YELP Recommendation System: Project Overview

Team 18: Uddhav Bhosle (ubhosle), Harika Malapaka (hsmalapa), Akshit Meghawat (ammeghaw), Deepak Patil (dpatil)

Description:

The project will recommend restaurants to users by analysing user's past activities and their requirement. Yelp's data on the restaurants domain can help modeling an algorithm that would output a list of desirable restaurants.

The data is distributed in various locations on S3. By using Apache Spark on Amazon EMR, we would learn batch processing of data in a distributed framework. This will enable us to process large amounts of data and subsequently train machine learning models for recommendations. This project would be a gateway for learning events of Big Data coupled with cloud computing.

Deliverables:

- A project report giving results of how Spark can be used in real time, it's execution environment and analyze whether in-memory computation has a fair performance.
- The results obtained from executing interactive queries like: (Mexican restaurant within 5 miles radius) will be shown. The accuracy of the algorithm in predicting may not be high since that's not the goal for this project.
- The complete implementation will be uploaded on Github.
- A presentation on our work in detail including model snippet.

Dependencies:

Dataset: reviews, businesses, users, tips and check-in data on Yelp regarding restaurants.

Storage: We need to store the data in a cloud storage so that distributed computing is feasible. S3 would be the storage platform.

Computing: EMR would be extracting data from S3 and would be handling all the computations parallely.

ML: MLlib which is built on top of Spark is backbone of our prediction algorithm.

Configuring **Python3** on EMR instance.

Issues:

- The size of dataset obtained from the source is just 3GB. In order to increase the size of the dataset, we will perform data synthesis (duplication with minor changes).
- Getting the desired predictions though parallel computations on distributed data
- Many more potential issues can be addressed after we deep dive into this project in next few weeks. Since we are not concerned with a pretty UI, the user may have to manually enter their requirements.