Practice Case Big Data

Ari Sulistiyo Prabowo

Comparing the Runtime Process using Spark and Pandas

TIMELINE

Spark Script (VM)

The whole process until load to Google Cloud Storage

Pandas Script (Local Computer)

The whole process until load to Google Cloud Storage

Pandas Script (VM)

The whole process until load to Google Cloud Storage

Spark Script in VM

https://github.com/densaiko/Apache_Spark_Simulation/blob/main/spark_data.py

There are **four steps** to do:

1. Read the data from S3
2. Data cleansing using query spark
3. Transform spark dataframe to pandas dataframe (it should be done to load the data into GCS)
4. Upload the data to Google Cloud Storage (GCS)

Runtime process Spark script in VM	
Time for reading data	134 s
Time for querying/cleansing data	0.53 s
Time for transforming data to pandas df	43.40 s
Time for uploading data to GCS	22.65 s
Overall time	200.67 s

```
Time for reading data:  134.08148597701802
Time for querying data:  0.5381860259803943
Time for transforming data to pandas dataframe:  43.40042050299235
Time to upload data:  22.659199767018436
Overall Time:  200.67984974200954
21/03/15 01:55:34 INFO SparkContext: Invoking stop() from shutdown hook
21/03/15 01:55:34 INFO SparkUI: Stopped Spark web UI at http://master.asia-southeast1-b.c.datafellowship-307406.internal:4041
21/03/15 01:55:34 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

Pandas Script in VM

https://github.com/densaiko/Apache_Spark_Simulation/blob/main/pandas_data.py

There are **three steps** to do:

1. Read the data from S3
2. Data cleansing using query spark
3. Upload the data to Google Cloud Storage (GCS)

Runtime process Pandas script in VM	
Time for reading data	62.14 s
Time for querying/cleansing data	0.22 s
Time for transforming data to pandas df	None
Time for uploading data to GCS	21.96 s
Overall time	84.34 s

Explanation between Spark and Pandas in VM

Runtime process Spark script in VM	
Time for reading data	134 s
Time for querying/cleansing data	0.53 s
Time for transforming data to pandas df	43.40 s
Time for uploading data to GCS	22.65 s
Overall time	200.67 s

Runtime process Pandas script in VM	
Time for reading data	62.14 s
Time for querying/cleansing data	0.22 s
Time for transforming data to pandas df	None
Time for uploading data to GCS	21.96 s
Overall time	84.34 s

Spark has four steps in its script, on the other hand, Pandas has three steps because **pandas does not need to transform** the data into dataframe.

In Spark script, the VM run **several process to call other scripts/process.** Therefore, it takes longer time than pandas.

Virtual machine runs script very well and faster than local computer. I show the different result of pandas script in local computer.

Pandas Script in Local Computer

https://github.com/densaiko/Apache_Spark_Simulation/blob/main/pandas_data.py

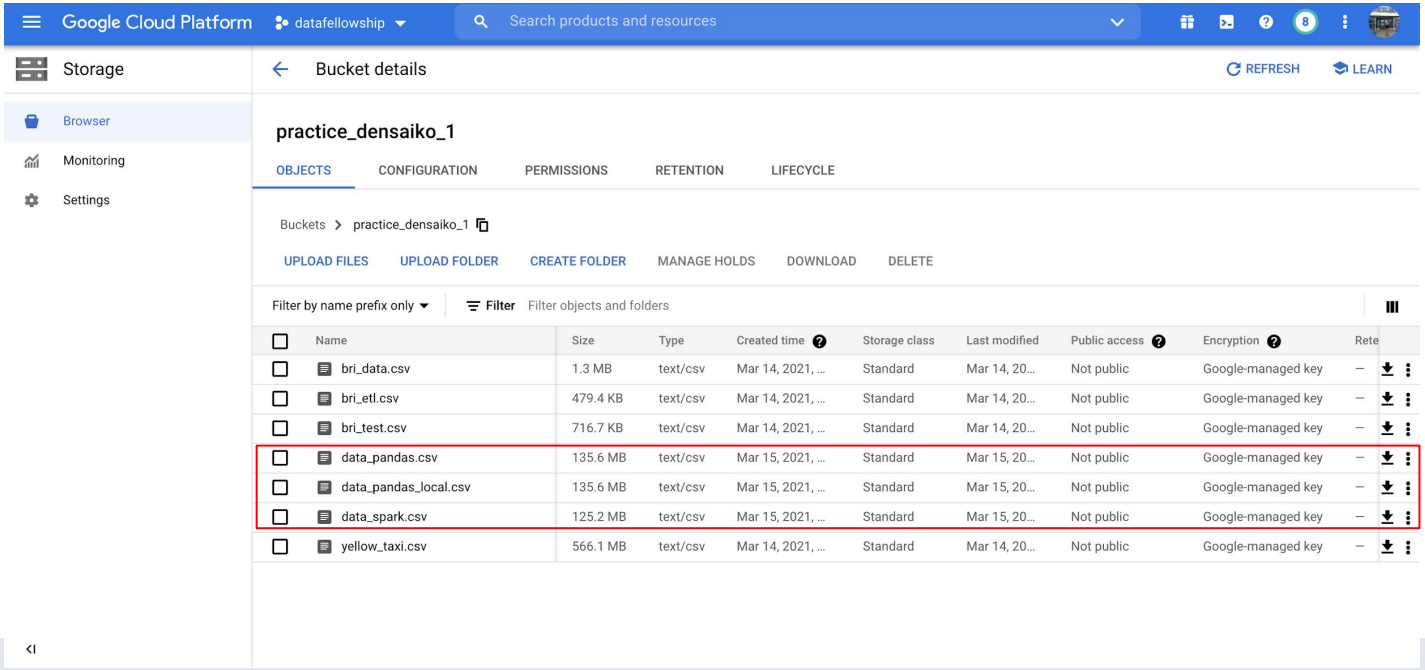
```
Time for reading data:  1731.234647712
Time for querying data:  1.8350863470004697
Time to upload data:  39.584837749000144
Overall Time:  1772.67463511
```

The script is similar like Pandas script that runs in VM. The result shows that pretty longer **around 30 minutes** for overall time.

My hypothesis is because the memory that I use in local is smaller than in virtual machine.

Runtime process Pandas script in Local Computer	
Time for reading data	1731 s
Time for querying/cleansing data	1.83 s
Time for transforming data to pandas df	None
Time for uploading data to GCS	39.58 s
Overall time	1772 s

Result in Loading Data to Google Cloud Storage



THANK YOU

I am looking for any feedbacks and collaborations 😇

https://www.linkedin.com/in/ariprabowo/

https://dataimpact.medium.com/

https://github.com/densaiko