We would like to offer you the opportunity to demonstrate your data engineering skills by creating an extract-transform-load application. The exercise should generally not take more than 3 or 4 hours, although you're free to take as much time as you'd like to work on it. If you don't finish within a few hours, that's okay; submit what you've got anyway.

Please read through the following instructions and return by end of day **<Send day + 1 week>**

**Directions**

1. Please write your application in Python, Java, Scala or GoLang. The application **must be buildable from the command line**; it should not require an IDE to build or run.
2. Download the data, a collection of 285,000 Yelp reviews on restaurants in the Phoenix area:

<https://dataengineeringexercise.s3-us-west-1.amazonaws.com/Yelp_dataengineering_dataset.zip>

1. The download contains files in 3 groups, based on the datatype of the downloaded objects (Json, Csv and Sqlite). Each group is associated with a different functional subject area of the Yelp dataset and might be split into multiple files
2. The application should extract data from the sources and perform the following computations on the extracted data:
   1. Joins the individual datasets into a base summarized dataset
   2. Calculate the mean reviews by business
   3. Calculate the mean reviews by zipcode for the Top 5 most business dense zipcodes
   4. Calculate the Top 10 most active reviewers
3. The application should format the output data set into 4 CSV files and upload them to a storage system of your choosing (e.g. S3) for download by consumers.
4. Other considerations:
   1. Deploy the application to some public location where we can interact with it (e.g. [Heroku](https://www.heroku.com/) or [AWS](https://aws.amazon.com/)), if time permits
   2. Include whatever documentation you believe is necessary for others to understand and maintain the application
   3. Include tests that verify the behavior of your application, if time permits
5. To submit the code, post the code to a private github and share the link to the code repository. We will provide github user names to give access to.

**Evaluation Criteria**

As part of your interview, we will do a joint review of your approach with the following criteria

1. Logical Reasoning and Analysis – Correct joins, valid joins and outputs
2. Enterprise Scale Engineering – Modular design, advanced programming techniques
3. Readability – Readable code, Readme file
4. Creativity & Innovation – Use your imagination

Take as much time as you need, but at least spend a 3-4 hours on this. Just do the best you can.

Let me know if you have any questions and I would be happy to help.