The Movement Cooperative

Engineering Hiring Exercise

Instructions

- You can use every engineer's best friend, the Internet, as well as any textbooks, notes, or outside materials. Don't ask your friend/mom/professor/stranger on the internet for help with this exam. We are working off the honor code here, plus we'll find out quickly if you can't do the things on this exam independently and efficiently.
- You should complete all three tasks in less than an hour.
- Please send a copy of this exam saved as a PDF to hr@movementcooperative.org

Tasks

- Those applying for the Support Engineer role should complete Tasks 1, 2 and 3
- Those applying for the Data Engineer role should complete Tasks 3 and 4

Hiring Exercise

Task 1

One of our ETL scripts failed while bringing new data into the sample_data.email_activity table in Redshift, producing an error message that references further diagnostic information available from the stl_load_errors system table.

When we query stl_load_errors for more details, this is what we see:

	filename	colname	type	col_length	raw_field_value	err_reason
1	s3://parsons-tmc/Parsons_RedshiftCopyTable/1525631758824885898.csv.gz	emailsubscribed	varchar	5	by-mail	String length exceeds DDL length

What Redshift/SQL query would you run to resolve the issue before re-running the failed portion of the data sync?

Task 2

Processing text data can be challenging, and our syncs often take steps to normalize data.

What Python code would you run to strip all formatting from a set of U.S. phone numbers and store them as a consistent 10-digit string?

Task 3

One of our syncs is broken, and we need to communicate the outage to TMC's member organizations in our #bugs-and-outages Slack channel.

Here's what we know:

- A vendor changed how they configure security on their side for a database mirror, and they will now require an SSL certificate
- We run the data sync via a Civis Platform job template, and Civis support staff need to talk to the vendor's Engineering team about how to handle the certificate
- TLDR; we're relying on outside entities to fix things that we don't directly control, so we do not have a clear timeline for resolution

Can you write a member-facing message about this outage?

Task 4

Build a pair of scripts that will:

1. pull voter file data from the Ohio Secretary of State website, and
2. match a provided input CSV file to that voter data, creating another CSV which looks like the input (including the row column) but has an extra column **matched_voterid** that specifies the matches.
3. include a README explaining your approach.

The input CSV file: https://drive.google.com/open?id=1o3SWFV1oJ4Z3hr6nFPAQPO8y8wWtlfgL

For gathering the voter file: Ohio is one of the few states that makes it really easy for anyone to download the voter file, which is the list of registered voters in the state.

- First check out https://www6.sos.state.oh.us/ords/f?p=111:1 to see where the data comes from, and what is the data format.
- Then you can download each county's data by using URLs like https://www6.sos.state.oh.us/ords/f?p=VOTERFTP:DOWNLOAD::FILE:NO:2:P2_PRODUCT_NUMBER:1, where that last number is a county number from 1 to 88. For performance reasons, you can limit your matching to the first 4 counties' data.

Matching is a "fuzzy" process that often doesn't have a clear right answer, so you'll have to weigh tradeoffs and make decisions, eg. how much to normalize strings, how much to weigh certain columns, what to do when there isn't a clear match, etc. Make decisions that seem reasonable to you, and document any interesting tradeoffs.

Your submission will be graded on the quality of matches and on the quality of the code. So make sure your code is readable. Imagine you might have to hand-off the project to another team member.

Spend no more than a few hours on this test, and feel free to wrap up sooner if you have a solid first pass. You will certainly have further improvements you would make with more time and with more input from users. As part of your README, explain those potential improvements and open questions. Those are as important as the code.