* The file is supposed to be run in a Google Cloud Shell, so authentication is not included in the script
* A GCS bucket is used to store the files
* The Apache Beam and Dataflow are partially used (only for concatenating and cleaning loans tables)
* The Bigquery engine is used to join and aggregate tables and exporting final csv files
* The loans, visits, customers, and final combined tables will also be present in a Bigquery dataset (dataset name is rocker\_pipeline)
* There are 3 mandatory field to be specified in command line while running (Google cloud project id, google cloud storage bucket name, and the runner {DirectRunner / DataFlowRunner}. Dataflow runner will produce a Dataflow job in Google Cloud and Direct runner will run Beam pipeline locally.
* A typical command line for running in google cloud shell terminal is:

python3 beam\_bigquery\_pipeline.py --bucket $BUCKET --project $DEVSHELL\_PROJECT\_ID –DataFlowRunner

where BUCKET is a global variable defined by me in the terminal.

* The base table is loans table, and all joins are left join, as it *loans* full information is requested
* For this task, ss I needed “gsutil URL” to pass to the beam pipeline, I uploaded them in my own bucket and read from there.
* Dataset “rocker\_pipeline” should be created in Bigquery before running the pipeline.
* Final csv file is presented in the directory