

YOUTUBE COMMENT ANALYSIS

SYSTEM REQUIREMENTS SPECIFICATION

PROJECT TOPIC

“YOUTUBE COMMENT ANALYSIS”

SUBMITTED BY:

RADHIKA MITTAL

VRUSHALI KADAM

AT

TECHDATA SOLUTIONS

SUBMITTED DURING INTERNSHIP PERIOD

NOV'16 - FEB'17

YOUTUBE COMMENT ANALYSIS

INDEX

Chapter No.	Topic	Page No.
1	INTRODUCTION 1.1 Abstract 1.2 Problem Definition 1.3 Purpose and Scope 1.4 Literature and Survey	5-13 6 7 8 9
2	SYSTEM ANALYSIS 2.1 Feasibility Study 2.2 Information Gathering	14-15 15 15
3	SYSTEM PLANNING AND SCHEDULING 3.1 Gantt Chart	16-18 18
4	SYSTEM DESIGN 4.1 Architecture 4.2 Use-Case Diagram 4.3 Sequence Diagram 4.4 Activity Diagram	19-24 20 22 23 24

YOUTUBE COMMENT ANALYSIS

5	SYSTEM IMPLEMENTATION 5.1 System Requirement 5.2 System Implementation procedure 5.3 Screenshots	25-44 26 27 40
6	SYSTEM TESTING 6.1 Methodology adopted for testing 6.2 Testing the system	45-47 46 47
7	CONCLUSION 7.1 Conclusion 7.2 Limitations of the system	48-50 49 50
8	FUTURE ENHANCEMENTS	51
9	ANNEXURE	53

INTRODUCTION

YOUTUBE COMMENT ANALYSIS

1.1 ABSTRACT:

"ANALYSIS OF YOUTUBE COMMENTS FOR YOUTUBERS"

- With the help of Big Data analytics and hadoop framework we intend to help the Youtubers having more than 5K subscribers by implementing a mechanism/software to serve as a solution to the above problem statement.
- We plan on pulling the comments related to a particular video using the video id and channel id to obtain highly requested video topics per youtuber.
- The result of the comment analysis will be stored into hive which is converting single queries into multiple jobs due to which faster processing of data is possible which will, then be displayed in the form of graphs/ charts.
- To design a web based application that makes every YouTuber's life easy going by providing them with a graphical representation of the analysis more than 10k comments for a particular video at a glance rather than going through them manually.
- Target Audience: Daily vloggers, Youtube Creators, Beauty bloggers, Fashion Bloggers, Comedy, Entertainment channel owners.
- These creators often ask their audience for suggestions based on what they would like to watch for their next video it can be a 'react to', 'collab with' or a 'q and a' based video. The most requested video idea by their respective viewers/subscribers will be brought to the creator's attention by giving them a graphical representation of the analytics performed on their comments.

YOUTUBE COMMENT ANALYSIS

1.2 PROBLEM DEFINATION

LIMITATIONS OF EXISTING SYSTEMS

Current Process:

Youtube provides a platform to more than 10 million youtube creators who create weekly videos and upload them. Over a period of time these videos quickly escalate to over a million views and comments.

With such a humongous amount of data it is nearly impossible and time consuming to read and contemplate each and every comment.

The current system provides analysis based on views, location of the subscribers, gender etc. However, there is no mechanism for YouTube as such to analyze the types of comments a particular video gets.

They can also keep a track of their subscribers but there is no mechanism to analyze the sentiments like sad, bad, good, worst, happy of the viewer and kind of comments at one glance.

A recent feature: Community Tab has been added by Youtube developers as an interactive platform for the creator and their respective subscribers. They can communicate with each other by using this feature yet, since there are huge number of subscribers for a creator's channel it is not possible for them to go through every single conversation manually and there is no analysis environment provided for the same.

YOUTUBE COMMENT ANALYSIS

1.3 PURPOSE AND SCOPE

Purpose:

Since YOUTUBE has become a profession for many as a way to connect to the audience with similar interests be it comedy, entertainment, education.

Youtubers put in a lot of effort to create videos and put forth their perspective in a visual manner. This system will make their life easy going and they will be able to connect with their audience even better based on their requests and choices of interest.

Going through more than 10k comments, manually consumes a whole lot of time and effort, providing analysis for the comments will solve the listed limitations.

Thus, youtubers will be able to invest their effort, time, resources in the correct direction and expect their upcoming videos to go viral because every youtuber would like to expand their community. Using big data analytics such huge chunk of data will be converted into a graphical representation which will in turn improve the efficiency for getting the results according to the most requested topics.

Scope:

Taking huge amount of data i.e. youtube comments then analyze it using big data by developing algorithms for pulling data from youtube channels which will generate reports in less time.

Give a better analytical view of comments to the youtubers owning a youtube channel.

Help the youtuber to create the most requested video idea by their respective viewers/subscribers.

YOUTUBE COMMENT ANALYSIS

1.4 LITREATURE SURVEY

Big Data

What is Big Data?

Big data is an item of jargon in today's world which is used to describe humongous amount of both structured and unstructured data which is nearly impossible or rather a tedious task to process by traditional databases and software techniques.

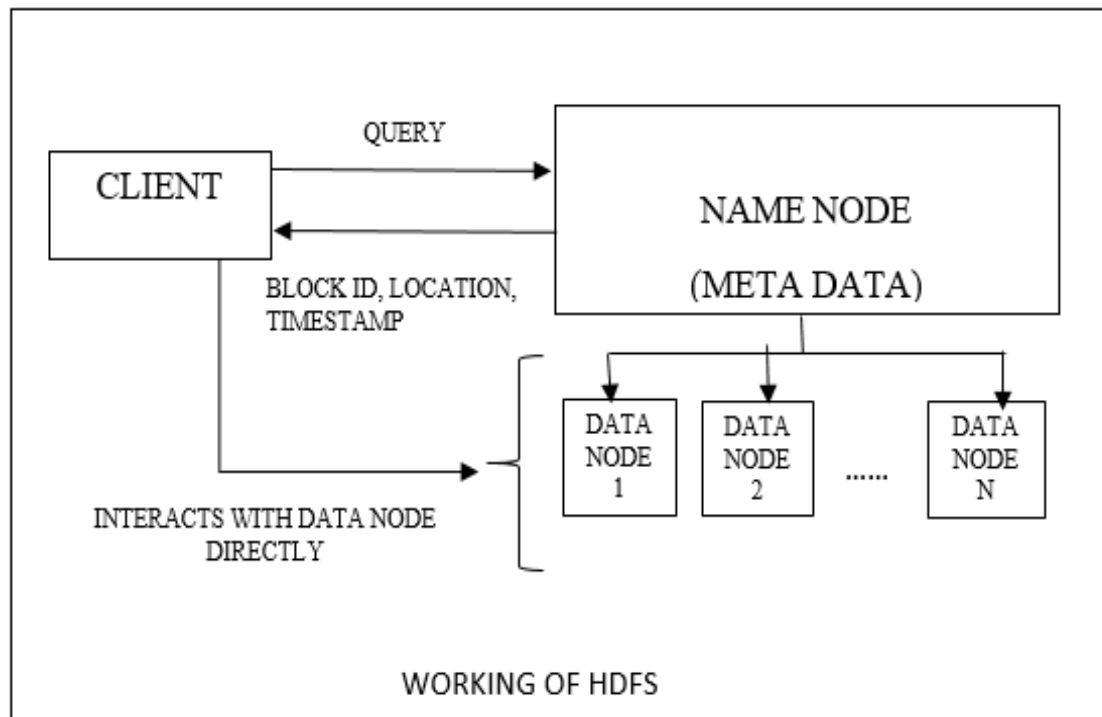
Big data is characterized by volume, velocity and variety, wherein velocity refers to the processing capacity within a considerable time, volume refers to huge chunks of data and variety refers to the data heterogeneity i.e. multiple datatypes, in which the data is stored.

A fourth characteristic considered by companies is value which refers to the meaningful insights provided by Big Data that can help make the analysis and predictive analysis an easy task. One can consider the data generated by sensors and intelligent systems as Big Data.

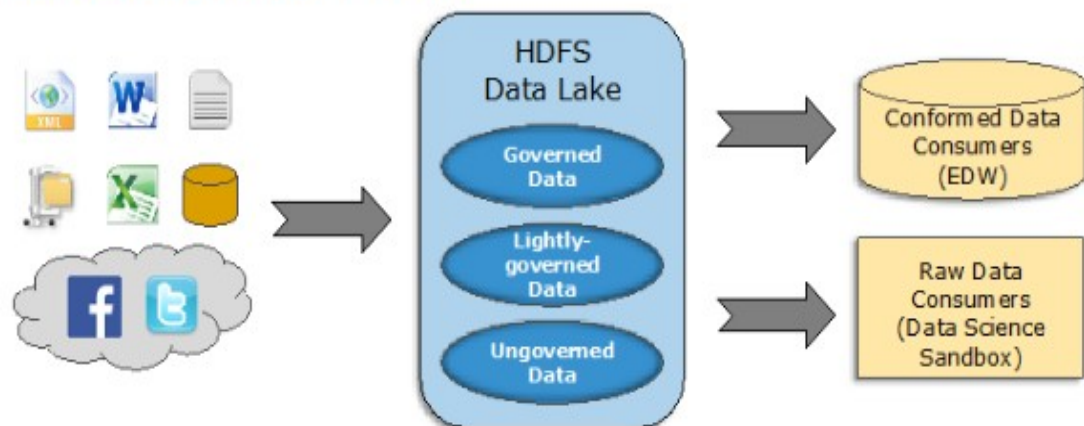
The key techniques that make BIG DATA processing in hardly sometime a possibility are:

- Data is distributed across multiple nodes.
- Big data systems moves the application to the processing nodes prior to the job execution.
- Each node processes independently (Local processing).
- The data is read sequentially rather than randomly.

YOUTUBE COMMENT ANALYSIS



Big Data: Data Flow



YOUTUBE COMMENT ANALYSIS

HTML

A web application developed with HTML(Hyper Text Markup Language) which links multiple pages and provides graphical representation of the analysis performed.

PYTHON

Python is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and first released in 1991.

An interpreted language, Python has a design philosophy which emphasizes code readability ; notably using whitespace indentation to delimit code blocks rather than curly braces or keywords, and a syntax which allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale.

CLOUDERA

Cloudera's open-source Apache Hadoop distribution, CDH (Cloudera Distribution Including Apache Hadoop), targets enterprise-class deployments of that technology. Cloudera says that more than 50% of its engineering output is donated upstream to the various Apache-licensed open source projects (Apache Hive, Apache Avro, Apache HBase, and so on) that combine to form the Hadoop platform.

CLOUDERA QUICKSTART VIRTUAL MACHINE (CDH 5.8)

Cloudera provides a scalable, flexible, integrated platform that makes it easy to manage rapidly increasing volumes and varieties of data in your enterprise. Cloudera products and solutions enable you to deploy and manage Apache Hadoop and related projects, manipulate and analyze your data, and keep that data secure and protected.

YOUTUBE COMMENT ANALYSIS

CDH 5.8 provides multiple integrated applications that can be used to analyze big chunks of data and present it graphically to the user. Our project uses two of these are listed as follows:

- HUE
- SOLR
- HIVE

HUE

Hue is an open source Web interface for analyzing data with any Apache Hadoop. The Hue team works with upstream Apache Hadoop and provides Hue releases on its website. HUE lets you easily interact with Apache Hadoop so you can focus on productivity. HUE is also present in some major Hadoop distributions (CDH, HDP, MapR) and demo VM.

SOLR

Apache SOLR is the open source platform for searches of data stored in HDFS in Hadoop. SOLR powers the search and navigation features of many of the world's largest Internet sites, enabling powerful full-text search and near real-time indexing. Whether users search for tabular, text, geo-location or sensor data in Hadoop, they find it quickly with Apache SOLR.

HIVE

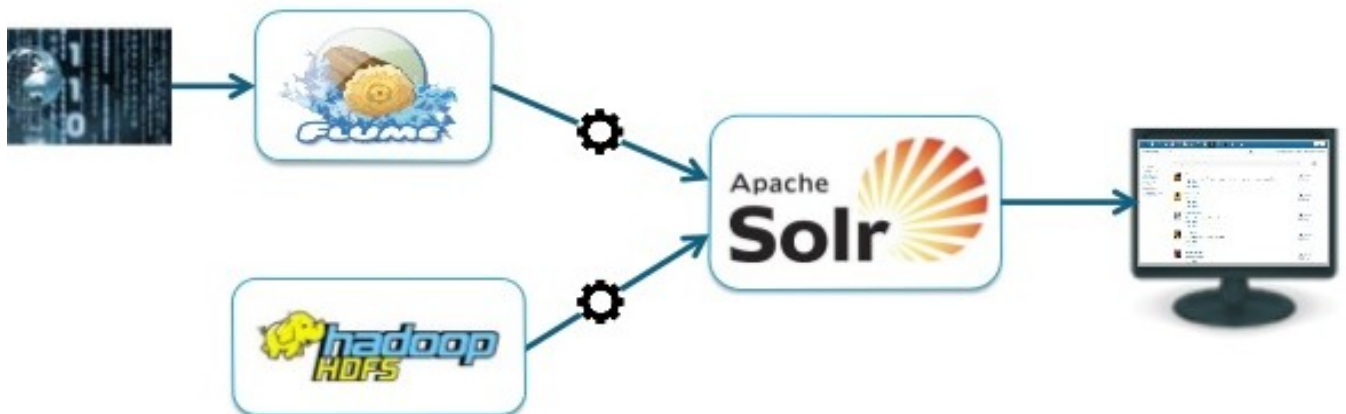
Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Hive gives an SQL -like interface to query data stored in various databases and file systems that integrate with Hadoop. Hive provides the necessary SQL abstraction to integrate SQL-like Queries (HiveQL) into the underlying Java API without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive supports easy portability of SQL-based application to Hadoop

YOUTUBE COMMENT ANALYSIS

Cloudera Search



- Real-time, scalable indexing
- Load any type of data
- Text and faceted searching



YOUTUBE COMMENT ANALYSIS

SYSTEM ANALYSIS

YOUTUBE COMMENT ANALYSIS

2.1 FEASIBILITY STUDY

The feasibility study proposes one or more conceptual solutions to the problem set of the project. The objective is assessing feasibility is to determine whether a development project has a reasonable chance of success. The following are the criteria that are considered to confirm project feasibility:

Technical Feasibility:

At first it is necessary to check whether the proposed system is technically feasible or not and to determine the technologies and skills necessary to carry out the implementation of the system. If they are not available, then find the solution to obtain them. After studying the technical requirements of the system the only problem was that the company did not have a machine with the capability to act as the database server. However, the company has outsourced this to third party data center for other internal projects and proposes to use the same server for this application.

Operational Feasibility:

The system has been specifically designed to be user friendly so that the system is easy to use. The end users of this system are administrator and user which do the necessary editing on the site depending upon the privileges allocated to them. The time will not be wasted in installation of compilers.

Economic Feasibility:

The system being having compilers already present in the website, the user doesn't have any cost of buying the compilers or any other IDE (Integrated Development Environment). Also it will be portable, so need to carry any storage devices.

2.2 INFORMATION GATHERING

The information is gathered from interviewing the Professor and from the websites. It took some days in doing the research for the project by using various websites listed in annexure and deciding about the making of it. Thus, the information was gathered of using various tools. All the information gathered is by interviewing and doing research.

SYSTEM PLANNING AND
SCHEDULING

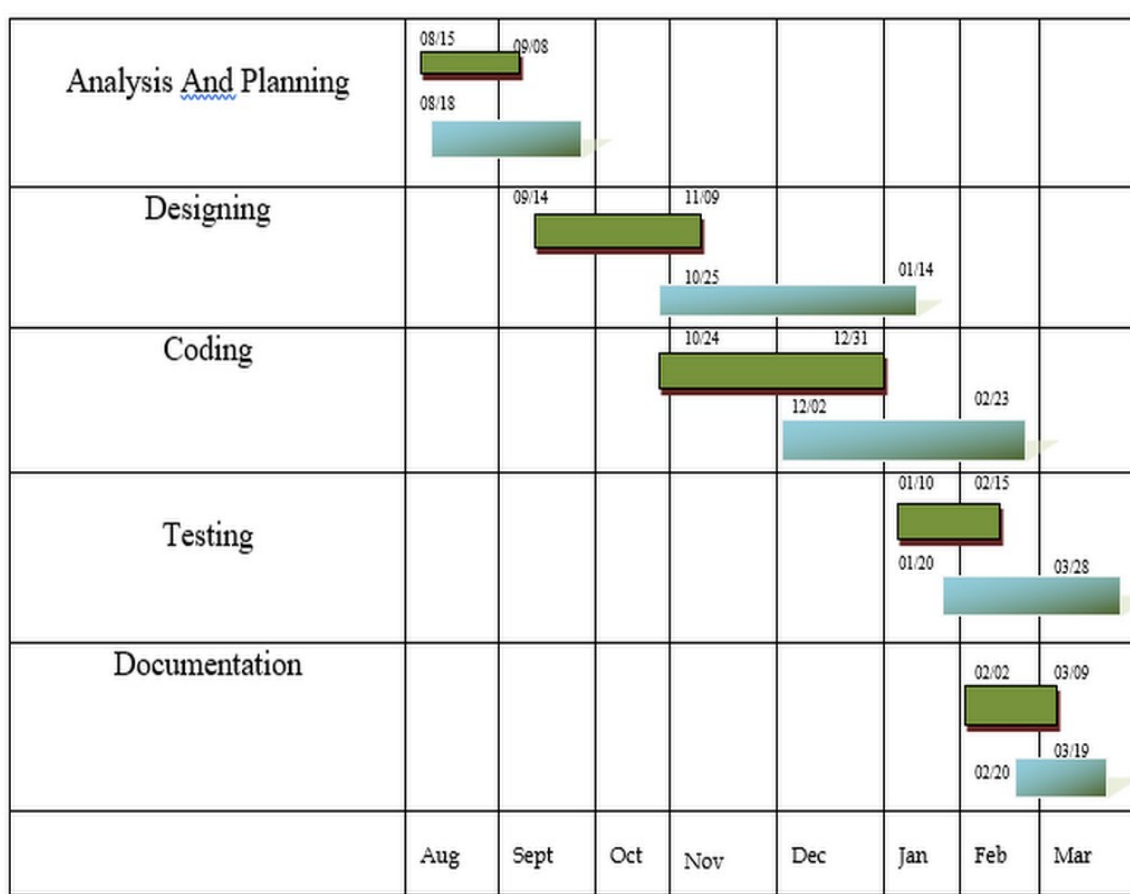
YOUTUBE COMMENT ANALYSIS

Project Schedule and Gantt Chart Project Schedule

Sr. No	Contents	Proposed Dates	Submission Dates	Teacher's sign	Remark
1.	Investigation	15/11/2016	15/11/2016	-	-
	Project fixing	20/12/2016	22/12/2016	-	
	Synopsis	04/01/2017	04/01/2017	-	
2.	Analysis	17/01/2017	17/01/2017	-	-
	Project History	15/03/2017	15/03/2017	-	
	Objective And Scope of project		16/03/2017		
	Problems with existing system		22/03/2017		
	Advantage of Proposed System		24/03/2017		

YOUTUBE COMMENT ANALYSIS

- Gantt chart is a project planning tool that can be used to represent the timing tasks required to complete the project.
- It provides a graphical illustration of a schedule that helps to plan, coordinate and track specific tasks in the project

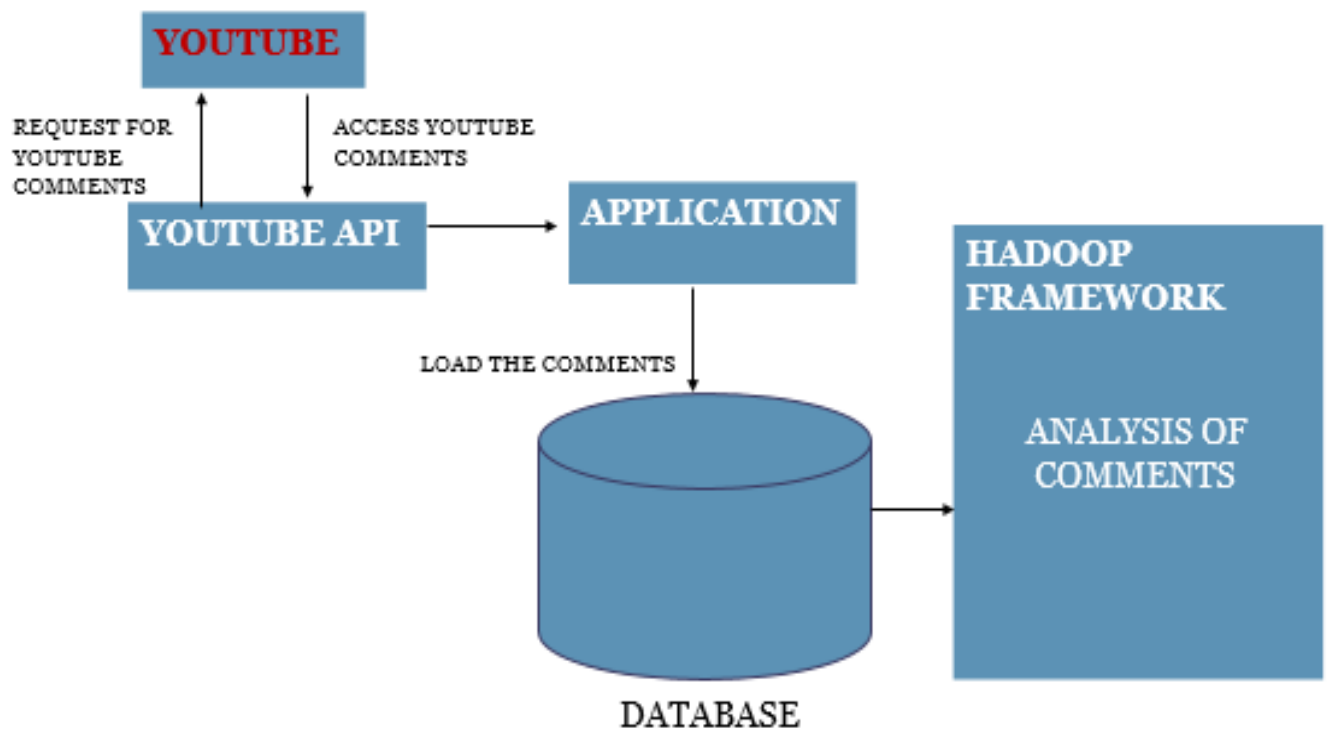


SYSTEM DESIGN

YOUTUBE COMMENT ANALYSIS

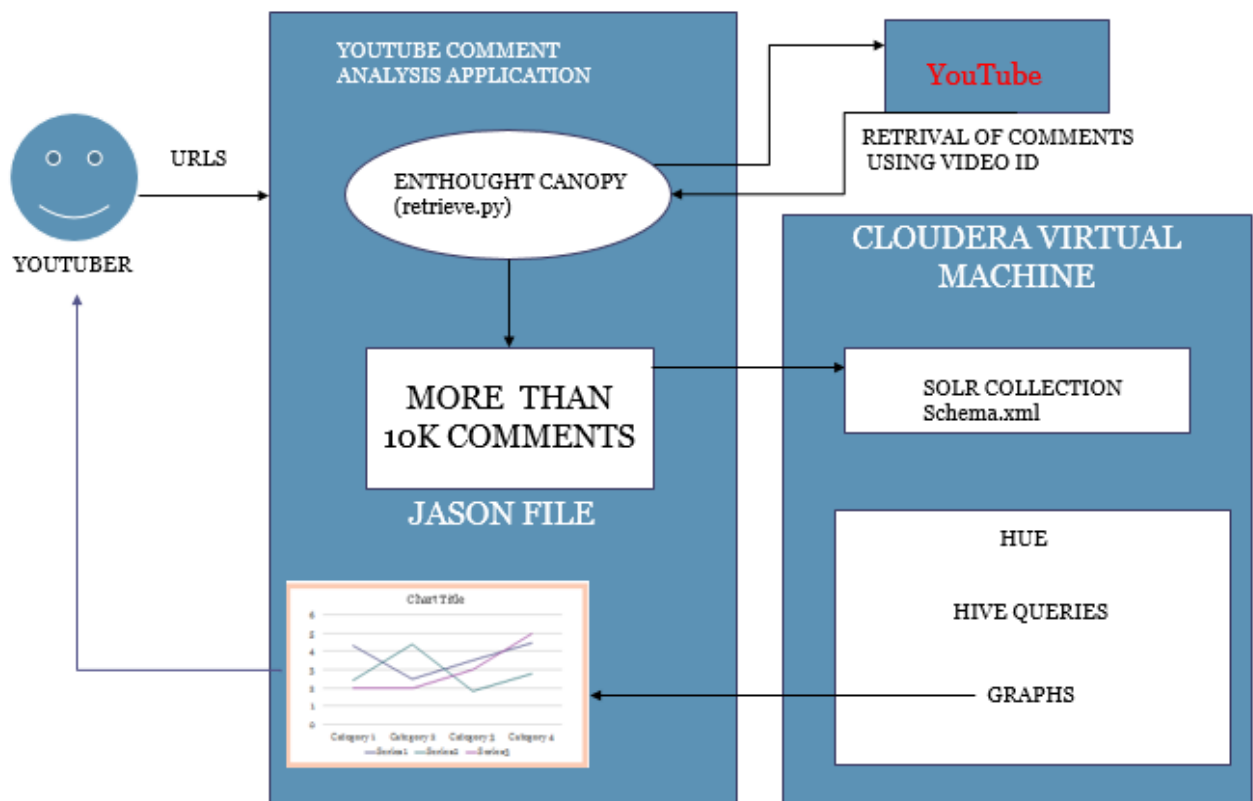
4.1 ARCHITECTURE DIAGRAM

THEORETICAL DIAGRAM



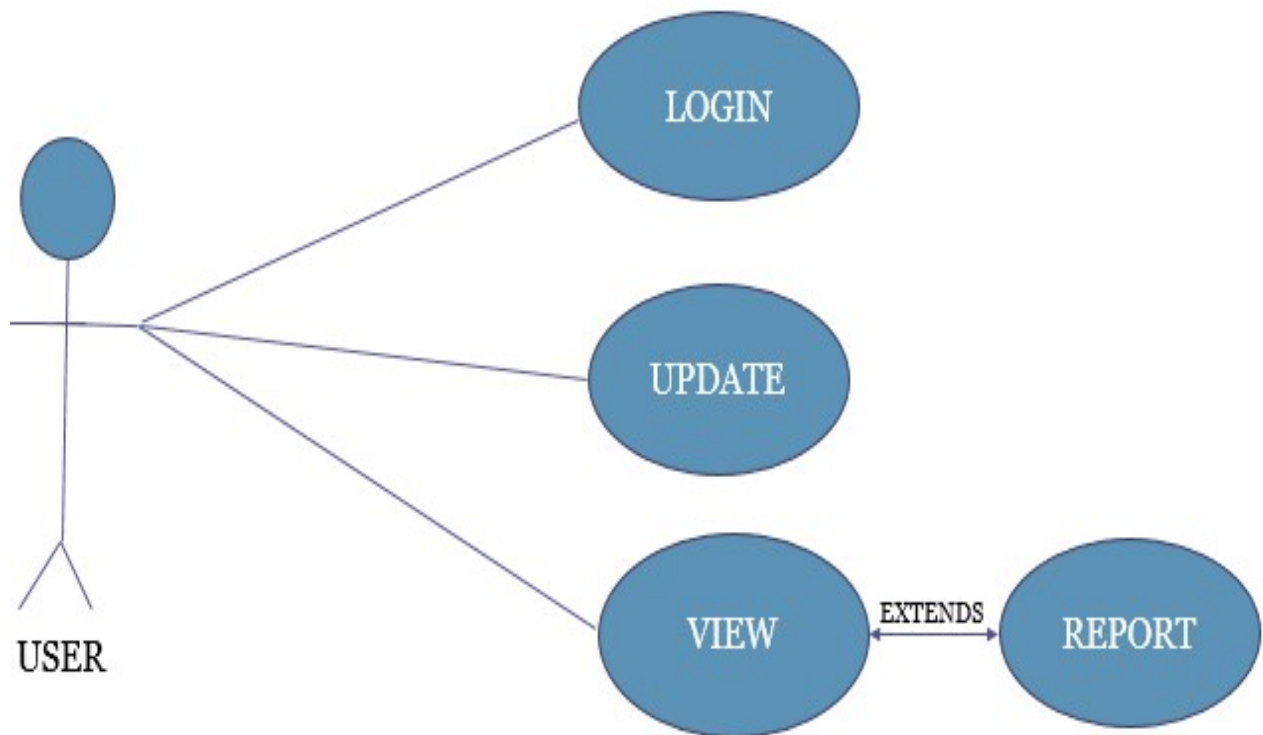
YOUTUBE COMMENT ANALYSIS

PRACTICAL DIAGRAM



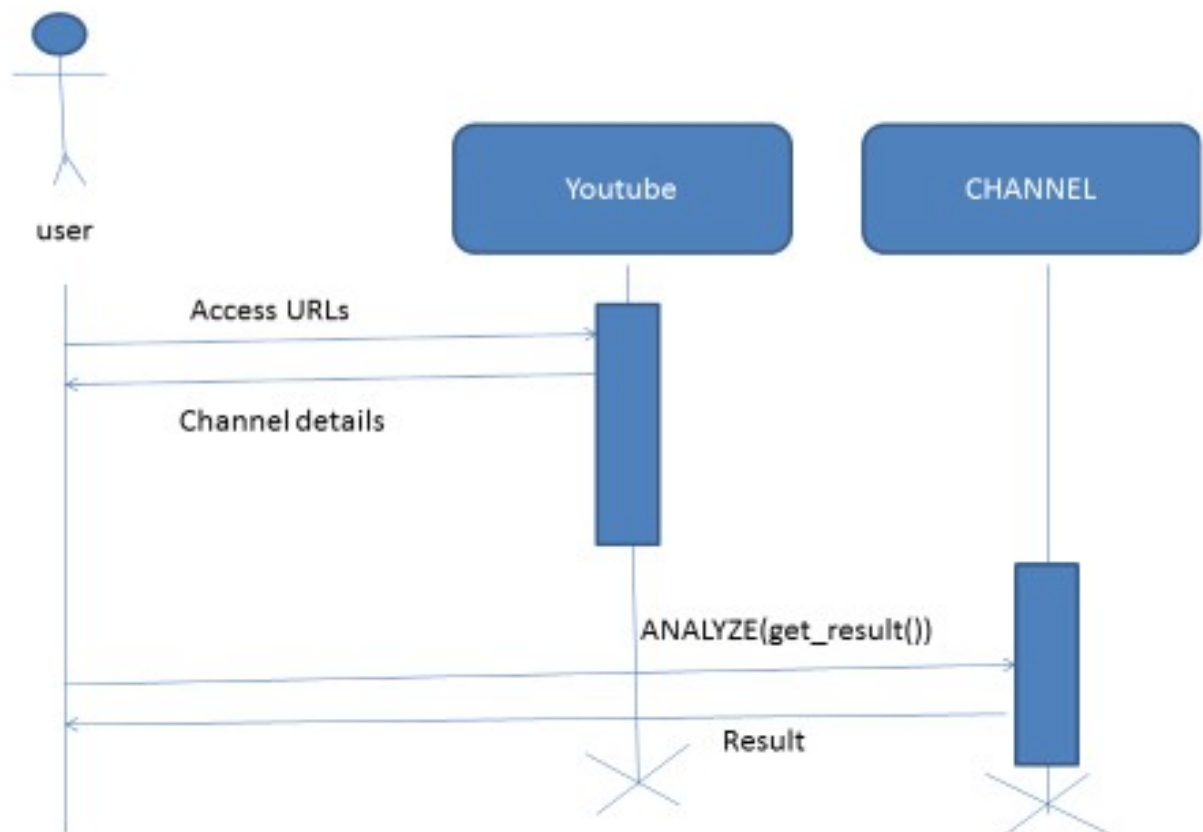
YOUTUBE COMMENT ANALYSIS

4.2 USE CASE DIAGRAM



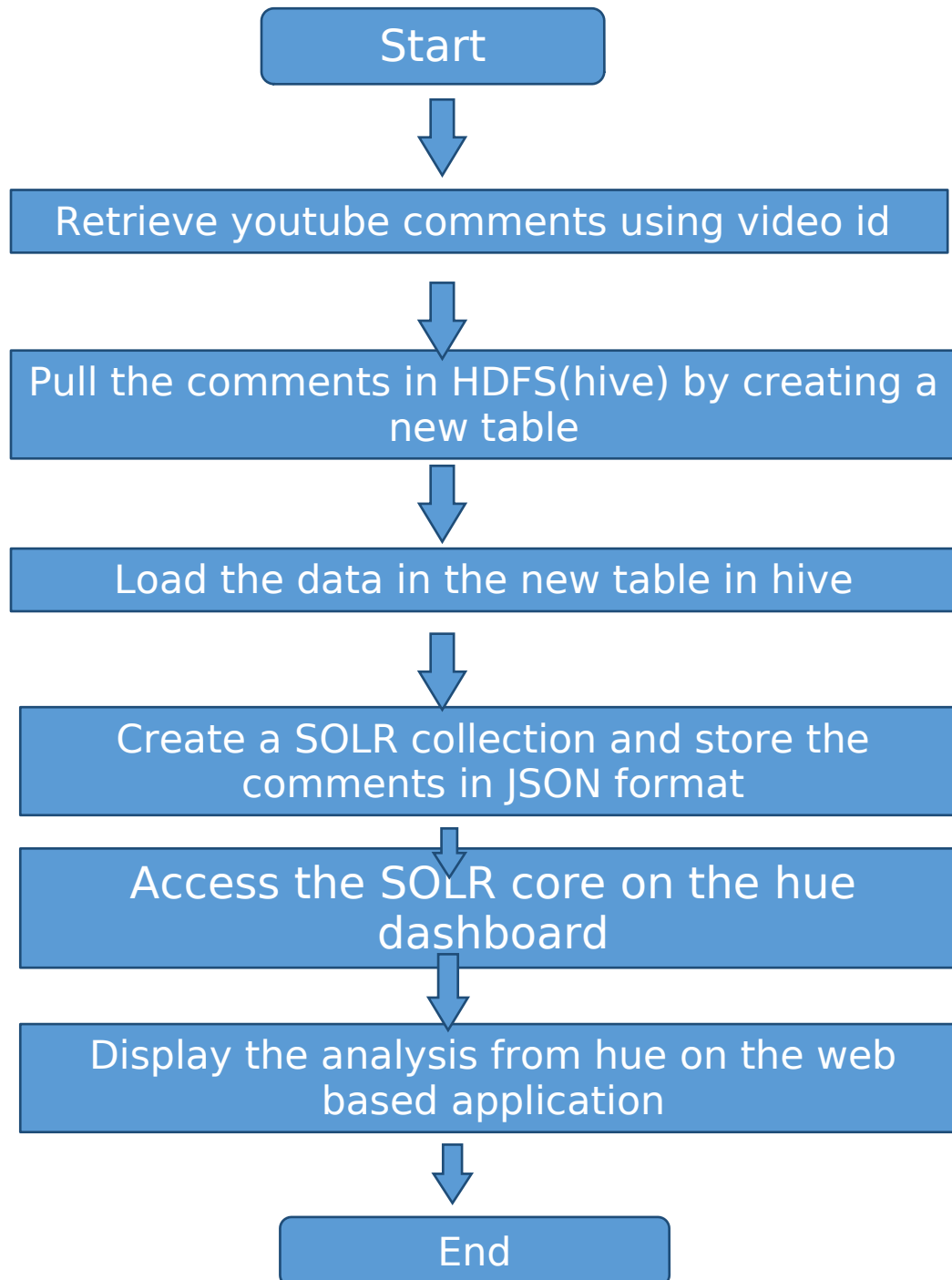
YOUTUBE COMMENT ANALYSIS

4.3 SEQUENCE DIAGRAM



YOUTUBE COMMENT ANALYSIS

4.4 ACTIVITY DIAGRAM



SYSTEM IMPLEMENTATION

YOUTUBE COMMENT ANALYSIS

5.1 SYSTEM REQUIREMENTS

Requirement Specification:

Hardware requirement specifications:

Processor: Intel core i5

Clock Speed: 1.8Ghz

RAM: 8Gb

HDD: 512 GB

CD/DVD Drive: 52x Reader

Pointing Device: Scroll Mouse

Keyboard: 101 Standard Keyboard

Software requirement specifications:

Operating System:

ClouderaCdh 5.8 (Virtual Machine)

Softwares used:

Enthought canopy

Hue

Solr

Language: Python/Java

Front end:

Web based application in html / Hue

Back end:

Hadoop Framework Cloudera

YOUTUBE COMMENT ANALYSIS

5.2 SYSTEM IMPLEMENTATION PROCEDURE

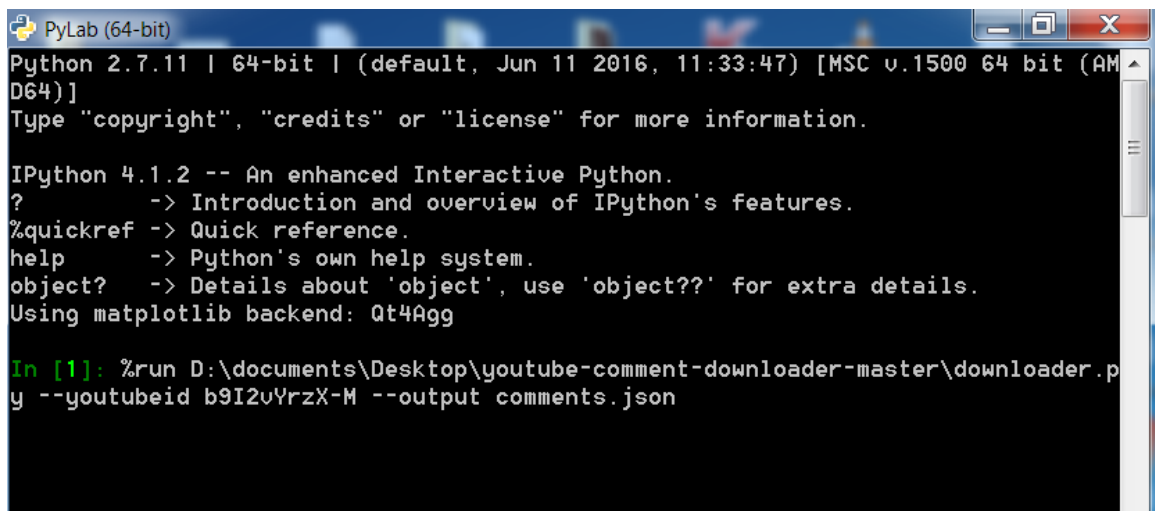
CODE:

STEP 1: USE VISUAL STUDIO CODE TO DOWNLOAD THE COMMENTS

File: downloader.py

To execute this code:

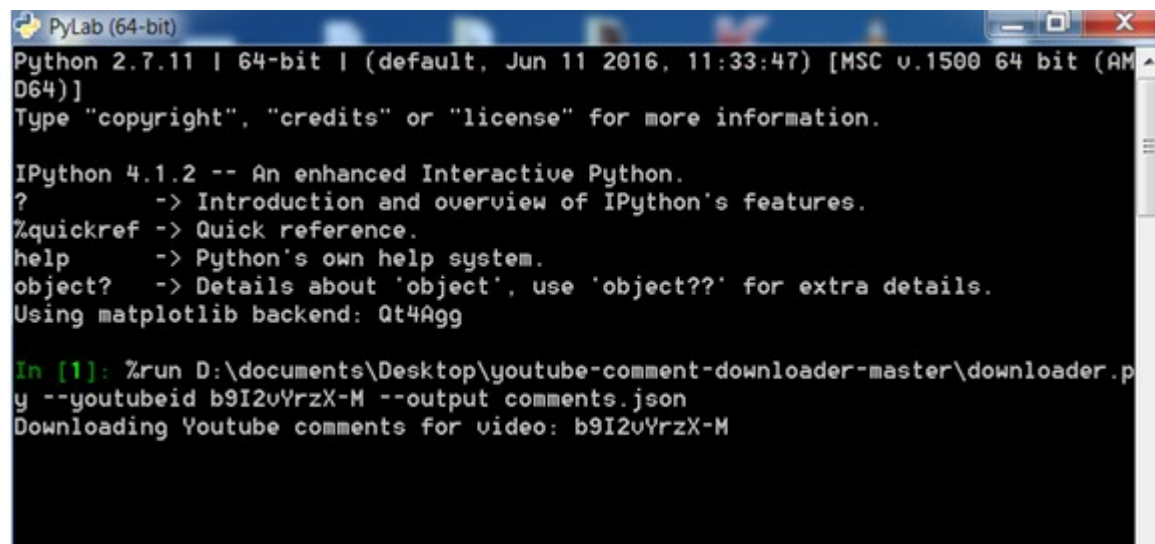
```
%run D:\documents\Desktop\youtube-comment-downloader-master\downloader.py --youtubeid _____ --output (name and format)
```



```
PyLab (64-bit)
Python 2.7.11 | 64-bit | (default, Jun 11 2016, 11:33:47) [MSC v.1500 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 4.1.2 -- An enhanced Interactive Python.
? -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help -> Python's own help system.
object? -> Details about 'object', use 'object??' for extra details.
Using matplotlib backend: Qt4Agg

In [1]: %run D:\documents\Desktop\youtube-comment-downloader-master\downloader.py --youtubeid b9I2vYrzX-M --output comments.json
```

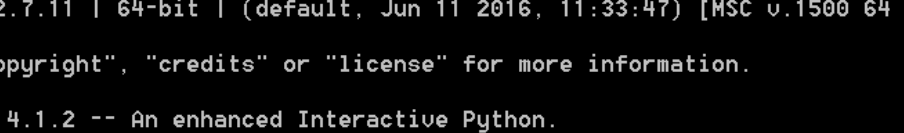


```
PyLab (64-bit)
Python 2.7.11 | 64-bit | (default, Jun 11 2016, 11:33:47) [MSC v.1500 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 4.1.2 -- An enhanced Interactive Python.
? -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help -> Python's own help system.
object? -> Details about 'object', use 'object??' for extra details.
Using matplotlib backend: Qt4Agg

In [1]: %run D:\documents\Desktop\youtube-comment-downloader-master\downloader.py --youtubeid b9I2vYrzX-M --output comments.json
Downloading Youtube comments for video: b9I2vYrzX-M
```

YOUTUBE COMMENT ANALYSIS



The screenshot shows a terminal window titled "PyLab (64-bit)". The terminal output includes the IPython version (4.1.2) and a list of help topics: "?", "%quickref", "help", and "object?". It also shows the matplotlib backend as "Qt4Agg". The user enters a command to run a Python script: `In [1]: %run D:\documents\Desktop\youtube-comment-downloader-master\downloader.py --youtubeid b9I2vYrzX-M --output comments.json`. The script output shows it is downloading YouTube comments for video b9I2vYrzX-M and has downloaded 1601 comment(s).

```
PyLab (64-bit)
Python 2.7.11 | 64-bit | (default, Jun 11 2016, 11:33:47) [MSC v.1500 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 4.1.2 -- An enhanced Interactive Python.
?                -> Introduction and overview of IPython's features.
%quickref        -> Quick reference.
help             -> Python's own help system.
object?          -> Details about 'object', use 'object??' for extra details.
Using matplotlib backend: Qt4Agg

In [1]: %run D:\documents\Desktop\youtube-comment-downloader-master\downloader.py --youtubeid b9I2vYrzX-M --output comments.json
Downloading Youtube comments for video: b9I2vYrzX-M
Downloaded 1601 comment(s)
```

OUTPUT: JSON FILE

[illegible]

YOUTUBE COMMENT ANALYSIS

STEP 2: CREATE SOLR COLLECTION

We need to create a core in solr to be able to retrieve the txt file that includes our comments in a JSON format. This json file is pulled in the solr collection which is then used in HUE to provide analysis (graphically) of the data.

We use various filters in hue to analyse the data based on keywords, frequency of words etc. This is done by querying the data for example the file 'output1.json' was loaded in the solr collection.

Once the collection was created named 'youtubeproj' we accessed the hue dashboard and started making a new dashboard. Where we passed the query "text:love" which gives us a result of all the comments including the keyword "love" and a graphical representation of the date when the comment was entered and its occurrence.

We need to execute the following commands to create a core in solr

```
[cloudera@quickstart ~]$ chmod 755 -R conf
```

```
[cloudera@quickstart ~]$ solrctl instancedir --create youtubeproj  
/youtubeproj/conf
```

```
[cloudera@quickstart ~]$ solrctl instancedir --update  
youtubeprojyoutubeproj/
```

```
[cloudera@quickstart ~]$ solrctl collection --create youtubeproj -s 1 -c  
youtubeproj
```

```
[cloudera@quickstart ~]$ solrctl instancedir --update  
youtubeprojyoutubeproj/
```

```
[cloudera@quickstart ~]$ curl -X POST  
'http://quickstart.cloudera:8983/solr/youtubeproj/update/json/docs?split=  
&f=/**&commit=true' -H 'Content-type:application/json' -d @output1.json
```

YOUTUBE COMMENT ANALYSIS

STEP 3: USE HIVE AND HUE TO PERFORM DATA ANALYSIS

Hive

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

Traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like Queries (HiveQL) into the underlying Java API without the need to implement queries in the low-level Java API.

Since most data warehousing applications work with SQL-based querying languages, Hive supports easy portability of SQL-based application to Hadoop. While initially developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority (FINRA).

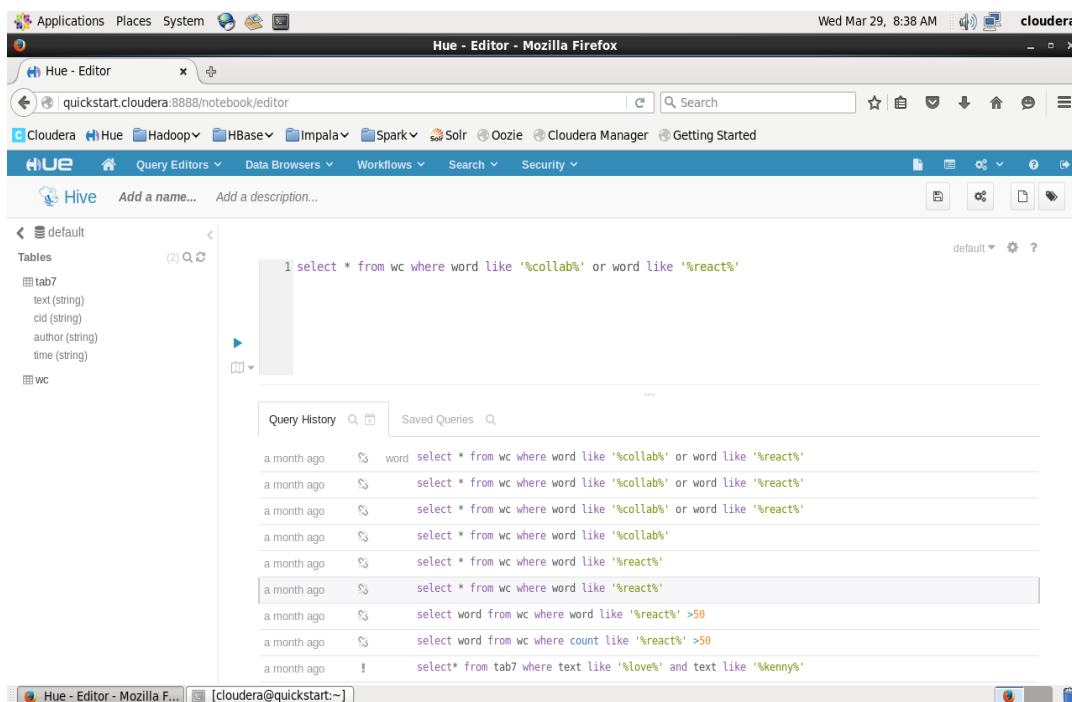
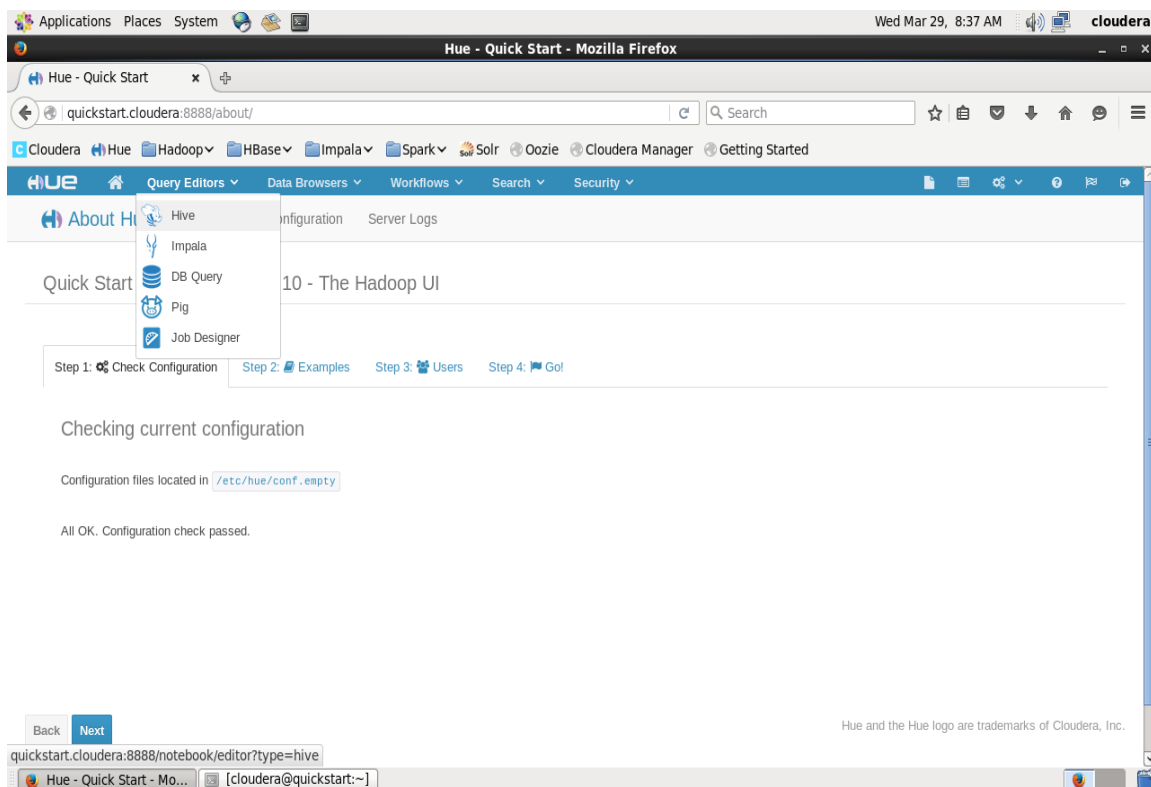
Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic MapReduce on Amazon Web Services.

We push data into hdfs using Hive queries.

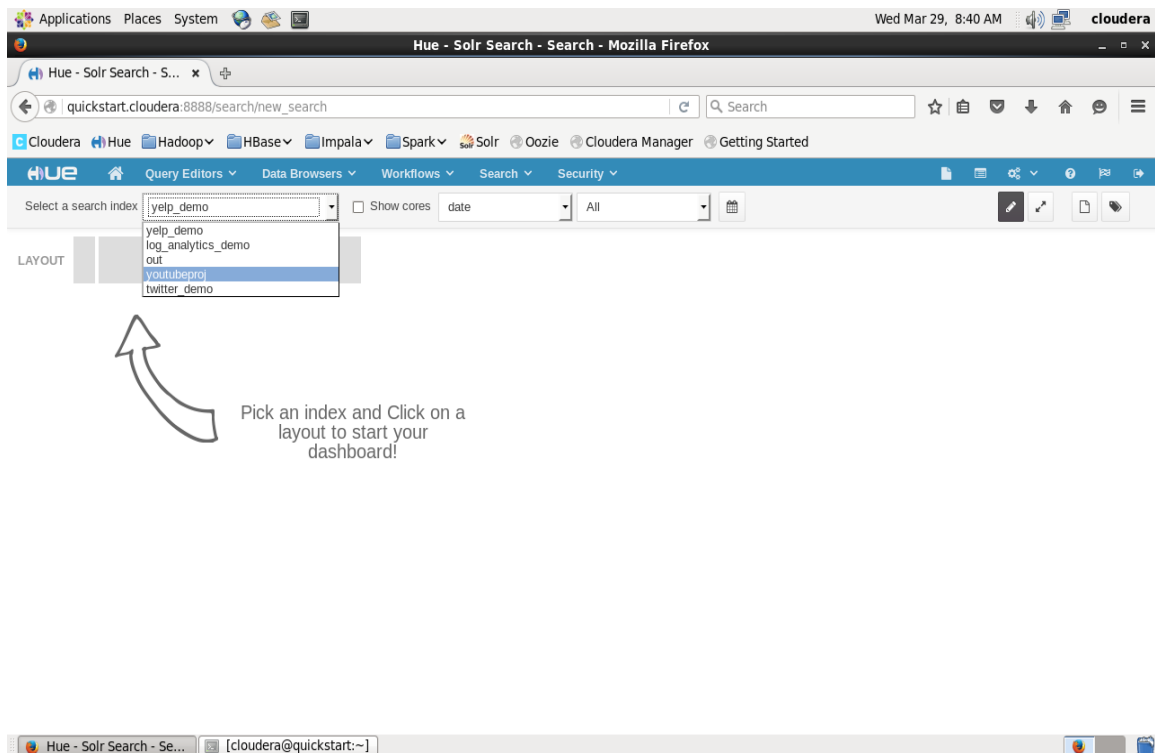
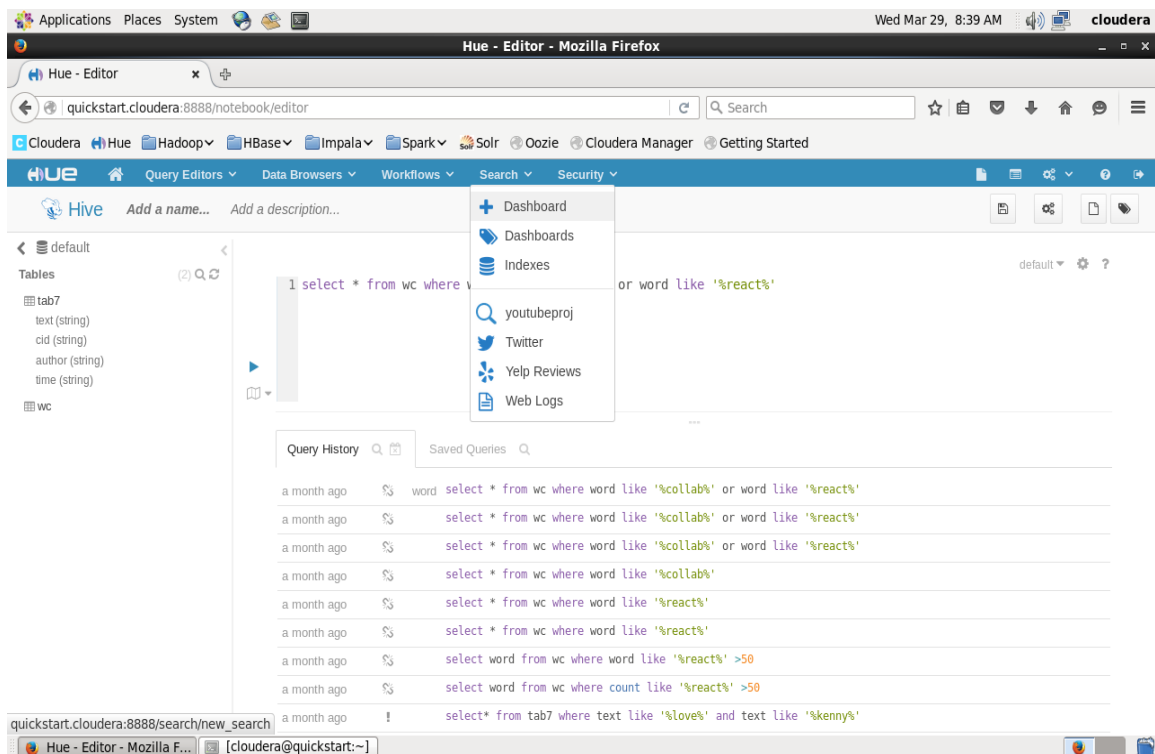
```
create table tab7(text string, cid string, author string, time string);  
load data inpath 'comments.txt' into tab7;
```

This will load the txt which includes our comments we retrieved from youtube using our python code.

YOUTUBE COMMENT ANALYSIS



YOUTUBE COMMENT ANALYSIS



YOUTUBE COMMENT ANALYSIS

The screenshot shows the Hue Solr Search interface in a Mozilla Firefox browser window. The search bar contains 'youtubeproject'. The results are displayed in a list format, showing 1 to 10 of 7082 results. The results are filtered by 'text' and 'cid'. The interface includes a sidebar with 'Filter fields' and a 'Field Name' section. The search results are displayed in a table format with columns for 'text' and 'cid'.

Search: youtubeproject

Showing 1 to 10 of 7082 results Show 10 results per page

Filter fields

All (6) / Current (1)

Field Name

- ☐ _version_
- ☐ author
- ☐ id
- ☒ text
- ☐ time
- ☐ cid

Results:

- Very cool,Very cool,1 hour ago,Карина Пентела,z12ys30rwoaqh4304cgfaeqzibqxczk,1560051279584034816
- I like your dance they are cool,I like your dance they are cool,9 hours ago,fapheng,z122hb3bkkygfdvcj22vdrqkuzfwblq404,1560051279672115200
- hagan el paso a paso plissssii,hagan el paso a paso plissssii,12 hours ago,gabi believerabrahamer,z12ygvkqwmmsubree232wdk40tqzrvwbh04,15600512796731637
- Do u listen to Ariana grande music?,Do u listen to Ariana grande music?,18 hours ago,Aisha AI,z13tjxnzga3eralv22zffbwph3fh1nwg04,1560051279674212352
- Heyy iedereen die nederlsnds kan spreken xD weet iemand misschien waar die dansschool is??,Heyy iedereen die nederlsnds kan spreken xD weet iemand missc
- You're cool!!!!!! ,You're cool!!!!!! ,21 hours ago,Баня Хам0р,z123u1diinqcejbdp224dt3qstbmf3tn1,1560051279675260929
- can u guise make a geographi of (Love me)?❤️,can u guise make a geographi of (Love me)?❤️,1 day ago,Diana Lobija,z13virj51ozw53jn04cil3g4vjv5wbtpc0k,156005127967538080
- Like si te gusta la canción,Like si te gusta la canción,1 day ago,Kuins 92,z12wpxnxtolc3tem23oshjbfwupjhnm04,1560051279677358081
- bailan perfectamente,bailan perfectamente,1 day ago,MARICHAT :3 lol,z121f51r1y3gylqic04cib0w0zucdmpdv40k,1560051279677358081
- I like the girl in the middle she is so good..Good job girls😊😊,I like the girl in the middle she is so good..Good job girls😊😊,1 day ago,Han Thanks,z13hufihtqdezv

The screenshot shows the Hue Solr Search interface in a Mozilla Firefox browser window. The search bar contains 'youtubeproject' and 'text:love'. The results are displayed in a list format, showing 1 to 10 of 7082 results. The results are filtered by 'text' and 'cid'. The interface includes a sidebar with 'Filter fields' and a 'Field Name' section. The search results are displayed in a table format with columns for 'text' and 'cid'.

Search: youtubeproject text:love

Showing 1 to 10 of 7082 results Show 10 results per page

Filter fields

All (6) / Current (6)

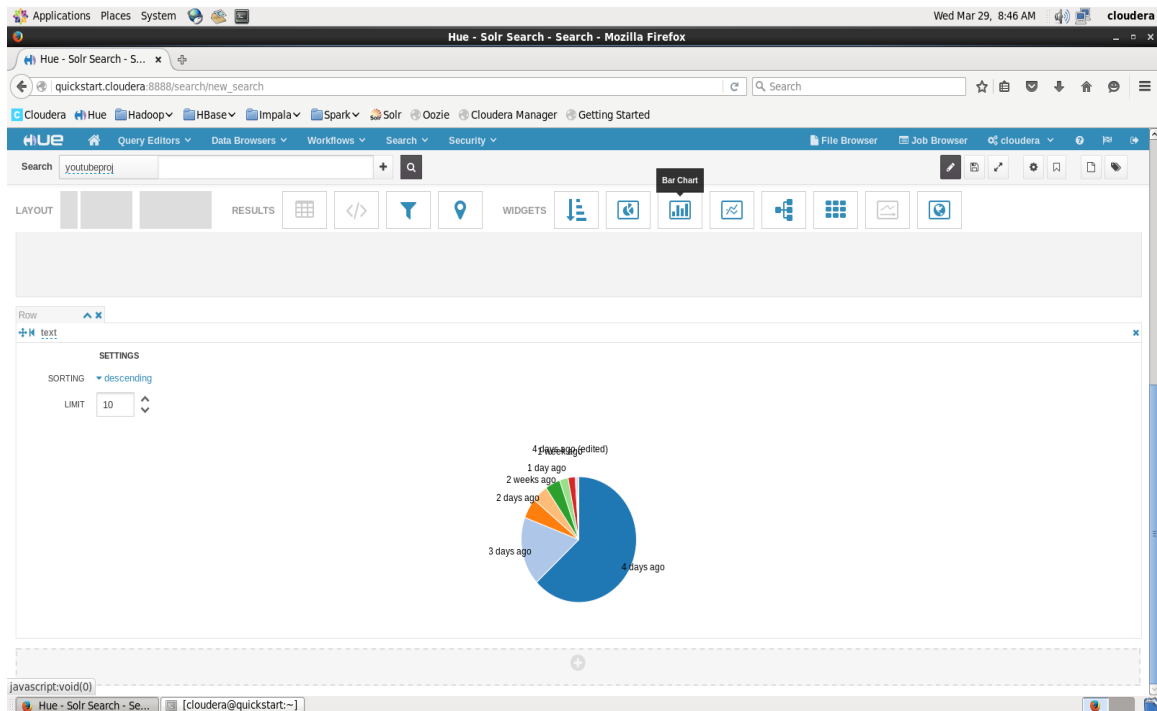
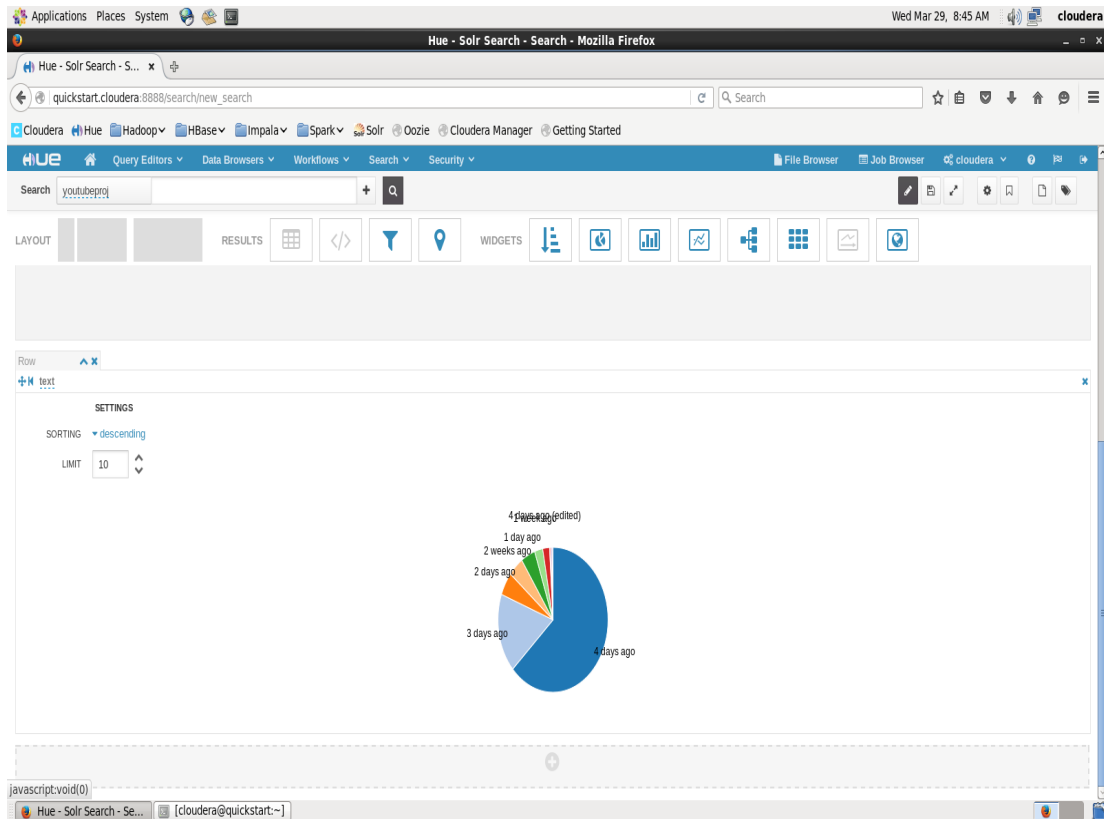
Field Name

- ☒ _version_
- ☒ author
- ☒ id
- ☒ text
- ☒ time
- ☒ cid

Results:

- Very cool,Very cool,1 hour ago,Карина Пентела,z12ys30rwoaqh4304cgfaeqzibqxczk,1560051279584034816
- I like your dance they are cool,I like your dance they are cool,9 hours ago,fapheng,z122hb3bkkygfdvcj22vdrqkuzfwblq404,1560051279672115200
- hagan el paso a paso plissssii,hagan el paso a paso plissssii,12 hours ago,gabi believerabrahamer,z12ygvkqwmmsubree232wdk40tqzrvwbh04,1560051279673163776
- Do u listen to Ariana grande music?,Do u listen to Ariana grande music?,18 hours ago,Aisha AI,z13tjxnzga3eralv22zffbwph3fh1nwg04,1560051279674212352
- Heyy iedereen die nederlsnds kan spreken xD weet iemand misschien waar die dansschool is??,Heyy iedereen die nederlsnds kan spreken xD weet iemand missc
- You're cool!!!!!! ,You're cool!!!!!! ,21 hours ago,Баня Хам0р,z123u1diinqcejbdp224dt3qstbmf3tn1,1560051279675260929
- can u guise make a geographi of (Love me)?❤️,can u guise make a geographi of (Love me)?❤️,1 day ago,Diana Lobija,z13virj51ozw53jn04cil3g4vjv5wbtpc0k,156005127967538080
- Like si te gusta la canción,Like si te gusta la canción,1 day ago,Kuins 92,z12wpxnxtolc3tem23oshjbfwupjhnm04,1560051279677358081
- bailan perfectamente,bailan perfectamente,1 day ago,MARICHAT :3 lol,z121f51r1y3gylqic04cib0w0zucdmpdv40k,1560051279677358081
- I like the girl in the middle she is so good..Good job girls😊😊,I like the girl in the middle she is so good..Good job girls😊😊,1 day ago,Han Thanks,z13hufihtqdezv

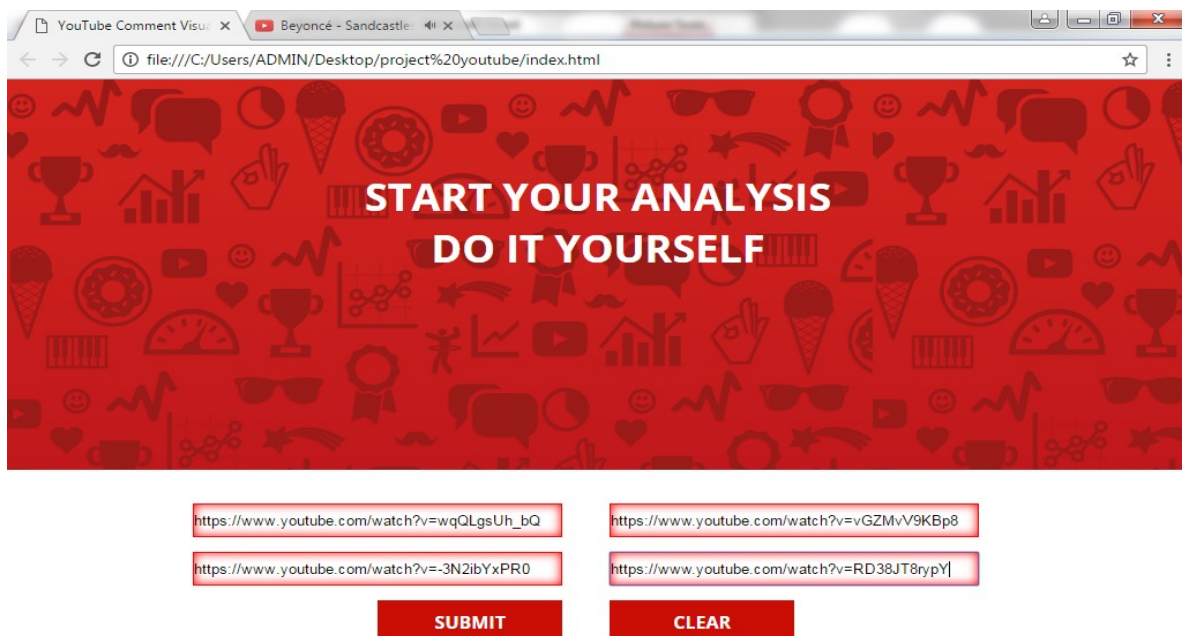
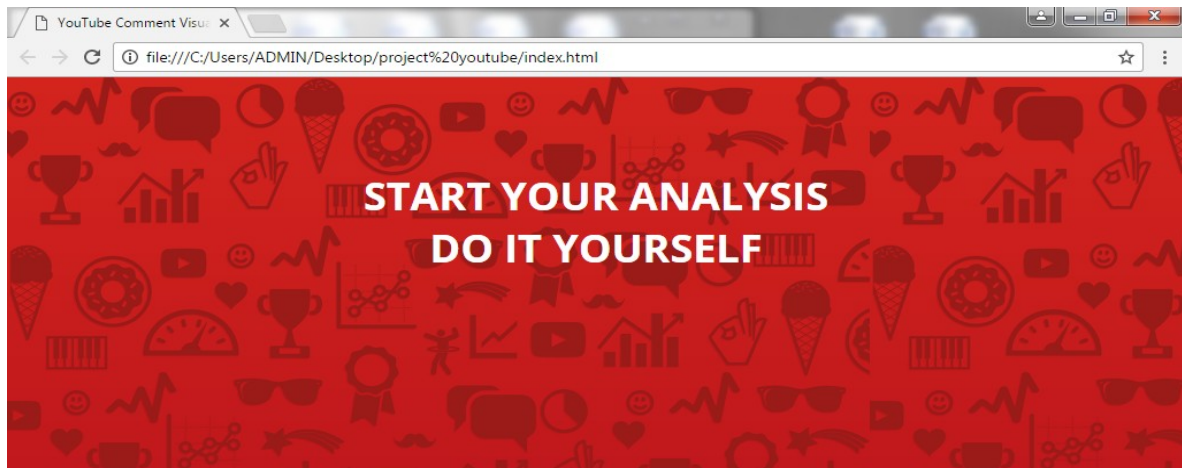
YOUTUBE COMMENT ANALYSIS



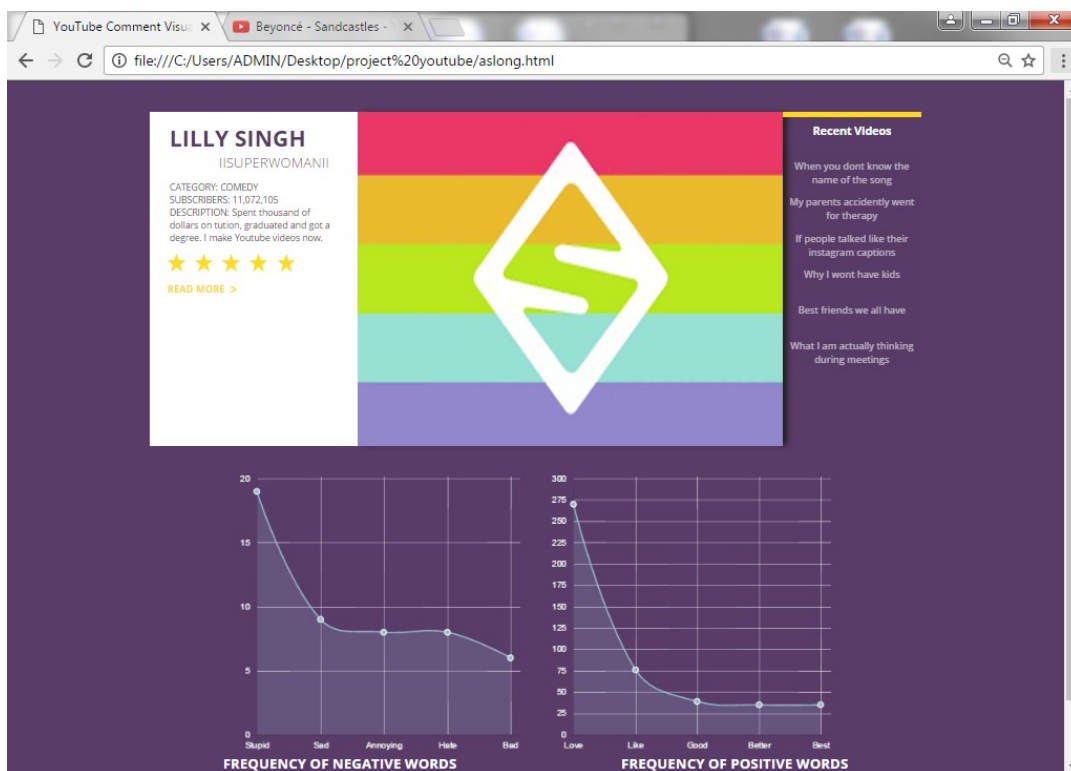
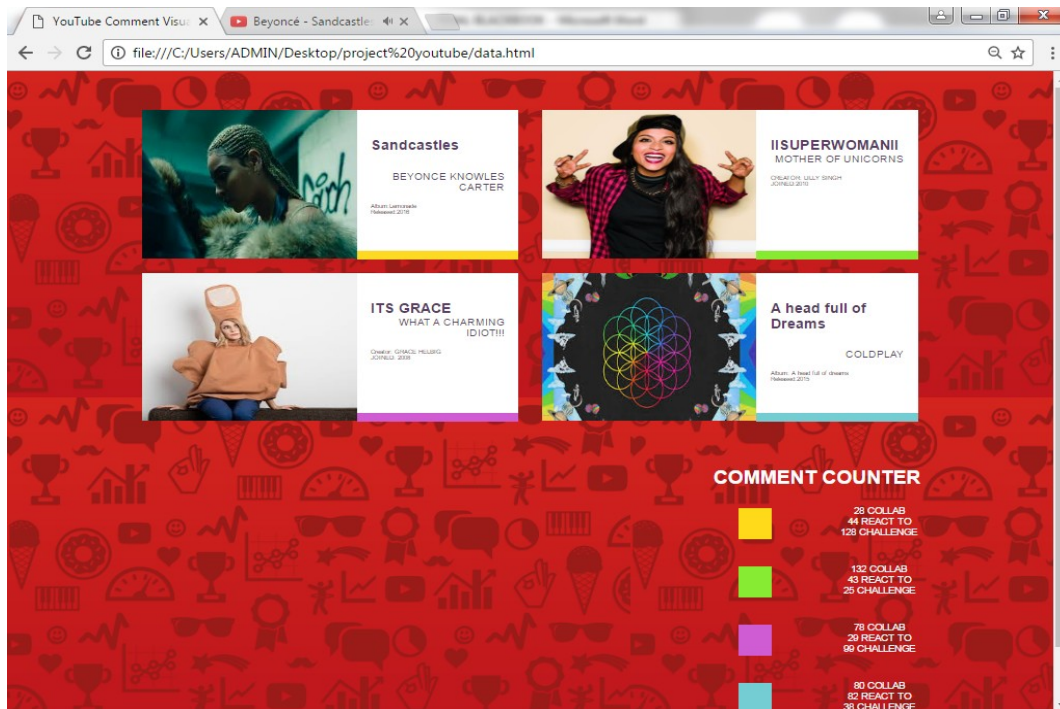
YOUTUBE COMMENT ANALYSIS

5.3 SCREEN SHOTS/ USER MANUAL/ INTERFACE

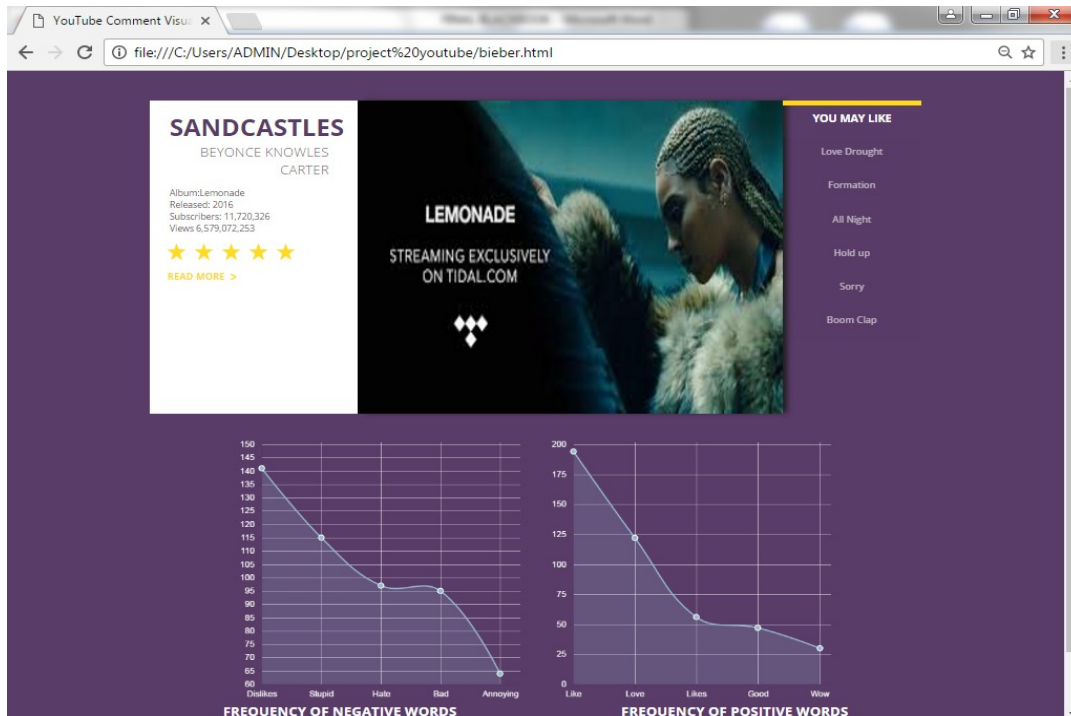
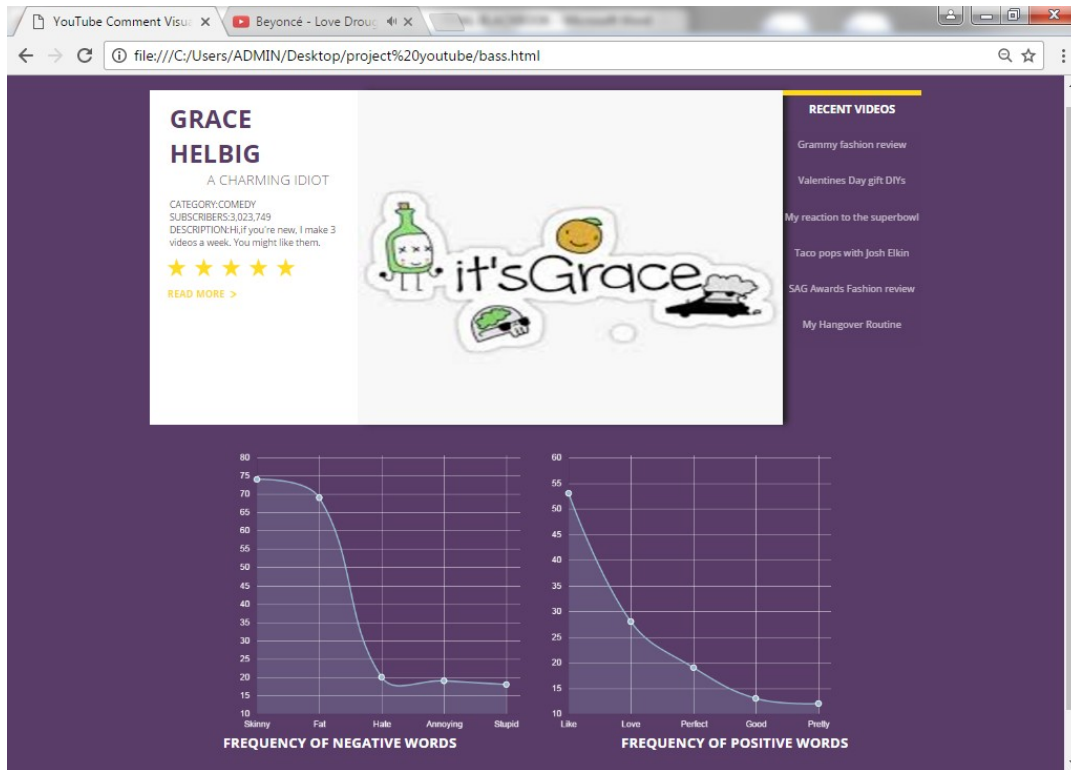
YOUTUBE COMMENT ANALYSIS



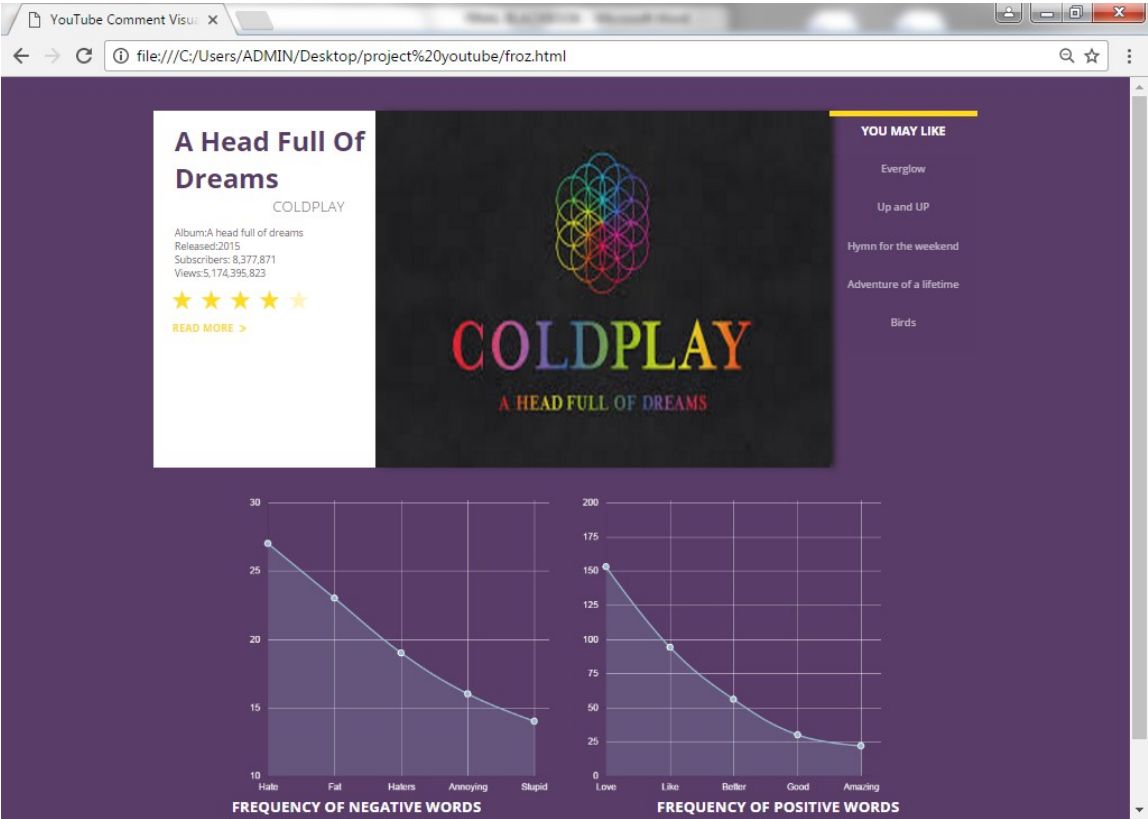
YOUTUBE COMMENT ANALYSIS



YOUTUBE COMMENT ANALYSIS



YOUTUBE COMMENT ANALYSIS



SYSTEM TESTING

YOUTUBE COMMENT ANALYSIS

6.1 METHODOLOGY ADOPTED FOR TESTING

Testing is the process of executing a program with intent of finding errors. A good test case is one that has a high probability of finding an as yet undiscovered error. If testing is conducted successfully it will uncover error in the software and testing demonstrates that software functions appears to be working according to specification, that behavior and performance requirements appear to have been met. In addition, data collected as testing provide a good indication of software reliability and some indication of software quality as a whole. But testing cannot show the absence of errors and defects, it can only show that errors and defects are present.

Testing Principles: All tests should be traceable to customer requirements.

- Test should be planned long before testing begins.
- Testing should begin in small scale and progress towards large scale.
- Exhaustive testing is not possible.
- To be most effective testing should be conducted by independent third party.

Testing Methods: Test must be designed with the highest likelihood to find possible errors in the system to avoid major problems before the system goes live. There are two methods to design the test cases.

Verification and Validation: Verification refers to the set of activities that ensure, software correctly implements a specific function. Validation refers to different set of activities that ensure, the software that has been built is traceable to customer requirements.

White Box Testing:

Glass Box Testing is a test case design method that uses the control structure of the procedural design to drive test cases. Using white box testing the software engineer can derive test cases that

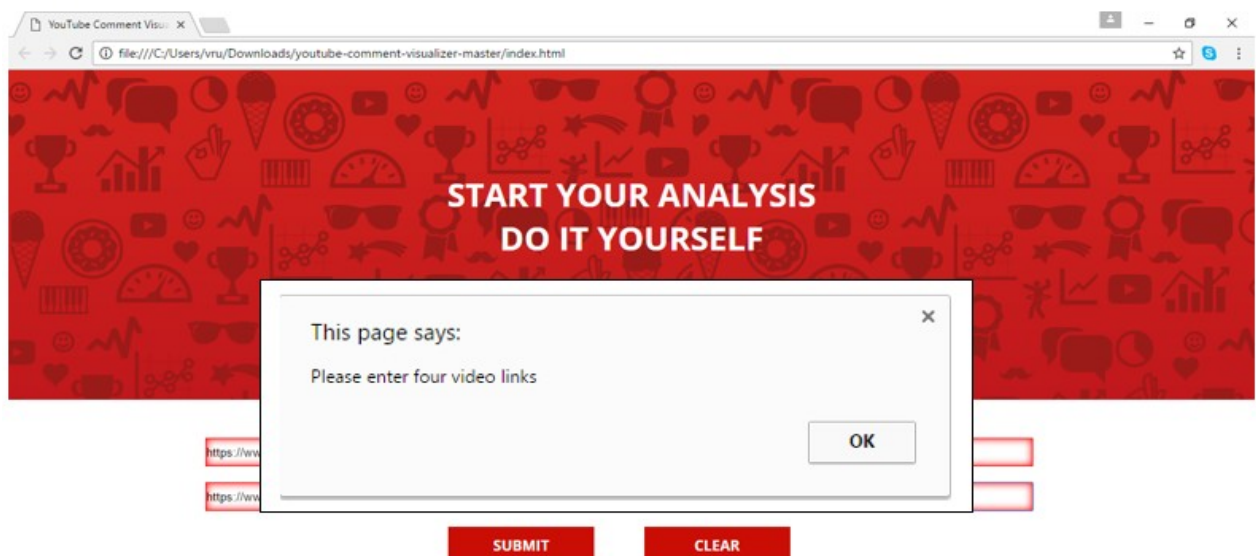
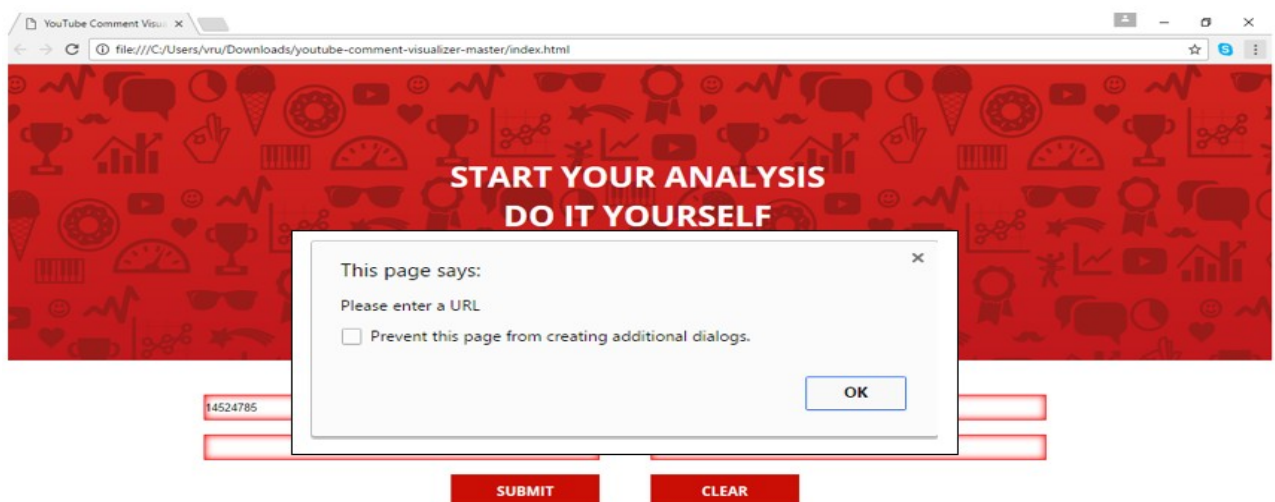
- Guarantee that all independent paths within the module have been executed at least once.
- Execute all logical decisions on their true and false side.
- Execute all loops at their boundaries and within their operational bounds.
- Execute internal data structures to ensure their validity.

YOUTUBE COMMENT ANALYSIS

Black Box Testing:

Black Box tests are used to demonstrate that software function are operational, that input is properly accepted and output is correctly produced. It is also used to demonstrate that integrity of external information is maintained. Black box test examines some fundamental aspects of system with little regard for the internal logic structure of the software.

TESTING SCREENSHOTS:



CONCLUSION

YOUTUBE COMMENT ANALYSIS

7.1 Conclusion

- This system will provide the analysis for the comments on every video. All the user has to do is copy and paste the URL of the desired video/ videos and submit.
- It will show the channel description of every channel who's URL was entered, the user can then choose the channel's video by clicking on the icon and view the analysis.
- It shows the frequency of positive and negative words based on their occurrence in the comments by the viewers.
- The youtuber can then at a glance analyze the comments rather than going through them manually.
- Big data is all about processing huge chunks of data within minimum possible time, which is why Big Data is implemented to represent systematic analysis of the comments.

7.2 Limitations of the proposed system:

➤ Graphical Representation:

The representation format is a basic graph which will exhibit the frequency of words related to sentiments. Basic line graphs and algorithms for implementing these line graphs.

➤ Report Storage:

There is no mechanism to store the reports generated, since there is no server involved. The reports are basically, represented as graphs there and then. The analysis has to be performed every time the report needs to be generated.

➤ Topic mining for community tab:

The proposed system is not performing any analysis on the community conversations.

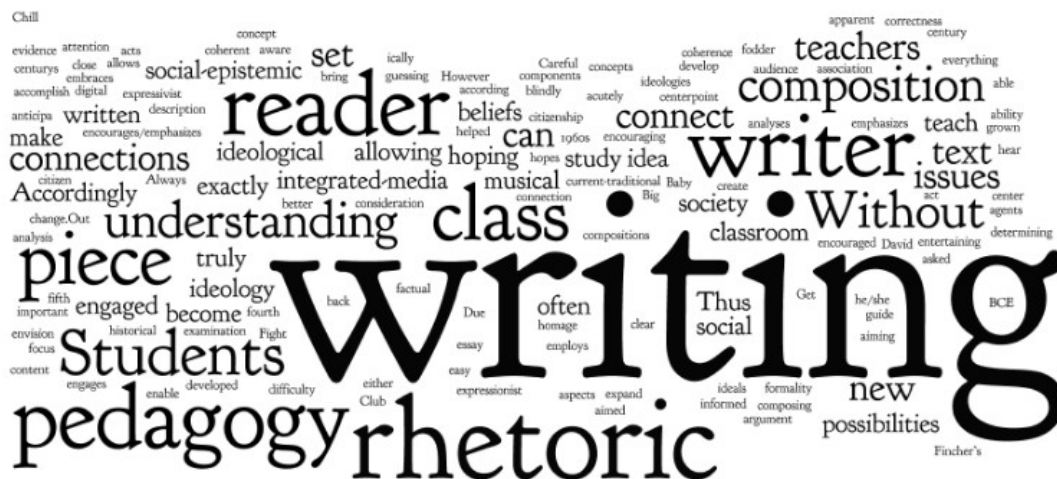
FUTURE ENHANCEMENTS

YOUTUBE COMMENT ANALYSIS

Future Enhancements

Solutions for the limitations:

➤ **Since** the current website provides a basic line graph or a pie chart or a bar graph, we will be working with Word Cloud to represent the analysis report. Word Cloud helps create an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance. Another interesting feature would be numerical count representation of the words in the word cloud by just hovering over any word.



➤ Another addition that would make this website even more interesting is by providing storage for each user. A user could thus be able to store his/her analysis reports for a limited period of time and access them. This will also help the Youtuber compare analytical results for the same video over a period of time.

➤ We look forward to also implement Topic Mining which is also referred to as **text data mining**, roughly equivalent to **text analytics**. It is the process of deriving high-quality information from **text**. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

ANNEXURE

YOUTUBE COMMENT ANALYSIS

REFERENCES:

BOOK:

Pro Apache Hadoop

Authors: Venner, Jason, Wadkar, Sameer, Siddalingaiah, Madhu

WEBSITES VISITED:

www.youtube.com

<https://www.cloudera.com/>

<http://gethue.com/>

<http://blog.cloudera.com/blog/2015/10/how-to-use-apache-solr-to-query-indexed-data-for-analytics/>