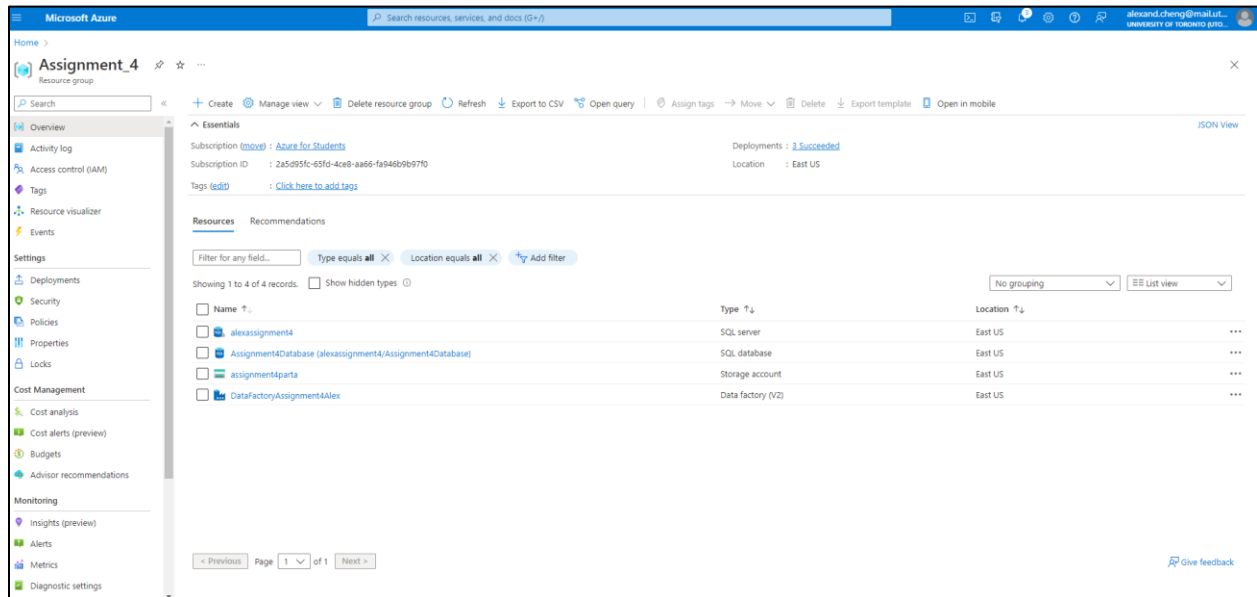
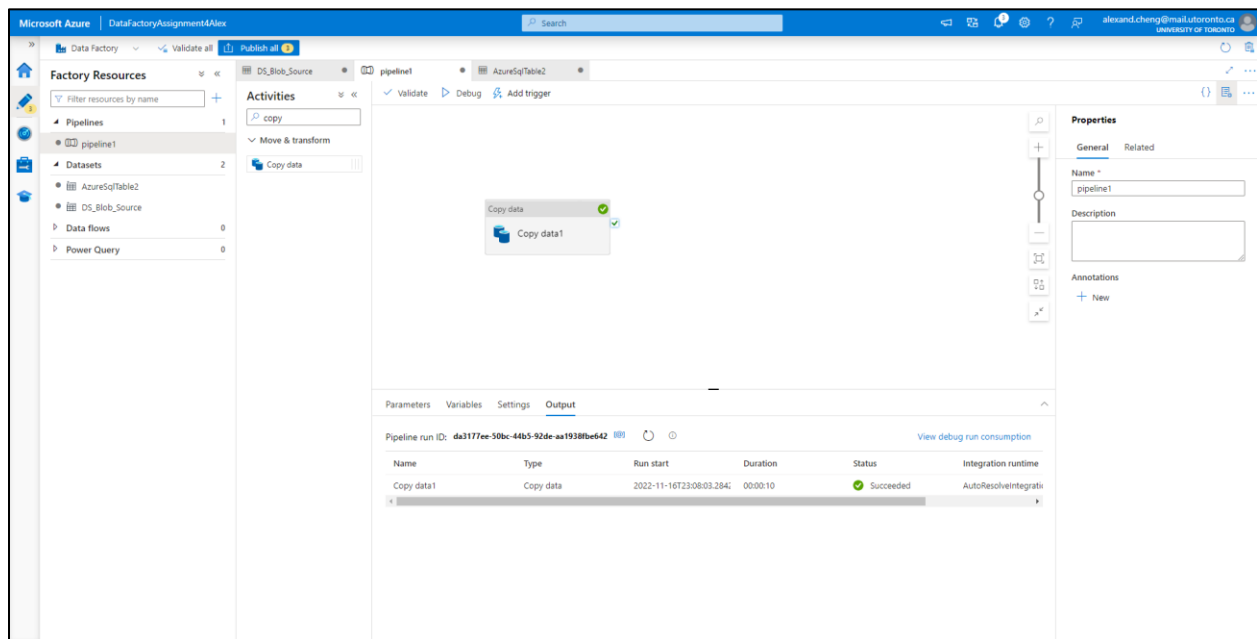


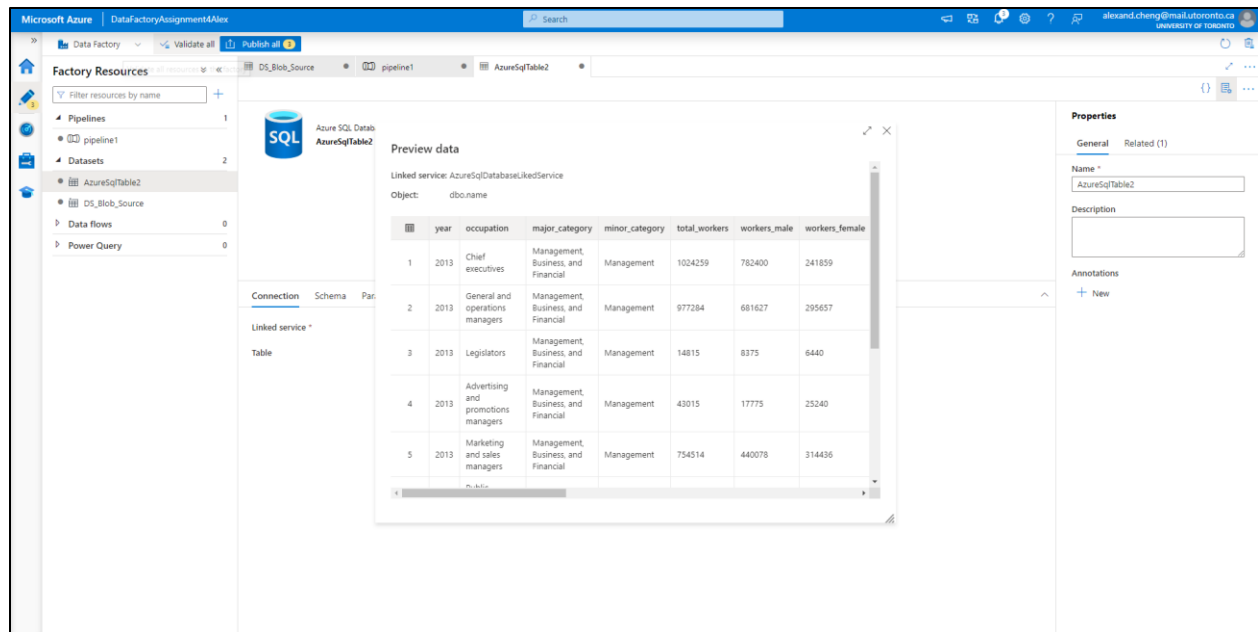
Part A:

1. Create a resource group in your Azure portal and deploy three resources. Azure Data Factory, Azure SQL DB and Blob storage account.



2. Now create a pipeline in Azure Data Factory and copy gender_jobs_data.csv file from the Blob storage account to Azure SQL DB. (First copy this file from your local machine to Blob Storage).

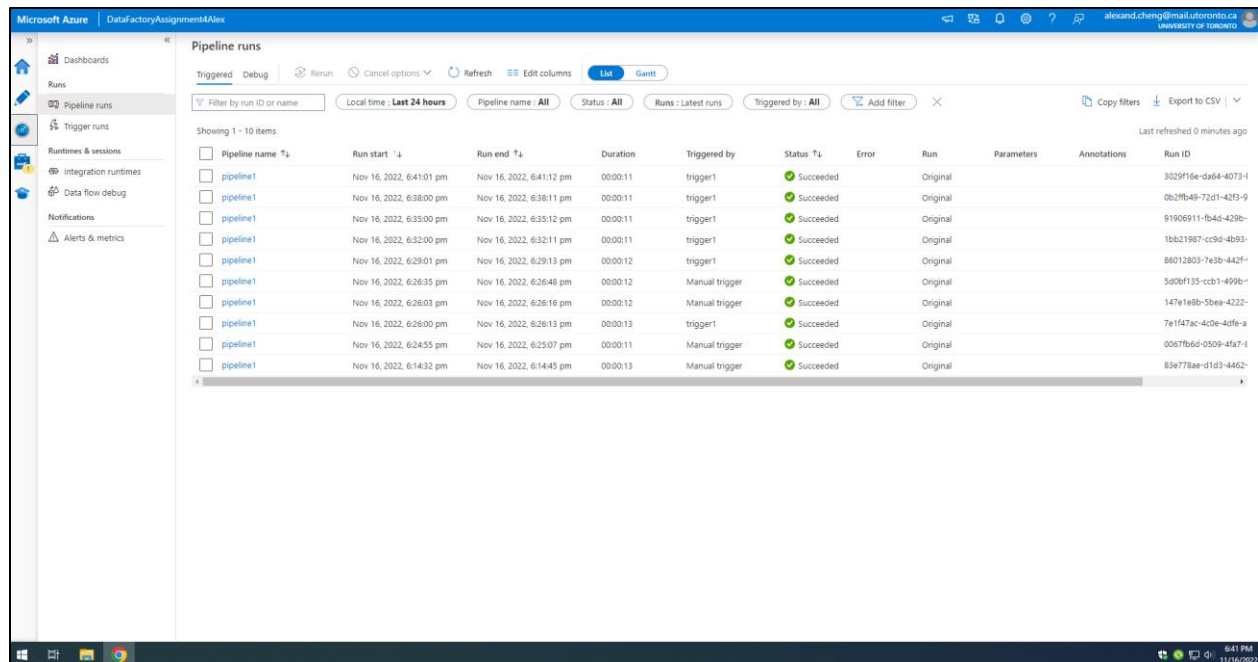




3. Explain the different types of triggers available in ADF. Now create a schedule trigger and run your pipeline every 3 minutes. Show 5 successful runs.

Trigger Options: Used to schedule a Data pipeline run without any interventions.

- Schedule: Can run a Data pipeline according to a predetermined schedule with different scheduling intervals. Can choose the start and end dates, and on specific future calendar dates.
- Tumbling Window: Executes Data pipelines at a specified time slice or predetermined periodic time interval. The tumbling window sends the start and end times for each time window in the Database, returning all data between those periods.
- Storage Event-based: Triggers occurs in response to blob-related events such as generating or deleting a blob event in an Azure blob storage. These are also compatible with Azure Data Lakes.



Pipeline name	Run start	Run end	Duration	Triggered by	Status	Error	Run	Parameters	Annotations	Run ID
pipeline1	Nov 16, 2022, 6:41:01 pm	Nov 16, 2022, 6:41:12 pm	00:00:11	trigger1	Succeeded		Original			3029f16e-9a64-4073-1
pipeline1	Nov 16, 2022, 6:38:00 pm	Nov 16, 2022, 6:38:11 pm	00:00:11	trigger1	Succeeded		Original			0b20fb49-72d1-42f5-9
pipeline1	Nov 16, 2022, 6:35:00 pm	Nov 16, 2022, 6:35:12 pm	00:00:11	trigger1	Succeeded		Original			91906911-fb4d-429b-
pipeline1	Nov 16, 2022, 6:32:00 pm	Nov 16, 2022, 6:32:11 pm	00:00:11	trigger1	Succeeded		Original			1bb21987-cc9d-4b93-
pipeline1	Nov 16, 2022, 6:29:01 pm	Nov 16, 2022, 6:29:13 pm	00:00:12	trigger1	Succeeded		Original			88012803-7e3b-442f-
pipeline1	Nov 16, 2022, 6:26:35 pm	Nov 16, 2022, 6:26:48 pm	00:00:12	Manual trigger	Succeeded		Original			5d0ef135-ccb1-499b-
pipeline1	Nov 16, 2022, 6:26:03 pm	Nov 16, 2022, 6:26:16 pm	00:00:12	Manual trigger	Succeeded		Original			147e1e8b-5bee-4222-
pipeline1	Nov 16, 2022, 6:26:00 pm	Nov 16, 2022, 6:26:13 pm	00:00:13	trigger1	Succeeded		Original			7e1f47ac-4c0e-4dfe-a
pipeline1	Nov 16, 2022, 6:24:55 pm	Nov 16, 2022, 6:25:07 pm	00:00:11	Manual trigger	Succeeded		Original			0067fb6d-0509-4fa7-e
pipeline1	Nov 16, 2022, 6:14:32 pm	Nov 16, 2022, 6:14:45 pm	00:00:13	Manual trigger	Succeeded		Original			83e778ae-d1d3-4462-

4. A client needs to replicate objects from ADLS Gen 2 in Canada Central to ADLS Gen 2 in West Europe. Let's say they want to do this in a bi-directional way. How can you set this up? [Hint: This probably can be done using Azure Data Factory and Event Triggers. For eg; every time there is a new Blob on one side, it needs to be replicated on the other one]

The process will be described on how to implement this but will not be shown.

Cross regional replication provides data recovery in cases of failure, allows staggered times for updates which will minimize downtimes, and reduce the chances of regional disaster network outage. Azure storage account are used to deploy storage resources such as blob containers, file shares, tables, or queues.

We want Geo-redundancy Storage (GRS) or Geo-Zone-redundant storage (GZRS), replication between different regions. Can change the replication setting using the portal "Storage -> Data Management -> Redundancy -> Update settings to Geo-redundant storage and choose "West Europe"".

Need to perform a manual migration, which allows us to move a storage account to another region, although this may have downtime in which a conversion option is a in-place migration with no downtime. However, with the GRS or GZRS options, the data in the secondary region isn't available for read/write access.

Using Azure portal will need to export the storage account template (resources -> automation -> export template), modify the template with the new storage account name and location in the JSON file, move/deploy the storage account to create a new storage in the target location. Configure the new storage account, copy the data, there are many tools to copy data such as AzCopy. To copy the data, the

ADLS Gen 2 Canada Central will be used as a source type and the ADLS Gen 2 West Europe will be used as a sink type.

Need to make sure that the storage in West Europe supports the desired replication settings (Also Geo-redundancy). Azure storage redundancy becomes more expensive when moving to geo-redundancy since it's a more sophisticated redundancy level.

Bi-directional sync is ideal for complex scenarios that involves many pipelines and dependencies. Need to create replication rules that determine the directory in the file system that will be replicated and the Zones that will be used in that replication. Without replication rules defined, each Zone's file system operates independently of the other. A tool such as WANdisco Fusion allows users control over how data is replicated between file systems and object stores.

Need to set up event triggers such that if there's a new blob in one Zone, it will replicate objects from ADLS Gen 2 Canada Central to ADLS Gen 2 in West Europe using the steps shown above. Another event trigger such that if there's a new blob in one Zone, will be used to replicate object from ADLS Gen 2 West Europe to ADLS Gen 2 in Canada Central using the steps shown above.

Part B:

1. In the gender_jobs_data table - Filter all the OCCUPATIONS in MAJOR_CATEGORY of Computer, Engineering, and Science for the YEAR 2013
 - See Part B-1.csv for output



The screenshot shows a SQL query editor with two tabs: 'Query 1' and 'Query 2'. The 'Query 1' tab is active. The query text is as follows:

```
1 SELECT occupation
2 FROM [dbo].[name]
3 WHERE major_category = 'Computer, Engineering, and Science' AND year = 2013
```

Below the query text, there are several icons and labels: a blue play button icon labeled 'Run', a square icon labeled 'Cancel query', a blue download icon labeled 'Save query', a blue download icon labeled 'Export data as' with a dropdown arrow, and a green grid icon labeled 'Show only Editor'.

occupation					
Web developers					
Computer support specialists					
Database administrators					
Network and computer systems administrators					
Computer network architects					
Computer , all other					
Actuaries					
Mathematicians					
Operations research analysts					
Statisticians					
Miscellaneous mathematical science					
Architects, except naval					
Surveyors, cartographers, and photogrammetrists					
Aerospace engineers					
Agricultural engineers					
Biomedical engineers					
Chemical engineers					
Civil engineers					
Computer hardware engineers					
Electrical and electronics engineers					
Environmental engineers					
Economists					
Computer and information research scientists					
Computer systems analysts					
Information security analysts					
Computer programmers					
Software developers, applications and systems software					
Industrial engineers, including health and safety					
Marine engineers and naval architects					
Materials engineers					
Mechanical engineers					
Mining and geological engineers, including mining safety engineers					
Nuclear engineers					
Petroleum engineers					
Engineers, all other					
Drafters					
Engineering technicians, except drafters					

Computer systems analysts					
Information security analysts					
Computer programmers					
Software developers, applications and systems software					
Industrial engineers, including health and safety					
Marine engineers and naval architects					
Materials engineers					
Mechanical engineers					
Mining and geological engineers, including mining safety engineers					
Nuclear engineers					
Petroleum engineers					
Engineers, all other					
Drafters					
Engineering technicians, except drafters					
Surveying and mapping technicians					
Agricultural and food scientists					
Biological scientists					
Conservation scientists and foresters					
Medical scientists					
Life scientists, all other					
Astronomers and physicists					
Atmospheric and space scientists					
Chemists and materials scientists					
Environmental scientists and geoscientists					
Physical scientists, all other					
Survey researchers					
Psychologists					
Urban and regional planners					
Miscellaneous social scientists and related workers, including sociologists					
Agricultural and food science technicians					
Biological technicians					
Chemical technicians					
Geological and petroleum technicians					
Nuclear technicians					
Social science research assistants					
Miscellaneous life, physical, and social science technicians					

2. In the gender_jobs_data table - How many OCCUPATIONS exist in the MINOR_CATEGORY of Business and Financial Operations overall?

The screenshot shows the Azure Data Studio interface with the 'Query editor (preview)' window open. The query editor displays the following SQL query:

```
1 SELECT COUNT(occupation)
2 FROM [dbo].[name]
3 WHERE minor_category = 'Business and Financial Operations'
```

The query is executed, and the results are displayed in the 'Results' tab. The results show a single row with the value 112.

Query succeeded | 0s

If asking for unique occupations:

The screenshot shows the Azure Data Studio interface with the 'Query editor (preview)' window open. The query editor displays the following SQL query:

```
1 SELECT COUNT(DISTINCT occupation)
2 FROM [dbo].[name]
3 WHERE minor_category = 'Business and Financial Operations'
```

The query is executed, and the results are displayed in the 'Results' tab. The results show a single row with the value 20.

Query succeeded | 0s

3. In the gender_jobs_data table - Get all relevant information for bus drivers across all years
 - See Part B-1.csv for output

Microsoft Azure | Assignment4Database (alexassignment4/Assignment4Database) | Query editor (preview)

SQL database

Query 1 × Query 2 × Query 3 × Query 4 × Query 5 ×

```
1 SELECT *
2 FROM [dbo].[name]
3 WHERE occupation = 'Bus drivers'
```

Results Messages

year	occupation	major_category	minor_category	total_workers	workers_male	workers_fem
2013	Bus drivers	Production, Transportation, and...	Transportation	275991	174930	101161
2014	Bus drivers	Production, Transportation, and...	Transportation	267775	161334	106441
2015	Bus drivers	Production, Transportation, and...	Transportation	288778	174214	114564
2016	Bus drivers	Production, Transportation, and...	Transportation	280228	178493	101735

Query succeeded | 0s

4. In the gender_jobs_data table - Summarize the total number of WORKERS_FEMALE in the MAJOR_CATEGORY of Management, Business, and Financial by each year

Microsoft Azure | Assignment4Database (alexassignment4/Assignment4Database) | Query editor (preview)

SQL database

Query 1 × Query 2 × Query 3 × Query 4 × Query 5 × Query 6 × Query 7 × Query 8 × Query 9 × Query 10 ×

```
1 SELECT Year, SUM(CAST(workers_female as int)) as 'Number of Female Workers'
2 FROM [dbo].[name]
3 WHERE major_category = 'Management, Business, and Financial'
4 GROUP BY year
5 ORDER BY year
```

Results Messages

Year	Number of Female Workers
2013	7748347
2014	8061480
2015	8381812
2016	8617853

Query succeeded | 0s

5. In the gender_jobs_data table - What were the total earnings of male (TOTAL_EARNINGS_MALE) employees in the Service MAJOR_CATEGORY for the year 2015?

The screenshot shows the Azure Data Studio interface with the 'Assignment4Database (alexassignment4/Assignment4Database)' open. The 'Query editor (preview)' is active, displaying a SQL query:

```
1 SELECT SUM(CAST(total_earnings_male as int))
2 FROM [dbo].[name]
3 WHERE major_category = 'Service' AND year = 2015
```

The 'Results' tab shows a single value: 2502426.

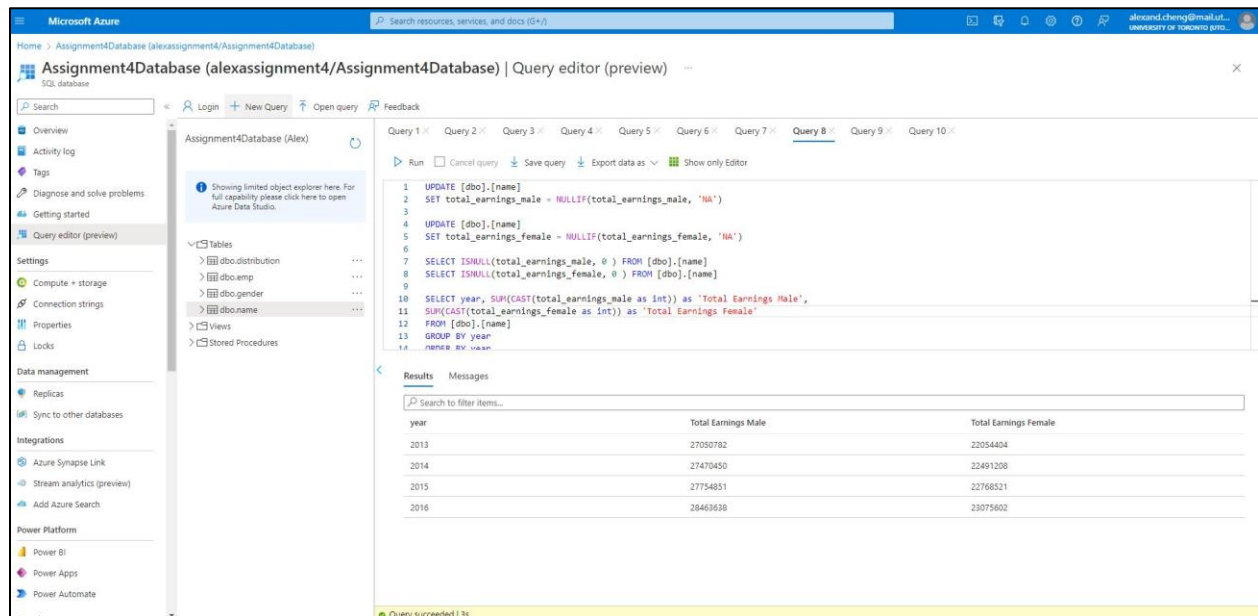
6. In the gender_jobs_data table - How many female workers were in management roles in the year 2015?

The screenshot shows the Azure Data Studio interface with the 'Assignment4Database (alexassignment4/Assignment4Database)' open. The 'Query editor (preview)' is active, displaying a SQL query:

```
1 SELECT SUM(CAST(workers_female as int))
2 FROM [dbo].[name]
3 WHERE minor_category = 'Management' AND year = 2015
```

The 'Results' tab shows a single value: 5160720.

7. In the gender_jobs_data table - Compare the TOTAL_EARNINGS_MALE and TOTAL_EARNINGS_FEMALE earnings irrespective of occupation by each year
- Filter out NA values as 0's



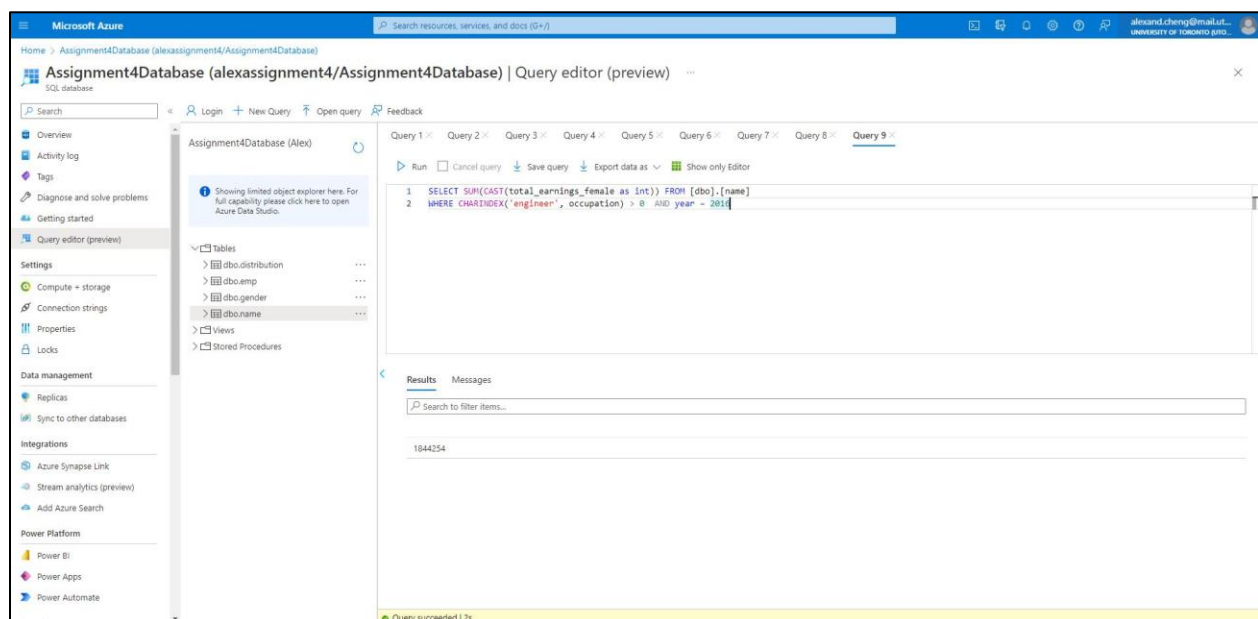
Assignment4Database (alexassignment4/Assignment4Database) | Query editor (preview)

```
1 UPDATE [dbo].[name]
2 SET total_earnings_male = NULLIF(total_earnings_male, 'NA')
3
4 UPDATE [dbo].[name]
5 SET total_earnings_female = NULLIF(total_earnings_female, 'NA')
6
7 SELECT ISNULL(total_earnings_male, 0) FROM [dbo].[name]
8 SELECT ISNULL(total_earnings_female, 0) FROM [dbo].[name]
9
10 SELECT year, SUM(CAST(total_earnings_male as int)) as 'Total Earnings Male',
11 SUM(CAST(total_earnings_female as int)) as 'Total Earnings Female'
12 FROM [dbo].[name]
13 GROUP BY year
14 ORDER BY year
```

year	Total Earnings Male	Total Earnings Female
2013	27050782	22054404
2014	27470450	22491208
2015	27754851	22768521
2016	28463638	23075602

Query succeeded | 3s

8. In the gender_jobs_data table - How much money (TOTAL_EARNINGS_FEMALE) did female workers make as engineers in 2016?
- Included every occupation that included the word 'Engineer', and summed up the total earnings



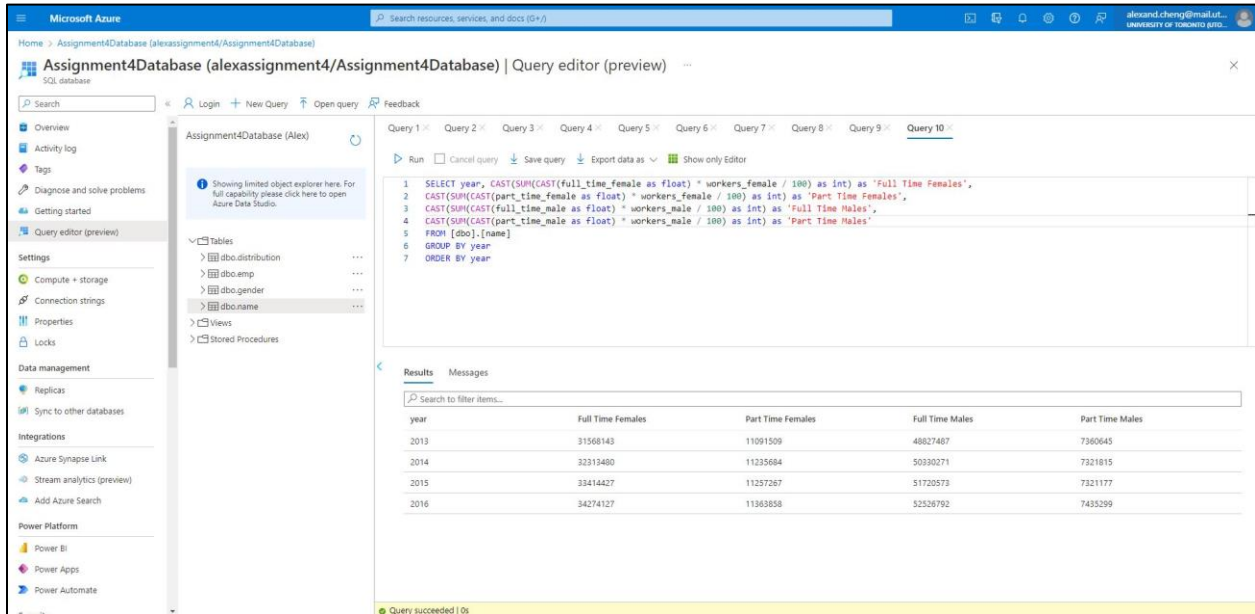
Assignment4Database (alexassignment4/Assignment4Database) | Query editor (preview)

```
1 SELECT SUM(CAST(total_earnings_female as int)) FROM [dbo].[name]
2 WHERE CHARINDEX('engineer', occupation) > 0 AND year = 2016
```

1844254

Query succeeded | 2s

9. What is the total number of full-time and part-time female workers versus male workers year over year?
- Divide by 100 since the numbers given are in percentages



The screenshot displays the Microsoft Azure portal interface for the 'Assignment4Database (alexassignment4/Assignment4Database)' in the 'Query editor (preview)'. The query editor shows a SQL query that calculates the total number of full-time and part-time female and male workers for each year from 2013 to 2016. The query is as follows:

```
1 SELECT year, CAST(SUM(CAST(full_time_female as float) * workers_female / 100) as int) as 'Full Time Females',
2 CAST(SUM(CAST(part_time_female as float) * workers_female / 100) as int) as 'Part Time Females',
3 CAST(SUM(CAST(full_time_male as float) * workers_male / 100) as int) as 'Full Time Males',
4 CAST(SUM(CAST(part_time_male as float) * workers_male / 100) as int) as 'Part Time Males'
5 FROM [dbo].[name]
6 GROUP BY year
7 ORDER BY year
```

The results table shows the following data:

year	Full Time Females	Part Time Females	Full Time Males	Part Time Males
2013	31568143	11091509	48827487	7360645
2014	32313480	11235684	50330271	7321815
2015	33414427	11257267	51720573	7321177
2016	34274127	11363858	52526792	7435299

The query succeeded, as indicated by the 'Query succeeded | 0s' status at the bottom.