



Setup instructions

Execution and Schedule



- Setup instructions - commands
- Schedule

- Setup with GitHub repository
 - All code and documents are available on GitHub
 - Link : https://github.com/ThiagoBarsante/DataEngineer_projects.git
 - Options to install (clone the github repository or download it manually)
 - Clone github repo with command
 - git clone https://github.com/ThiagoBarsante/DataEngineer_projects.git
 - Option to just download and install manually
 - Download the zip file and unzip and run it
- Execution – detailed commands - file in this repo at /doc dir
 - Doc3_Setup_and_execution_DataPipeline_local_GCP_and_AWS_commands.TXT

Note - The solution provide:

- Daily execution or Execution for specific day
- Problems and exceptions registered in the logs

jupyter Execution_info Last Checkpoint: an hour ago (autosaved)



Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted

Python 3

Run Code

Mix Panel daily execution - simulation

```
In [1]: %run mixpanel_daily_datapipeline.py local
```

```
print - Process start...
| Creation of log file ...
| ----- PROGRAM EXECUTION
| ----- Data pipeline start
| Python program: mixpanel_daily_datapipeline.py
| local/cloud parameter: local
| Execution date: 2020-03-18
| Data dir: ../data/
| Log dir: ../log/
| Json download file dir: ../dir_json_stg/
| Log file name: ../log/20200319_114341_mixpanel_daily_datapipeline.log
|
2020-03-19 11:43:41 | ----- Starting execution -----
2020-03-19 11:43:41 | Download JSON file and create one dataframe for each EVENT
2020-03-19 11:43:41 | curl https://data.mixpanel.com/api/2.0/export/ -u XXXYYY680770fe99b03e4631ba22fAPI: -d from_date
="2020-03-18" -d to_date="2020-03-18" -d event='["Company First Start","Conversation Started","Conversation Complete
d","A/B - Onboard - Voice","Subscription Confirmed"]' >> ../dir_json_stg/20200319_114341_mixpanel_ALL_events_STG.json
b''
2020-03-19 11:43:41 | JSON API DOWNLOAD - OK
2020-03-19 11:43:41 | Merge all Data frames and filter rows and columns ...
2020-03-19 11:43:41 | Label Encode valueabonbvoi ...
2020-03-19 11:43:41 | One Hot Encode paths ...
2020-03-19 11:43:41 | Calculate number of hours...
2020-03-19 11:43:41 | Processing target - number of conversations ...
```

File: bin/_Notebook_execution_info.ipynb

- The schedule could be done using simple cron-job (basic sample below) or schedule with cloud provider such as GCP or AWS for example
 - Edit the cronjob with the command with best time schedule
 - `crontab -e`
`XX YY * * * cd /home/<bin direcotry>/bin && ./cronjob_mixpanel_daily_datapipeline.sh`
 - Add the daily execution or run the shell script with Path variable already setup in the shell with sample
 - `cronjob_mixpanel_daily_datapipeline.sh`
- Schedule options with GCP, AWS or on-premise/local environment could be:
 - Airflow (datapipeline orchestration) to run local / on-premise
 - GCP Composer (Airflow) or Cloud Scheduler
 - AWS Step Functions or AWS Lambda triggered by Cloud Watch events