Department of Computer Science and Engineering
Walchand College of Engineering, Sangli

REPORT

T.Y in Computer Science and Engineering

Project Title

Big Data Analysis on Discussion Forums

Project Members

Name	Exam Seat No.	Roll No.	Email-Id	Contact
Vedant Sharma	2014BCS045	045	vedant.shama200@gmail.com	9422252590
Anish Joshi	2014BCS049	049	joshianish18@gmail.com	9403068497
Mukund Sudharsan	2014BCS078	078	mukund.sudharsan@walchandsangli.ac.in	9920347210

Project Guide

Prof. A. R. Surve

Academic Year

2016-2017

APPENDIX 2

BONAFIDE CERTIFICATE

This is to certify that this project report entitled **“*Big Data Analysis on Discussion Forums*”** submitted to **Walchand College of Engineering**, **Sangli** is a bonafide record of work done by Vedant Sharma, Anish Joshi and Mukund Sudharsan under my supervision from **“05/01/2017**” to “**14/04/2016”**

Prof. A. R. Surve	Dr. SMRITI BHANDARI
(Project Guide)	(HOD, Dept. Of C.S.E.)

Place: Sangli

Date: 17/11/2016

APPENDIX 3

Declaration by Authors

This is to declare that this report has been written by us. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. We aver that if any part of the report is found to be plagiarized, we are shall take full responsibility for it.

Vedant Sharma	**2014BCS045**
Anish Joshi	**2014BCS049**
Mukund Sudharsan	**2014BCS078**

Place: Sangli

Date: 15/04/2017

APPENDIX 4

Table Of Content

Technical Area(s) Explored

- Big Data Analysis.
- Hadoop System.
- Hue System.
- Hive System.
- Impala System.

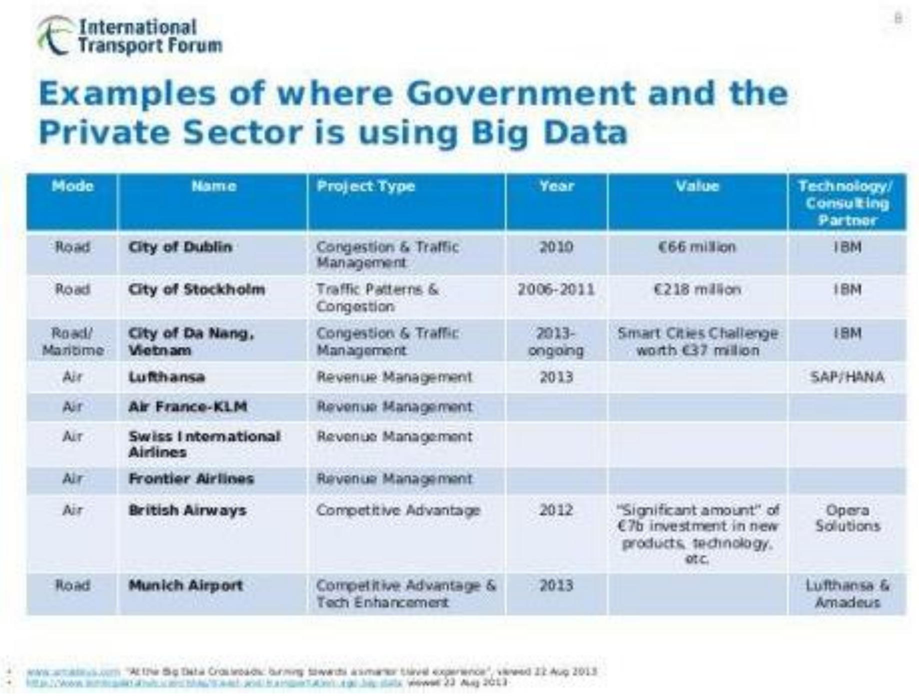
Application Domain

- Big Data Analysis
- Hadoop System.

Applications:

- A Gartner Survey for 2015 shows that more than 75% of companies are investing or are planning to invest in big data in the next two years.
- Banking and Securities
  - The Securities Exchange Commission (SEC) is using big data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity in the financial markets.
  - Amazon Prime, which is driven to provide a great customer experience by offering, video, music and Kindle books in a one-stop shop also heavily utilizes big data.
  - Spotify, an on-demand music service, uses Hadoop big data analytics, to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users.

- Education:
    - Time spent by a student when he logs onto a system
    - Student grade analysis
- Applications of big data in manufacturing and natural resources:
    - Big data allows for predictive modeling to support decision making that has been utilized to ingest and integrate large amounts of data from geospatial data, graphical data, text and temporal data.
- Applications of big data in Government
    - The Food and Drug Administration (FDA) is using big data to detect and study patterns of food-related illnesses and diseases. This allows for faster response which has led to faster treatment and less death.
- Construction Planning
    - Governments use of big data: traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)



International Transport Forum

8

Examples of where Government and the Private Sector is using Big Data

Mode	Name	Project Type	Year	Value	Technology/ Consulting Partner
Road	**City of Dublin**	Congestion & Traffic Management	2010	€66 million	IBM
Road	**City of Stockholm**	Traffic Patterns & Congestion	2006-2011	€218 million	IBM
Road/ Maritime	**City of Da Nang, Vietnam**	Congestion & Traffic Management	2013-ongoing	Smart Cities Challenge worth €37 million	IBM
Air	**Lufthansa**	Revenue Management	2013		SAP/HANA
Air	**Air France-KLM**	Revenue Management			
Air	**Swiss International Airlines**	Revenue Management			
Air	**Frontier Airlines**	Revenue Management			
Air	**British Airways**	Competitive Advantage	2012	"Significant amount" of €7b investment in new products, technology, etc.	Opera Solutions
Road	**Munich Airport**	Competitive Advantage & Tech Enhancement	2013		Lufthansa & Amadeus

- www.amadeus.com "At the Big Data Crossroads: turning towards a smarter travel experience", viewed 22 Aug 2013
- [illegible] viewed 22 Aug 2013

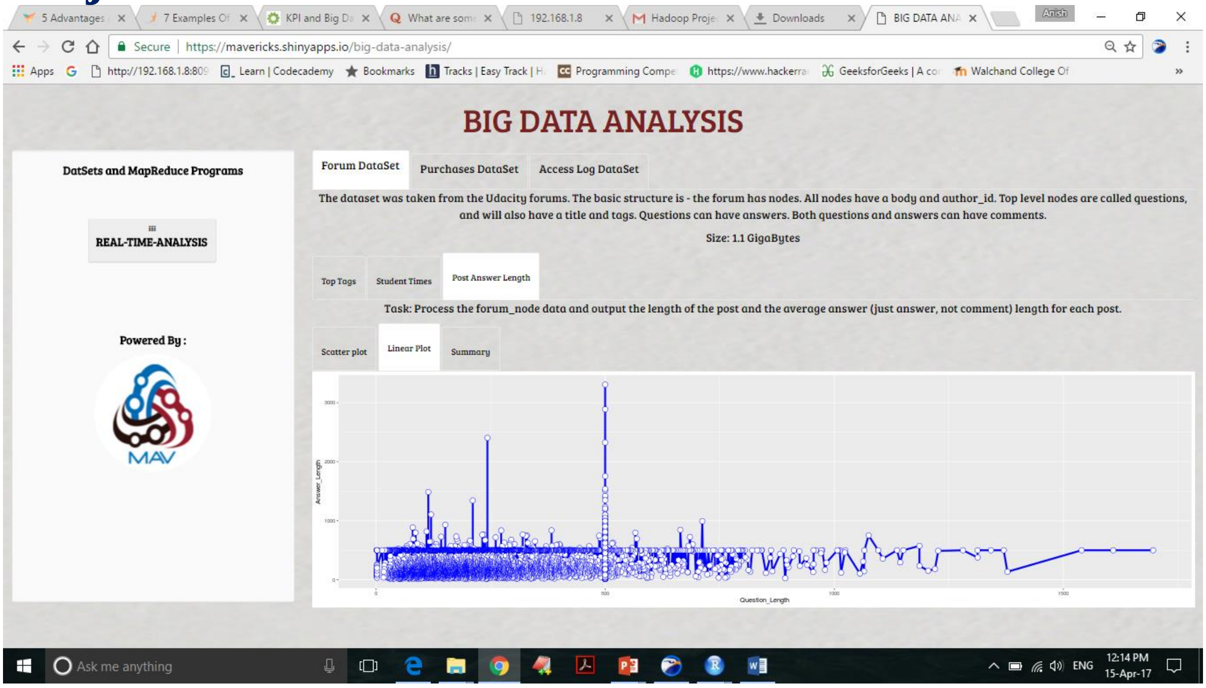
Abstract

We live in on-demand, on-command Digital universe with data prolifering by Institutions, Individuals and Machines at a very high rate. This data is categories as "Big Data" due to its sheer Volume, Variety and Velocity. Most of this data is unstructured, quasi structured or semi structured and it is heterogeneous in nature. The volume and the heterogeneity of data with the speed it is generated, makes it difficult for the present computing infrastructure to manage Big Data. Traditional data management, warehousing and analysis systems fall short of tools to analyse this data. Due to its specific nature of Big Data, it is stored in distributed file system architectures. Hadoop and HDFS by Apache is widely used for storing and managing Big Data. Analysing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. Map Reduce is widely been used for the efficient analysis of Big Data. Traditional DBMS techniques like Joins and Indexing and other techniques like graph search is used for classification and clustering of Big Data. These techniques are being adopted to be used in Map Reduce. In this paper we suggest various methods for catering to the problems in hand through Map Reduce framework over Hadoop Distributed File System (HDFS). Map Reduce is a Minimization technique which makes use of file indexing with mapping, sorting, shuffling and finally reducing. Map Reduce techniques have been studied in this paper which is implemented for Big Data analysis using HDFS.

System Configurations

Platform: Hadoop File System.
Programming Languages: Python.
External Devices Used: None.
Standards Followed: Python Coding Standards.

Project Overview

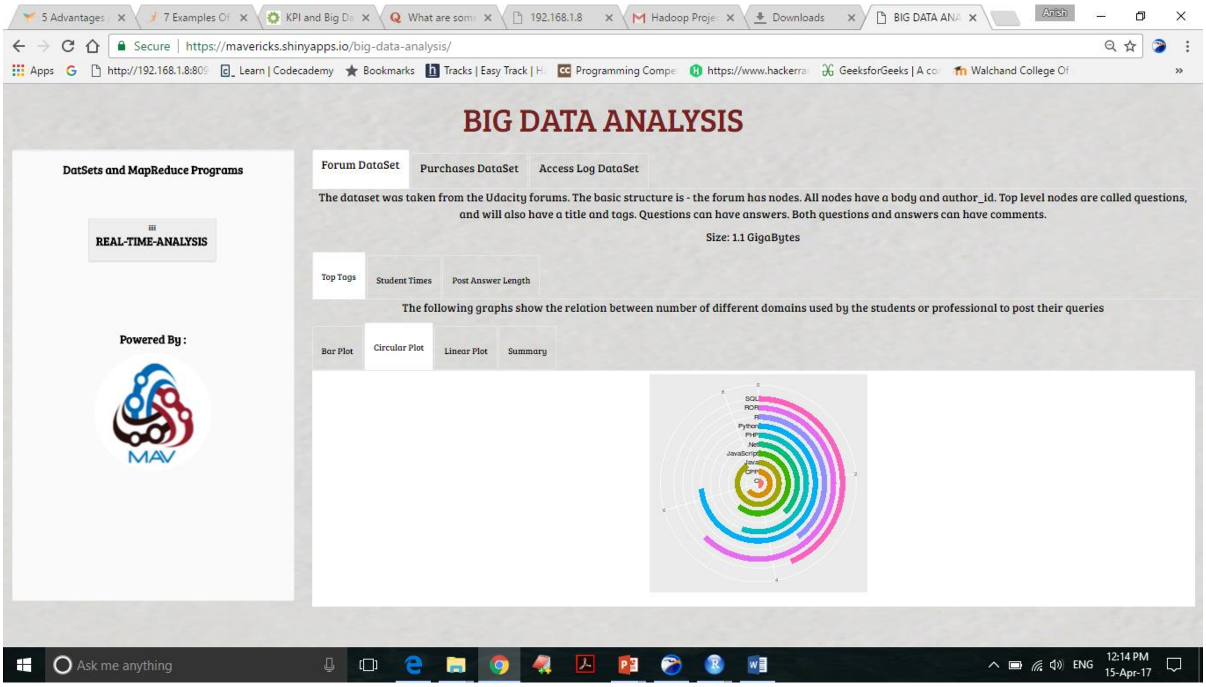


BIG DATA ANALYSIS

The dataset was taken from the Udacity forums. The basic structure is - the forum has nodes. All nodes have a body and author_id. Top level nodes are called questions, and will also have a title and tags. Questions can have answers. Both questions and answers can have comments.

Size: 1.1 GigaBytes

Task: Process the forum_node data and output the length of the post and the average answer (just answer, not comment) length for each post.

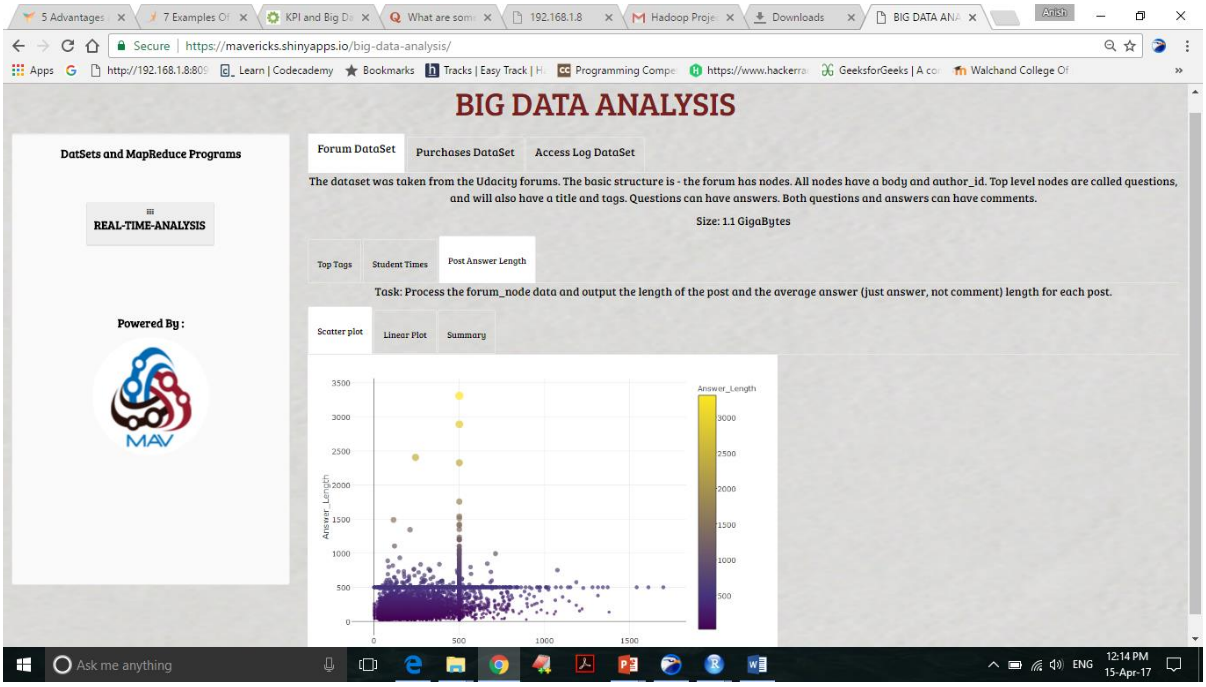


BIG DATA ANALYSIS

The dataset was taken from the Udacity forums. The basic structure is - the forum has nodes. All nodes have a body and author_id. Top level nodes are called questions, and will also have a title and tags. Questions can have answers. Both questions and answers can have comments.

Size: 1.1 GigaBytes

The following graphs show the relation between number of different domains used by the students or professional to post their queries



BIG DATA ANALYSIS

DatSets and MapReduce Programs

REAL-TIME-ANALYSIS

Powered By :

MAV

Forum DataSet | Purchases DataSet | Access Log DataSet

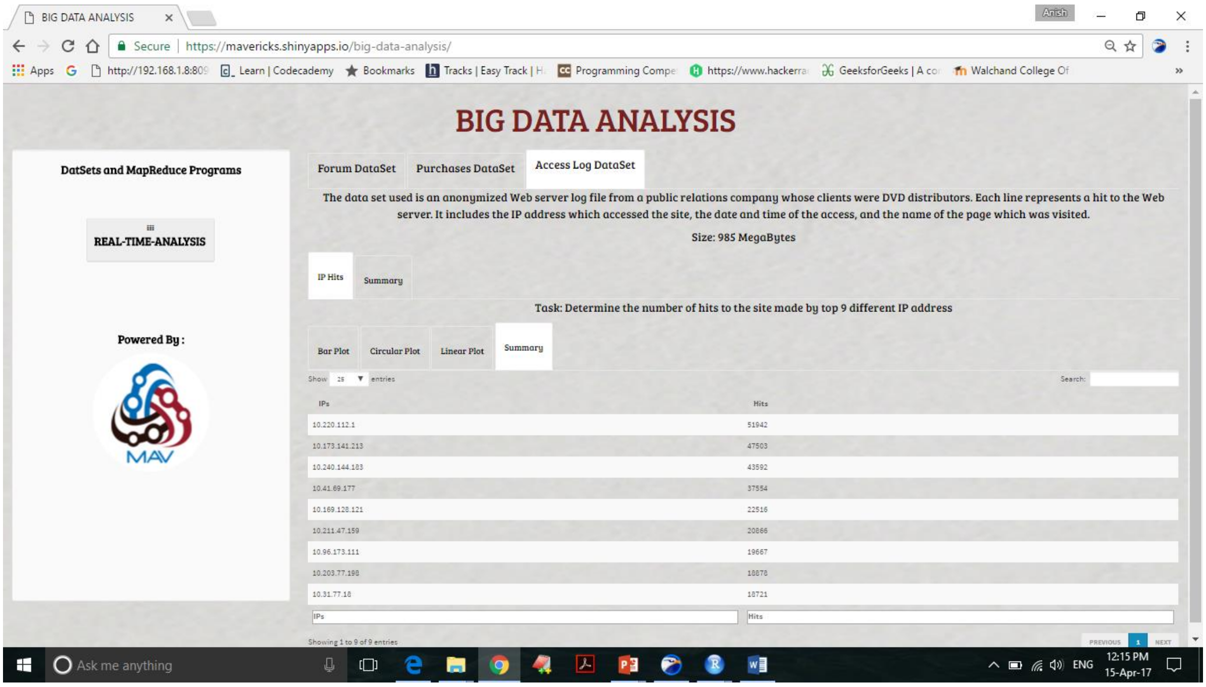
The dataset was taken from the Udacity forums. The basic structure is - the forum has nodes. All nodes have a body and author_id. Top level nodes are called questions, and will also have a title and tags. Questions can have answers. Both questions and answers can have comments.

Size: 1.1 GigaBytes

Top Tags | Student Times | Post Answer Length

Task: Process the forum_node data and output the length of the post and the average answer (just answer, not comment) length for each post.

Scatter plot | Linear Plot | Summary



BIG DATA ANALYSIS

DatSets and MapReduce Programs

REAL-TIME-ANALYSIS

Powered By :

MAV

Forum DataSet | Purchases DataSet | Access Log DataSet

The data set used is an anonymized Web server log file from a public relations company whose clients were DVD distributors. Each line represents a hit to the Web server. It includes the IP address which accessed the site, the date and time of the access, and the name of the page which was visited.

Size: 985 MegaBytes

IP Hits | Summary

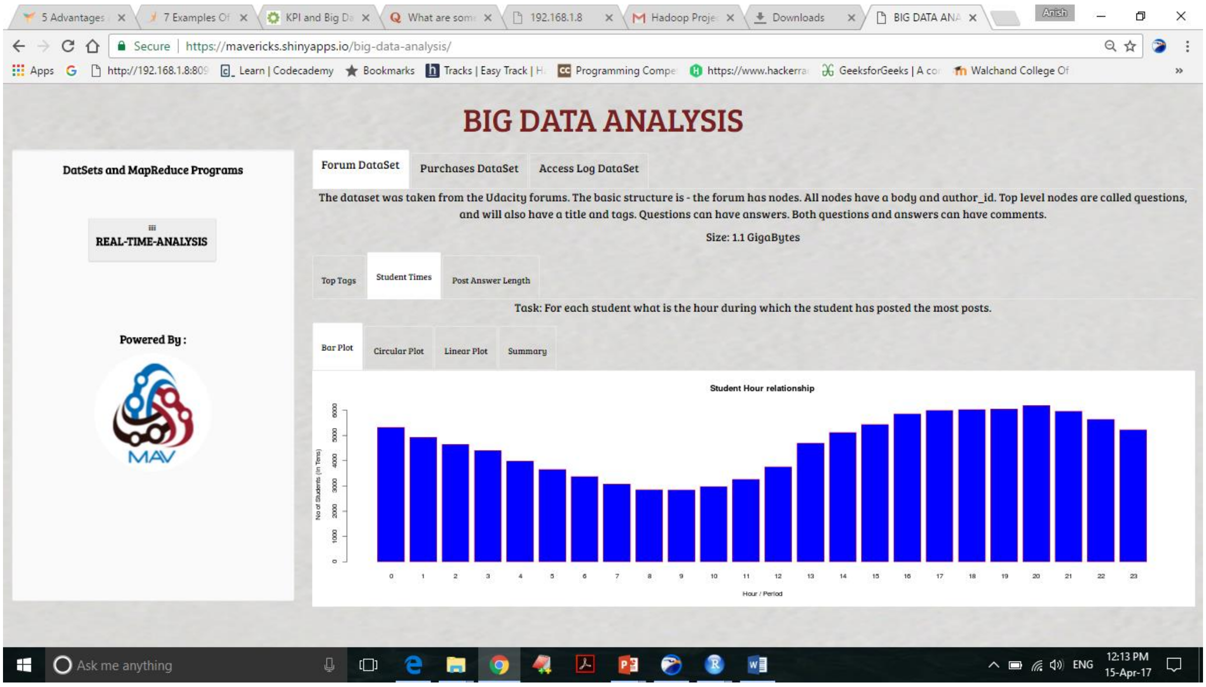
Task: Determine the number of hits to the site made by top 9 different IP address

Bar Plot | Circular Plot | Linear Plot | Summary

Show 25 entries Search:

IPs	Hits
10.220.112.1	51942
10.173.141.213	47503
10.240.144.183	43592
10.41.69.177	37554
10.169.128.121	22516
10.211.47.159	20866
10.96.173.111	19667
10.203.77.198	18878
10.31.77.18	18721
IPs	Hits

Showing 1 to 9 of 9 entries PREVIOUS 1 NEXT



BIG DATA ANALYSIS

DatSets and MapReduce Programs

REAL-TIME-ANALYSIS

Powered By :

Forum DataSet | Purchases DataSet | Access Log DataSet

The dataset was taken from the Udacity forums. The basic structure is - the forum has nodes. All nodes have a body and author_id. Top level nodes are called questions, and will also have a title and tags. Questions can have answers. Both questions and answers can have comments.

Size: 1.1 GigaBytes

Top Tags | Student Times | Post Answer Length

Task: For each student what is the hour during which the student has posted the most posts.

Bar Plot | Circular Plot | Linear Plot | Summary

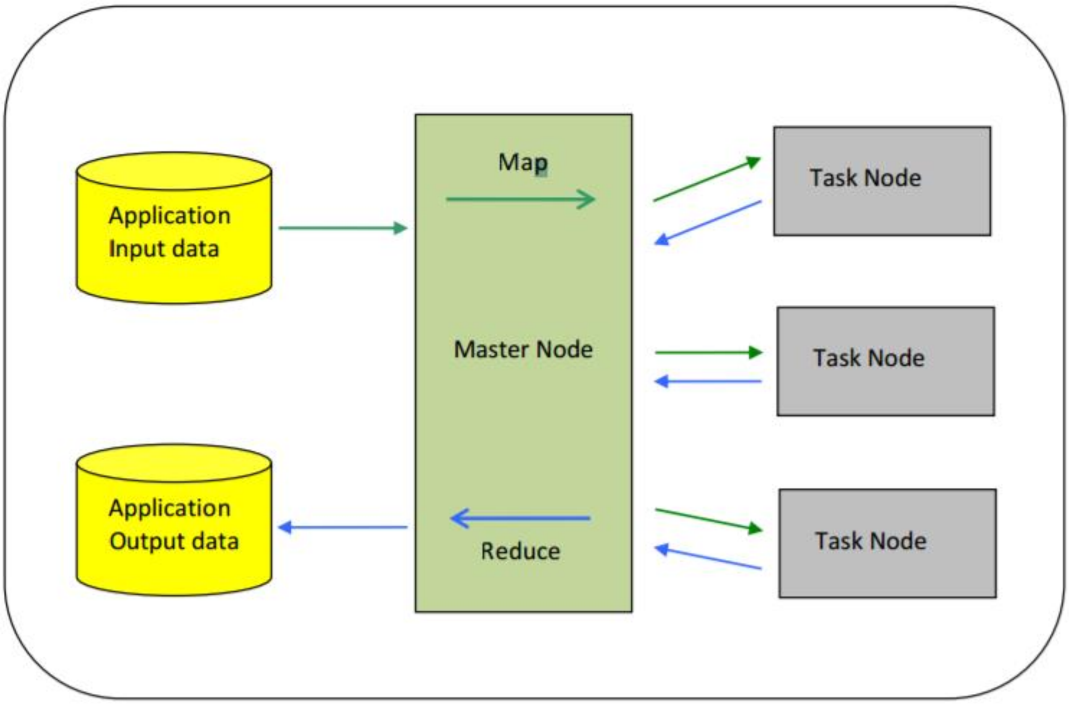
Objective

- To implement algorithm for automatic classification of text into positive, negative or neutral.

- Big data Analysis to determine the attitude of the mass is positive , negative towards the subject of interest.

- Graphical Representation of the sentiment in the form of various charts.

Algorithm Implemented



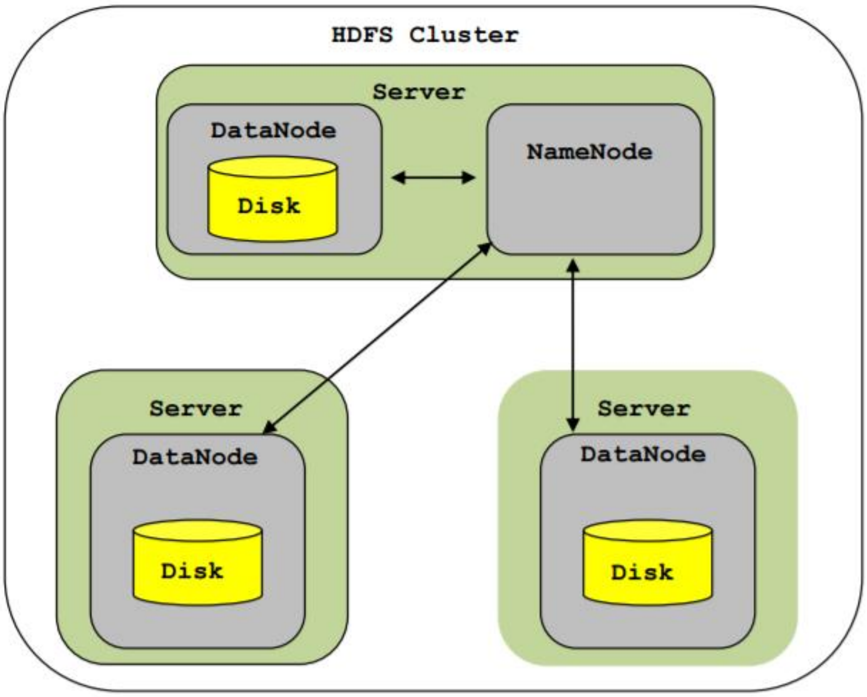
Hadoop Distributed File System (HDFS) HDFS is a subproject of the Apache Hadoop project. Hadoop uses HDFS to achieve high data throughput access. HDFS is built using Java and runs on top of local file system. This was designed to process, read and write large data files with size ranging from Terabytes to Petabytes. An ideal file size is a multiple of 64 MB. HDFS stores large files across multiple commodity machines. Using HDFS you can easily access and store large data files split across multiple computers, as if you were accessing or storing local files. High reliability is gained by replicating the data across multiple nodes and hence does not require expensive hardware infrastructure like RAID storage on the nodes. The default replication value is 3 and hence data is replicated on three nodes

One of the advantages of using HDFS is data awareness between JobTracker and TaskTracker. The JobTracker schedules the map and reduce jobs to TaskTrackers with an awareness of data location. For example: Assume that node A contains data (a, b, c, d) and node B contains data (x, y, z). The JobTracker will schedule node A to perform map/reduce tasks on (a, b, c, d) and node B would be scheduled to perform map/reduce tasks on (x, y, z).This will greatly reduce the amount of traffic that goes over the network and prevents unnecessary data transfer. Brining the data to the place where map function resides is more expensive and time-consuming than letting the map function execute at the place where the data resides. This advantage is not available in any other file systems. Hadoop uses several types of nodes to form a proper reliable cluster.

The NameNode is the major part of the HDFS file system. Its main goal is to maintain the directory tree of all the files in the file system and tracks where across the cluster the file data is stored. It does not store the data of these files itself. Applications interact with NameNode to create copy, move and delete a file in the HDFS file system. Apart from this, a DataNode stores data in the HDFS file system.

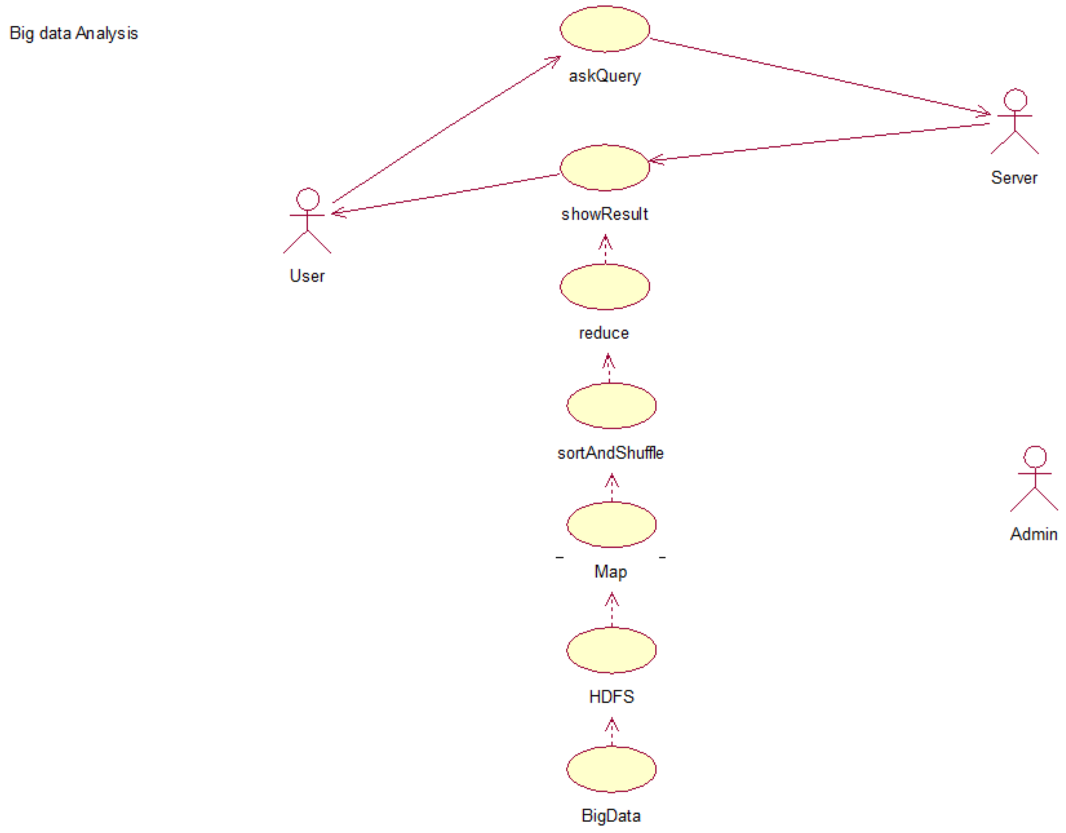
On Hadoop system startup, a DataNode connects to the NameNode and waits until the service is up and running. The DataNode will respond to the request from the NameNode for file system operations. Applications can directly talk to a DataNode once the NameNode has provided sufficient information about the location of the data. In this process, the map/reduce tasks are performed by TaskTracker node near a DataNode. One of the important performance tunings is to have the TaskTracker instance deployed on the same server where the DataNode instance exists. This will allow MapReduce operations to be performed close to the data. Typically, the HDFS file system uses TCP/IP layer for communication.

HDFS Architecture

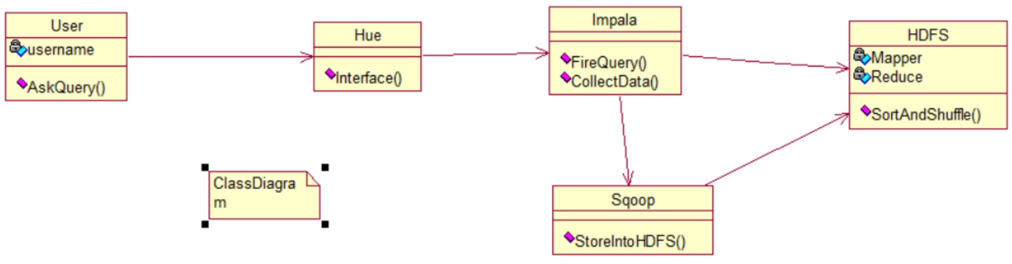


Project Design

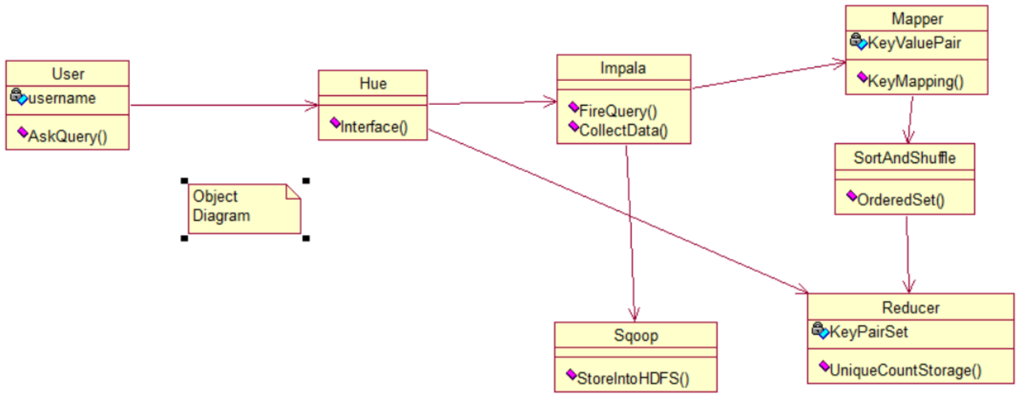
Use Case Diagram:



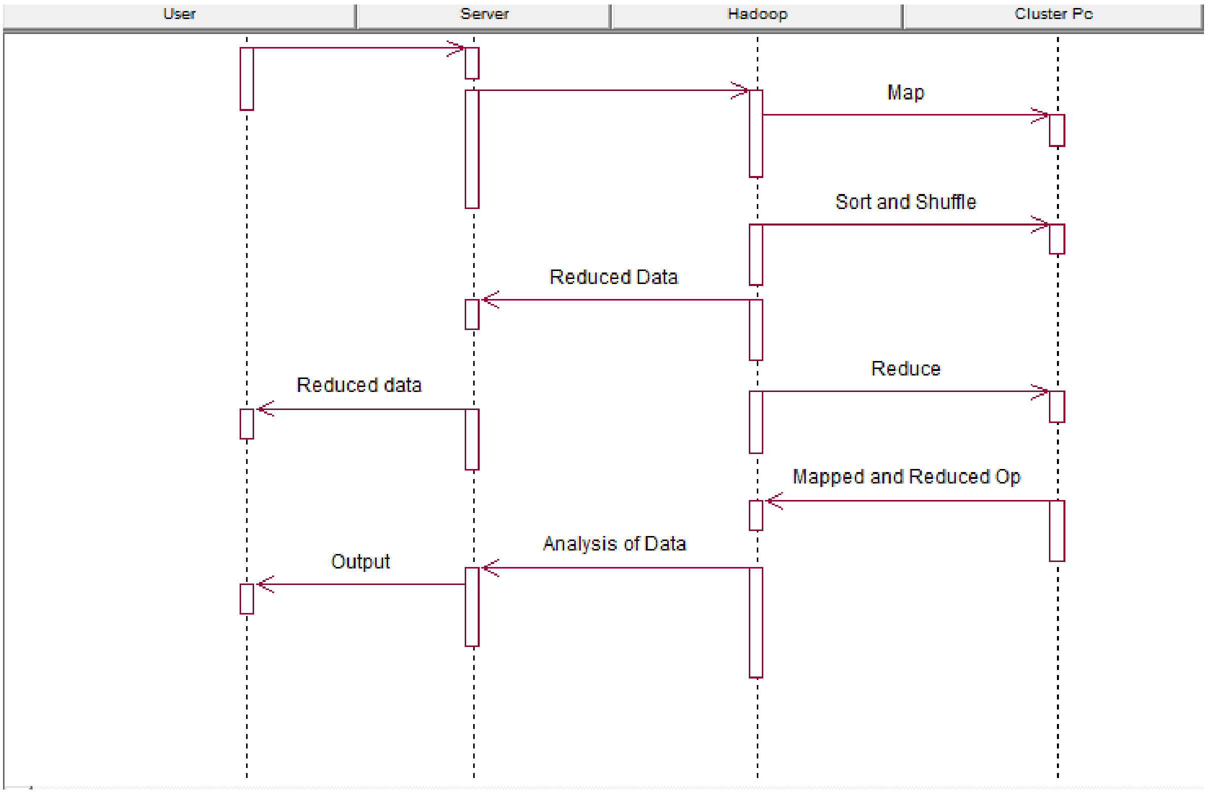
Class Diagram:



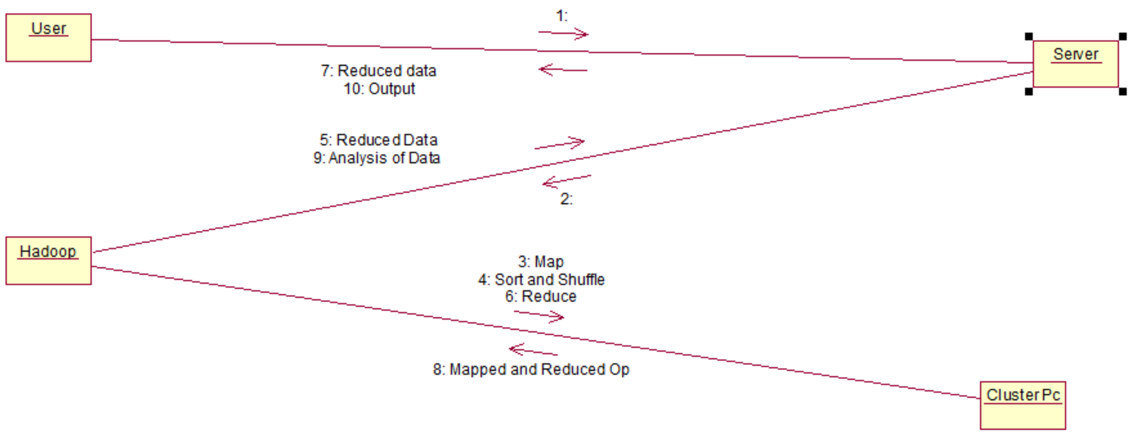
Object Diagram:



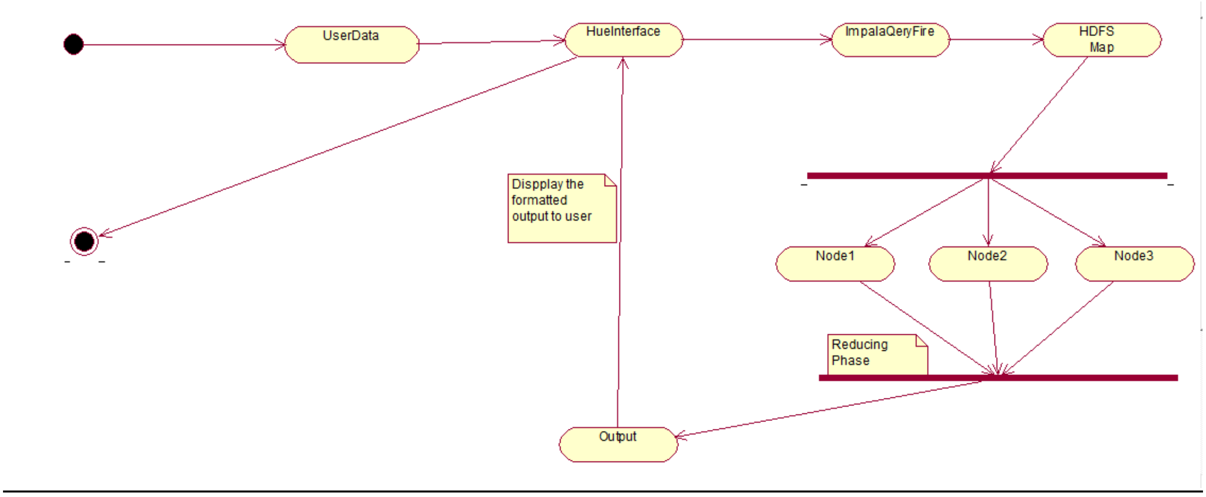
Sequence Diagram:



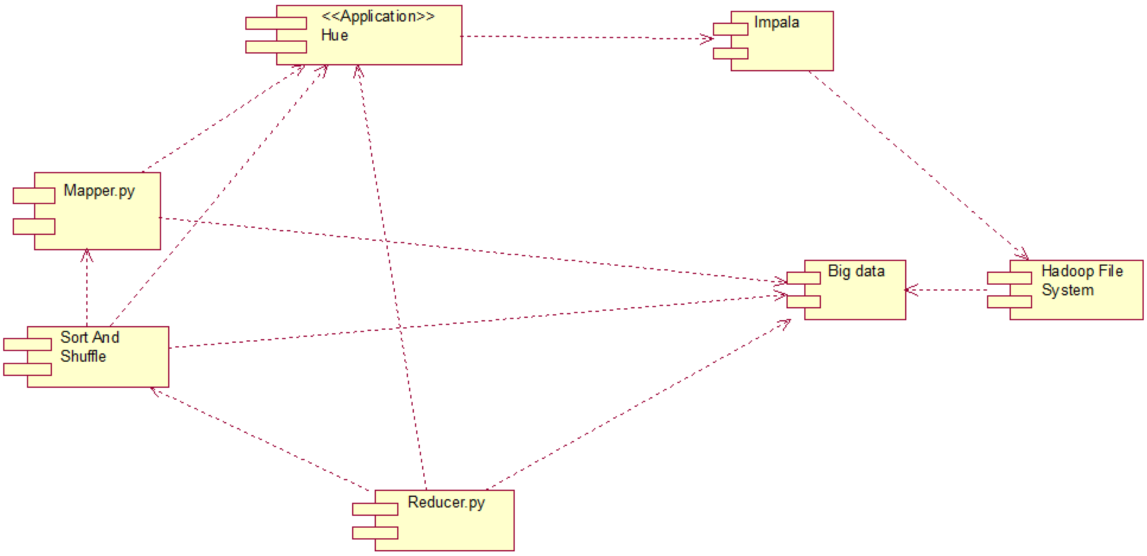
Collaboration Diagram:



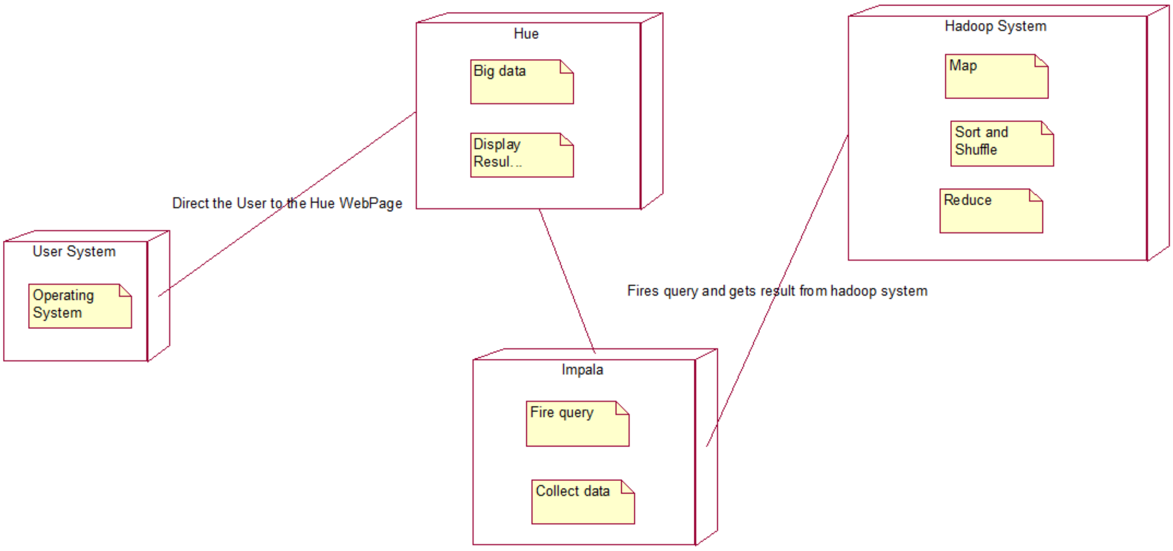
Activity Diagram:



Component Diagram:



Deployment Diagram:



PROJECT MANAGEMENT:

Gantt Chart:

Number	Milestone Name	Milestone Description	Timeline Number of weeks required to complete the milestone
1	Requirement Specification	A requirement specification document should be delivered.	1 week
2	Technology Familiarization	Understanding of technology. Each person should get themselves as expert in each of the technology and should arrange a half day session to share the info and come up with a document for reference	Working 3 week
3	System Setup	Setup up dev environment with the database servlet engine, also setup a test environment	1 week
4	Design	A high level architecture diagram and detailed design of all the modules. Also a datadictionary document should be delivered	2 week