**ScrumConnect – Data Engineer – Bijen Shah – Test Solution**

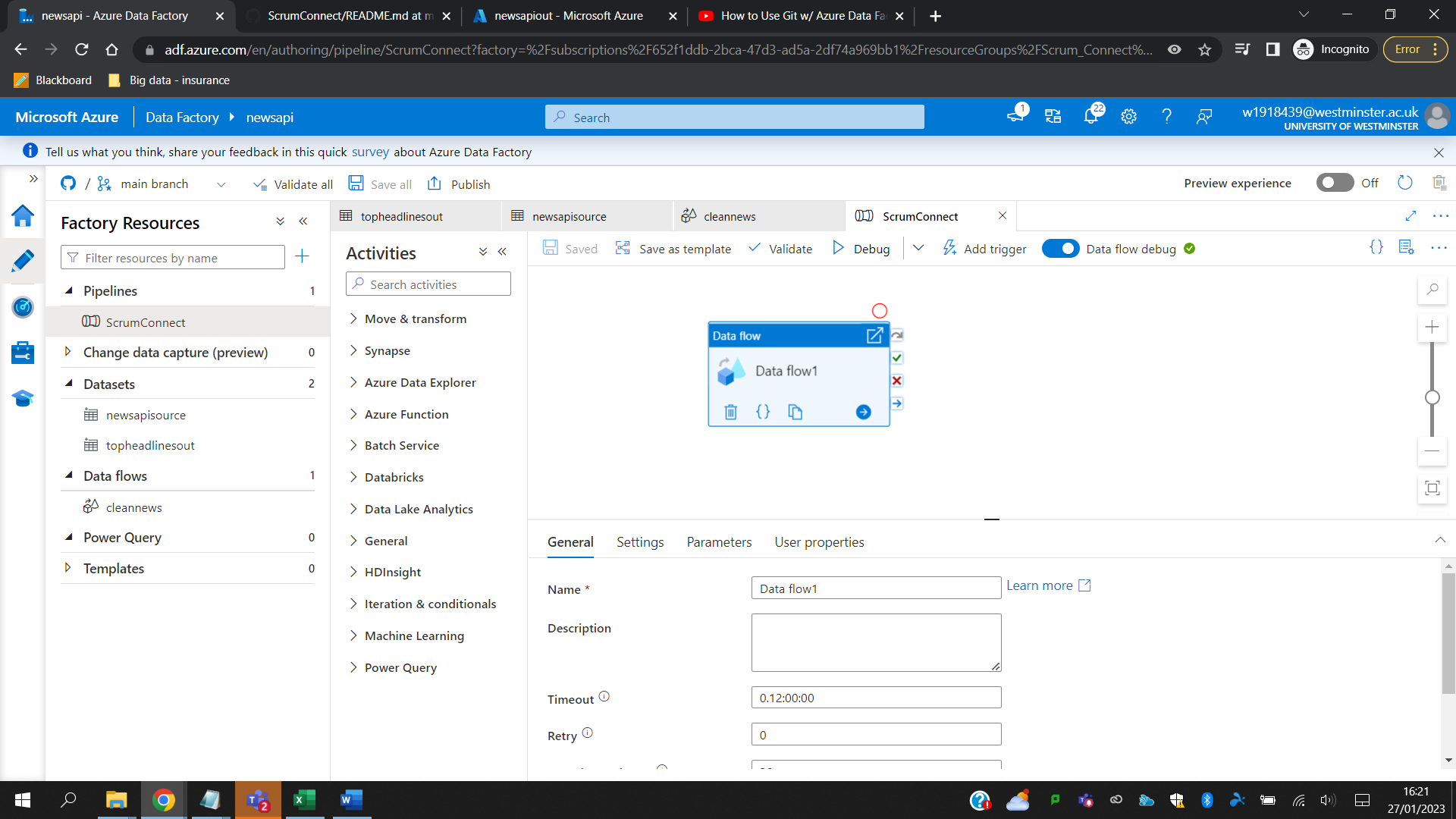
Solution:

1. The api.txt file stores the api key from newsapi.
2. The newsapi.py file extracts the data from Newsapi and stores the data into a CSV format file.
3. This file is then stored into Azure Storage Container.
4. A pipeline is built under Azure Data Factory to pull the CSV file clean the data and save it into a different CSV file.

Limitations: There were some limitations to my Azure account. To run the python code as a Batch inside ADF pipeline, hence the code was run for pilot testing on Google Colab.

Doubts/Improvements: The data is cleaned and some columns are split for better understanding

The adf pipeline contains one dataflow.



Screenshots of the dataflow.

A screenshot of a computer

Description automatically generated

Containers to store source data and output of dataflow

A screenshot of a computer

Description automatically generated

File stored in the output directory after running dataflow.

A screenshot of a computer

Description automatically generated