**Data download:**

Lending club dataset has pre and post login option to download data. Post login, you can get the entire data which is tagged secure.

Mechanical soup was used to automatically login and download data. Mechanical soup has the combined features of request and BeautifulSoup.

The links for the secure files were found in a hidden div, pipe separated.

Once the data is downloaded, it is timestamped after reading and creating dataframes.

ts = time.time()

date\_time = datetime.datetime.fromtimestamp(ts).strftime('%Y-%m-%d %H:%M:%S')

These dataframes are merged into a single csv for feature engineering.

**Missing data Analysis:**

**Feature engineering:**

**Pipeline:**

The same steps were followed for Rejected/Declined Loan data except that it did not require any login for complete data

The credentials of the Lending Club login and the access keys for s3 upload are being taken as parameters during docker run image command. After this, no human intervention is requied. The pipeline ends with the processed csv uploaded to s3.

**Dockerizing:**

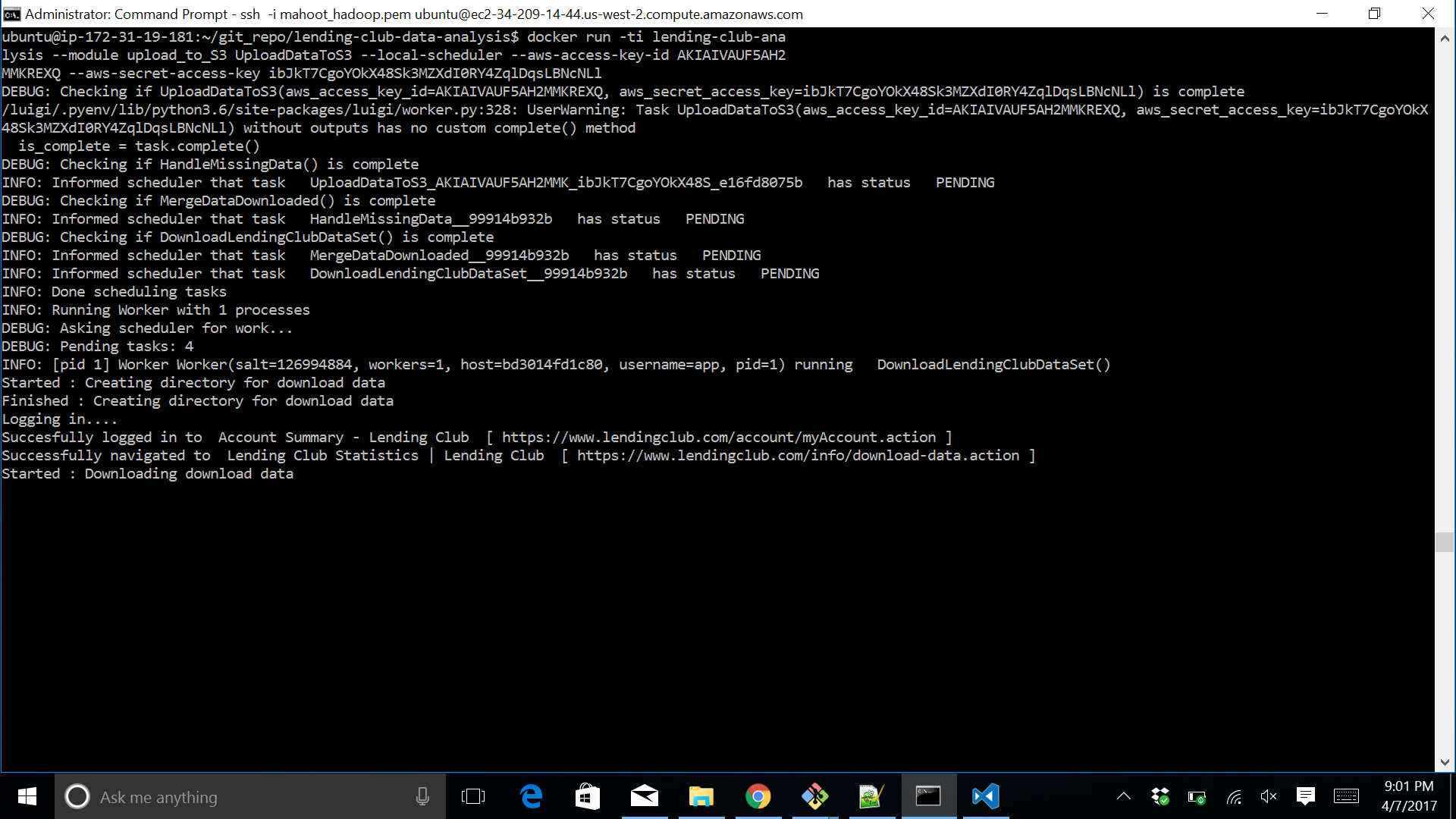
The docker image is

**Scheduling the pipeline:**

**CRON???**

**Exploratory Data Analysis :**

**Outputs on EC2**



References

[1] [hickford/MechanicalSoup]: A Python library for automating interaction with websites (https://github.com/hickford/MechanicalSoup)

[2]