Setup instructions

Execution and Schedule

Agenda

- Setup instructions - commands

- Schedule

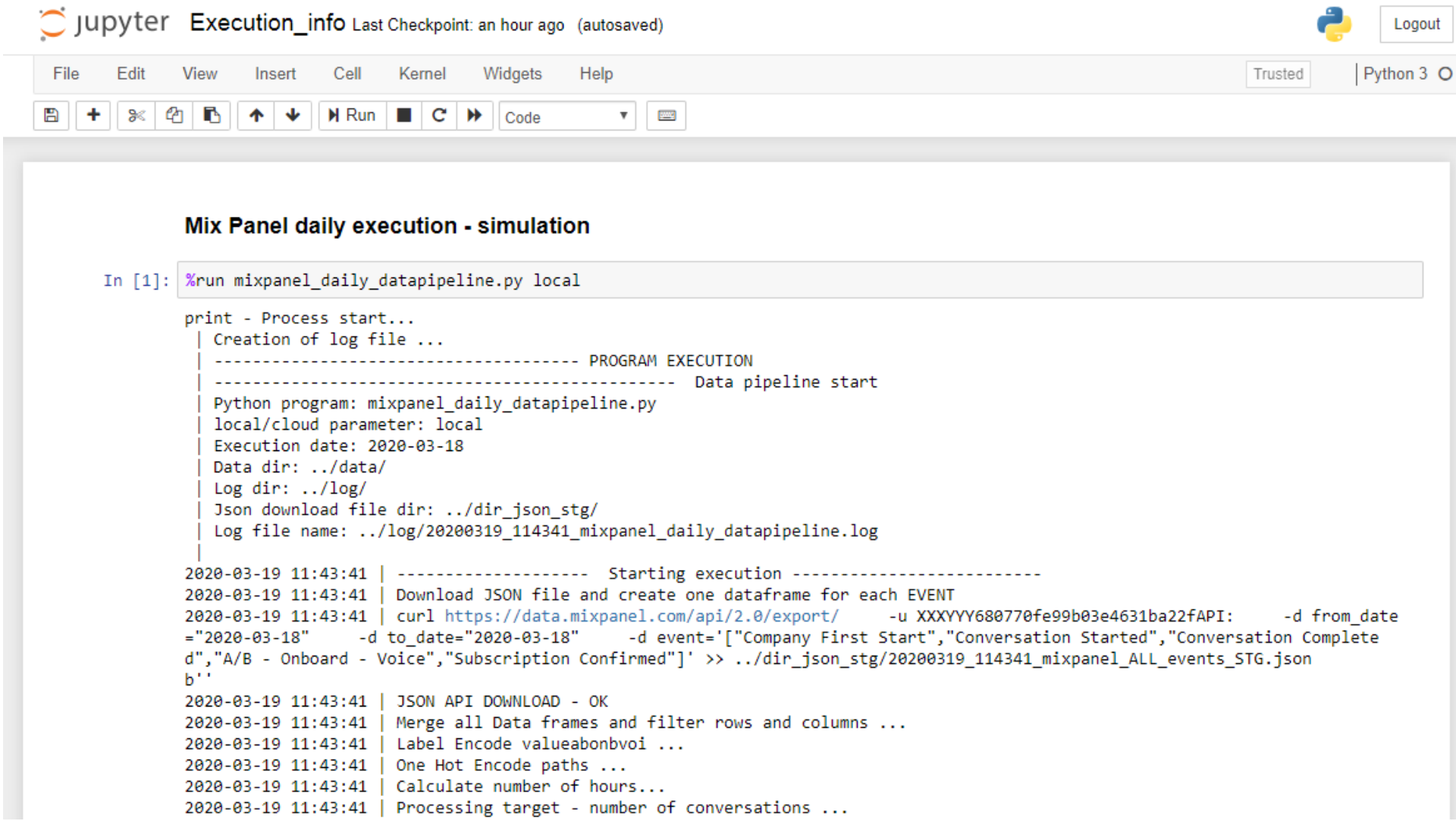
Setup instructions

- 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 mannualy)
  - 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

Notebook with project solution - sample



File: bin/_Notebook_execution_info.ipynb

Schedule

- 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:
  - Aifrlow (datapipeline orchestration) to run local / on-premise
  - GCP Composer (Airflow) or Cloud Scheduler
  - AWS Step Functions or AWS Lambda triggered by Cloud Watch events