

INFO7374 AlgorithmicDigitalMarketing

Yifu Liu: 001083896

Pengfei He: 001426963

Presentation PDF version

Assignment1_Team7.pdf

Codelab Link

<https://codelabs->

[preview.appspot.com/?file_id=1ONiLiGPDp8zpd5HHdtn1OZ0sPoUrdtI64xQmsEx3wms#0](https://codelabs-preview.appspot.com/?file_id=1ONiLiGPDp8zpd5HHdtn1OZ0sPoUrdtI64xQmsEx3wms#0)

About Dataset

There are eight csv files we can use. Some of them are exceeding the 100M limitation, as well as having missing values or invalid values:

campaign_desc.csv

campaign_table.csv

causal_data.csv

coupon_redempt.csv

coupon.csv

hh_demographic.csv

product.csv

transaction_data.csv

Data Wrangling

We have used Trifacta for joining the tables, in order to transform and map data from one "raw" data form into another format. The charts created by Trifacta also provides us a better view of the data as a whole and provides the hints on how to join them together.

Data Preprocessing

We use pandas to preprocess the data. The usage of the two python file is as follows:

1 Digital marketing HW1(1).ipynb

Handled missing data and gotten sample from casual data which is 600MB (free tier limitation is 100MB)

2 Digital marketing HW1(2).ipynb

Applied group-by and aggregation function to organize dataset. Plotted the pictures and merged some tables.

Analysis and visualization

We used Snowflake to stage and query the data modified from pandas. We have created our custom warehouse, schema and table.

We generated serval custom views to query the dataset. Most of our business insights come from those queries.

Visualization is done by dashboard after import data into EA.

Insights

Through this assignment, we have a better understanding on process we as data scientists can do before handing to market team, as professor described on the class. The related screenshots can be found in the folders named by the tools we use.

Some insights about the tools we use are as follows:

1 Trifacta

Advantage: 1 It is very handy to see all the data in all table. 2 Trifacta is a good tool in the big data tool category of a tech stack. 3 Trifacta is an open source tool with GitHub stars and GitHub forks.

Disadvantage: 1 Dataset cant exceed 100MB. 2 Dataset load becomes slow when loading large dataset.

2 Pandas

Advantage: 1 Easy to use when conduct quick development. 2 Plotting method is handy to visualize data.

Disadvantage: 1 Hard to find out the joining relationship among tables.

3 Snowflake

Advantage: 1 User-friendly UI, especially for large dataset. 2 Strongly computing capabilities in handling huge datas.

Disadvantage: 1 Dataset has a limitation, so we have to sample the data to use it.