

1628 Assignment 4

Assignment on Azure Cloud Platform

Due by Jul 23, 2023

Ziruo Song

Part A:

For Part A: Submit all screenshots showing deployed resources in your Azure portal, Azure Blob Storage, Azure Data Factory, ADLS Gen 2 and Azure SQL DB including your account information at the top right corner of the webpage. Include the successful pipeline runs screenshots with triggers.

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.

ANS:

The resource group I created was named ‘1628hw4’, and the 3 deployed resources were grouped as shown below.

The screenshot shows the Azure portal's 'All resources' view. At the top, there are filter options: 'Subscription equals all', 'Resource group equals all' (set to '1628hw4'), 'Type equals all' (set to 'Storage account'), and 'Location equals all'. Below the filters, there are two buttons: 'Recommendations' and 'Unsecure resources'. The main table lists three resources:

Name	Type	Resource group	Location	Subscription
bshw4	Storage account	1628hw4	East Asia	Azure for Students
df-hw4	Data factory (V2)	1628hw4	East Asia	Azure for Students
sqldb	SQL database	1628hw4	East Asia	Azure for Students

2. [Marks: 15] 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). See this <https://docs.microsoft.com/en-us/azure/data-factory/tutorial-copy-data-portal> for reference.)

ANS:

Code for Sink table:

(*total_earnings_male*, *total_earnings_male*, and *total_earnings_male* contain NAs. Here I choose the method 2 according to the methods on Piazza.)

```
Query 1 ×   Query 2 ×   Query 3 ×
<  ▶ Run   Cancel query  ⏪ Save query  ⏪ Export data as  ⏪ Sho
1  CREATE TABLE info_table(
2    year int,
3    occupation varchar(255),
4    major_category varchar(255),
5    minor_category varchar(255),
6    total_workers int,
7    workers_male int,
8    workers_female int,
9    percent_female float,
10   total_earnings int,
11   total_earnings_male varchar(255),
12   total_earnings_female varchar(255),
13   wage_percent_of_male varchar(255),
14   total_full_time float,
15   total_part_time float,
16   full_time_female float,
17   part_time_female float,
18   full_time_male float,
19   part_time_male float
20 )
21 GO
22
23
```

Preview of output data:

Preview data

Linked service: AzureSqlDatabase1

Object: dbo.info_table

#	year	occupation	major_category	minor_category	total_workers	workers_male	workers_female	percentage
1	2013	Chief executives	Management, Business, and Financial	Management	1024259	782400	241859	2
2	2013	General and operations managers	Management, Business, and Financial	Management	977284	681627	295657	3
3	2013	Legislators	Management, Business, and Financial	Management	14815	8375	6440	4
4	2013	Advertising and promotions managers	Management, Business, and Financial	Management	43015	17775	25240	5
5	2013	Marketing and sales managers	Management, Business, and Financial	Management	754514	440078	314436	4
6	2013	Public relations and fund raisers	Management, Business, and Financial	Management	1000000	600000	400000	5

Success of pipeline:

Parameters Variables Settings Output ^

Pipeline run ID: 98cb167b-88f3-4b05-91f6-f9b11f313bf8 [@](#) [↻](#) [ⓘ](#) [View debug run](#)

All status ▾ [Export to CSV](#) ▾

Showing 1 - 1 of 1 items

Activity name ↑↓	Status ↑↓	Activity type ↑↓	Run start ↑↓	⋮
CopyFromBlobToSql	✓ Succeeded	Copy data	7/24/2023, 4:22:48 AM	⋮

Details		Refresh									
Learn more on copy performance details from here.											
Activity run id: ee07fb2e-b102-4312-9f9a-fed7a938a7ec											
	Azure Blob Storage Region: East Asia	Succeeded									
	Azure SQL Database Region: East Asia										
Data read: ⓘ	388.834 KB	Data written: ⓘ	628.39 KB								
Files read: ⓘ	1	Rows written: ⓘ	2,088								
Rows read:	2,088	Peak connections: ⓘ	2								
Peak connections: ⓘ	1										
Copy duration	00:00:07										
Throughput: ⓘ	194.417 KB/s										
✓ Azure Blob Storage → Azure SQL Database											
Start time	7/24/2023, 4:22:49 AM										
Used DLU's ⓘ	4										
Used parallel copies ⓘ	1										
Duration	00:00:07										
<table border="1"> <thead> <tr> <th>Details</th> <th>Working duration</th> <th>Total duration</th> </tr> </thead> <tbody> <tr> <td>Queue ⓘ</td> <td>00:00:00</td> <td>00:00:04</td> </tr> <tr> <td>Transfer ⓘ</td> <td> Listing source ⓘ Reading from source ⓘ Writing to sink ⓘ </td> <td>00:00:02</td> </tr> </tbody> </table>			Details	Working duration	Total duration	Queue ⓘ	00:00:00	00:00:04	Transfer ⓘ	Listing source ⓘ Reading from source ⓘ Writing to sink ⓘ	00:00:02
Details	Working duration	Total duration									
Queue ⓘ	00:00:00	00:00:04									
Transfer ⓘ	Listing source ⓘ Reading from source ⓘ Writing to sink ⓘ	00:00:02									
Data consistency verification ⓘ	Not verified										
	How satisfied or dissatisfied are you with the performance of this copy activity?										

3. [Marks: 10] 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.

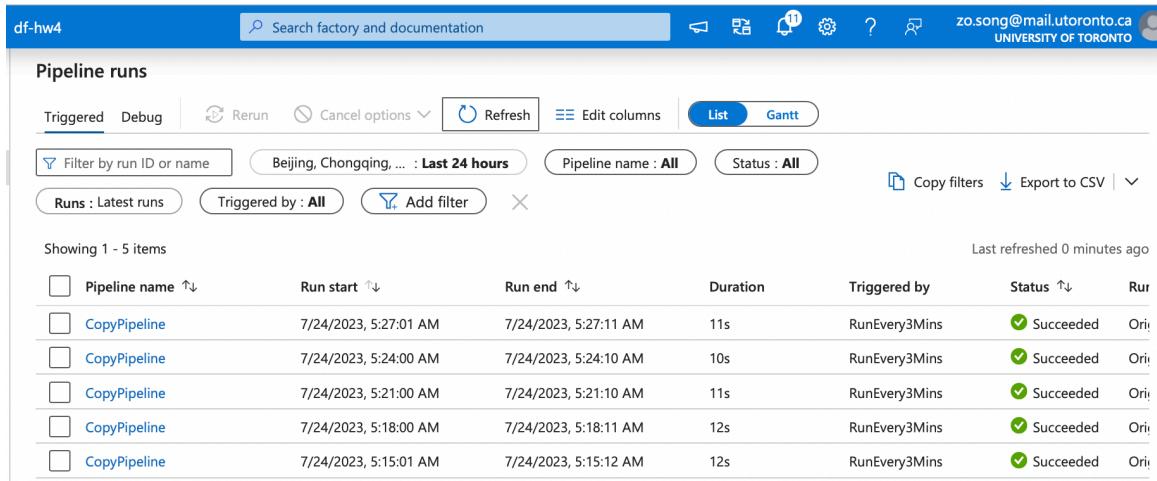
ANS:

Explanation: Based on the official document of Azure and other resource online, there are three types of trigger offered by Azure Data Factory (ADF) right now:

- Schedule trigger, it runs pipelines on a predefined schedule with flexible time intervals, like daily, or every 5 mins.
 - Tumbling Window Trigger: Executes Data Pipelines at specific time slices or intervals, which mean this is quite useful for historical data scenarios. Compared with schedule trigger above, Tumbling Window Triggers support Backfill scenarios, Retry capability, concurrency, and but only one pipeline can be triggered.

- Event-based Trigger: It responds to blob-related events, triggering pipelines based on changes in data sources.

5 successful runs:



The screenshot shows the 'Pipeline runs' page in the Azure Data Factory interface. The top navigation bar includes a search bar, user information (zo.song@mail.utoronto.ca, UNIVERSITY OF TORONTO), and various icons. The main header is 'Pipeline runs' with tabs for 'Triggered' (selected), 'Debug', 'Rerun', 'Cancel options', 'Refresh', 'Edit columns', 'List' (selected), and 'Gantt'. Below the header are filters: 'Filter by run ID or name' (Beijing, Chongqing, ... : Last 24 hours), 'Pipeline name : All', 'Status : All', 'Runs : Latest runs', 'Triggered by : All', 'Add filter', and 'Copy filters', 'Export to CSV'. The table below shows 5 items, each with a checkbox, pipeline name, run start, run end, duration, triggered by, status, and run ID. All runs are succeeded.

<input type="checkbox"/> Pipeline name ↑↓	Run start ↑↓	Run end ↑↓	Duration	Triggered by	Status ↑↓	Ru
<input type="checkbox"/> CopyPipeline	7/24/2023, 5:27:01 AM	7/24/2023, 5:27:11 AM	11s	RunEvery3Mins	✓ Succeeded	Ori
<input type="checkbox"/> CopyPipeline	7/24/2023, 5:24:00 AM	7/24/2023, 5:24:10 AM	10s	RunEvery3Mins	✓ Succeeded	Ori
<input type="checkbox"/> CopyPipeline	7/24/2023, 5:21:00 AM	7/24/2023, 5:21:10 AM	11s	RunEvery3Mins	✓ Succeeded	Ori
<input type="checkbox"/> CopyPipeline	7/24/2023, 5:18:00 AM	7/24/2023, 5:18:11 AM	12s	RunEvery3Mins	✓ Succeeded	Ori
<input type="checkbox"/> CopyPipeline	7/24/2023, 5:15:01 AM	7/24/2023, 5:15:12 AM	12s	RunEvery3Mins	✓ Succeeded	Ori

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?

ANS:

By hint, to establish bi-directional replication between ADLS Gen 2 storage accounts located in Canada Central and West Europe, we can utilize Azure Data Factory with Event Triggers. Firstly, create ADLS Gen 2 storage accounts in both Canada Central and West Europe regions, and then set up Linked Services and defining datasets, like what we did for the above sections. However, we can choose to build two separate pipelines, instead of one, to do it bi-directionally. Also configure Copy Activities to effectively transfer the data between the two storage accounts, and we can create Event Triggers to trigger the respective pipelines based on changes in data sources.

PART B:

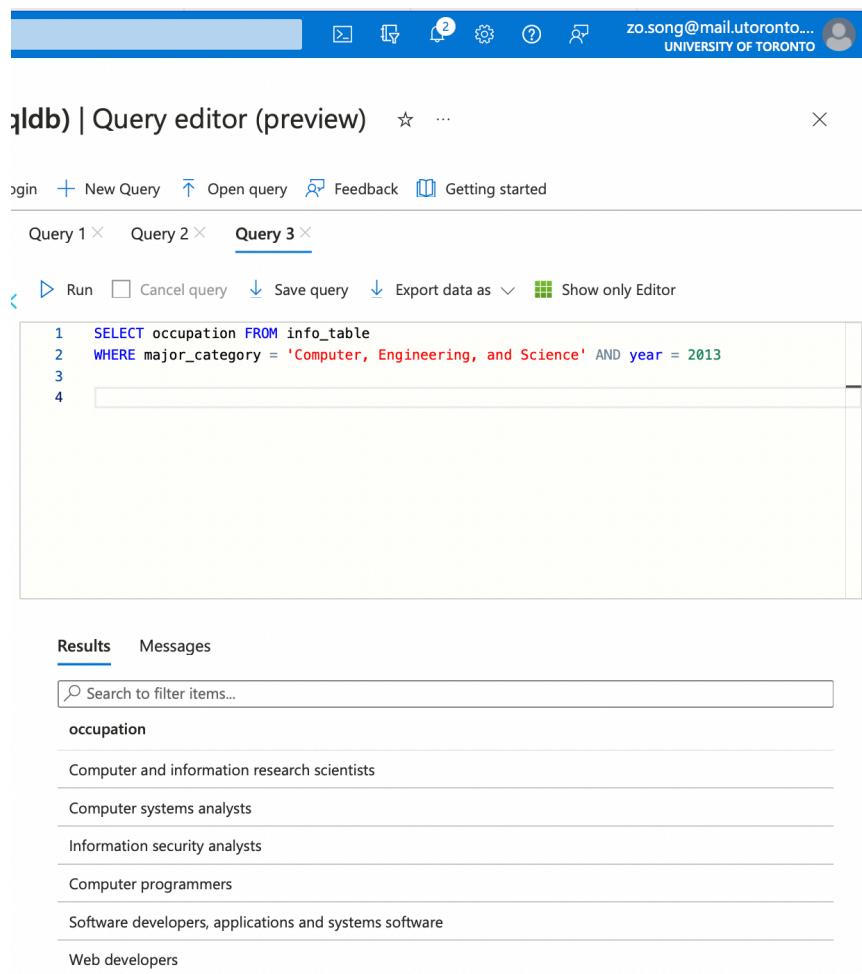
For Part B: Please submit all screenshots showing deployed resources in your Azure portal, Azure SQL DB and Query Editor screenshots where you run your code with output.

In this part, you will use Query Editor in Azure SQL DB and use the *gender_jobs_data.csv* table to perform the below queries.

Implementation

You need to use Azure SQL Database for this part.

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



The screenshot shows the Azure SQL Database Query Editor interface. The top navigation bar includes 'Login', 'New Query', 'Open query', 'Feedback', and 'Getting started'. The user is signed in as 'zo.song@mail.utoronto... UNIVERSITY OF TORONTO'. Below the navigation is a toolbar with 'Run' (highlighted in blue), 'Cancel query', 'Save query', 'Export data as', and 'Show only Editor'. There are three tabs: 'Query 1', 'Query 2', and 'Query 3' (selected). The query editor window contains the following SQL code:

```
1  SELECT occupation FROM info_table
2  WHERE major_category = 'Computer, Engineering, and Science' AND year = 2013
3
4
```

The results pane at the bottom shows the output of the query:

occupation
Computer and information research scientists
Computer systems analysts
Information security analysts
Computer programmers
Software developers, applications and systems software
Web developers

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

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface for a query editor. The title bar reads ':sqldb | Query editor (preview)'. The top navigation bar includes links for Login, New Query, Open query, Feedback, and Getting started. Below the navigation is a tab bar with three tabs: Query 1, Query 2, and Query 3, where Query 3 is currently selected. A toolbar below the tabs includes Run, Cancel query, Save query, Export data as, and Show only Editor. The main query window contains the following T-SQL code:

```
1  SELECT COUNT(DISTINCT occupation) FROM info_table
2  WHERE minor_category = 'Business and Financial Operations'
3
4
5
6
```

Below the query window, there are two tabs: Results and Messages, with Results being the active tab. A search bar labeled 'Search to filter items...' is also present.

28

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

The screenshot shows a query editor interface with the following details:

- Header:** and docs (G+/-) with a user icon, zoom, refresh, gear, help, and sign-in options.
- Title:** sqldb | Query editor (preview)
- Toolbar:** Login, New Query, Open query, Feedback, Getting started.
- Query Tab:** Query 1, Query 2, **Query 3** (selected).
- Query Editor:** Contains the following SQL code:

```
1 SELECT * FROM info_table
2 WHERE occupation = 'Bus drivers'
3
4
5
6
```
- Run Options:** Run, Cancel query, Save query, Export data as, Show only Editor.
- Results Tab:** Results (selected), Messages.
- Search:** Search to filter items...
- Table Results:** A table showing data for bus drivers across years. The columns are: year, occupation, major_category, minor_category, total_workers, and workers_fem. The data is as follows:

year	occupation	major_category	minor_category	total_workers	workers_fem
2013	Bus drivers	Production, Transportation, an...	Transportation	275991	17
2014	Bus drivers	Production, Transportation, an...	Transportation	267775	16
2015	Bus drivers	Production, Transportation, an...	Transportation	288778	17
2016	Bus drivers	Production, Transportation, an...	Transportation	280228	17
2013	Bus drivers	Production, Transportation, an...	Transportation	275991	17
2014	Bus drivers	Production, Transportation, an...	Transportation	267775	16

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.



View) ⭐ ... ×

Feedback Getting started

Query 1 × Query 2 ×

Run Cancel query Save query Export data as Show only Editor

```
1 SELECT year, SUM(workers_female) FROM infor_table
2 WHERE major_category = 'Management, Business, and Financial'
3 GROUP BY year
4 ORDER BY year
5
6
```

Results Messages

Search to filter items...

year

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?

A screenshot of a Microsoft SQL Server Management Studio (SSMS) window. The title bar shows the user's email address: zo.song@mail.utoronto.... UNIVERSITY OF TORONTO. The main interface has a toolbar with icons for file, edit, and help. Below the toolbar, there are tabs for 'Getting started' and 'Query 2'. The 'Query 2' tab is selected. Underneath the tabs are buttons for 'Run', 'Cancel query', 'Save query', 'Export data as', and 'Show only Editor'. The query editor area contains the following SQL code:

```
1  SELECT SUM(COALESCE(TRY_CAST(total_earnings_male as int), 0))
2  FROM infor_table
3  WHERE major_category = 'Service' AND year = 2015
4
5
```

The results pane below the editor shows a single row of data: 2502426. The 'Results' tab is selected.

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

A screenshot of a Microsoft SQL Server Management Studio (SSMS) window. The title bar shows the user's email address: zo.song@mail.utoronto.... UNIVERSITY OF TORONTO. The main interface has a toolbar with icons for file, edit, and help. Below the toolbar, there are tabs for 'Feedback', 'Getting started', 'Query 1', 'Query 2', 'Query 3', and 'Query 4'. The 'Query 2' tab is selected. Underneath the tabs are buttons for 'Run', 'Cancel query', 'Save query', 'Export data as', and 'Show only Editor'. The query editor area contains the following SQL code:

```
1  SELECT SUM(workers_female) FROM infor_table
2  WHERE minor_category = 'Management' AND year = 2015
3
4
5
```

The results pane below the editor shows a single row of data: 5166720. The 'Results' tab is selected.

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

The screenshot shows a database interface with a blue header bar. On the right side of the header, there are icons for a user profile (zo.song@mail.utoronto.... UNIVERSITY OF TORONTO) and a search function. Below the header, there are tabs for 'Query 1' and 'Query 2', with 'Query 2' currently selected. Below the tabs are buttons for 'Run', 'Cancel query', 'Save query', 'Export data as', and 'Show only Editor'. The main area contains a code editor with the following SQL query:

```
1  SELECT year,
2    SUM(COALESCE (TRY_CAST (total_earnings_male as int), 0)) as sum_male,
3    SUM(COALESCE (TRY_CAST (total_earnings_female as int), 0)) as sum_female
4  FROM infor_table
5  GROUP BY year
6  ORDER BY year
7
8
```

Below the code editor, there are tabs for 'Results' and 'Messages', with 'Results' currently selected. A search bar is present above the results table. The results table has three columns: 'year', 'sum_male', and 'sum_female'. The data is as follows:

year	sum_male	sum_female
2013	27050782	22054404
2014	27470450	22491208
2015	27754851	22768521
2016	28463638	23075602

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

The screenshot shows a database query editor interface. At the top, there is a navigation bar with icons for file, settings, and user information (zo.song@mail.utoronto.... UNIVERSITY OF TORONTO). Below the navigation bar, the title "view)" is displayed along with a star and three dots. A feedback link and a getting started link are also present.

The main area contains four tabs: "Query 1", "Query 2", "Query 3", and "Query 4". The "Query 4" tab is currently selected. Below the tabs, there are buttons for "Run" (with a play icon), "Cancel query" (with a cancel icon), "Save query" (with a save icon), "Export data as" (with a download icon), and "Show only Editor" (with a grid icon).

The query code in the editor is:

```
1 SELECT SUM(COALESCE(TRY_CAST(total_earnings_female as int), 0)) FROM info_tables
2 WHERE occupation LIKE '%engineer%' AND year = 2016
3
4
```

Below the editor, there are two tabs: "Results" and "Messages". The "Results" tab is selected. A search bar is available to filter items. The results section displays the output of the query:

1844254

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

- calculate by multiply each rate with the number.



W) ⭐ ... X

Feedback Getting started

Query 1 X Query 2 X **Query 3 X** Query 4 X

Run Cancel query Show only Editor

```
1  SELECT year,
2    sum(full_time_female * workers_female*0.01) as full_time_female_sum,
3    sum(full_time_male * workers_male*0.01) as full_time_male_sum,
4    sum(part_time_female * workers_female*0.01) as part_time_female_sum,
5    sum(part_time_male * workers_male*0.01) as part_time_male_sum
6  FROM info_tables
7  GROUP BY year
8  ORDER BY year
```

Results Messages

Search to filter items...

year	full_time_female_sum	full_time_male_sum	part_time_female_s...	part_time_male_s...
2013	31568143.22	48827487.577	11091509.78	7360645.423
2014	32313480.43	50330271.951	11235684.57	7321815.049
2015	33414427.86	51720573	11257267.14	7321177
2016	34274127.486	52526792.592	11363858.514	7435299.408

Query succeeded | 0s