**Stack Exchange Dataset Analysis**

**Introduction** :

Stack Exchange is a network of question and answer Web sites on topics in varied fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process

**Goal** :

We need to maintain instructive information on all site to provide user with useful information.

Hence useful information can be achieved by below things.

1. Identify expert users depending upon different factor like age, upvote, downvote, reputation

And assigning them moderator access to maintain reliable data/

2. Identify spammers who inject dirty data and useful information in lost or identifying users who are answering their own question and generating reputation.

**Dataset** :

* Ask Ubuntu
* Stack Overflow

Initial operations performed on Ask Ubuntu and tested on Stack Overflow for testings

scalability

* Data for six years processed

Earliest Post: 2009-01-08

Latest Post: 2016-03-06

**Technologies used :**

Pig

HBase

Map Reduce

Hive

Apache Mahout

Azure ML Clustering

Power-BI

Amazon EC2

Java

**Pig :**

We have used the pig scripts to extract data from the dataset and apply for filter to clean the data and feed into HBase.

Pig was used for extraction and it has piggybank library which has XPath() functions to extract the xml tags attributes.

Pig was used to perform ETL on our dataset that includes cleaning and filtering.

**HBase:**

HBase is used to store the data.

This choose HBase since it uses since it can work on real time data since hive does not have features of updated.

**MapReduce**

Since achieving/retrieving the one to many mapping is difficult to store and store in HBase

Using pig or hive, we implemented Map reduce programs.

We loaded the data from hive and feed to HDFS and it one to many mapping for userid.itemid and interaction level which was feeded to mahout for Mahout recommendation.

**Hive**

Hive is SQL based and we integrated hive with hbase to retrieve data from hbase .

We extracted all the data using hive queries and feeded that to Power BI Visual Tool

**Apache Mahout**

We used apache mahout to recommend question to user who are already expert in that particular technology.

We can many map reduce to achieve the input required by mahout.

**Use-case** : there are many cases where we don’t have answers to question and hence this recommendation will be helpful for users to answer the question which they may like based on other users taste.

**Azure Machine Learning**

We achieved to cluster all users having similar taste or expertize in same technology.

**use-case** : This helped us to cluster the users with similar tags and then find out question and notify expert users to answer similar question and hence helps to make this site more instructive

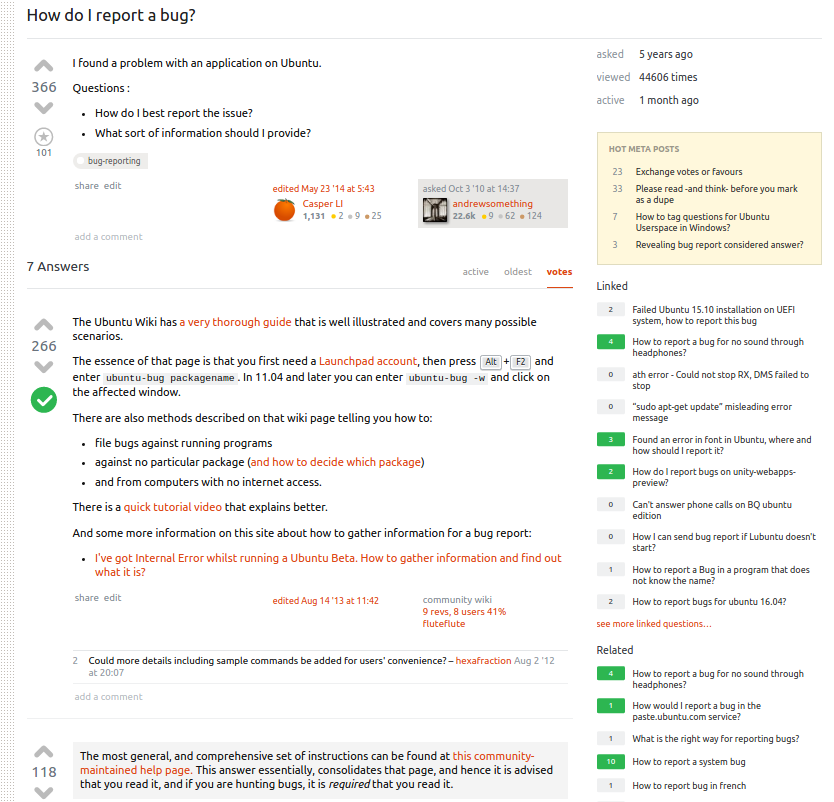
**Amazon EC2 :**

We set up Hadoop and hive in distributed mode on Amazon EC2 with cluster of one master and 6 slave nodes.

We are able to run large dataset around 8GB on Hadoop cluster.

C

## PageRank



As is visible in the snapshot above, every post has linked links and related posts links.

The related posts are out-links from a page whereas the linked posts are the in-links to a page.

When a webpage A refers to a webpage B, webpage B appears in the related post list of the webpage A and webpage A appears in the linked post list of webpage B.

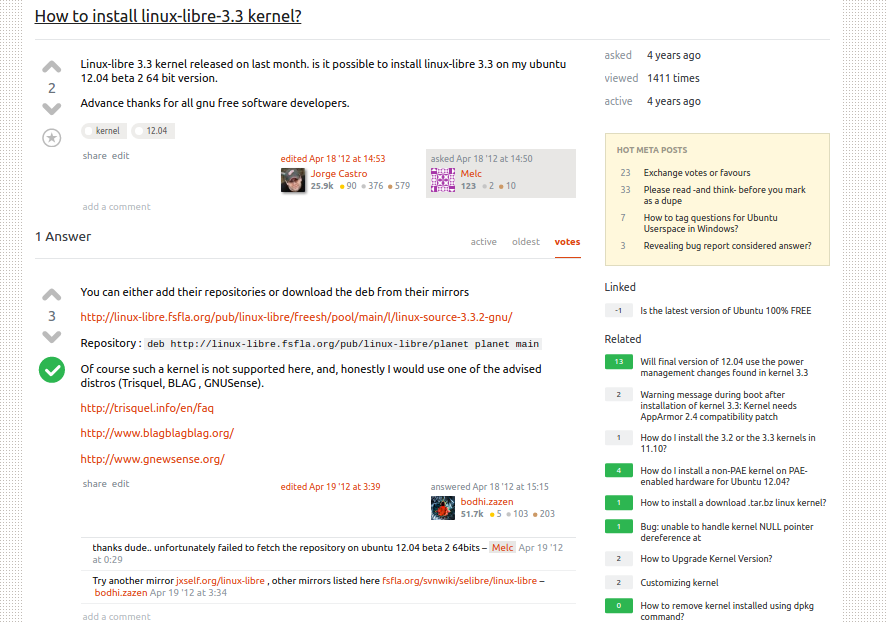
The PageRank algorithm was implemented with the damping factor set to 0.25 and the initial value of 0.1. Several iterations were performed and after several runs, we concluded that 0.25 was the best damping factor value where the values converged the fastest.

The outputs provided a glance into the popularity of webpages.

\*screenshot of output\*

The higher ranked webpages are the ones that are being linked to by a number of posts while the lower ranked ones aren’t so. But at this point, a threshold value of 0.023 indicated that a number of webpages exist which are up-voted, have numerous views and at times an accepted answer as well but are still being referred to by other pages.

All such pages become candidates for to be linked to other pages. Please refer to the image below for one such example.



We implemented Mahout Recommendation, Similarity co-occurrence and plain simple wordcount to find similar questions by tags and hot words.

The list of questions returned by this algorithm become candidates for becoming linked posts.

The moderators and the post owners can then be notified of all such posts and adding these posts as related posts would be at their discretion.