Mini-project for Data Engineering Job at Unison Investment Management

Introduction:

You will be setting up a micro-service on a server that executes a simple extraction, transformation and loading sequence.

We have launched a MySQL database that stores raw data on home characteristics. You are tasked with pulling the data into a server, compressing the data into a set of summary statistics, and pushing those statistics back to the MySQL database into a new table.

Next week, we will begin periodically adding new data to this table and your server should be able to automatically detect when new data is added to the input table and add to your output table.

Details:

Database configuration:

Host: de-application.cef17qxjlavg.us-west-2.rds.amazonaws.com

Port: 3306

Username: <check email>

Password: <check email>

Input Table: “ApplicationData.raw\_data”

Output Table: “[YOUR SCHEMA].output\_table”

1. Set up a server (using a free tier AWS EC2 server is sufficient). This server will allow you to keep your code running so that it can detect when data is added to the input table “raw\_data”.
2. Write a python script that is capable of pulling all of the data from the “ApplicationData.raw\_data” table. You will generate the following output table of summary statistics for a few of the attributes:

**Column Name | Average | Standard Deviation | Median | Count | timestamp**

Description of Columns:

-Column Name: There should be one row for each of the following column headers: “lot\_size\_sqft”, “total\_building\_sqft”, “yr\_built”, “bedrooms”, “total\_rooms”, “bath\_total”, “final\_value”. The “Column Name” column should store these headers.

-Average: Should include the average of non-zero/non-null values of the associated header. (e.g. for the row with the “Column Name”=”lot\_size\_sqft” the average of “lot\_size\_sqft” should be posted into this column)

-Standard Deviation: The sample standard deviation of non-zero/non-null values should be stored in this column

-Median: The median of non-zero/non-null values should be stored in this column

-Count: The count of non-zero/non-null values should be stored in this column

-Timestamp: The timestamp should be a “yyyy-mm-dd-hh-mm-ss” timestamp of when the upload occurs

1. The script should be able to continuously add rows to your output table as more data is added to the input table “raw\_table” (append the summary statistics with a new timestamp, do not delete your old data). Your mandate is to have an update within 1 hour of new data being added to the “raw\_table” table. Extra points if you ONLY update the table when new data is added to the “raw\_table” table.