TECHNOLOGY TREND ANALYSIS OF STACK OVERFLOW USING HADOOP

Saksham Chaudhary[1], Sakshi Bhalla[2], Mohd Faizan[3,] Rishabh Agrawal[4], Mohd Ilyas[5]
[1,2,3,4,5]Assistant Professor, Moradabad Institute of Technology , Moradabad

ABSTRACT

The project utilizes the Stack Exchange API (Application Programming Interface) that allows the applications/websites to incorporate functions that are used by Stack Overflow application to fetch and view information. Stack Exchange uses IIS, SQL Server, and the ASP.NET framework, all from a single code base for every Stack Exchange site to generate a unique access key which is further required to fetch Stack Overflow data. Once the API key is generated, a PHP (PHP Hypertext Preprocessors) based application is designed to use the Stack Exchange API for fetching information based on a search criteria. The text file output generated from the console application is then loaded to HDFS (Hadoop Distributed File System). HDFS (Hadoop Distributed File System) is a primary Hadoop application and a user can directly interact with HDFS using various shell-like commands supported by Hadoop. This project uses MapReduce operations that are later run on data using Hadoop to extract the meaningful output which can be used by the management for analysis. The main objective of this project is to show how we can analyze Stack Overflow data using Stack Exchange API to make targeted real time and informed decisions. This project will help in understanding changing trends among people by analyzing Stack Overflow data and fetching meaningful results. For example, users can find the topmost trending programming languages so that they can take decisions to learn the new languages/technologies. This application can analyze how many people asked the questions about a particular programming languages and how many people answered them, to understand the popularity of a particular programming language. This way users can take decisions which programming languages they should learn to maximize their recognition in the corporate sector.

Keywords: Hadoop, Big Data, Cloudera, MapReduce, HDFS

1. INTRODUCTION

With rapid innovations and surge of internet companies like Google, Yahoo, Amazon, eBay and a rapidly growing internet savvy population, today's advanced systems and enterprises are generating data in a very huge volume with great velocity and in a multi-structured formats including videos, images, sensor data, weblogs etc. from different sources. This has given birth to a new type of data called Big Data which is unstructured sometime semi structured and also unpredictable in nature. This data is mostly generated in real time from social media websites which is increasing exponentially on a daily basis.

Stack Overflow is one of the most popular and engaging tool and an amazing platform that reveals the community feedback through questions and answers, number of votes, tags, number of ratings for a particular question.

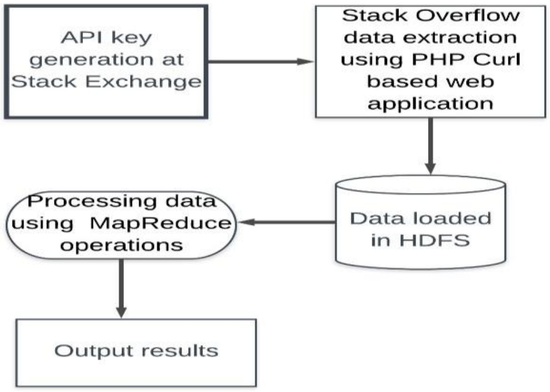

API key generation at Stack Exchange
Stack Overflow data extraction using PHP Curl based web application
Data loaded in HDFS
Processing data using MapReduce operations
Output results

Fig 1.1 High level Dataflow

Overflow collects a wide variety of traditional data points including questions, answers, Votes, and tags. The analysis of the above listed data points constitutes a very interesting data source to mine for obtaining implicit knowledge about users, technologies, questions and community interests.

2.LITERATURE REVIEW

"Stack Overflow has over a billion users and every day developers post thousands of questions and gets answered." Every day, people across the world are uploading thousands of questions and videos to Stack Overflow and this number is ever increasing. In May 2010, Stack Overflow (as its own new company) raised US$6 million in venture capital from Union Square Ventures and other investors, and it switched its focus to developing new sites for answering questions on specific subjects, Stack Exchange 2.0. Users vote on new site topics in a staging area called "Area 51", where algorithms determine which suggested site topics have critical mass and should be created. In November 2010, Stack Exchange site topics in "beta testing" included physics, mathematics, and writing. Stack Exchange publicly launched in January 2011 with 33 Web sites; it had 27 employees and 1.5 million users at the time, and it included advertising. At that time, it was compared to Quora, founded in 2009, which similarly specializes in expert answers.Other competing sites include WikiAnswers and Yahoo! Answers.

3. PROPOSED SYSTEM

The main objective of this project is to focus on how data generated from Stack Overflow can be mined and utilized by developers and beginners to make targeted, real time and informed decisions about trending programming languages in the market and to learn them. This can be done by using Hadoop concepts. The given project focuses on how data generated from Stack Overflow can be mined and utilized. There are multiple applications of this project. Developers can use this project to understand which programming languages are popular in the corporate world. In addition to the questions, answers and votes, we can also evaluate posts and comments according to date range. Applications for Stack Overflow data can be endless. For example, Companies can analyze how much a technology is liked by people. This project can also help in analyzing new emerging trends and knowing about people's changing behavior with time. Also people in different countries have different preferences. By analyzing the questions, answers, upvotes, downvotes, tags etc. of the programming languages, companies can understand what are the likes/dislikes of people around the world and work on their preferences accordingly.

4. REQUIREMENT ANALYSIS

After analyzing the requirements certain amount of tasks that have to be performed, the next step is to analyze the problem and understanding the context. There are majorly two phases that are present. The first activity in the phase is studying the existing system and the other is to understanding the requirements of the new system.

Understanding the properties and requirements of a system is more difficult and requires creative thinking and understanding of existing systems.

Requirements which we are using in this project include both Hardware and software requirements for the process.

(a) Hardware Requirements

- Dual Quad-core CPU
- 4-8 GB of memory per processor core
- 1 Gigabit Ethernet

(b) Software Requirements:

- Hadoop 2.x
- MySQL
- VMware or Ubuntu
- HDFS
- PHP Curl
- PHP Server
- Chart.js
- Python

5. METHODOLOGY

Most of the technology geeks are uploading their questions on Stack Overflow and they anxiously await for the answers. Major number of developers and IT professionals ask questions on Stack Overflow and answer the questions of others. This further shows that how much a technology is being used in the Computer science and IT sector worldwide. Stack Overflow collects a wide variety of traditional data points including questions, answers, Votes, and tags. Hence the above listed data points can be used for the analysis and understand the developers' views about a particular technology/programming language.

1. This project will help people in understanding how to fetch Stack Overflow's data using Stack Exchange API.
2. This project requires access to Stack Exchange API and generate a unique access key. That unique key is required to fetch Stack Overflow data.
3. With the help of the unique access key, the required data is fetched from Stack Exchange using a .PHP application.
4. The extracted data is first stored in .csv file and then the data is loaded into HDFS. The queries are run into HDFS to execute the MapReduce operation so that the Stack Exchange data can be mined intelligently and the findings can be shared with the users and management.

6. IMPLEMENTATION

Collection of Data sets according to customer complaints like questions, tags, view count, answer count, score, last activity date, creation date, question id and so on.

These are different types of attributes which are collected from the Stack Exchange API.

To implement Trending Technology analysis the following steps are to be followed:

Step1: Create an authentication key at Stack Exchange.

Step 2: Fetch the questions data using PHP Curl app.

Step 3: Convert the database into CSV format and copy the file into Cloudera/Ubuntu.

Step4: Move the data from local storage to HDFS.

Step 5: Execute the MaprReduce operations on the data using shell like commands in HDFS **.**

Step 6: Get the result in TSV format.

Step 7: Plot the charts in web based UI using Chart.js.

7. RESULTS

After execution of MapReduce operations we have designed three modules:

First one is, find top technologies. In this module, user can enter a number n and he will get n number of topmost trending technologies sorted in decreasing order.

The image given below, shows the top ten technologies given in the form of a graph. In the graph, x-axis represents the names of the technologies and y-axis represents the numbers of questions asked on the respective technologies.

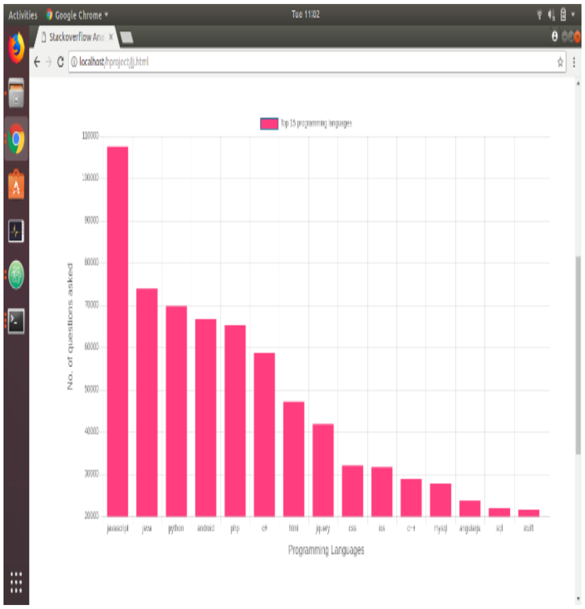


Fig.7.1 Results in Bar graphs

Second module is, compare two technologies. In this module, the user can compare the trends of two technologies. In this module we have used the data of 15 days of March 2019. The image given below shows the comparison chart of two technologies entered by the user. User enters the two technologies in the given input boxes.

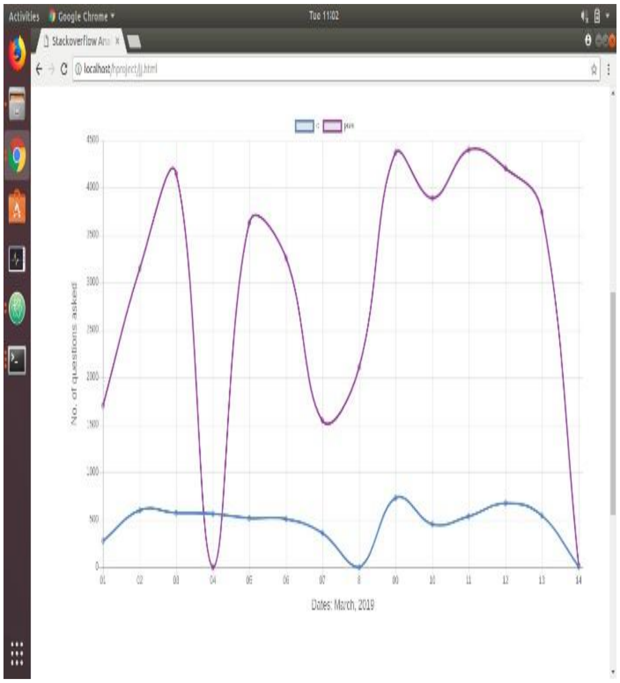
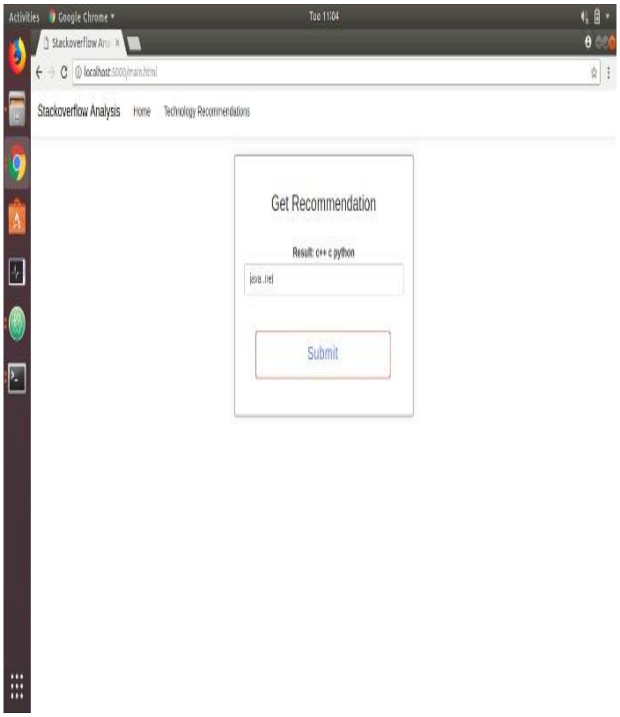


Fig.7.2 Results in line graphs

Third module is Recommendation module. In this module, user can get the recommendation to learn new technologies according to the market trend and his previous knowledge. The image given below shows the recommendation system showing the output to the user in sorted order.



Result: c++ c python

java .net

Fig 7.3 Recommendation System

8. CONCLUSION

The task of big data analysis is not only important but also a necessity. In fact many organizations that have implemented Big Data are realizing significant competitive advantage compared to other organizations with no Big Data efforts. The project is intended to analyze the Stack Overflow Data and come up with significant insights which cannot be determined otherwise. The output results of Stack Overflow data analysis project show key insights that can be extrapolated to other use cases as well. One of the output results describes that developers have asked the questions on the JavaScript in the highest number. Thus we can conclude that JavaScript is the most popular programming language in the present scenario, since it provides a number of open source libraries and frameworks that are capable to develop web applications, mobile applications as well as system software. Thus the developed system is a successful application of Hadoop for the analysis of Stack Overflow's data.

REFERENCES

[1]. Evans, Chris (2013-10-25). "Big data storage: Hadoop storage basics". "HDFS is not a file system in the traditional sense and isn't usually directly mounted for a user to view".

[2]. Malak, Michael (2014-09-19). "Data Locality: HPC vs. Hadoop vs. Spark". datascienceassn.org. Data Science Association.

[3]. Murthy, Arun (2012-08-15). "Apache Hadoop YARN – Concepts and Applications".

[4]. Shafer, Jeffrey; Rixner, Scott; Cox, Alan (2010-05-17). "The Hadoop Distributed Filesystem: Balancing Portability and Performance" (PDF). Rice University.

[5]. Vance, Ashlee (2009-03-17). "Hadoop, a Free Software Program, Finds Uses Beyond Search". The New York Times.

[6]. Wickham, Hadley (2011-08-16). "The split- apply-combine strategy for data analysis". Journal of Statistical Software.