CSE487/587: DATA INTENSIVE COMPUTING

Instructor: Bina Ramamurthy

Most Followed User on GitHub

Submitted By:

Aman Bhayana
50290968
Pratik Agarwal
50290570

Abstract

GitHub is a web-based hosting service for version control using Git. It serves as a platform for hosting codes which provides Source Code Management as well as Distributed Version Control mechanism. Different users have their own programming preferences in terms of style, language, framework etc. There are many open-source as well as organizational level projects hosted on GitHub. There are many users who have many followers, whose projects have been forked the most etc. In short, this information can be used to determine as to which user has most influence on other GitHub users in terms of project forks, following. We aim to demonstrate this feature using Apache SPARK. The user data i.e. the followers of a given user can be procured using GitHub API or using the public dataset made available by GitHub on Google Big Query. The user details on this dataset can act as the nodes of a directed graph which would be our input. This input when visualized in form of a directed graph, can be simulated to obtain the result using the concept of Connected Components. It is often used in the most real-life social networking models. We use Apache SPARK as it uses Lazy Computation. It means that the transformations are not processed until some action is encountered or is encountered. It uses RDD as the lowest level of abstraction and data frames or datasets are based on RDD only. We can use any data visualization tool like Tableau or d3.js to demonstrate the results as per the problem proposition.