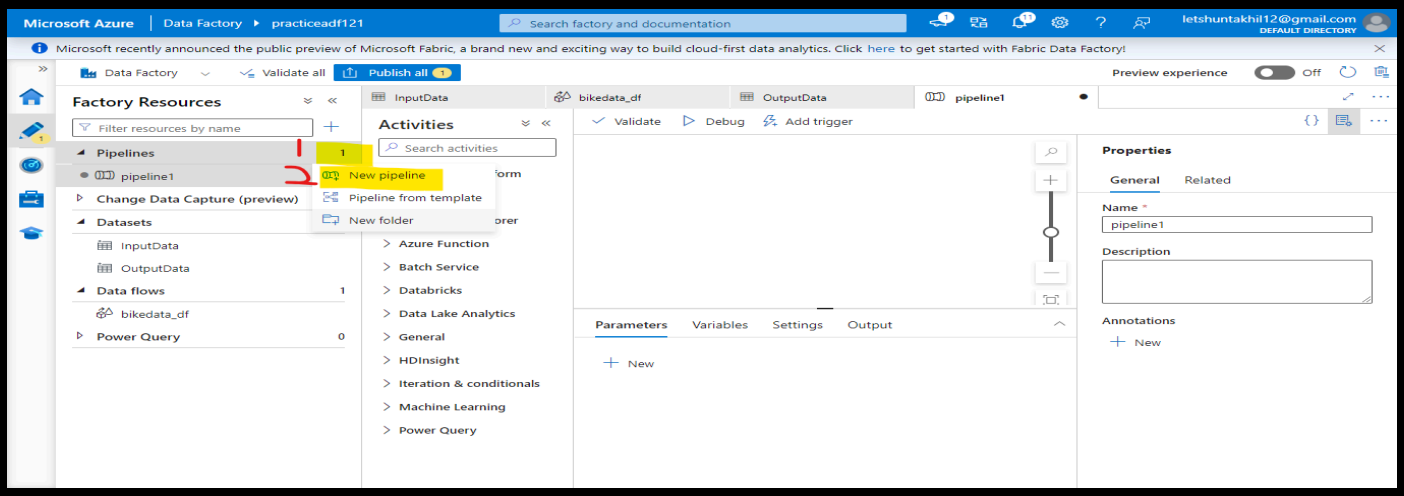




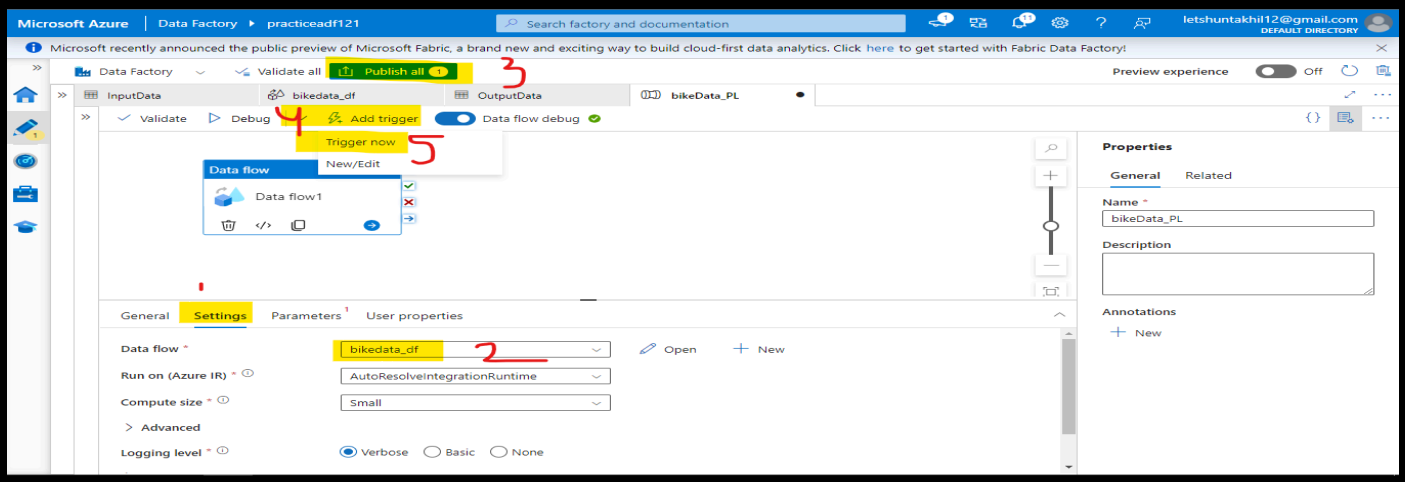
- Save the Target dataset path by publish all → publish



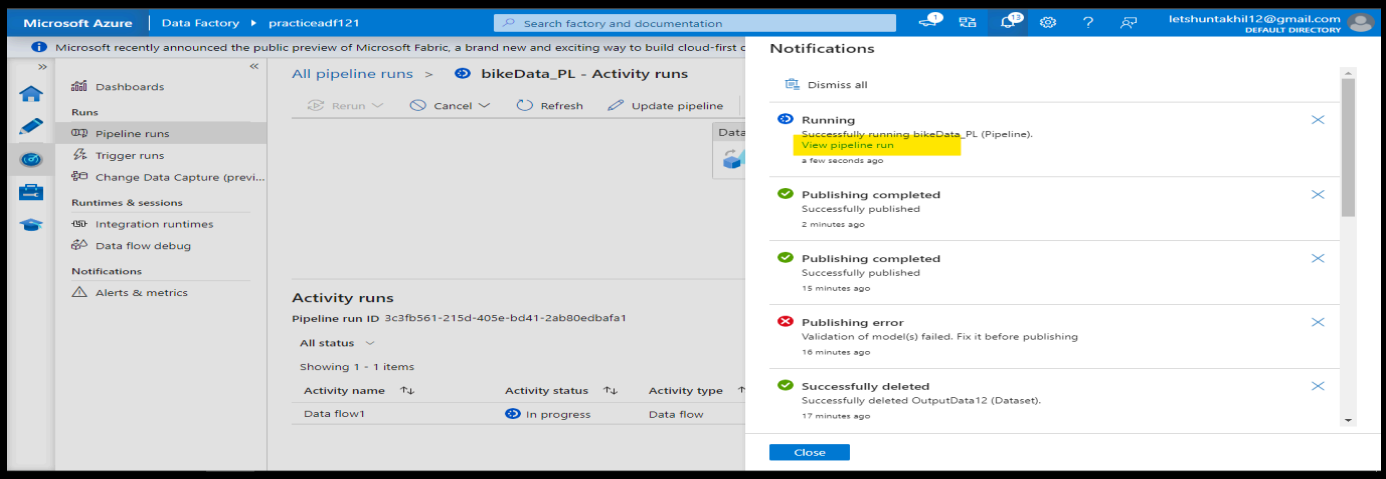
- Save the transformation by publish all → publish. Let's execute the dataflow
- Pencil icon (✏️) → new pipeline



- Drag and drop the data flow activity to the workspace and click publish all → publish.



- Let's run the pipeline —> Add Trigger → Trigger now → ok.
- See the status of the pipeline —> view pipeline run .



- Now after successful pipeline execution , see the table in the azure sql database and Query the Data.