Yelp Dataset

Overview

The Yelp dataset is a dataset which contains 5 JSON datafile named business.json, review.json, user.json, checkin.json, tip.jso. The dataset contains 6,990,280 reviews, 150,346 businesses and 11 metropolitan areas. The dataset is sourced from Kaggle. The total size of the dataset is around 8 GB which falls into the category of "Big Data".

Problem Statement

The dataset will be used to answer the following questions: **How does cuisine preference change with location? How does business attributes such as Wi-Fi, etc. change the review ratings and popularity of the business?**

Scope

The scope of the project would be to (1) Acquire the data from the data source (2) Perform ETL process (3) Automate the process using automated data pipeline in a cloud-based environment (4) Query and analyse the data to find solutions of the problem

Data Source(s)

The data is sourced from https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_business.json .

Architecture of the Solution

1. The data would be acquired and stored in a data lake
2. The data from data lake would be then put through ETL process to form a data warehouse. Cloud solutions will be used to scale the database
3. The data warehouse would be used to analyse the dataset

Technology

The choice of technology would depend on the stage of project. The initial choice of technologies is:

1) Data Acquisition: Python, API modules
2) Data Exploration: Python, Pandas, SQL
3) ETL: Python, SQL
4) Analysis: Python, SQL