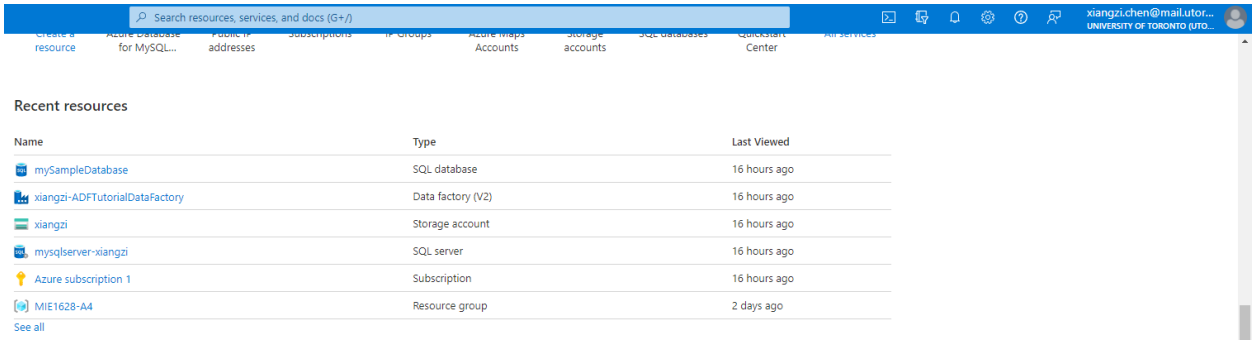
MIE1628 – Assignment 5

Xiangzi Chen (1003818915)

Part A:

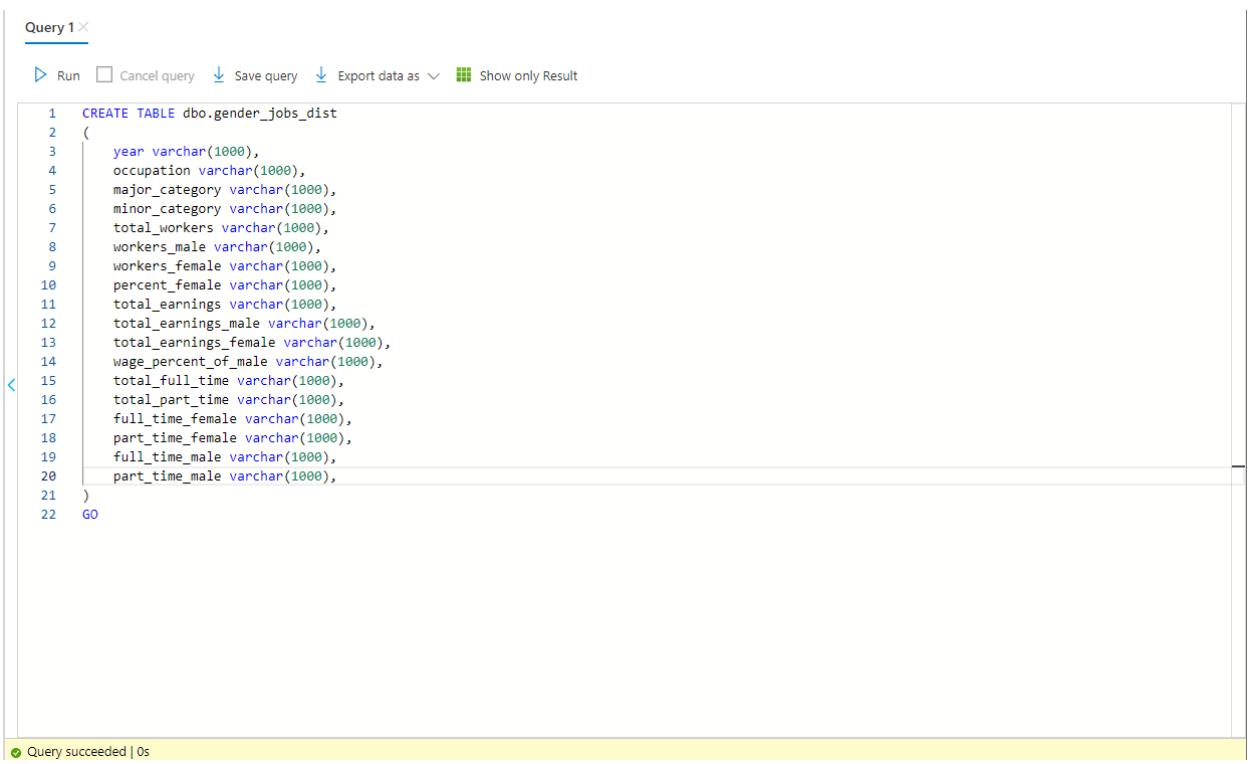
1. [Marks: 5] Create a resource group in your Azure portal and deploy three resources. Azure Data Factory, Azure SQL DB and Blob storage account.



All the resources are contained in resource group 'MIE1628-A4'.

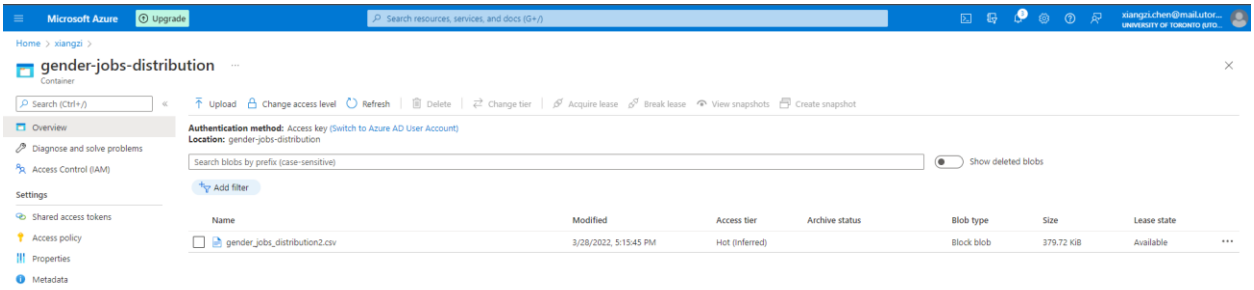
2. [Marks: 15] Now create a pipeline in Azure Data Factory and copy *gender_jobs_data.csv* file from Blob storage account to Azure SQL DB. (First copy this file from your local machine to Blob Storage). See this https://docs.microsoft.com/en-us/azure/data-factory/tutorial-copy-data-portal for reference.

1) Use SQL query to create the schema for file *gender_jobs_data.csv*.



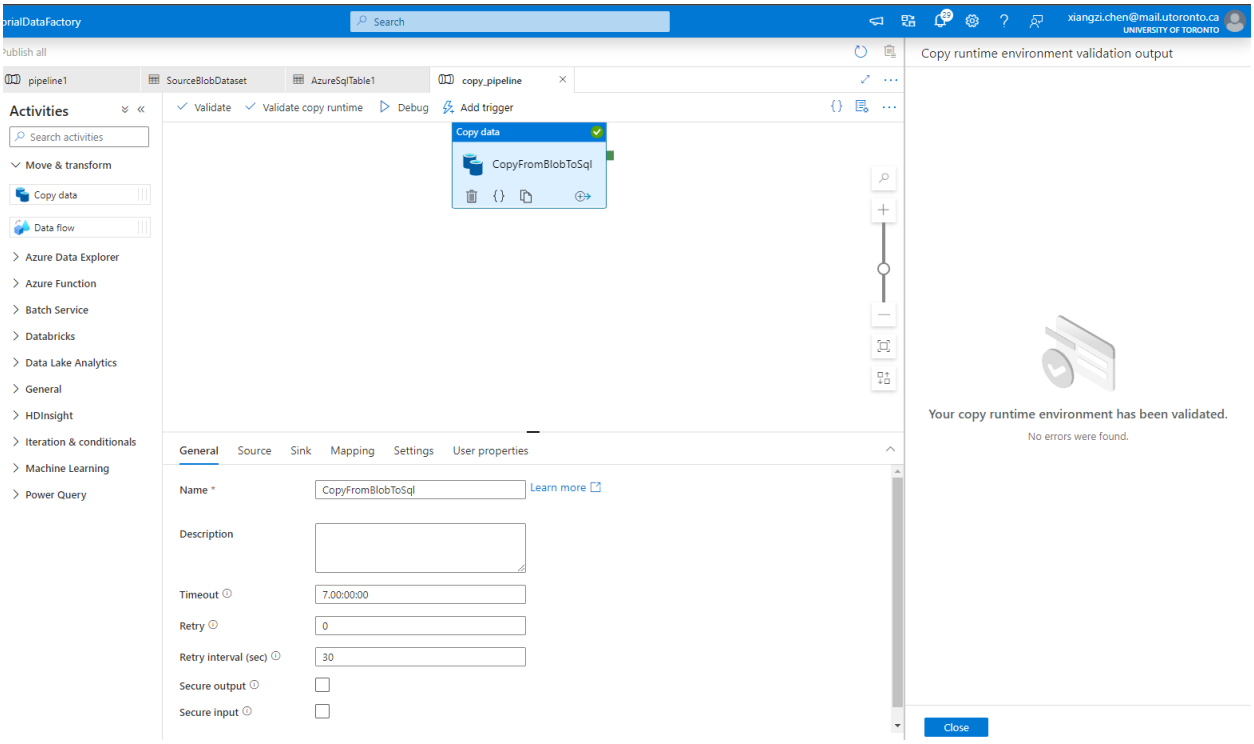

```
CREATE TABLE dbo.gender_jobs_dist
(
    year varchar(1000),
    occupation varchar(1000),
    major_category varchar(1000),
    minor_category varchar(1000),
    total_workers varchar(1000),
    workers_male varchar(1000),
    workers_female varchar(1000),
    percent_female varchar(1000),
    total_earnings varchar(1000),
    total_earnings_male varchar(1000),
    total_earnings_female varchar(1000),
    wage_percent_of_male varchar(1000),
    total_full_time varchar(1000),
    total_part_time varchar(1000),
    full_time_female varchar(1000),
    part_time_female varchar(1000),
    full_time_male varchar(1000),
    part_time_male varchar(1000),
)
GO
```

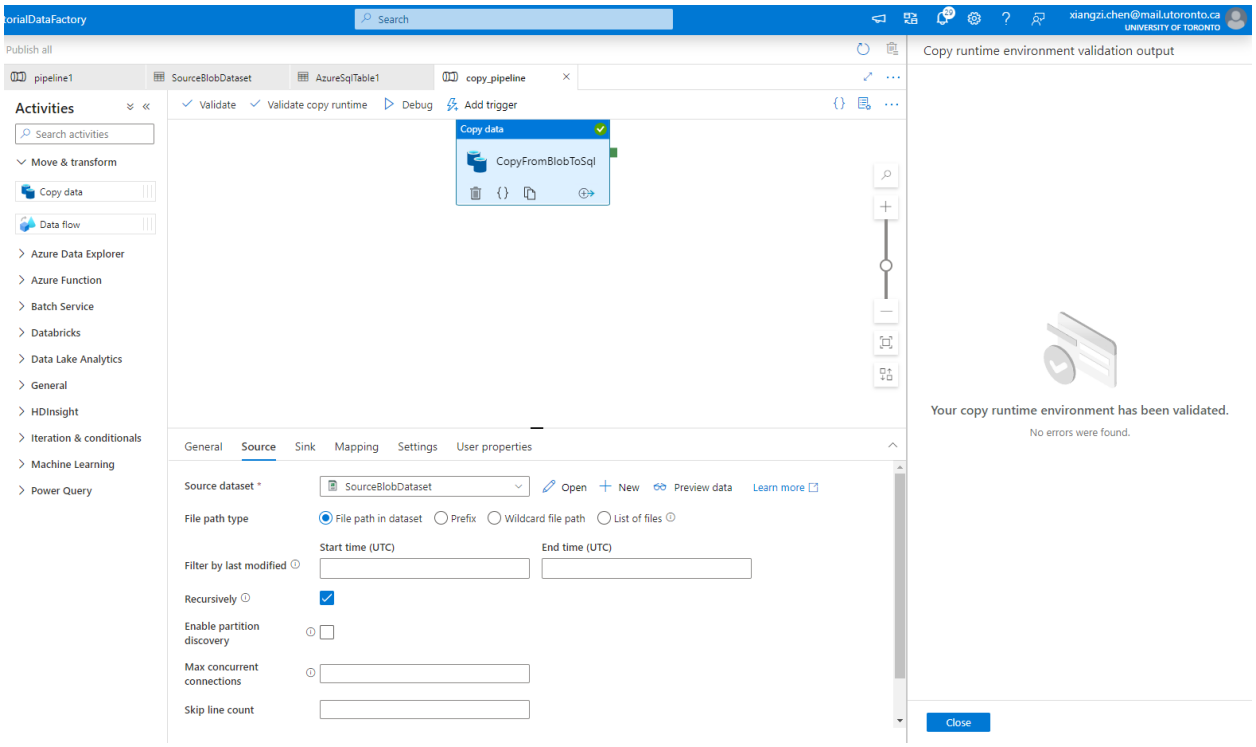
2) Upload the csv file from local to a blob container in the storage account.

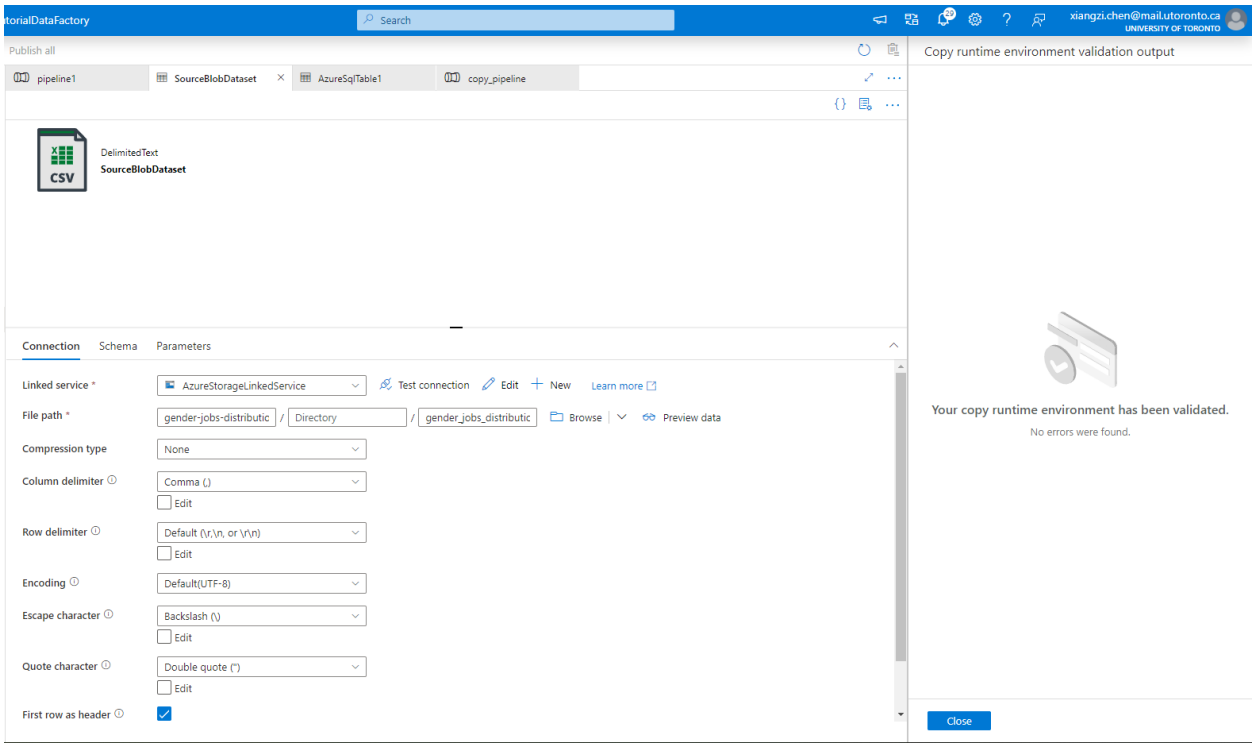


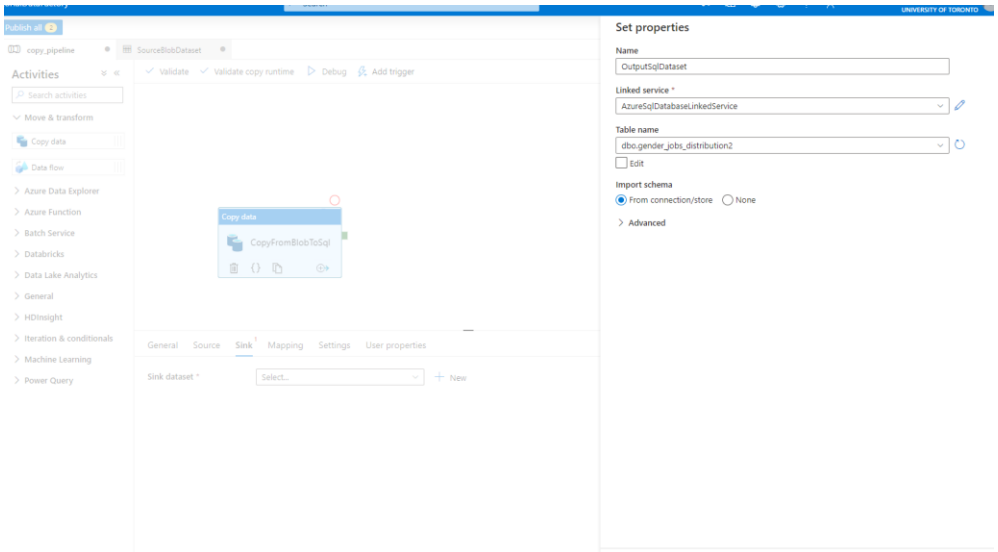
3) Create a copy pipeline to copy data from Blob storage to SQL database in the data factory.

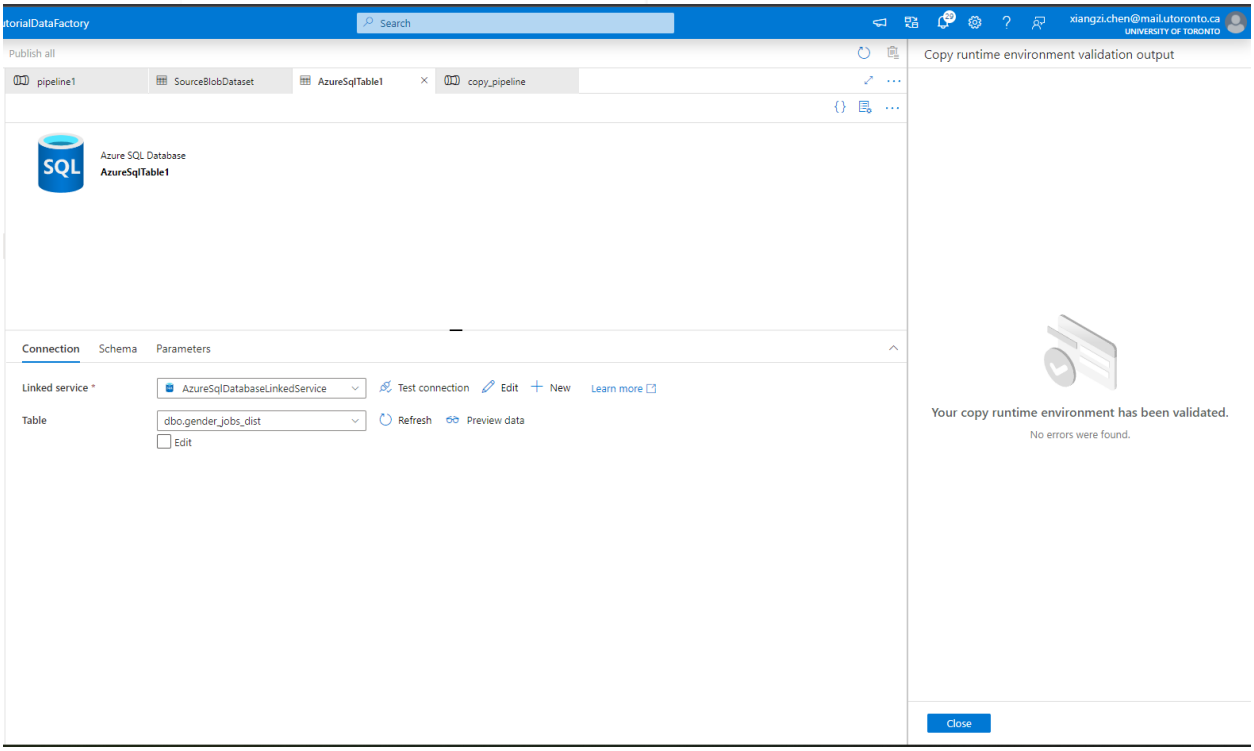
(Forgot to include the account information at first, some of the following figures are screenshotted after being successfully executed)

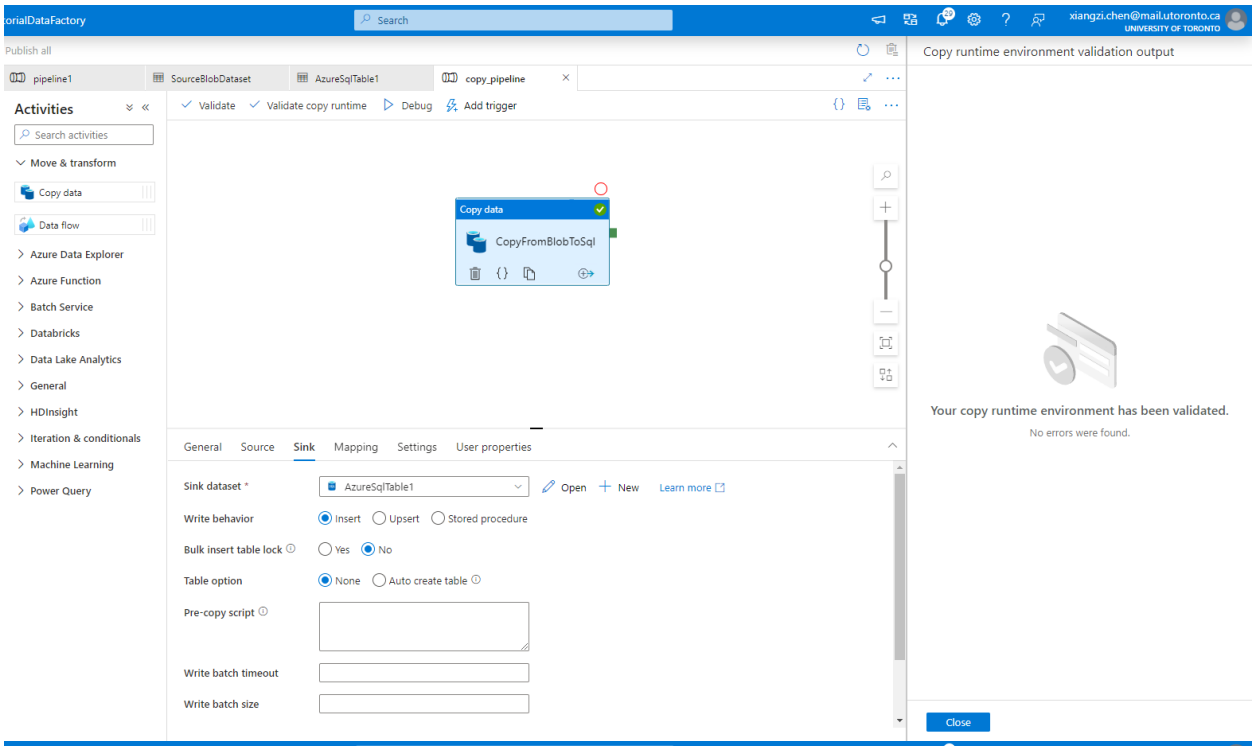


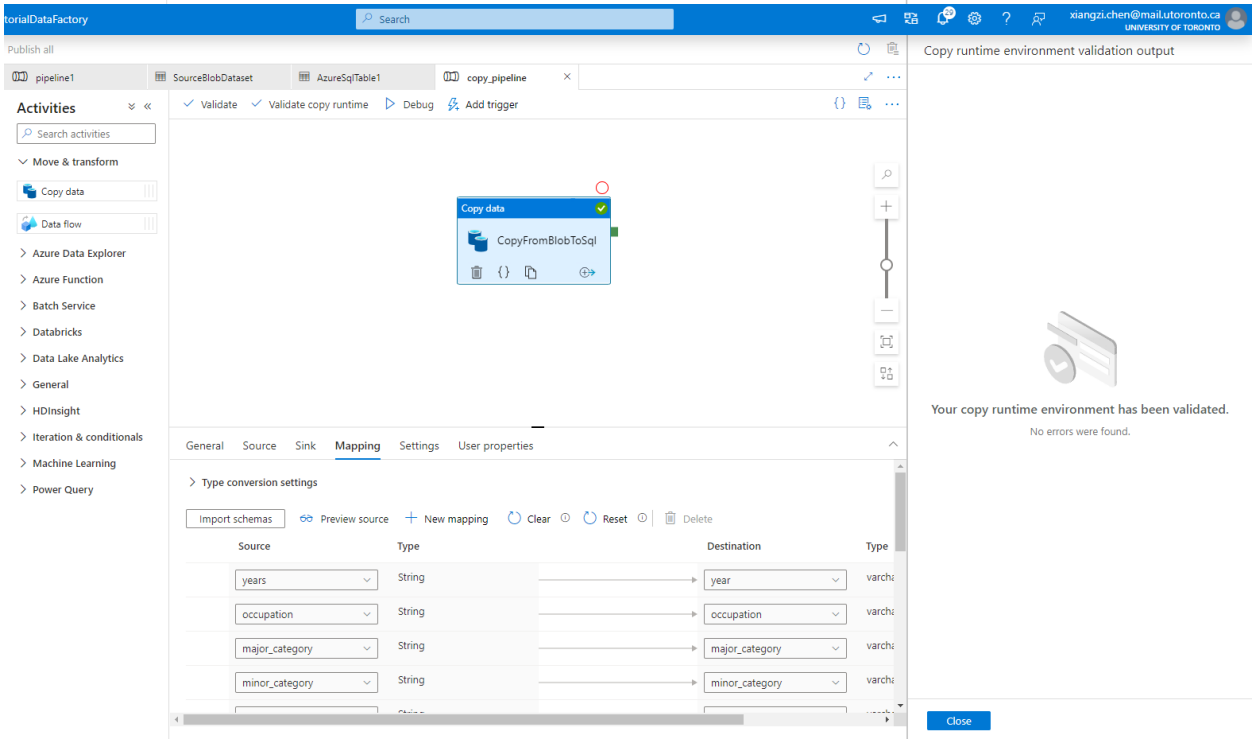

torialDataFactory
Search
xiangzi.chen@mail.utoronto.ca
UNIVERSITY OF TORONTO
Publish all
pipeline1
SourceBlobDataset
AzureSqlTable1
copy_pipeline
Activities
Search activities
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
CopyFromBlobToSql
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
SourceBlobDataset
Open
New
Preview data
Learn more
File path type
File path in dataset
Prefix
Wildcard file path
List of files
Start time (UTC)
End time (UTC)
Filter by last modified
Recursively
Enable partition discovery
Max concurrent connections
Skip line count
Copy runtime environment validation output
Your copy runtime environment has been validated.
No errors were found.
Close


torialDataFactory
Search
xiangzi.chen@mail.utoronto.ca
UNIVERSITY OF TORONTO
Publish all
pipeline1
SourceBlobDataset
AzureSqlTable1
copy_pipeline
CSV
DelimitedText
SourceBlobDataset
Connection
Schema
Parameters
Linked service *
AzureStorageLinkedService
Test connection
Edit
New
Learn more
File path *
gender-jobs-distributic
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
Copy runtime environment validation output
Your copy runtime environment has been validated.
No errors were found.
Close

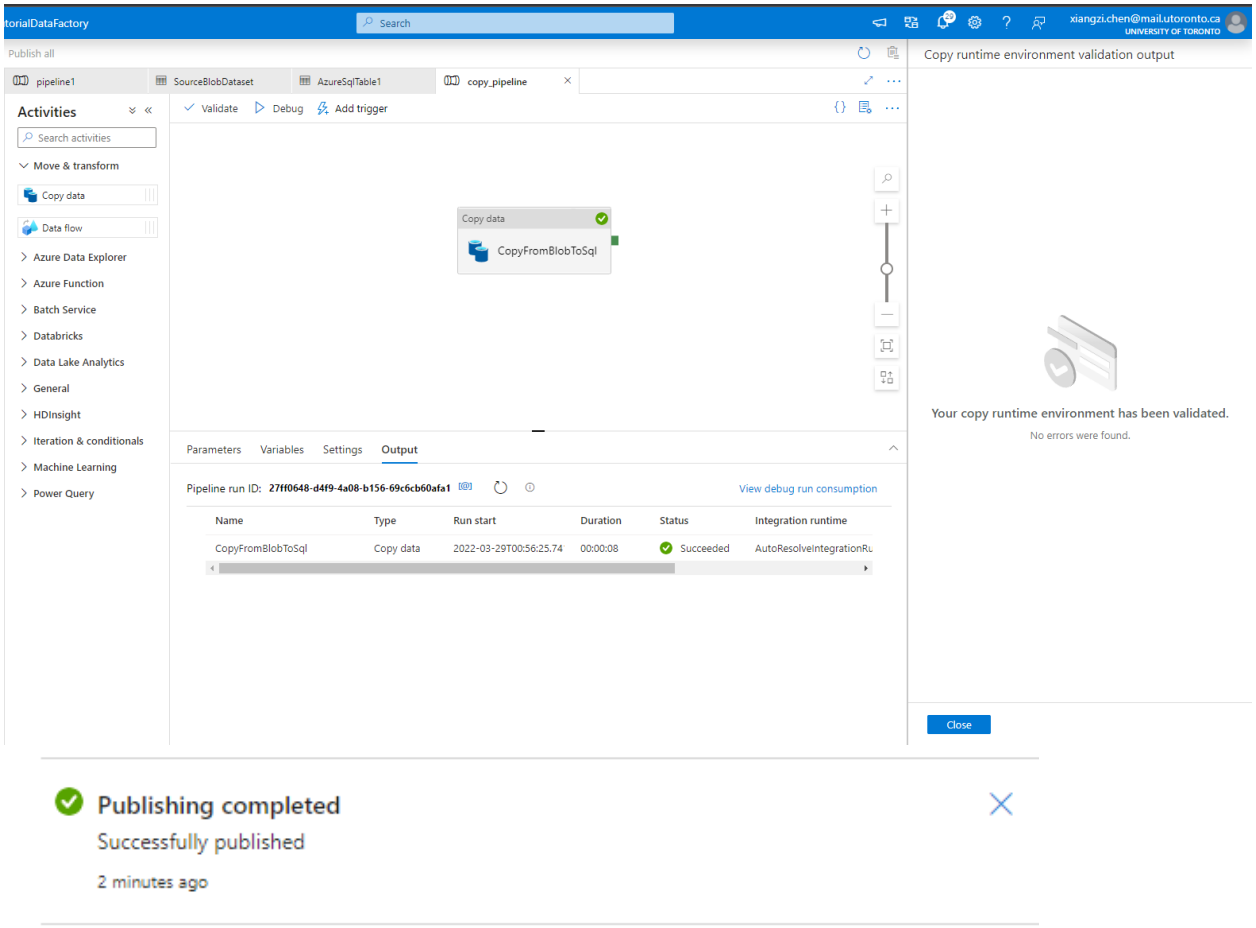

Publish all
copy_pipeline
SourceBlobDataset
Activities
Validate
Validate copy runtime
Debug
Add trigger
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Copy data
CopyFromBlobToSql
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
Select...
New
Set properties
Name
OutputSqlDataset
Linked service *
AzureSqlDatabaseLinkedService
Table name
dbo.gender_jobs_distribution2
Edit
Import schema
From connection/store
None
Advanced


torialDataFactory
Search
xiangzi.chen@mail.utoronto.ca
UNIVERSITY OF TORONTO
Publish all
pipeline1
SourceBlobDataset
AzureSqlTable1
copy_pipeline
Azure SQL Database
AzureSqlTable1
Connection
Schema
Parameters
Linked service *
AzureSqlDatabaseLinkedService
Test connection
Edit
New
Learn more
Table
dbo.gender_jobs_dist
Refresh
Preview data
Edit
Copy runtime environment validation output
Your copy runtime environment has been validated.
No errors were found.
Close


torialDataFactory
Search
xiangzi.chen@mail.utoronto.ca
UNIVERSITY OF TORONTO
Publish all
pipeline1
SourceBlobDataset
AzureSqlTable1
copy_pipeline
Activities
Search activities
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
CopyFromBlobToSql
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
AzureSqlTable1
Open
New
Learn more
Write behavior
Insert
Upsert
Stored procedure
Bulk insert table lock
Yes
No
Table option
None
Auto create table
Pre-copy script
Write batch timeout
Write batch size
Copy runtime environment validation output
Your copy runtime environment has been validated.
No errors were found.
Close


torialDataFactory
Search
xiangzi.chen@mail.utoronto.ca
UNIVERSITY OF TORONTO
Publish all
pipeline1
SourceBlobDataset
AzureSqlTable1
copy_pipeline
Activities
Search activities
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
CopyFromBlobToSql
General
Source
Sink
Mapping
Settings
User properties
Type conversion settings
Import schemas
Preview source
New mapping
Clear
Reset
Delete
Source
Type
Destination
Type
years
String
year
occupation
String
occupation
major_category
String
major_category
minor_category
String
minor_category
Copy runtime environment validation output
Your copy runtime environment has been validated.
No errors were found.
Close

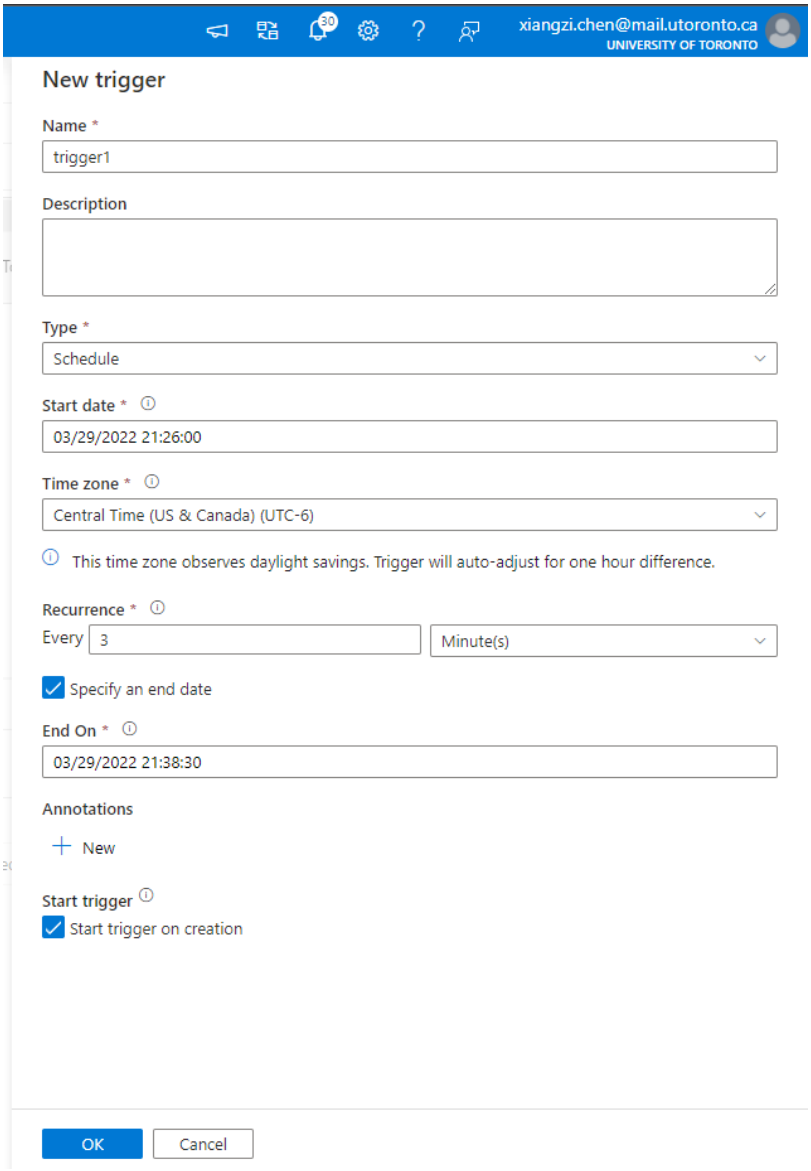
4) Publish the copy pipeline along with the input and output.

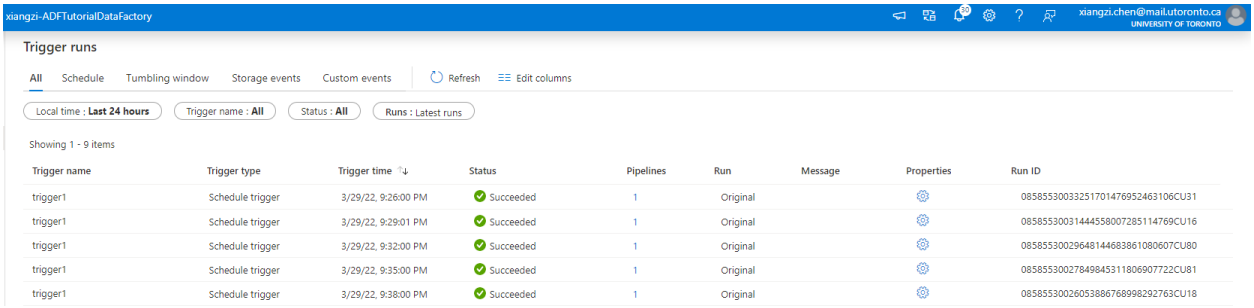


3. [Marks: 10] Explain different type of triggers available in ADF. Now create a schedule trigger and run your pipeline every 3 minutes. Show 5 successful runs.

Azure data factory support three types of triggers:

1) Schedule trigger: The triggers can execute pipelines on a certain time interval we set. We can define the start and end times for activating the triggers.
2) Tumbling window trigger: Tumbling window triggers run at a periodic interval from a specified start time.
3) Event-based trigger: Using Event-based triggers, we can schedule to execute pipelines when a certain event occurs in azure blob storage, such as deleting files in blob storage.


xiangzi.chen@mail.utoronto.ca
UNIVERSITY OF TORONTO
New trigger
Name *
trigger1
Description
Type *
Schedule
Start date *
03/29/2022 21:26:00
Time zone *
Central Time (US & Canada) (UTC-6)
This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.
Recurrence *
Every
3
Minute(s)
Specify an end date
End On *
03/29/2022 21:38:30
Annotations
New
Start trigger
Start trigger on creation
OK
Cancel


xiangzi-ADFTutorialDataFactory
xiangzi.chen@mail.utoronto.ca
UNIVERSITY OF TORONTO
Trigger runs
All
Schedule
Tumbling window
Storage events
Custom events
Refresh
Edit columns
Local time : Last 24 hours
Trigger name : All
Status : All
Runs : Latest runs
Showing 1 - 9 items
Trigger name
Trigger type
Trigger time
Status
Pipelines
Run
Message
Properties
Run ID
trigger1 Schedule trigger 3/29/22, 9:26:00 PM Succeeded 1 Original 085855300332517014769524631 06CU31
trigger1 Schedule trigger 3/29/22, 9:29:01 PM Succeeded 1 Original 085855300314445580072851147 69CU16
trigger1 Schedule trigger 3/29/22, 9:32:00 PM Succeeded 1 Original 085855300296481446838611080 607CU80
trigger1 Schedule trigger 3/29/22, 9:35:00 PM Succeeded 1 Original 085855300278498453118069077 22CU81
trigger1 Schedule trigger 3/29/22, 9:38:00 PM Succeeded 1 Original 085855300260538867689982927 63CU18

4. [Marks: 20] 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 in one side, it needs to be replicated to the other one]

We can accomplish this process using data factory in Azure.
Firstly, we need to build two copy pipelines since we want to do this in a bi-directional way. One is to copy data from West Europe, the source, to Canada Central, the sink. The other one is to copy data from Canada Central, which is the source, to West Europe, the sink.

Then, we have to create an Azure Data Lake Storage Gen2 linked service in the Azure portal UI. This linked service will be used on both pipelines. To create the linked service, we need to browse the manage tab in the Azure Data Factory and create a new linked service. By searching Azure Data Lake Storage Gen2 and selecting Azure Data Lake Storage Gen2 connector, configuring all the service details, testing the connection, creating the new linked service, the new linked service has been conducted. For the authentication, we need to use 'account key', Service Principal authentication, System-assigned managed identity authentication and User-assigned managed identity authentication. Then, we need to set the copy activity properties. For the source, we need to copy from ADLS Gen2, such as from a specific path of the folder to be transferred, set a wildcard filter against the folder path, etc. For the sink, we need to set the properties supported for ADLS Gen2 via storeSetting. Then, we can debug and publish the copy pipelines. Lastly, we need to set the event-based trigger for both pipelines. The pipelines will be executed whenever there are new blobs created.

Part B

1. [Marks:5] In the *gender_jobs_data* table - Filter all the OCCUPATIONS in MAJOR_CATEGORY of Computer, Engineering, and Science for the YEAR 2013

Run | Cancel query | Save query | Export data as | Show only Editor

```
select distinct occupation
from [dbo].[gender_jobs_distribution1]
where year = 2013
and major_category = 'Computer, Engineering, and Science'
```

Results | Messages

Search to filter items...

occupation
Actuaries
Aerospace engineers
Agricultural and food science technicians
Agricultural and food scientists
Agricultural engineers
Architects, except naval
Astronomers and physicists

Atmospheric and space scientists
Biological scientists
Biological technicians
Biomedical engineers
Chemical engineers
Chemical technicians
Chemists and materials scientists
Civil engineers
Computer , all other
Computer and information research scientists
Computer hardware engineers
Computer network architects
Computer programmers
Computer support specialists
Computer systems analysts
Conservation scientists and foresters
Database administrators
Drafters
Economists
Electrical and electronics engineers
Engineering technicians, except drafters
Engineers, all other
Environmental engineers
Environmental scientists and geoscientists
Geological and petroleum technicians
Industrial engineers, including health and safety
Information security analysts
Life scientists, all other
Marine engineers and naval architects
Materials engineers
Mathematicians
Mechanical engineers
Medical scientists
Mining and geological engineers, including mining safety engineers
Miscellaneous life, physical, and social science technicians
Miscellaneous mathematical science
Miscellaneous social scientists and related workers, including sociologists
Network and computer systems administrators
Nuclear engineers
Nuclear technicians
Operations research analysts
Petroleum engineers
Physical scientists, all other
Psychologists
Social science research assistants
Software developers, applications and systems software
Statisticians
Survey researchers
Surveying and mapping technicians
Surveyors, cartographers, and photogrammetrists
Urban and regional planners
Web developers

There are 59 occupations.

2. [Marks:5] In the *gender_jobs_data* table - How many OCCUPATIONS exist in the MINOR_CATEGORY of Business and Financial Operations overall?

```
select count(distinct occupation)
from [dbo].[gender_jobs_distribution1]
where minor_category = 'Business and Financial Operations'
```

Results Messages

Search to filter items...

28

3. [Marks:5] In the *gender_jobs_data* table - Get all relevant information for bus drivers across all years

```
select *
from [dbo].[gender_jobs_distribution1]
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	174830	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

4. [Marks:5] 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?

```
select year, sum(cast(workers_female as int)) as 'Total number of female workers'
from [dbo].[gender_jobs_distribution1]
where major_category='Management, Business, and Financial'
group by year
```

Results Messages

Search to filter items...

year	Total number of female workers
2013	7748347
2014	8061480
2015	8381812
2016	8617853

5. [Marks: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?

```
select sum(cast(total_earnings_male as int)) as 'Total earnings male in 2015'
from [dbo].[gender_jobs_distribution1]
where major_category = 'Service' and year = 2015
```

Results Messages

Search to filter items...

Total earnings male in 2015
2502426

6. [Marks:5] In the *gender_jobs_data* table - How many female workers were in management roles in the year 2015?

```
select sum(cast(workers_female as int)) as 'Total female workers in management in 2015'
from [dbo].[gender_jobs_distribution1]
where minor_category = 'management' and year = 2015
```

Results Messages

Search to filter items...

Total female workers in management in 2015
5166720

7. [Marks:5] In the *gender_jobs_data* table - Compare the TOTAL_EARNINGS_MALE and TOTAL_EARNINGS_FEMALE earnings irrespective of occupation by each year

```
select year, sum(cast(total_earnings_male as int)) as 'male total earnings', sum(cast(total_earnings_female as int)) as 'female total earnings'
from [dbo].[gender_jobs_distribution1]
group by year
```

Results Messages

Search to filter items...

year	female total earnings	female total earnings
2013	22054404	27050782
2014	22491208	27470450
2015	22768521	27754851
2016	23075602	28463638

8. [Marks:5] In the *gender_jobs_data* table - How much money (TOTAL_EARNINGS_FEMALE) did female workers make as engineers in 2016?

```sql
select year, sum(cast(total_earnings_female as int)) as 'Total female engineer earnings in 2016'
from [dbo].[gender_jobs_distribution1]
where total_earnings_female != 'NA' and (occupation like '%%engineers' or occupation like '%%engineer' or occupation like '%%engineers%%'
or occupation like '%%engineer%%') and year = 2016
GROUP BY year

```

Results Messages

Search to filter items...

year	Total female engineer earnings in 2016
2016	1844254

9. [Marks:10] What is the total number of full time and part time female workers versus male workers year over year?

```sql
select year,
round(sum((cast(workers_female as float)) * cast(full_time_female as float)), 0) as 'Full time female workers',
round(sum((cast(workers_female as float)) * cast(part_time_female as float)), 0) as 'Part time female workers',
round(sum((cast(workers_male as float)) * cast(full_time_male as float)), 0) as 'Full time male workers',
round(sum((cast(workers_male as float)) * cast(part_time_male as float)), 0) as 'Part time male workers'
from [dbo].[gender_jobs_distribution1]
group by year
order by year DESC
```

Results Messages

Search to filter items...

year	Full time female workers	Part time female workers	Full time male workers	Part time male workers
2016	3427412749	1136385851	5252679259	743529941
2015	3341442786	1125726714	5172057300	732117700
2014	3231348043	1123568457	5033027195	732181505
2013	3156814322	1109150978	4882748758	736064542