Understanding and Implementing Big Data Concepts using Stack Overflow Data

---

FA17-IN-CSCI-59000-35973
Big Data Management
Project Report

Under the guidance of: Dr. Yuni Xia

By

Adithya Morampudi

Manjusha Kottala

Yashwanth Kuruganti

Aakarsh Nadella

CONTENTS

1. Abstract

Stack overflow is a Q/A website for Developers, where a huge number of developers from all over the world interact with each other on open source as well as enterprise questions. The number of questions posted per day in this website is high and the time required to answer these questions is considerably low, which means any user can post a question and get the response within minutes. We are trying to analyze all the information relating to the data present in this dataset. The Dataset is a collection of the information about the posts posted, the information about the users, the geographic locations of the users and all such related information. The main aim of this project is to do a in depth analysis using the big data concepts discussed in the class.

2. Motivation

The dataset which we are working on is a public dataset available at the stack Exchange open source data dumps website [2], the size of this dataset is around 75GB, which is a good source for us to learn and implement the concepts of big data. The main reason behind selecting this dataset is the fact that it is a computer science Q/A forum and the analysis would help us finding great insights about the present trends in the computer science industry. One use case is when I posted a question in stack overflow the response time in which I got the answer was very small, like literally 14 mins, which is impressive, when we were searching for the datasets for this project, we came across this huge data dump and then started analyzing this dataset after necessary data preprocessing.

3. Introduction

Traditional databases can store and handle a reasonable amount of data very well. But as the size and complexity of data increases, these systems fail to deal with such huge data. This is where Big data comes into play. Big data deals with vast amount of data through programs that can run on multiple machines by performing robust computations and provide a means of extracting useful information. For a practical implementation of big data concepts, we have chosen Stack Overflow data dumps, a subset of Stack Exchange [1]. Stack over flow is a Question and Answer website covering diverse range of Computer science topics.

4. Goals of the Project

1. To Efficiently handle the huge dataset without significant delays or errors
2. To do a in depth analysis on the dataset and provide insights and recommendations
3. Learn different big data concepts.
4. To provide a good visual representation of the results
5. To get a deeper insight in the languages Pig [3] and Spark [4]

5. Dataset Description

The dataset [2] which we are working on has a total of 8 tables, each table has different attributes which can provide a deep insight into things like which tags are increasing over a period,

which tags are decreasing over time, who are the top most active users in the website, so on and so forth. These are the tables present in the Dataset

1. Posts Table
   - Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, Body, OwnerUserId, OwnerDisplayName, LastEditorUserId, LastEditorDisplayName, LastEditDate, LastActivityDate, Title, Tags, AnswerCount, CommentCount, FavoriteCount, ClosedDate, CommunityOwnedDate
2. Users Table
   - Id, Reputation, CreationDate, DisplayName, LastAccessDate, WebsiteUrl, Location, AboutMe, Views, UpVotes, DownVotes, ProfileImageUrl, EmailHash, Age, AccountId,
3. Tags Table
   - Id, TagName, Count, ExcerptPostId, WikiPostId
4. Post History
   - Id, PostHistoryTypeId, PostId, RevisionGUID, CreationDate, UserId, UserDisplayName, Comment, Text
5. Comments
   - Id, PostId, Score, Text, CreationDate, UserDisplayName, UserId
6. Votes
   - Id, PostId, VoteTypeId, UserId, CreationDate, BountyAmount
7. Badges
   - Id, UserId, Name, Date, Class, TagBased
8. PostTag
   - PostId, TagId

6. Data Processing Techniques

The first and foremost thing which needs to be performed in any sort of analysis is the data preprocessing, since the data collected is not in the required format or even if it is in the required format it may contain noisy data, null values. Our data was in the xml format and it was not so convenient for us to work with this data, so we have written a XML parser in PySpark [4].

7. Technologies Used

For this project we have used a combination of big data technologies to process, query on data and visualize the results.

- Analyzing and Querying
  - Pig Latin using AWS
  - Spark using Python: Aspen Cluster provided by SOIC, IUPUI containing 1 Master and 19 Slaves
- Visualization using Tableau

8. Implementation

The first thing which we had to do was to load the dataset into the Aspen cluster as well as AWS [5]. We had setup a cluster with 1 Master node and 2 core nodes which was the default configuration of the AWS cluster [6], the software configuration of our cluster was Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.3.1, Hue 4.0.1, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4, we executed all our queries with this configuration. We split the parts to be executed onto AWS as well as ASPEN cluster of the informatics department, for saving time as well as cost. All the parser queries were implemented using PySpark [4] and these were run on the ASPEN cluster, the data-block size of this cluster was just 32MB. The main challenge was when the queries of the parser were run, the resulting dataset which was in the form of csv were split into small blocks of 32MB each and exporting this data from the ASPEN cluster to AWS took a bit of time, as we were new to this kind of stuff. Once the dataset was exported to the AWS s3 [5], we started writing the queries in PIG as well as HIVE.

Tags File

The queries which we have implemented on the Tags file are

1. Tags which are increasing over the past 10 years, we have implemented this using join on different tables, for instance to achieve this task we have used joins on Tags file, Posts file and the comments file, we matched the Tag Id from the tags file to the Tags attribute in the Posts file and count this tag has appeared in the posts file, grouped by the Tag Id.
2. Which tags are decreasing over the past 10 years?
3. The correlation of scores to the tags, by that we mean which tags are in demand or which tags most of the people are interested in viewing. We have done this using the tag id from the tags table and then joining it with the posts table to find the score the tag is receiving, and then relating the percentage of score each tag is receiving. One important observation here is that we are also finding the top tags inherently
4. The speed at which the questions are being answered. This is the most important thing any Q/A website needs to think upon, because the response time is what decides the fate of the website, if the questions are answered at a smaller time, then more number of users would be interested in visiting this website and will be interested in posting more questions, thus keeping the website up and running smoothly.

Comments File

1. Finding the top ten users who have posted the maximum number of comments, this was done by joining three different files which are the posts file, Users file and the comments file.
2. Finding the comments which were posted more than 8 months ago and still active. This gives us insight on which posts are still active and why is there still a discussion going on? Which means a adequate answer has not been posted yet or the discussion is leading into a new discovery.
3. Top ten comments with highest scores, we found this to know the comments posted by which users have received the highest scores.

4. We also found the top questions which have the maximum comments, which fall in the 90^{th} percentile, this gives the insight on which questions and question topics people have most queries in.

Posts File

1. This is where we did the main analysis, the data in this file was huge and all the other files which were a reference to this file, there is a attribute in this file called PostTypeId which corresponds to the post types posted in the website, a PostTypeId of 1 correspond to a post of the Type Question and PostTypeId 2 corresponds to the post of type answer.
2. The first analysis which we have done on this file is finding the top questions posted in each category of the tag. This is done by joining the posts file, tags file and the votes file, the question is said to be a top question if it has the maximum number of UpVotes.
3. Questions which have the maximum number of answers, and the user whose answer received the highest UpVotes, this is done by using the posts file, tags file and the users file. This gives the knowledge of which user has a good knowledge on which type of question or category, the result of this analysis can be used to build a feature recommendation system which is presented in the Future work of this application.
4. Questions with no answers or not at all addressed per year, as with any website there are some posts which are not at all addressed in the stack Overflow website, the number of total posts which had no answers kept on decreasing year by year, which is a good indication that the users are becoming more active and social to help the community. As of 2016 the number of posts with no answers were less than 5%(4.83%). This is very good indication that any question posted, has a 95% chance of being answered.
5. The bounty count which was given to a question and the number of responses received after adding the bounty to the post. The posts which have received no responses even after providing adequate bounty amount, this means that the question posted is either something which not many people are interested in or this also means that there are very little people with this little expertise.
6. We also found the information from the dataset such as who are the top active users in the website, this is determined based on different factors such as the Reputation of the user, the badges which the user has, the number of answers which the user has answered so far, total number of upvotes the user has received so far, and also the view count of the user, this query had to deal with joins on different tables such as users, posts, comments, badges table's.
7. We also tried to find the number of dead accounts in the website which means the number of accounts which are inactive since the past 3 – 4 years.
8. The number of accounts which are being opened within a span of 6 months, and surprisingly this number is found to be a big one in the past two years.
9. The most viewed questions in each category were also found. This is done using the view count attribute. This gives us information about the top posts which are most helpful to other users.

Users File

1. Top Most active users in the stack Overflow community. This is calculated based on the user reputation, the number of badges the user has received so far, the number of posts posted by the user which are of type 1 and the score each post has received, the number of posts posted by the user which are of type 2 and the scores received and the difference of the UpVotes and DownVotes.
2. Finding the users who can solve a category of the question.
3. Number of Dead accounts, we found this to analyze the potential of the website.
4. We found the users with the Top 10 Reputation, these are not necessarily the same users who are top most active users.
5. The top number of badges a user has. This is just to know the trend of the badges being assigned to different users, to find what kind of badges a user with a reputation has. This gives us insight on how the badges are assigned and what is required to achieve these badges.
6. The users with no silver or Gold badges, we found a very interesting fact that the users with only Bronze badges are more than 40% of the total users, which means that the users who have silver and gold are less compared to bronze medals.

9. Results

All the above analysis output many figures from which its tough to grasp useful insights. Visualization helps a human eye to better process and understand the variation and recognize similarities. For live interactions, please refer https://cs.iupui.edu/~mkottala/BigdataAnalysis.html

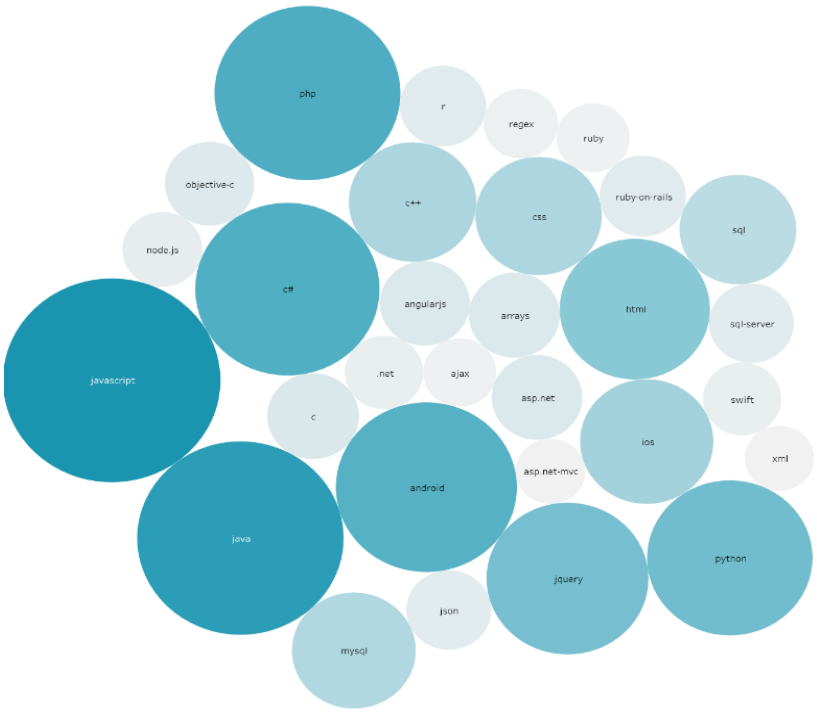

Figure 1: Top tags over last 5 years.

1. Figure 1 illustrates the top 30 tags that have been observing an increase over the last five years. As we observe JavaScript, Java and PHP are the most used tags in descending order. The size of the bubble indicates popularity of the tag.
   Tables used: PostHistory, PostTag, Tags.
   Attributes Used: PostId, TagId, TagName, CreationDate.

2. Based on user reputation, their posts and answers, each user will be awarded badges: Gold, Silver and Bronze. Stacked bar graph in Figure 2, shows the users with most number of badges. The top part of the stack represents total number of gold medals, whereas the middle and bottom part of graph represents the total silver and bronze medals per user respectively. Jon Skeet, is the user with highest number of badges.
   Tables Used: Users, Badges
   Attributes Used: UserId, DisplayName, Class, Reputation

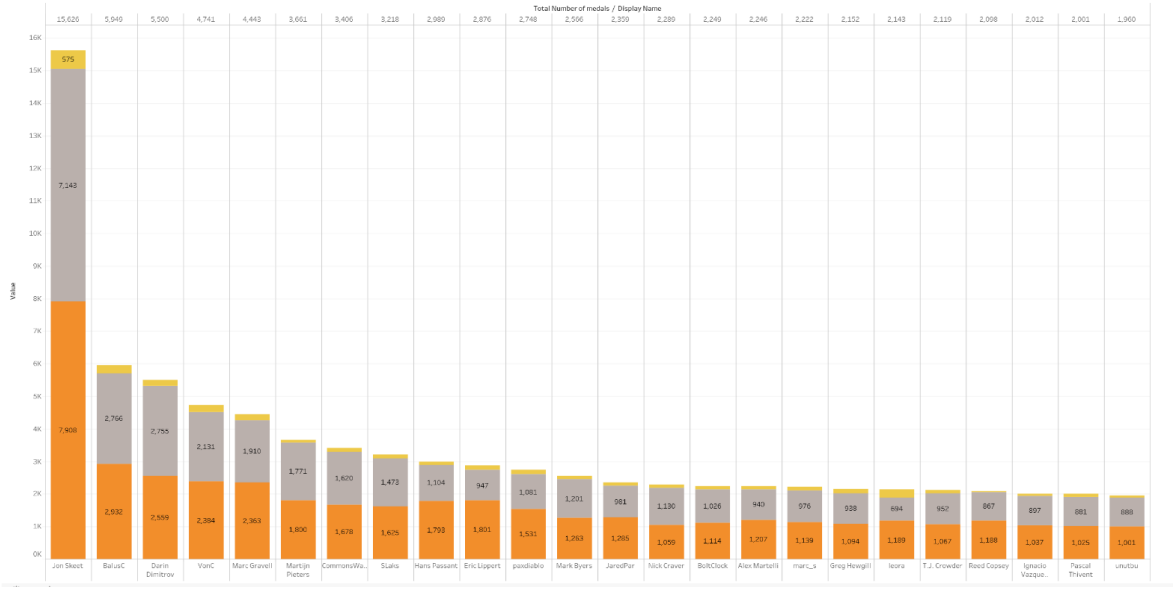

Figure 2: Top Users based on badges

3. For each question posted in stack overflow, users can upvote or like the question and answer posted, if they feel it is helpful to them. As all the posts last for long period, every time when a new user arises with same question, he will be directed to that post. So, each time when a new user visits this post, the number of views increases. The tabular form in Figure 3 depicts the top 10 posts based on highest number of upvotes(scores) and views. From the table below, “How to undo the last commits in Git?” is the post with highest views and upvotes.
   Table Used: Posts
   Attributes Used: Title, ViewCount, Score

Title	Score	View Count
How to undo the last commits in Git?	15,916	5,686,549
How do I delete a Git branch both locally and remotely?	12,079	4,967,749
How to redirect to another webpage?	7,276	4,329,452
How to check whether a string contains a substring in JavaScript?	6,937	4,347,422
How to revert Git repository to a previous commit?	5,203	3,585,257
How do I remove a particular element from an array in JavaScript?	5,114	3,949,196
How to check whether a checkbox is checked in jQuery?	3,613	3,323,884
How do I find all files containing specific text on Linux?	2,912	3,577,735
How to convert a String to an int in Java?	2,245	4,710,459
How to create an HTML button that acts like a link?	1,140	3,672,555

Figure 3: Top 10 Posts

4. Users post answers to a wide range of questions with which they are familiar with. Users who posted highest number of answers are analyzed and the results are visualized as shown in Figure 5. The users who posted highest number of answers is depicted with darkest color and the intensity of the color decreases with number of answers posted, when compared to most valuable user. From the visualization Gordon Linoff, Jon Skeet and Darin Dimitrov are the most valuable users in decreasing order.
   Tables Used: Posts, User
   Attributes Used: PostId, UserId, OwnerUserId, UpVotes, DownVotes

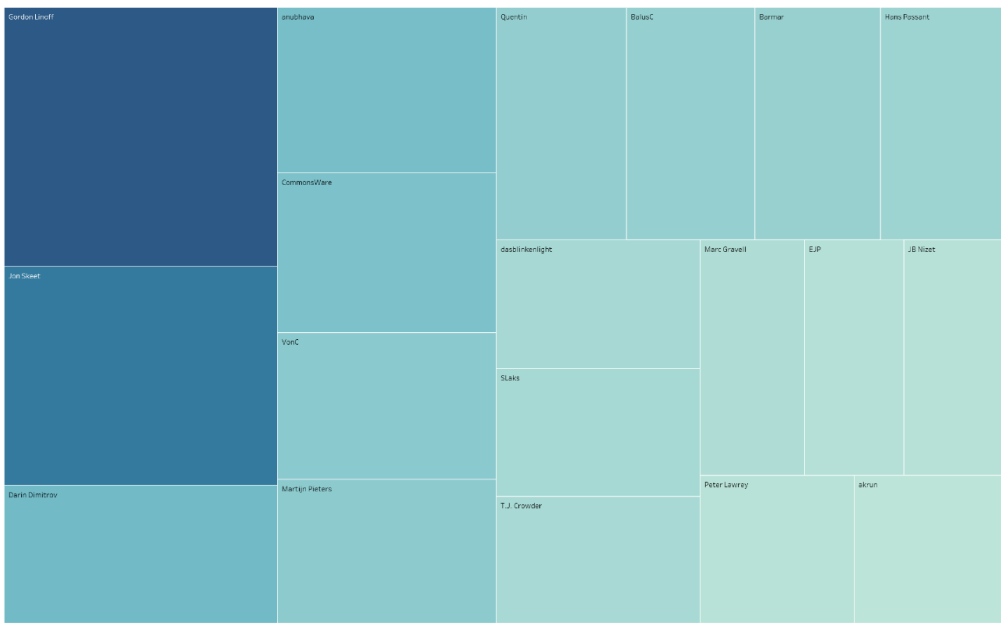

Figure 4. Top 30 valuable users.

5. There are good number of users who are associated with Stack Overflow. Some users create account and shall not login again. There are many such inactive accounts over the years. The number of accounts which are not used for a time frame of say one year, two years, three years, etc. are as shown in the line chart on the left side in Figure 5. On the right-hand side of Figure 5 represents, the number of new accounts created in each year starting from 2009 to 2017.
   Tables Used: User
   Attributes Used: UserId, LastAccessDate, CreationDate

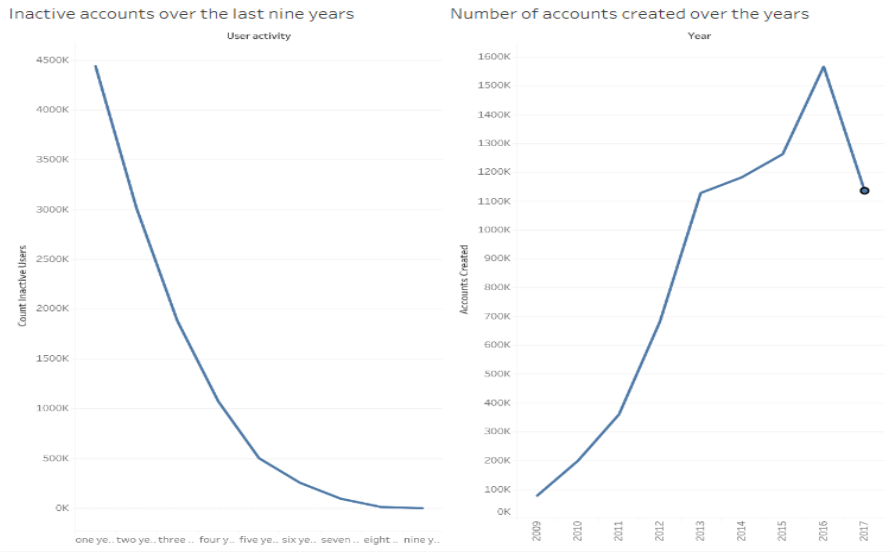

Figure 5: Inactive accounts and New accounts created.

6. Each question, answer and various other topics that a user does is considered as a Post.

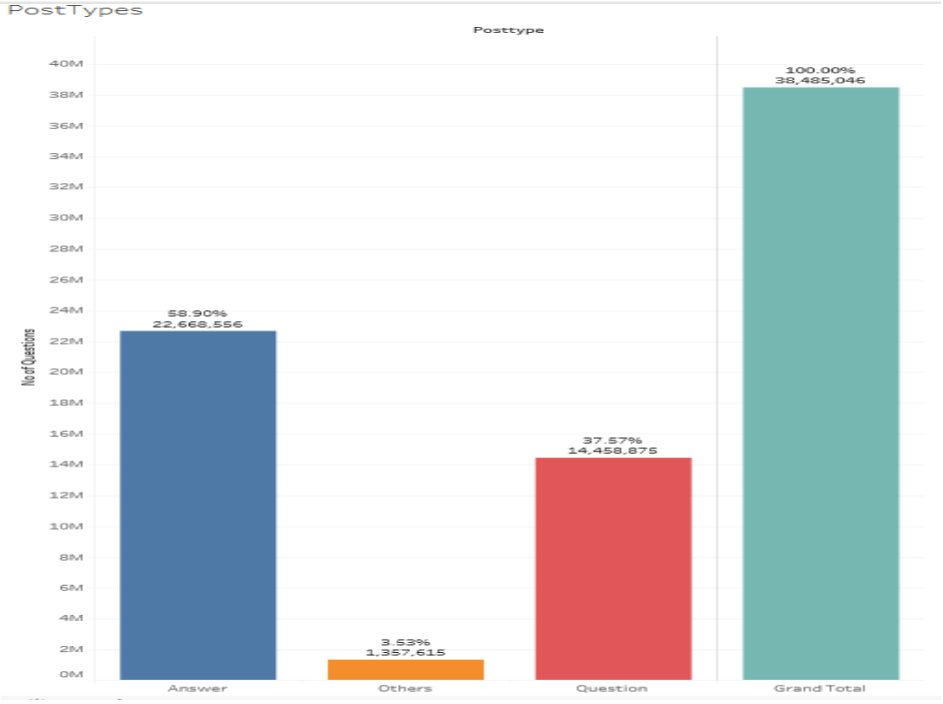

Figure 6: Post Types

Breaking down each of these categories, we got total number of answers, questions and post which are neither a question nor an answer. All these results are illustrated in Figure 6.
Table Used: Posts
Attributes Used: Id, PostTypeId

7. Every question posted in Stack Overflow doesn't get answer. But users who are interested in these questions can mark them as favorite, so that when another user posts answers to it, notifications will be sent to the users who marked them as favorite. The bar graph below in Figure 7 represents the number of unanswered questions in each tag, while the bar graph above represents number of users who have marked these questions as favorite.
Table Used: Posts
Attributes Used: Id, PostTypeId, AnswerCount, FavoriteCount.

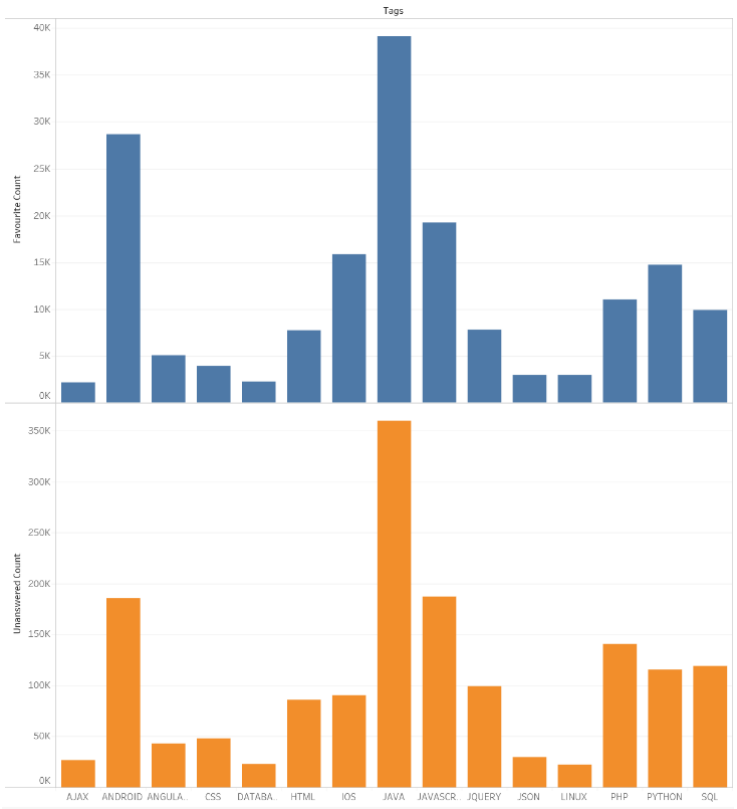

Figure 7: Unanswered Questions but marked as Favorite in each Tag

10. Challenges

Since this kind of project is new to us, learning how to use the technologies involved was the main challenge. The dataset is in the XML format which was not very easy to run queries on, so we had to write a data parser which converts the xml format to csv, this was not a cake walk. There was so much redundant data that had to be cleaned for a proper analysis. Writing pig queries

and running them on Aws platform was another major challenge, it's not because of the complexity of the pig code, but it's because of the runtime which it takes to run the queries, some queries took almost an hour to finish and then it might sometimes result in an unknown error, which left us with no clue where the error has happened. This made us re-run the queries several times. For the visualization part, we initially started off with Amazon's Quicksight, but we later found that it has not much of a support for visualizing multi-dimensional data or maybe the learning curve of Quicksight was high, so we then shifted to Tableau as it has a good support for multidimensional data. We also faced some issues while running the queries which had to deal with sorting of the data. While sorting the data, along with the data files many temporary files were also created and then it became hard to determine which were the data files.

11. Future Work

To develop a system, where the whole process can be automated, by whole process I mean pulling out the information from the data dumps and do all the preprocessing by its own, this might seem a little over-whelming, but it is achievable. An application where users can register and find information relating to the top users in each category and can contact them if any important question needs some help, only with the consent of the users. This way it reduces the time required to answer a query.

Role of Team Members

As the dataset is very huge, it is difficult if the whole work is concentrated on only few members of the team. So, each member of the team has contributed equally towards the project. The dataset contains 8 individual files. Each member of the group worked on 2 files each. Below is the list of tasks that are contributed equally by all the team members:

- ✓ Total of 8 xml data dumps
- ✓ Data Preprocessing – Data Cleaning, Sampling
- ✓ XML Parser – to CSV
- ✓ Querying using Spark and Pig
- ✓ Visualization
- ✓ Report

References

1. https://en.wikipedia.org/wiki/Stack_Overflow
2. https://archive.org/download/stackexchange
3. https://pig.apache.org
4. http://spark.apache.org
5. https://aws.amazon.com/documentation/s3/
6. https://aws.amazon.com/documentation/emr/