DANA DATA QUALITY ENGINEER
TEST CASE

Project : How Weather Affect Restaurant Ratting

In this project, you will merge two massive, real-world datasets in order to draw conclusions about how weather affects Yelp reviews.

The first step is to obtain the data we will use for the project.

Make sure you have around 10 GB free disk space. We've provide dataset from yelp (https://www.yelp.com/dataset/download) and U.S Weather data (CSV attached)

You will be creating and saving several documents, as well as taking numerous screenshots as evidence of completion. Keep track of the files.

In the final, you need to analyze the data and create a query to determine how weather affects yelp reviews.

What you need to do?

1. Create submission document using word and convert into pdf when it done and ready to submit. All the question will be need to put in this document. This is the document that you'll need to submit.

2. Create a data architecture diagram to visualize how you will ingest the dataset into Staging, Operational Data Store (ODS), and Data Warehouse environments, so as to ultimately query the data for relationships between weather and Yelp reviews. Please provide a screenshot and put in the submission document.

3. Extract the tar file. Take a capture and put in the submission document.

4. Convert yelp dataset from json to csv using your preferred code.

   Make your project stand out by Dockerize your project. We also will value your code and file structure. So please use good file structure and convention. Commit your code into your GitHub account, and put the GitHub link in the submission document.

5. Analyze the datasheet and draw an entity-relationship (ER) diagram to visualize the data structure in ODS. Provide the screen capture in submission document.

6. In this step you need to build star schema data warehouse layer based on the data you have. Analyze the data, draw star schema diagram and create DWH layer table represent your star schema diagram.
   Provide the screen capture in submission document.

7. Implement your data architecture diagram using your preferred datastore (local & cloud). Create staging and ODS table for each dataset. Commit the DDL statement for each table to your GitHub and put the link in the submission document

8. Migrate the data from ODS to DWH layer. Commit your queries into your GitHub and put link in the submission document.

9. Query the data to determine how weather affects yelp reviews. Commit your queries into your GitHub and put link in the submission document.

10. In your opinion, let say you want to do data quality before delivering your final data product to business user, be it report, model or dashboard. What kind of metrics you want to check first and when will you do data cleansing and in which part. Create your own version of analysis in mindmap or flowchart.

11. Based on your design in Question #10, please add at least 1 script rule for each table (ODS) and DWD (consistency).

Deliverables :

1. Submission document (pdf) containing all the screenshot, diagram and methodology explanation. **(especially the data quality mindmap or diagram)**

2. Share your GitHub account that you use as Code and SQL Repository

If there is something that not clear enough, feel free to contact us via email
yoanna.artha@dana.id