Proposal

Abstract

Firstly, I would like to thank all of the trainees and mentors for providing such a wonderful intern session, and helping me to learn many more things and providing an opportunity to work on such a wonderful project.

I will go for the YELP Business dataset for the final project , because I think it will be a challenging task for me and I will learn more from this project.

Introduction about YELP

Yelp,Inc. Develops , hosts and markets the Yelp.com website and the Yelp mobile app,which publish crowd-sourced reviews about businesses.It also operates Yelp Reservations, a table reservation service,It is headquartered in San Francisco, California.

Source: Wikipedia

Objectives

* Develop ETL pipeline to successfully load the datasets provided by the YELP
* To make a conceptual and physical model of the Data warehouse for online analytical processing(OLAP)
* To validate the date to ensure the accuracy and quality of data in the warehouse.
* To visualize the data to get the business insight from it.

Methodology

I will go with the waterfall model as it is a linear-sequential life cycle model and very simple to understand and use.

Tools and Technology Used

* Python Programming Language
* PostgreSQL for DBMS
* SQL
* Git for version controlling
* GitHub for hosting online repository
* Power BI for data visualization
* Google Docs For documentation
* Draw.io For graphs
* Google slides for presentation
* VS code for writing code
* Dbeaver-ce

Project task and time schedule

As I have only 10 days for the completion of the project proper task management and time scheduling is also a key process.

| Task | Number of days |
| --- | --- |
| Understanding YELP websites ,and exploring its datasets | 1 |
| Creating raw database and understanding more about the datasets | 1 |
| Designing the Data warehouse | 1 |
| Physical implementation of warehouse | 1 |
| Making ETL pipeline which is a continuous process throughout the project. | 3 |
| Data validation | 1 |
| Data Visualization | 1 |
| Making documentation which is also a continuous process throughout the project | 1 |
| Final Presentation and overall review of project | 1 |

Bibliography

1. <https://www.yelp.com/dataset/documentation/main> , for datasets.
2. <https://docs.python.org/3.8/> -Python 3.8
3. <https://www.postgresql.org/docs/12/release-12-8.html> -postgresql 12.8 DBMS
4. <https://git-scm.com/> , for version controlling
5. <https://github.com/> , for hosting online repository
6. <https://powerbi.microsoft.com/en-us/> , for data visualization
7. <https://code.visualstudio.com/> , for writing codes
8. <https://app.diagrams.net/> , for diagrams
9. <https://www.google.com/docs/about/> , for documentations
10. <https://www.google.com/slides/about/> , for presentation
11. <https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm> ,SDLC