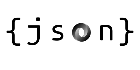
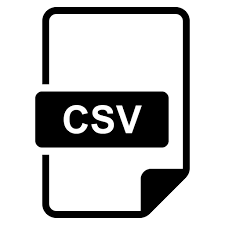
Interview Challenge (Data Engineer) – Document

**Apache Airflow Architecture**



GCSToBigQueryOperator

PythonOperator

(convert)

PostgresToGCSOperator

google bigquery

(output)

csv file

json file

postgresql

(input)

**Dags**

* postgresql\_to\_gcs\_user\_log.py – Dag file that query ‘user\_log’ data from postgresql and upload to GCS
* postgresql\_to\_gcs\_users.py – Dag file that query ‘users’ data from postgresql and upload to GCS
* user\_log\_to\_bigquery.py – Dag file that transform ‘user\_log’ postgresql data to csv file and load to Google BQ
* users\_to\_bigquery.py – Dag file that transform ‘users’ postgresql data to csv file and load to Google BQ

-----------------------------------------------------------------------------

**Google Cloud Platform Service Usage**

* Google Compute Engine
* Google Storage
* Google BigQuery
* Google Dataproc

-----------------------------------------------------------------------------

airflow-vm - GCE instance

IP: 35.213.169.173

Machine Type: e2-highcpu-4 (4 vCPUs, 4 GB memory)

Description: Apache Airflow Single Node

Airflow URL: <http://35.213.169.173:8080/home>

Airflow Account:

* Admin
  + Username: jon
  + Password: password
* Viewer
  + Username: bluepi
  + Password: password

Airflow PATH: /srv/airflow

Airflow DAG(s) Directory: /srv/airflow/dags

Services:

1. airflow-webserver.service – autostart
2. airflow-scheduler.service - autostart

Usage Guide:

1. sudo su airflow
2. source /srv/airflow/bin/activate
3. sudo systemctl status airflow-webserver.service
4. sudo systemctl status airflow-scheduler.service

jupyter-bigquery-m - GCE instance

IP: 35.213.131.149

Machine Type: n1-standard-4 (4 vCPUs, 15 GB memory)

Description: Jupyter Notebook master node for Google Dataproc Cluster

jupyter-bigquery-w-0 - GCE instance

IP: 35.213.135.153

Machine Type: n1-standard-4 (4 vCPUs, 15 GB memory)

Description: Jupyter Notebook worker node 1 for Google Dataproc Cluster

jupyter-bigquery-w-1 - GCE instance

IP: 35.213.143.219

Machine Type: n1-standard-4 (4 vCPUs, 15 GB memory)

Description: Jupyter Notebook worker node 2 for Google Dataproc Cluster

-----------------------------------------------------------------------------

airflow-postgres - GCS instance

Description: Contain JSON and CSV file for Airflow Pipeline

bigquery\_bluepi\_output - GCS instance

Description: Contain files that use in Google Dataproc Cluster

dataproc-temp\* - GCS instance

Description: Contain files when Google Dataproc Cluster get processing

-----------------------------------------------------------------------------

sirapob-bluepi-de-exam:airflow\_gcs\_to\_bigquery – Google BigQuery

Description: Google BigQuery Dataset

sirapob-bluepi-de-exam:airflow\_gcs\_to\_bigquery.user\_log\_to\_bigquery – Google BigQuery

Description: Google BigQuery ‘user\_log’ Table

sirapob-bluepi-de-exam:airflow\_gcs\_to\_bigquery.users\_to\_bigquery – Google BigQuery

Description: Google BigQuery ‘users’ Table

-----------------------------------------------------------------------------

jupyter-bigquery – Google Dataproc

Jupyter Notebook URL: [Jupyter Notebook](https://oqw2nil25vfwtchkittgjsw6ii-dot-asia-southeast1.dataproc.googleusercontent.com/jupyter/lab/)

Type: Dataproc Cluster

Cluster Detail:

1. jupyter-bigquery-m – Master
2. jupyter-bigquery-w-0 – Worker
3. jupyter-bigquery-w-1 – Worker

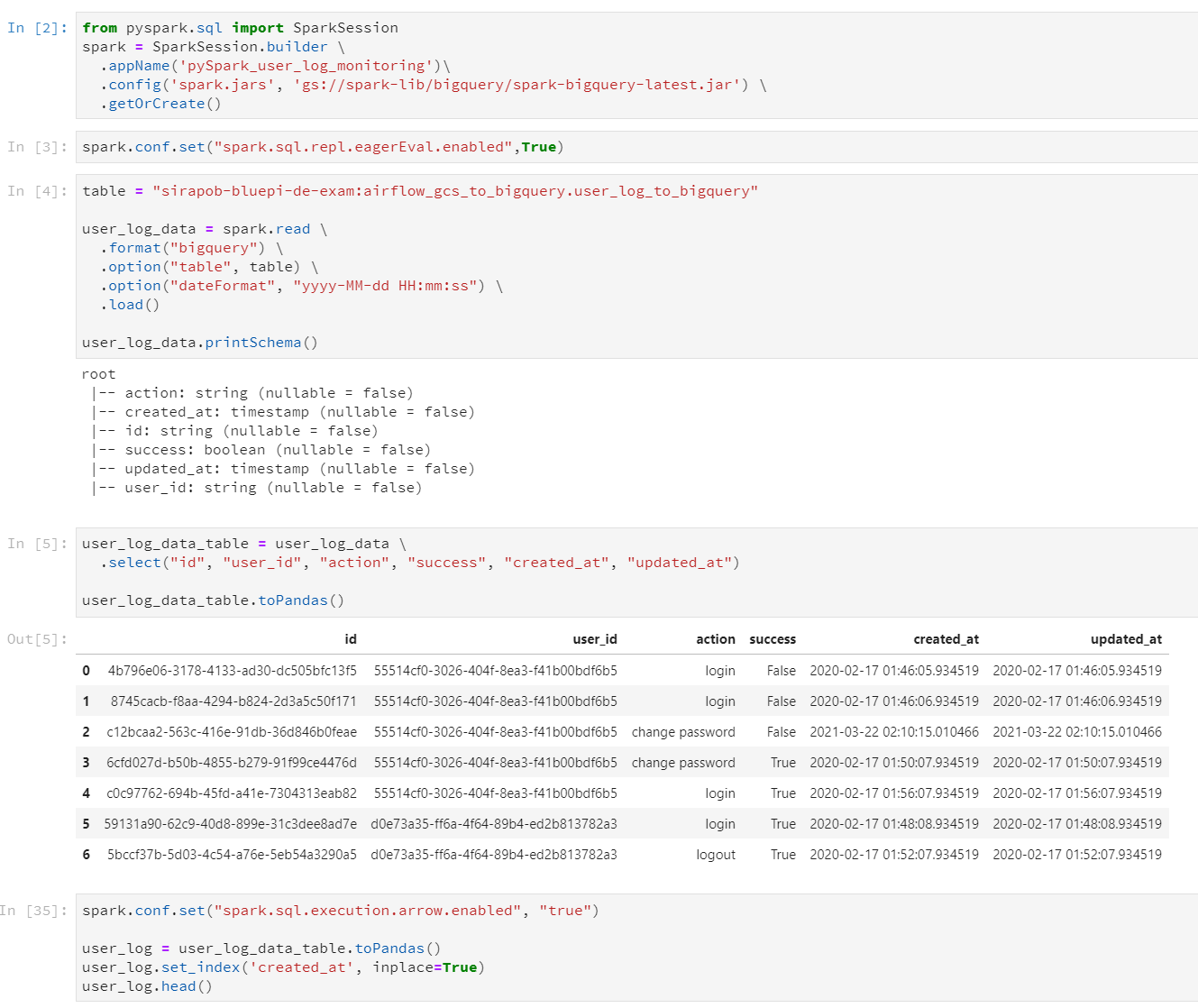
-----------------------------------------------------------------------------

**Jupyter Notebook**

Output Monitor Files

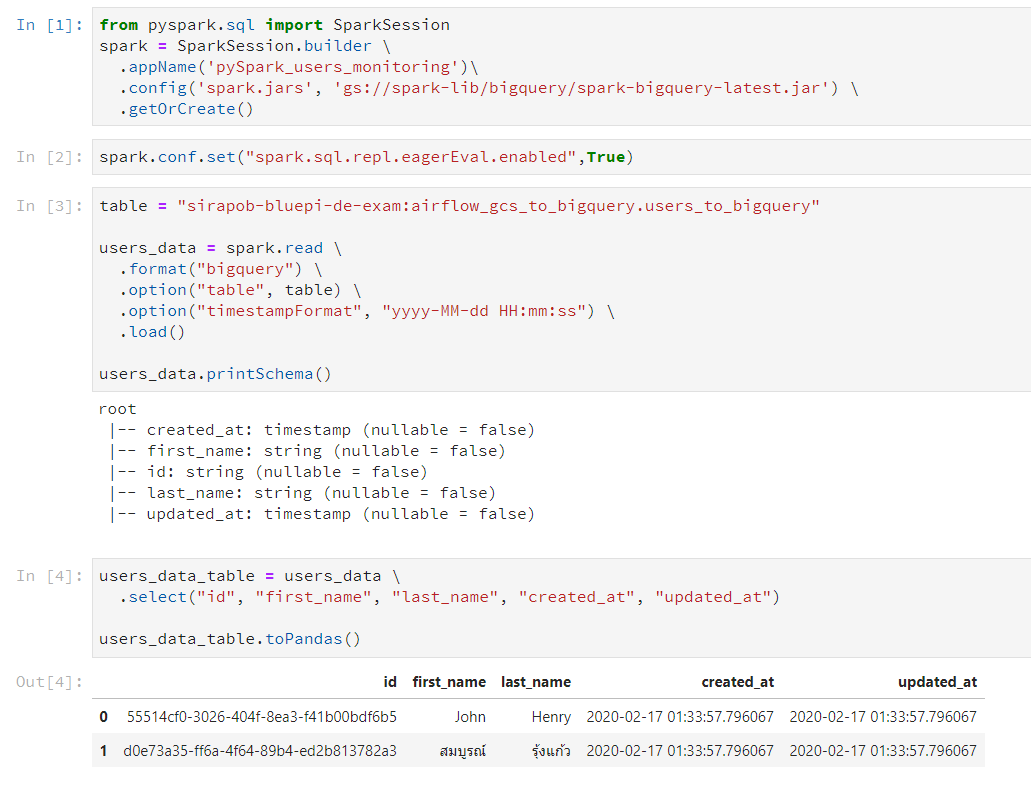
* pySpark\_user\_log\_monitor.ipynb
* pySpark\_users\_monitor.ipynb

pySpark\_user\_log\_monitor.ipynb – Jupyter Notebook IPYNB

****

****

pySpark\_users\_monitor.ipynb – Jupyter Notebook IPYNB



-----------------------------------------------------------------------------