Scotiabank Technical Test

For this technical test, I decided to use the Azure cloud architecture. The cloud architecture consisted of the following resources:

Graphical user interface, text, application, email

Description automatically generated

The Datafactory was used to design the data ingestion pipeline.

The BlobStorage was used as the storage location for raw and processed data.

The Key Vault was used to store sensitive information such as credentials.

The Databricks was used to process the data and carry out the different exercises asked in the test.

Diagram

Description automatically generatedHere is an overview of a simple pipeline created to ingest and process the raw data:

The following Linked Services were created to accomplish the pipeline ingestion:

Graphical user interface, text, application

Description automatically generated

1)

The first step of the test was to ingest two raw data sources from the Stat Can website. To do this step, I used the Copy data feature in the Datafactory to directly connect to the link and download the datasets.

Graphical user interface, text, application, email

Description automatically generated

Both of the datasets were ingested into the BlobStorage as it can be seen here:

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generated

2)

Once the datasets were ingested into the blob storage, it was time to process the data and calculate different metrics. For this I used Databricks which takes advantage of Spark to process the data. All of the metrics calculations can be found inside the notebook.

<https://github.com/ManpreetNanreh/scotiabanktechnicaltest/blob/main/dataprocessing.ipynb>

3)

This task is also done inside the databricks notebook.

4)

I did this task in Tableau public and hosted it on their website.

<https://public.tableau.com/views/StatCanTest/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link>

5)

This task is also done inside the databricks notebook.

Challenges:

It was not fully clear how some of the fields in the dataset were calculated, so assumptions were. I have listed them in the databricks notebook.

Another issue I had was properly deploying the tableau dashboard on the tableau public host and share the link.

The GitHub repository for this project can be found here:

<https://github.com/ManpreetNanreh/scotiabanktechnicaltest>

I would be delighted to go over the entire pipeline with you in person or over a conference call. I think a demo would truly showcase all the work went into this.

Please let me know if you have any questions and I would be happy to answer them.