GitHub Pull Requests

Overview

The GitHub Pull Requests is a Kaggle dataset which contains 5 CSV datafile for 5 days of Pull Requests on GitHub. The dataset has data for a different non-holiday Monday from January to May of 2019. The dataset is sourced from GHTorrent project. The total size of the dataset is around 42 GB which falls into the category of "Big Data".

Scope

The scope of the project would be to (1) Acquire the data from the data source and store it into data lake (2) Perform ETL process (3) Analyse the data using the data warehouse.

Data Source(s)

The data is sourced from https://www.kaggle.com/datasets/stephangarland/ghtorrent-pull-requests?select=ghtorrent-2019-01-07.csv.

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. Cloud solutions will be used to scale the database
3. The data lake would be used to analyse the dataset using concepts such as Text Analysis, etc.

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) Machine Learning and Analysis: Python, SQL