Data Engineer Case Study

Brief

This test is designed to minimally replicate a couple of real world ETL scenarios:

1. Using the BigQuery public dataset bigquery-public-data:noaa_gsod you need to produce a Google Cloud SQL table of minimum and maximum daily temperatures in degrees Celsius by state between the years 1990 and 2000
2. Download 3 files from a Google Cloud Storage bucket and produce 3 Google Cloud SQL tables, making sure you describe and explain any relationships you feel would be appropriate between the sets of data

You should demonstrate:

- Writing a python module to interact with BigQuery. This can be a wrapper around the standard Google API library. (bq_module)
- Writing a python module to interact with Google Cloud SQL database. (sql_module)
- Writing a python module to interact with Google Cloud Storage. (gcs_module)
- For #1, write an application script that calls bq_module to extract data and calls sql_module to load it into the SQL database. Database schema is to be created as part of this exercise.
- Similarly, for #2, write an application script that calls gcs_module to extract data and calls sql_module to load it into the SQL database. Database schema is to be created as part of this exercise.
- To demonstrate that data has been inserted correctly, create an application script that logs into the DB using a view (i.e. non-root) user which takes a city as an argument and then returns all the information known about the city from the three tables

You can gain access to BigQuery public datasets with a free Google Cloud account that can be created here. You are allowed to perform 1TB of queries per month without incurring any charges.

Once you have created your Google Cloud account, give us the account email and we can provide access to Google Cloud Storage as well as email the database credentials separately.

Requirements

- Write code using Python 3
- Demonstrate knowledge of unit testing and be able to justify anywhere that it hasn't been used
- Submit code as public github repository, ensuring there are no credentials included
- Don't write more code than you need to

- Use comments appropriately, including highlighting where you would do things differently if you had more time
- **Please provide github links to us no less than 24 hours before the interview.**

We expect each part of the case study to take around 2 days - 1 day for research and 1 day for coding. It may take you more time if you are unfamiliar with the systems described. Please ask questions if you get stuck - it's more important that you produce working code for some of the case study fully than attempt all of it and be unable to show a working program.

Google Cloud Details

- Storage Bucket: im-training
- Project ID: rare-basis-686
- Project Number is: 1079477301385
- Database IP: <will be emailed once account has been created>
- Database Root User Password: <will be emailed once account has been created>

Support

This exercise is meant to reflect how working in an office on a day-to-day basis would function. To this end, using any form of external support is ok, including StackOverflow and other repositories of knowledge on the web.

Downloading/copying code is fine, as long as you cite your sources. You will not be marked down if most of your code is externally sourced, instead we are looking for how your apply resources to a problem. Often, less is more, as long as it's still readable and makes sense.

You may request help or clarification on any parts of this case study at any time by emailing daniel.bowman@infectiousmedia.com or dan@infectiousmedia.com and we encourage you to do so. There are no stupid questions. Please also use these addresses for sending links to your github repository.