MissionWired Hiring

Data Engineer Exercise

The purpose of this exercise is to evaluate your level of skill when it comes to manipulating and aggregating a large dataset through code. We'll evaluate the quality, output and readability of your code as well as the efficacy of provided documentation.

We recommend using Python and Pandas to complete this exercise. Most of our production data engineering work is done using Python, Pandas and PySpark (a "big data" alternative to Pandas).

We recommend submitting your code by way of a personal GitHub repository. Directly submitting code files is also acceptable.

Draft documentation describing how a reviewer can run your app locally. Be sure to include steps like installing dependencies or other "pre-flight" configurations necessary for your code to run.

Dataset

A dataset simulating CRM data is available in some public AWS S3 files:

- Constituent Information: https://als-hiring.s3.amazonaws.com/fake_data/2020-07-01_17%3A11%3A00/cons.csv
- Constituent Email Addresses: https://als-hiring.s3.amazonaws.com/fake_data/2020-07-01_17%3A11%3A00/cons_email.csv
    - Boolean columns (including `is_primary`) in all of these datasets are 1/0 numeric values. 1 means True, 0 means False.
- Constituent Subscription Status: https://als-hiring.s3.amazonaws.com/fake_data/2020-07-01_17%3A11%3A00/cons_email_chapter_subscription.csv
    - We only care about subscription statuses where `chapter_id` is 1.
    - If an email is not present in this table, it is assumed to still be subscribed where `chapter_id` is 1.

Use these files to complete the exercises below.

MissionWired

Exercises

1. Produce a “people” file with the following schema. Save it as a CSV with a header line to the working directory.

Column	Type	Description
email	string	Primary email address
code	string	Source code
is_unsub	boolean	Is the primary email address unsubscribed?
created_dt	datetime	Person creation datetime
updated_dt	datetime	Person updated datetime

2. Use the output of #1 to produce an “acquisition_facts” file with the following schema that aggregates stats about when people in the dataset were acquired. Save it to the working directory.

Column	Type	Description
acquisition_date	date	Calendar date of acquisition
acquisitions	int	Number of constituents acquired on acquisition_date