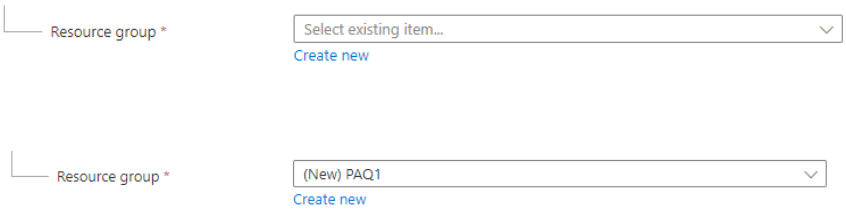
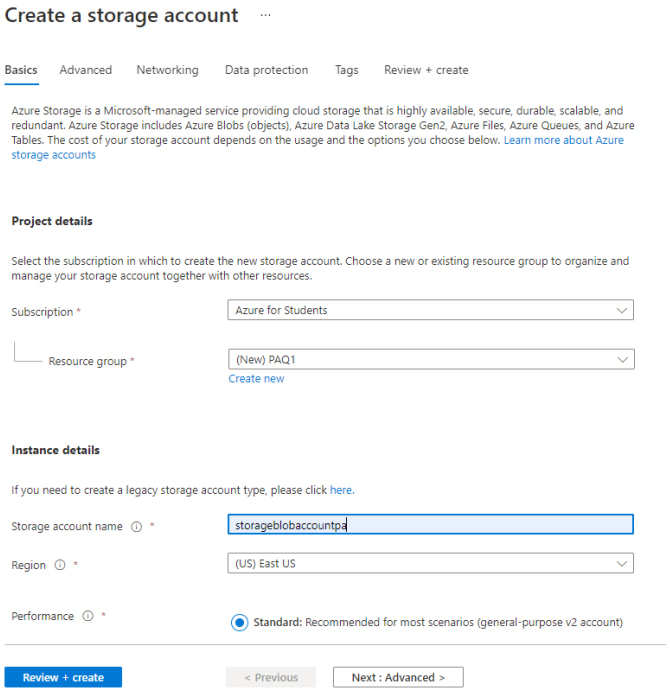
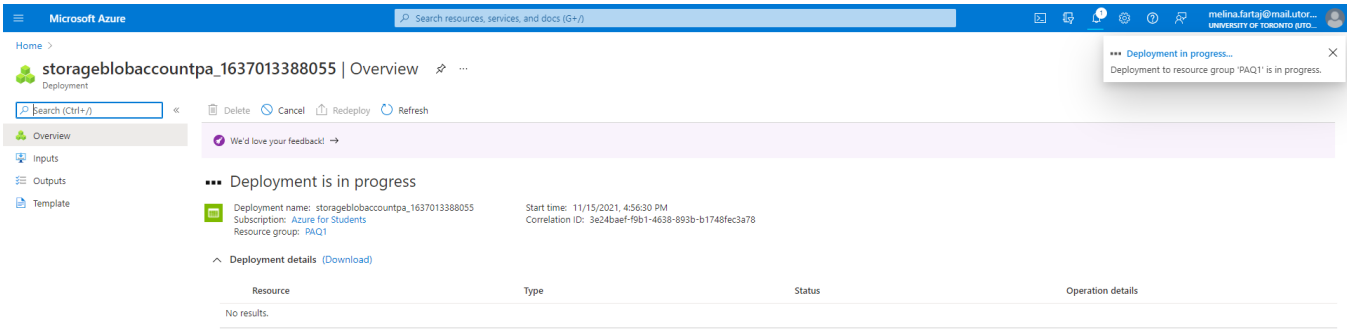
Part A:

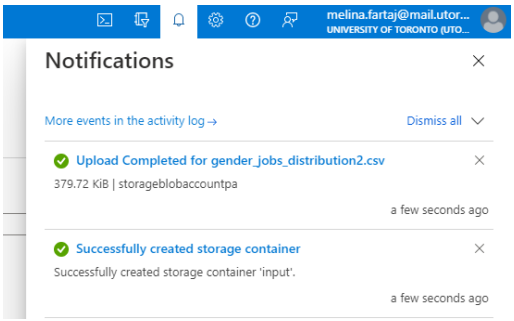
1. Resources
    a. Creating a resource group: click create any resource and then create new under resource group.



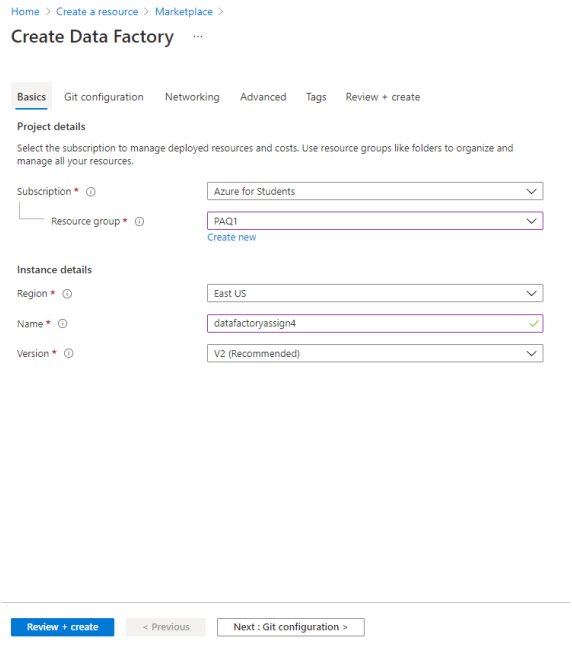
    b. Creating Blob Storage Account + container:
        i. Click create a resource
        ii. Create name
        iii. Click Review + Create
        iv. Wait for Validation to Pass
        v. Click Create
        vi. Deploy Resource
        vii. Click on resource
        viii. Click Blob Service
        ix. Create Container
        x. Name input
        xi. Upload .csv file to input container

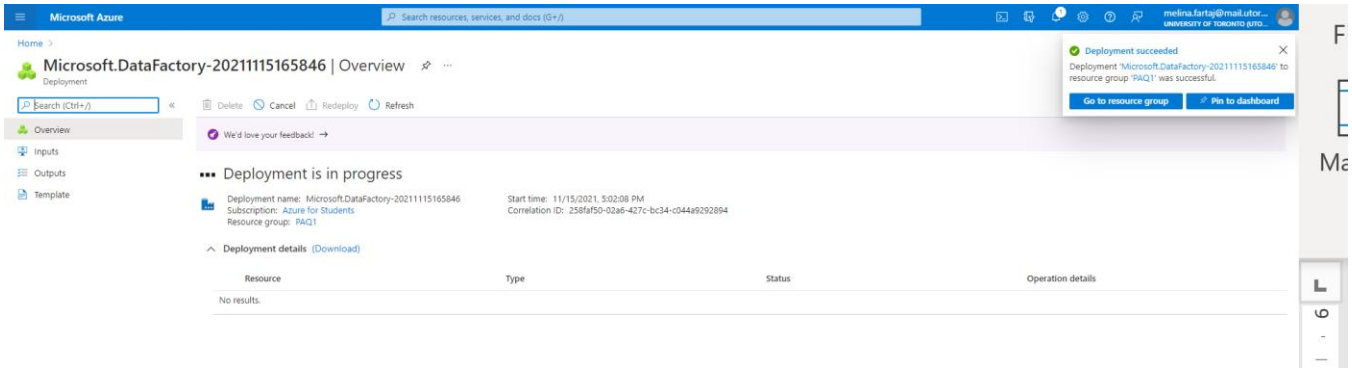




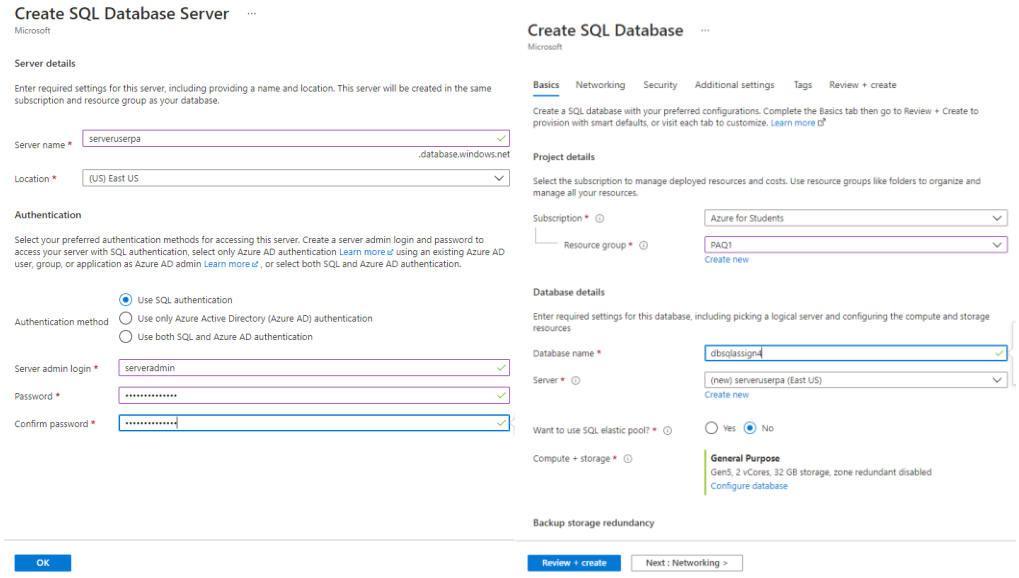


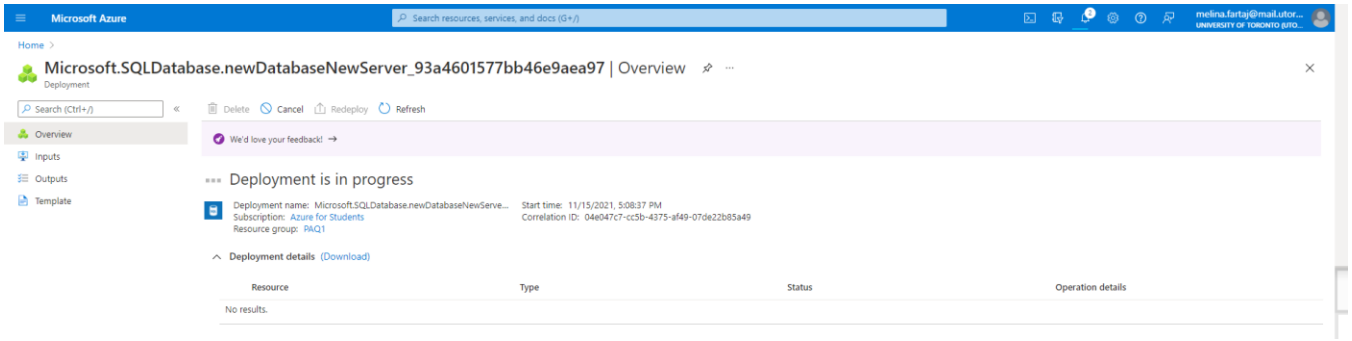
c. Create Data Factory
   i. Click create a resource
   ii. Search for data factory
   iii. Click create
   iv. Select resource group
   v. Create name
   vi. Click configure git later on ‘git configuration’
   vii. Click Review + Create
   viii. Wait for validation to pass
   ix. Click Create
   x. Deploy Resource



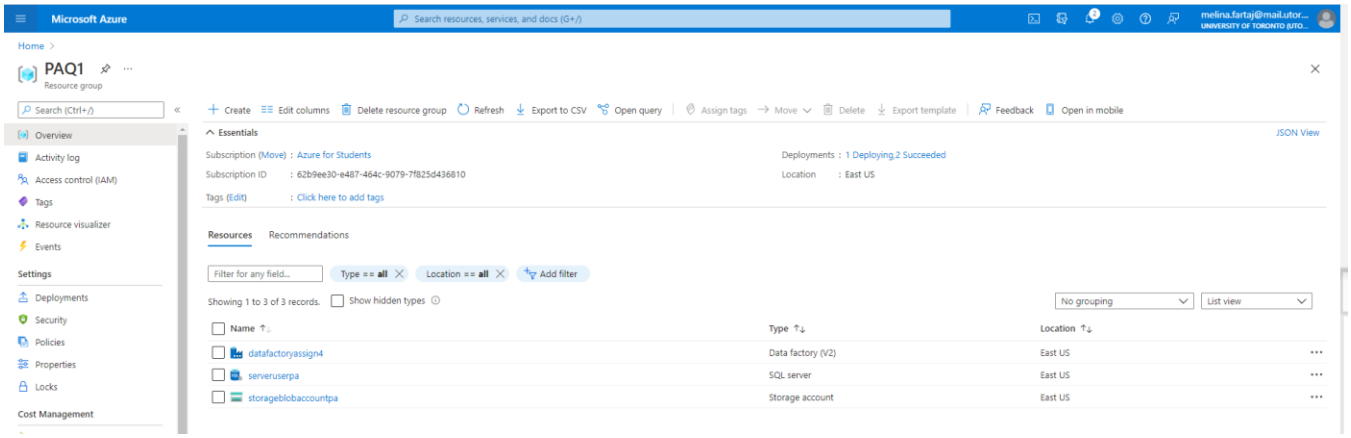


d. Create SQL Database
   i. Click create a resource
   ii. Find SQL database
   iii. Click create
   iv. Select resource group
   v. Create name
   vi. Create server
   vii. Click review + create
   viii. Wait for validation to pass
   ix. Click create
   x. Deploy resource

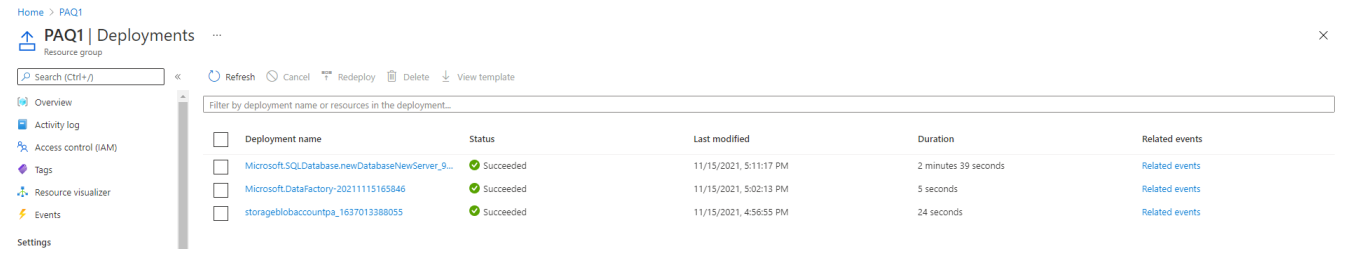




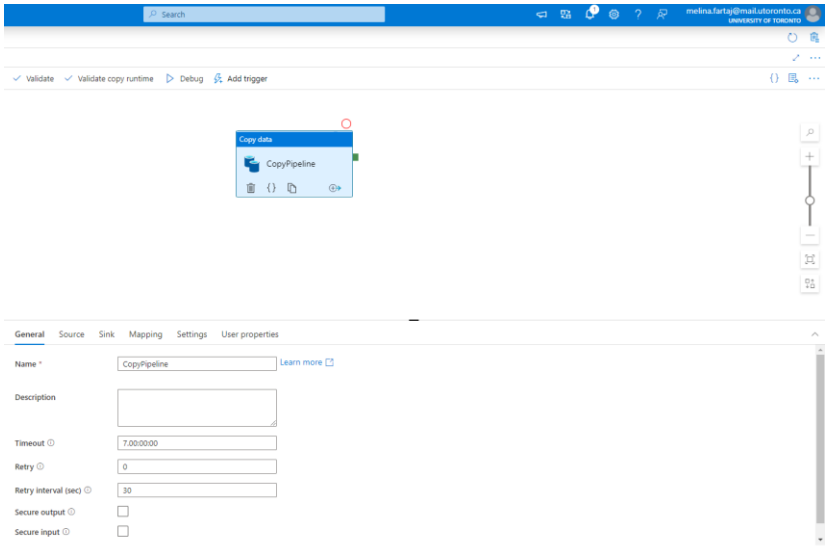
e. Overview of Resource group:



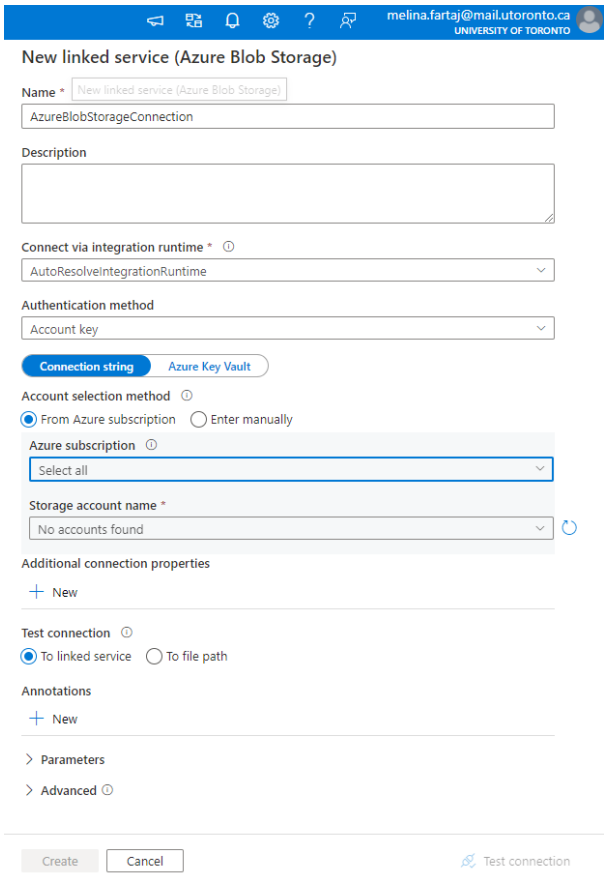
f. Deployed resources

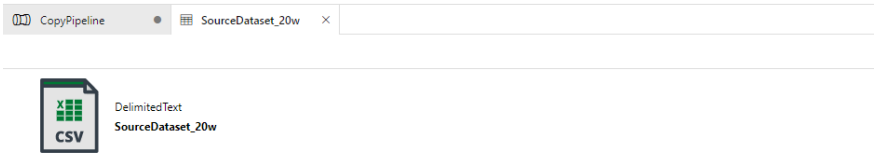


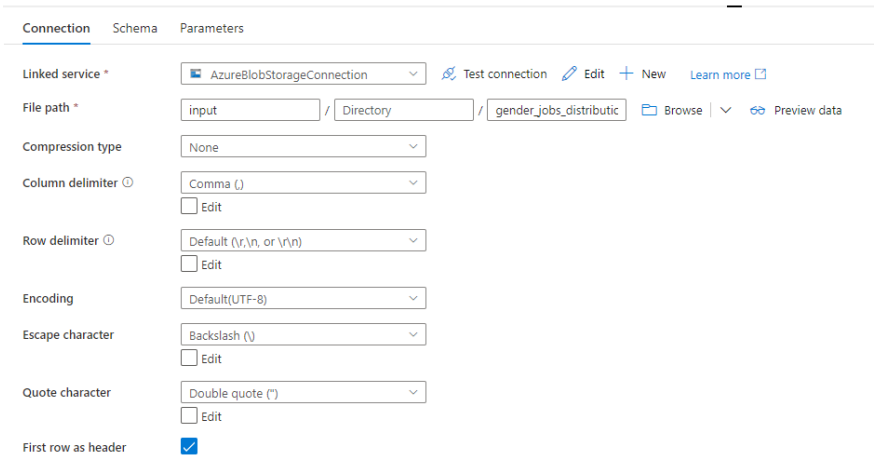
2. Pipeline
    a. Click on datafactory (datafactoryassign4) in resource group
    b. Click open Azure Data Factory Studio
    c. Orchestrate
    d. Move copy data to pane
    e. Change name under general
    f. Create new source dataset
        i. Click New
        ii. Azure Blob Storage
        iii. DelimitedText
        iv. Name Source and create Linked service
    g. Create Sink dataset
        i. Create sql table in sql database
        ii. Click New
        iii. Azure SQL Database
        iv. Name Sink and create Linked service
    h. Validate Pipeline
    i. Debug Pipeline
    j. Publish

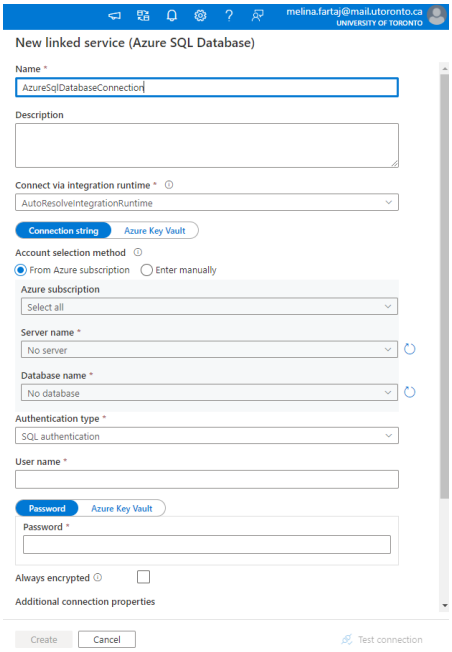

Search
melina.fartaj@mail.utoronto.ca
UNIVERSITY OF TORONTO
Validate
Validate copy runtime
Debug
Add trigger
Copy data
CopyPipeline

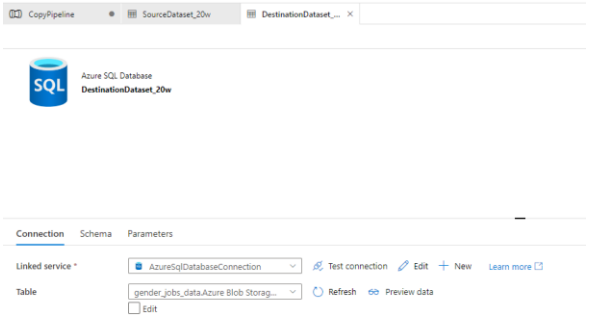
General
Source
Sink
Mapping
Settings
User properties
Name *
CopyPipeline
Learn more
Description
Timeout
7.00:00:00
Retry
0
Retry interval (sec)
30
Secure output
Secure input

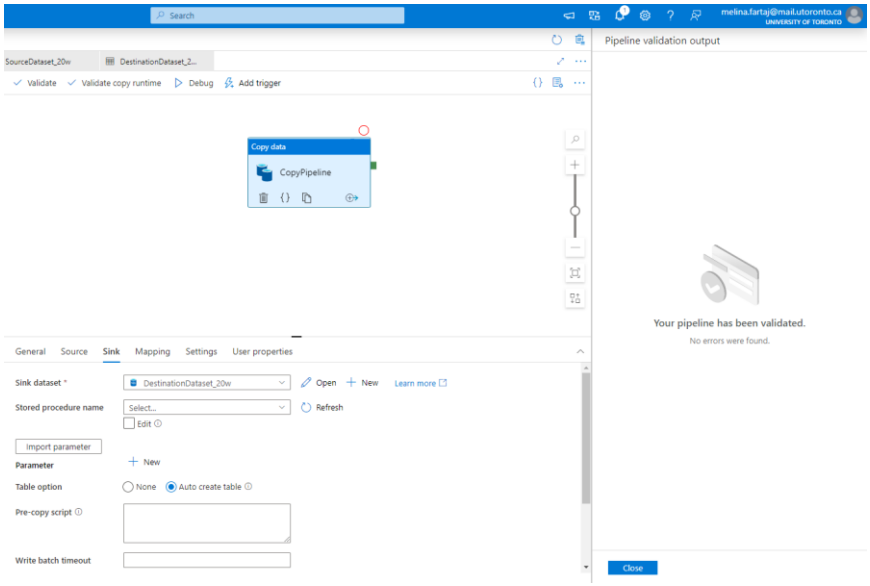

melina.fartaj@mail.utoronto.ca
UNIVERSITY OF TORONTO
New linked service (Azure Blob Storage)
Name *
New linked service (Azure Blob Storage)
AzureBlobStorageConnection
Description
Connect via integration runtime *
AutoResolveIntegrationRuntime
Authentication method
Account key
Connection string
Azure Key Vault
Account selection method
From Azure subscription
Enter manually
Azure subscription
Select all
Storage account name *
No accounts found
Additional connection properties
New
Test connection
To linked service
To file path
Annotations
New
Parameters
Advanced
Create
Cancel
Test connection

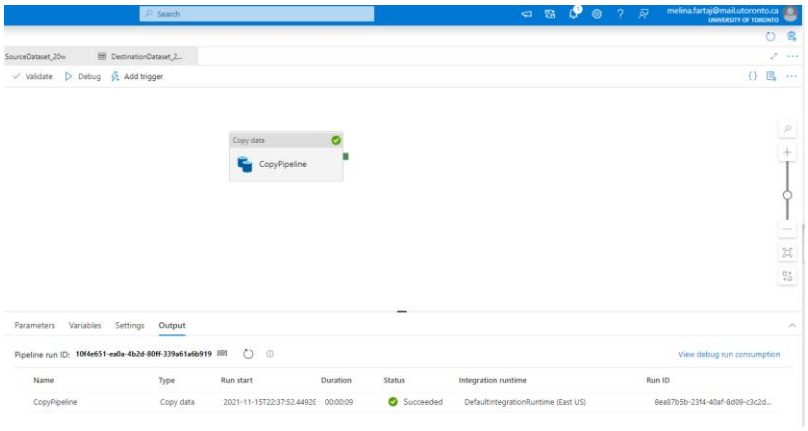

CopyPipeline
SourceDataset_20w
DelimitedText
SourceDataset_20w

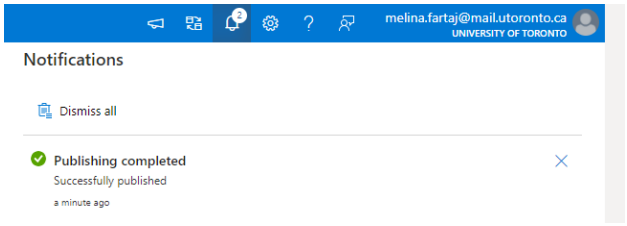

Connection
Schema
Parameters
Linked service *
AzureBlobStorageConnection
Test connection
Edit
New
Learn more
File path *
input
Directory
gender_jobs_distributic
Browse
Preview data
Compression type
None
Column delimiter
Comma (,)
Edit
Row delimiter
Default (\r,\n, or \r\n)
Edit
Encoding
Default(UTF-8)
Escape character
Backslash (\)
Edit
Quote character
Double quote (")
Edit
First row as header

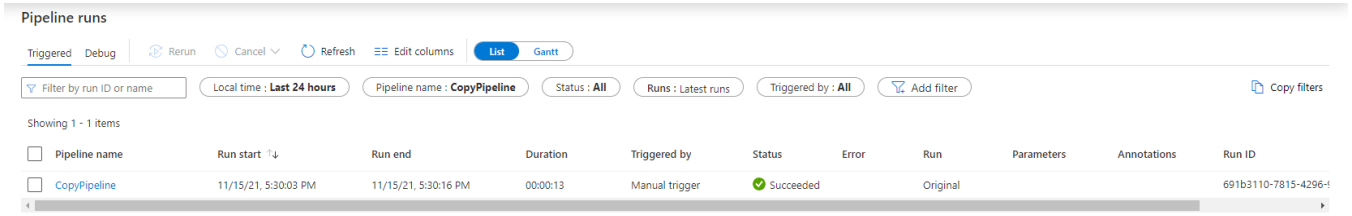

melina.fartaj@mail.utoronto.ca
UNIVERSITY OF TORONTO
New linked service (Azure SQL Database)
Name *
AzureSqlDatabaseConnection
Description
Connect via integration runtime *
AutoResolveIntegrationRuntime
Connection string
Azure Key Vault
Account selection method
From Azure subscription
Enter manually
Azure subscription
Select all
Server name *
No server
Database name *
No database
Authentication type *
SQL authentication
User name *
Password
Azure Key Vault
Password *
Always encrypted
Additional connection properties
Create
Cancel
Test connection


CopyPipeline
SourceDataset_20w
DestinationDataset_...
Azure SQL Database
DestinationDataset_20w
Connection
Schema
Parameters
Linked service *
AzureSqlDatabaseConnection
Test connection
Edit
New
Learn more
Table
gender_jobs_data.Azure Blob Storag...
Refresh
Preview data
Edit





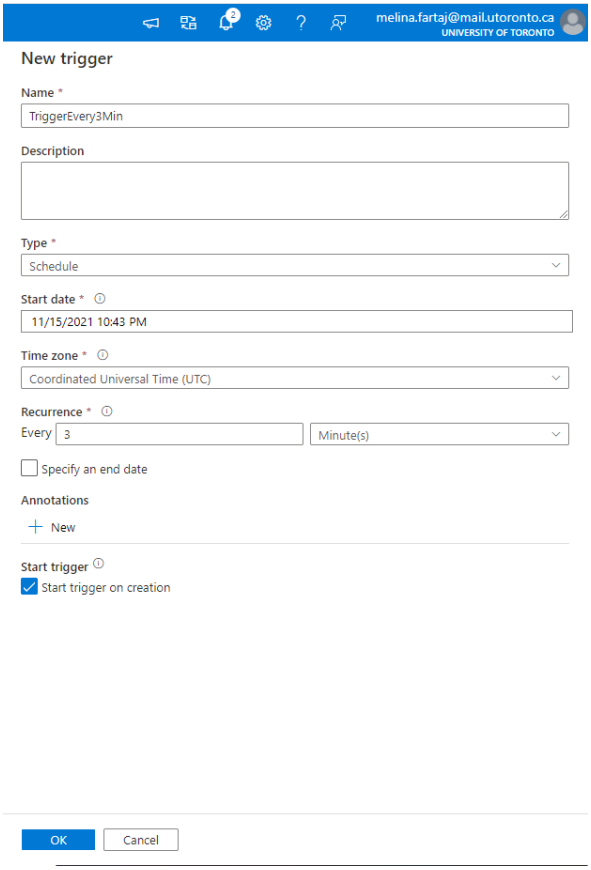


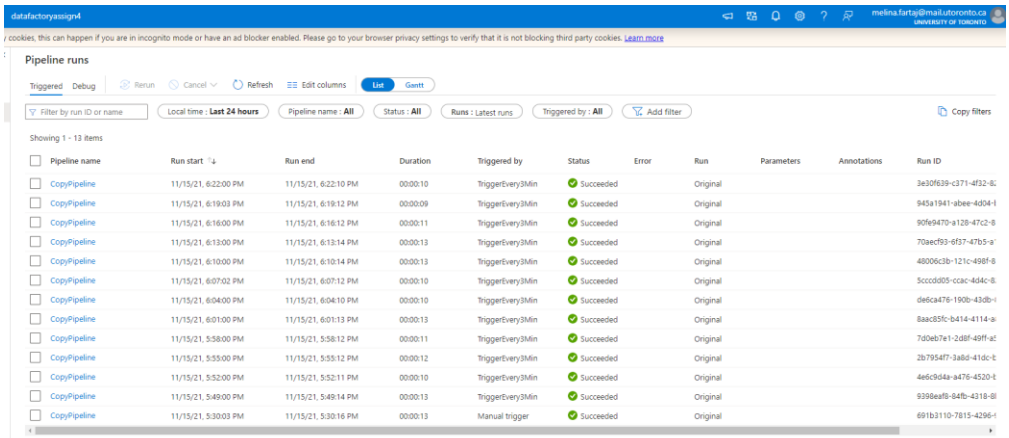


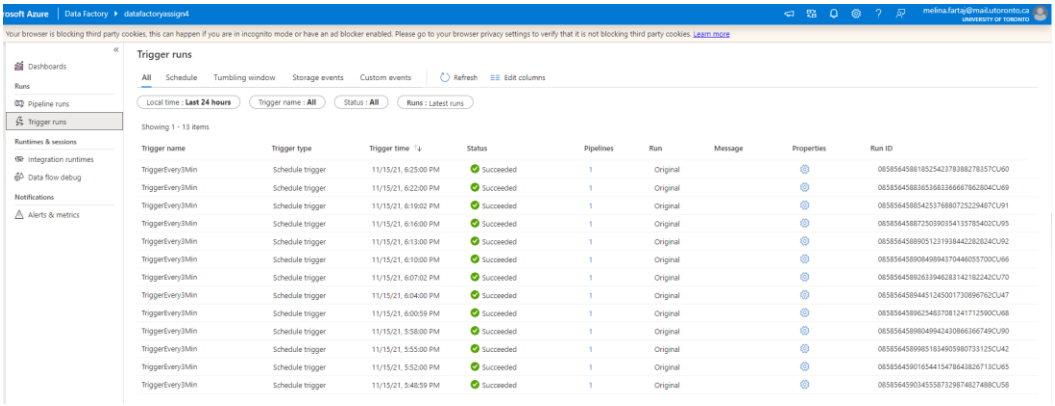
3. Triggers (screenshot + explain)
    a. Different Types of Triggers
        i. Manual – This is on demand execution. The user goes into data factory designer studio and start this pipeline manually.
        ii. Schedule – Scheduling the time your pipeline gets triggered based on a wall clock time.
        iii. Tumbling Window- User defines a time and says that each time interval becomes a slice and that 1 slice will execute the data at that point of time and so on. This goes well with time series data.

iv. Event-based – Trigger based on a new file arrival or file deletion. Trigger is based on an event.
v. Logic App Calls – Actions within our logic apps to create and run pipelines.

b. Schedule Trigger every 3 min
   i. Click on Author tab
   ii. Click add trigger
   iii. New/edit trigger
   iv. Schedule for type
   v. Every 3 min
   vi. OK
   vii. Publish





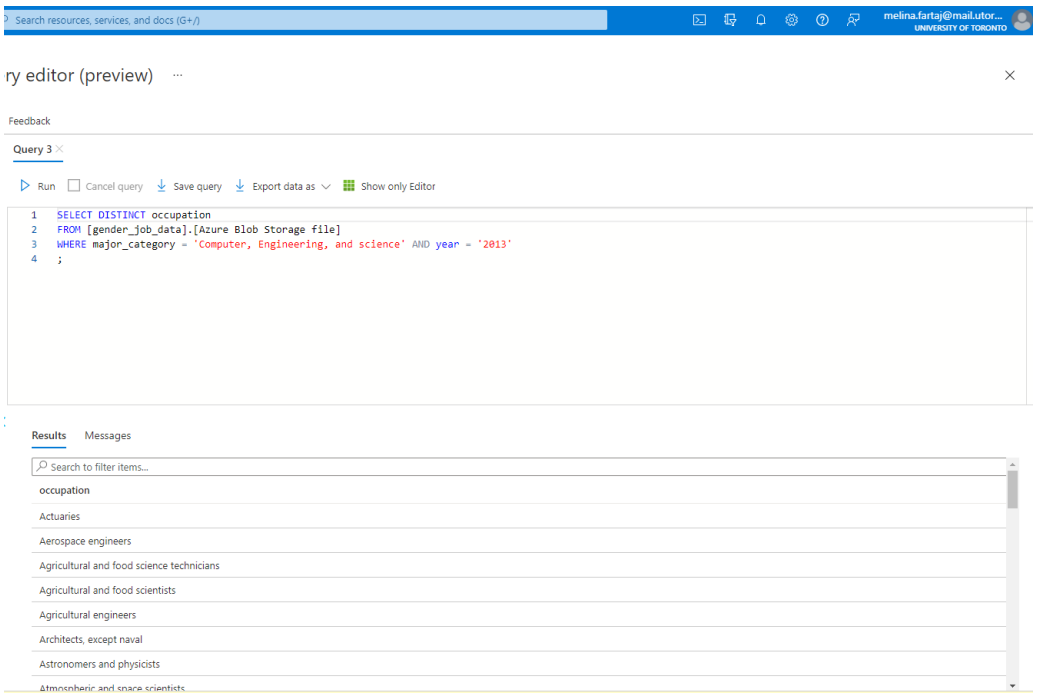


4. Replicate objects
    a. Create two ADLS Gen 2 Storage accounts
        i. Click create resource → storage account
        ii. Put in resource group, storage name and location (Canada central for one and west Europe for another)
        iii. Go to advanced tab
        iv. Click ‘Enable hierarchical namespace ’ for Data Lake Storage Gen2
        v. Review + Create
        vi. Create
        vii. Create Containers in both storage account
    b. Use AzCopy to copy objects
        i. AzCopy is a command-line utility that can be used to copy data to and from storage accounts.
        ii. Open cloud shell
        iii. Type azcopy - - version to check downloaded (it should be already there)
        iv. Receive the SAS token
            1. Click on storage resource
            2. Shared access signature under security + networking
            3. Check mark container and object under allowed resource types
            4. Click Generate SAS and connection string
            5. Copy the SAS token
            6. Repeat for source and destination storage accounts
        v. Use format below command to copy all storage account contents to another storage account.
        vi. Fill in the required names and paste in the SAS tokens and enter into the bash command

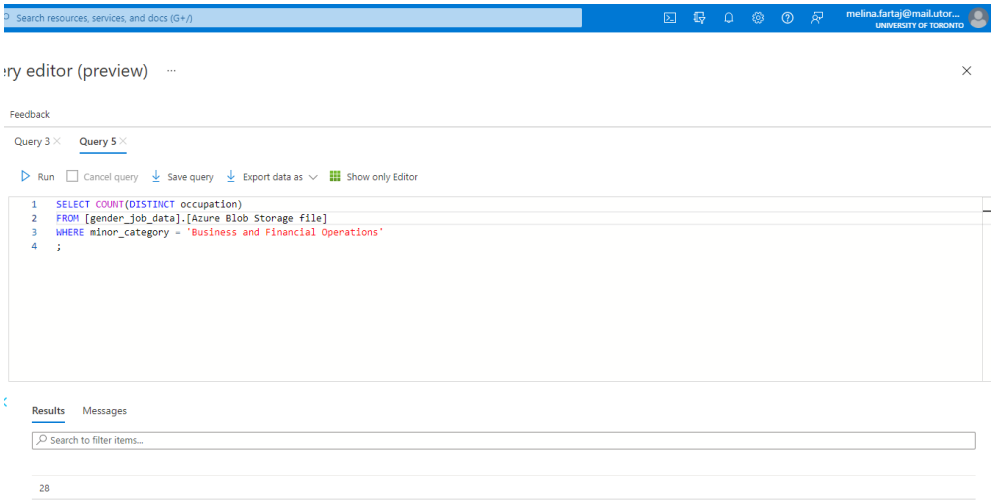
azcopy copy 'https://<source-storage-account-name>.file.core.windows.net/<SAS-token>' 'https://<destination-storage-account-name>.file.core.windows.net/<SAS-token>' --recursive'

Part B:

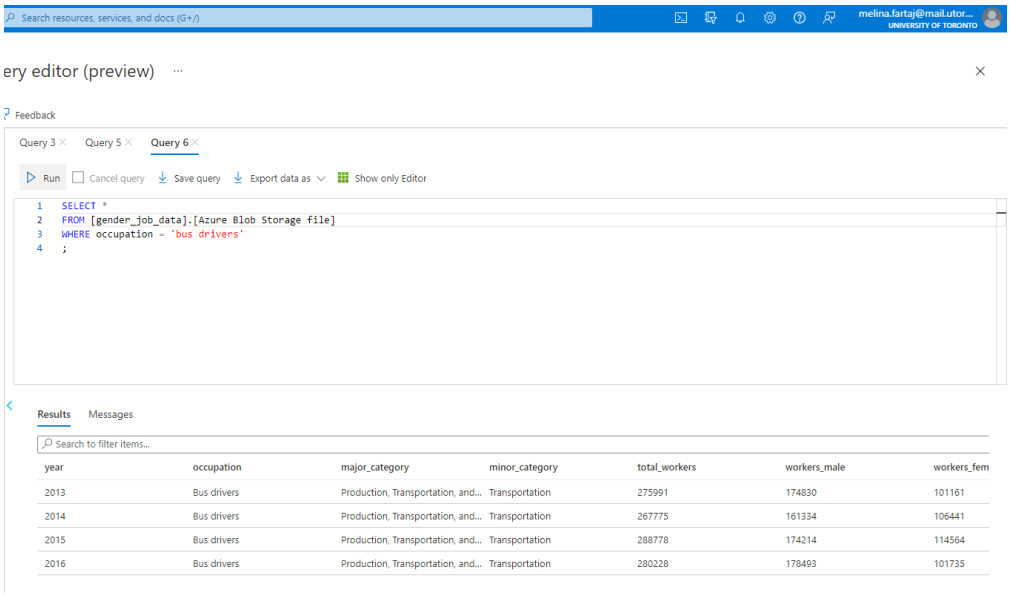
1. Occupations



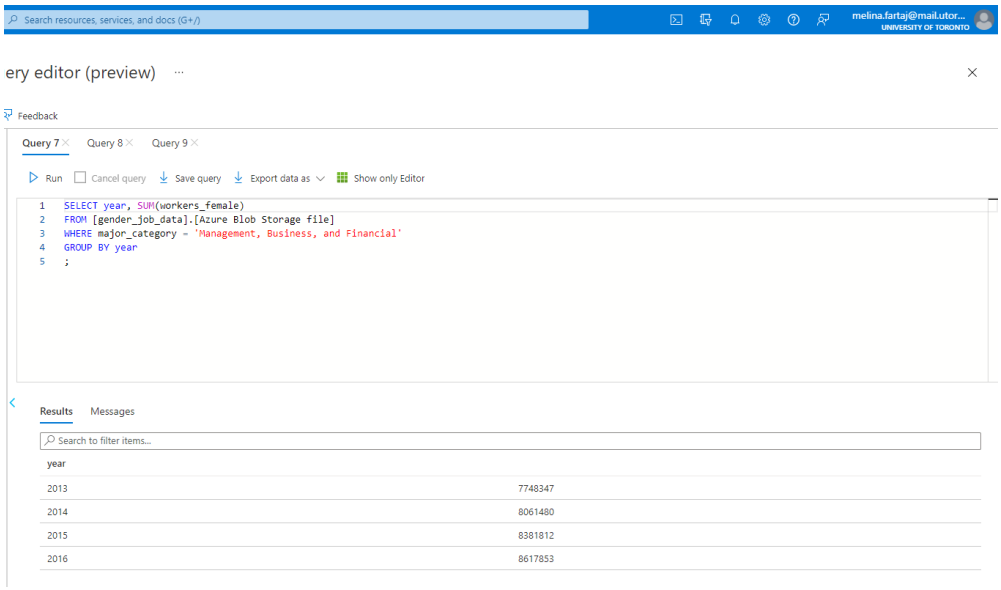
2. Count of Occupations



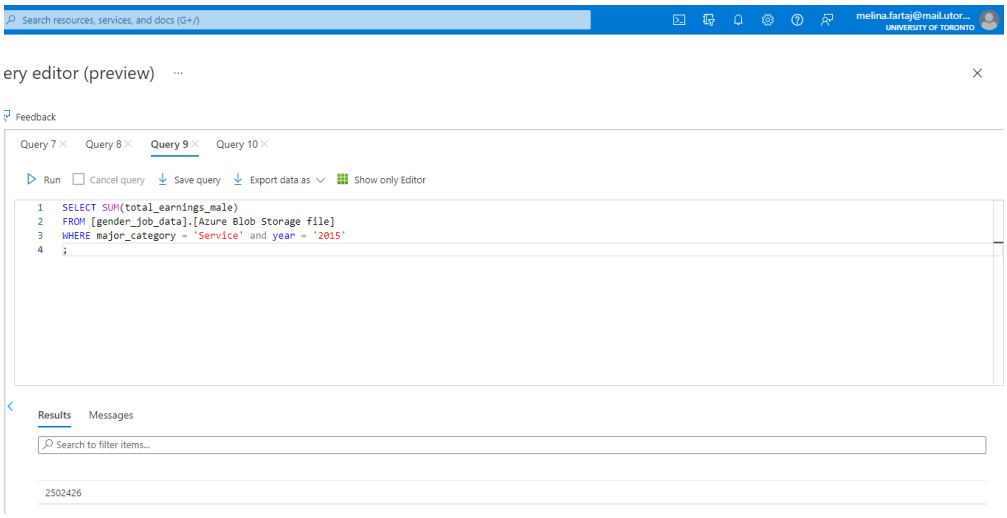
3. Relevant Information



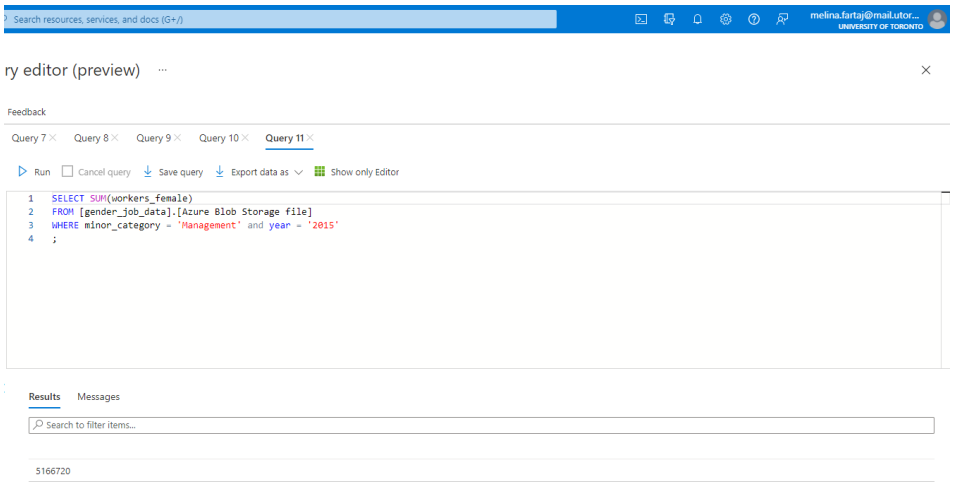
4. Number of Female Workers



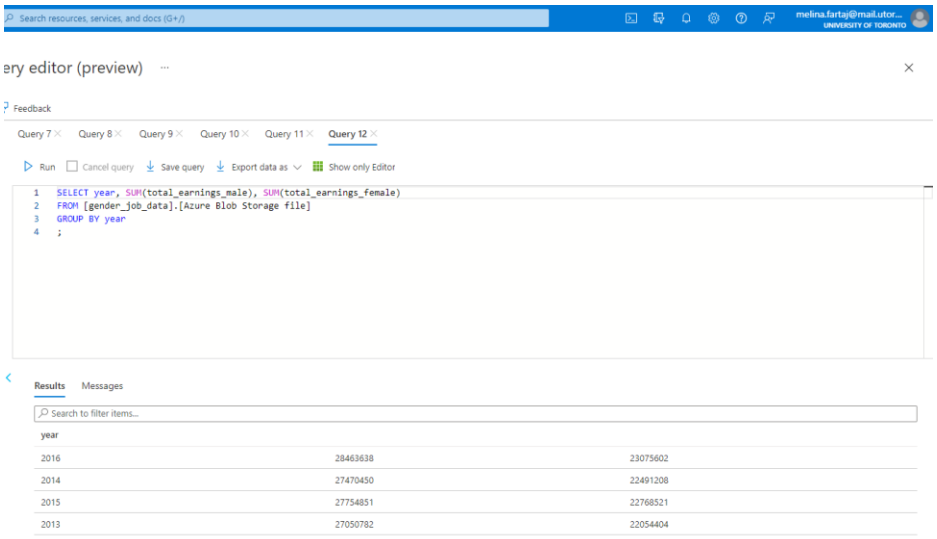
5. Total earnings in Service



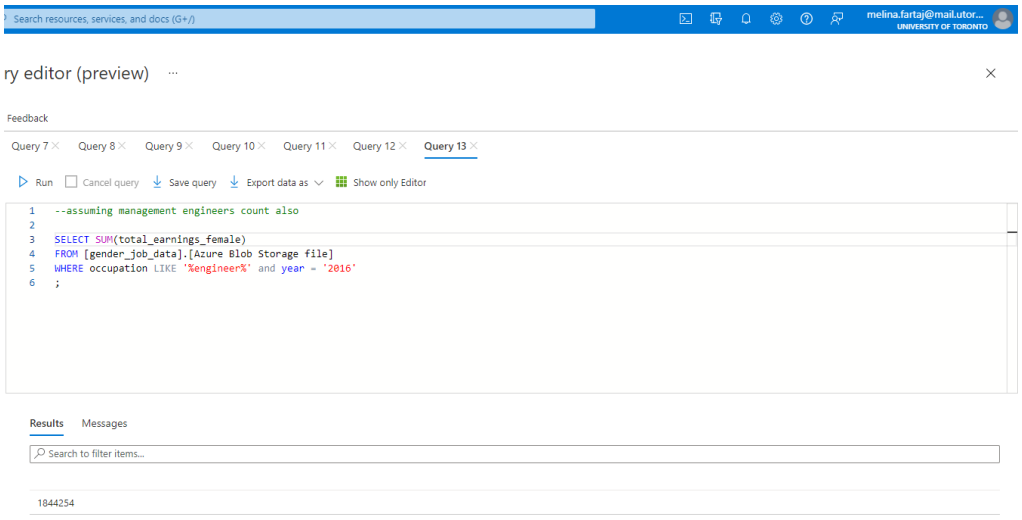
6. Management roles



7. Total earnings



8. Total earnings as Engineers



9. Number of Workers