**GOVERNMENT POLYTECHNIC**

**SOLAPUR**

PROJECT REPORT ON

**YOUTUBE DATA ANALYSIS USING HADOOP**

**Submitted To Board of Technical Education, Mumbai.**

In partial fulfillment for the Diploma in Computer Technology.

**Submitted By**

CM6G (Group No.5)

Miss Gotipamul.B.V Roll No.14 Exam seat no.188938

Miss Kankurti.T.NA Roll No.20 Exam seat no.188935

Miss Boga.S.S Roll No.09 Exam seat no.188974

Miss Gulbargawale.T.RA Roll No.15 Exam seat no.188950

**Under the Guidance of**

**Prof. Mr. Gangundi R.Y**

**DEPARTMENT OF COMPUTER TECHNOLOGY**

**(2018-2019)**

**Government Polytechnic, Solapur**

**Department of Computer Technology**

****

**CERTIFICATE**

It is to certify that

Mrs/Miss:**Bhavana.V.Gotipamul**

**Tasmiya.NA.Kankurti**

**Sneha.S.Boga**

**Tabassum.RA.Gulbargawale**

Class: CM6G

has satisfactorily completed the project titled as

**YOUTUBE DATA ANALYSIS USING HADOOP**

as a partial fulfillment

For awarding the Diploma in Computer Technology

by Maharashtra State Board of Technical Education, Mumbai

For the year **2018-2019**.

Project Guide Head of Department Principal

Internal Examiner Examiner External

Date:

Place: Solapur

**Acknowledgement**

The Success of any project is never limited to the individual undertaking the project. There are some personalities involved whose role is very vital for successful completion of project.

We take this Opportunity to extend our sincere thanks & gratitude them. We wish to thank **Prof.Gangundi R.Y** our guide who helped us time to time during the project.

At last we are thankful to all Teachers & other staff members for their directly & indirectly helped in project.

**Index**

**Chapter 1***:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Planning………………………………………………………

Feasibility Study………………………………………………

Cost Estimation……………………………………………….

Scheduling …………………………………………………

**Chapter 2**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Analysis……………………………………………………….

Problem Statement…………………………………………….

Objectives of project…………………………………………..

Software/hardware requirements………………………………

Constraints on project development…………………………..

Functional Requirements………………………………………

Performance Requirements……………………………………

Acceptance Criterion………………………………………….

**Chapter 3***:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

UI Design………………………………………………………

Reports design…………………………………………………

File Design, if any……………………………………………..

New algorithms, if any, Design………………………………..

Database Design………………………………………………..

ER Diagram……………………………………………………

Table Design…………………………………………………...

Normalization up to 3NF………………………………………

Error Design……………………………………………………

User Manual Design……………………………………………

Test Design…………………………………………………….

**Chapter 4***:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Coding………………………………………………………….

Languages / tools used…………………………………………

Justification for Selection of languages/tools………………….

Important modules……………………………………………..

**Chapter 5***:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Testing………………………………………………………

Test Plan…………………………………………………….

Test Reports………………………………………………...

References…………………………………………………..

**List of Diagrams**

Characteristics of big data……………………………………………………………………..10

Hadoop Architechture……………………………………………………………………………..12

Data processing using MapReduce framework**…………………………………………………………..15**

UI Design………………………..................................................................................................28

Output of category count…………………………………………………………………………29

Output of most likes videos……………………………………………………………………..29

Output of most dislikes videos…………………………….…………………………………..30

Output of Top rated videos………………………………….…………………………………..30

**Chapter 1**

**Planning:**

  In our Project we analyze the data to identify the Top rated videos, most liked videos, most disliked videos, Top views of videos , Average ratings of videos which are uploaded.

The dataset is gathered using the **websites** and stored **in Hadoop** Distributed ... configuring Master-Slave architecture.

**Feasibility Study:**

Depending on the results of the initial investigation the survey is now expanded to a more detailed feasibility study. “**FEASIBILITY STUDY**” is a test of system proposal according to its workability, impact of the organization, ability to meet needs and effective use of the resources. It focuses on these major questions:

1. What are the user’s demonstrable needs and how does a candidate system

Meet them?

2. What resources are available for given candidate system?

3. What are the likely impacts of the candidate system on the organization?

4. Whether it is worth to solve the problem?

During feasibility analysis for this project, following primary areas of interest are to be considered. Investigation and generating ideas about a new system does this.

**Steps in feasibility analysis**

* Eight steps involved in the feasibility analysis are:
* Form a project team and appoint a project leader.
* Prepare system flowcharts.
* Enumerate potential proposed system.
* Define and identify characteristics of proposed system.
* Determine and evaluate performance and cost effective of each proposed system.
* Weight system performance and cost data.
* Select the best-proposed system.
* Prepare and report final project directive to management.

**Cost Estimation:**

It checks whether the system can be developed with the available funds. The **YOUTUBE DATA ANALYSIS** does not require enormous amount of money to be developed. This can be done economically if planned properly, so it is economically feasible. The cost of project depends upon the number of manhours required.

Hadoop is an open source framework so cost is not required.

**Scheduling:**

Time evaluation is the most important consideration in the development of project. The time schedule required for the developed of this project is very important since more development time effect machine time, cost and cause delay in the development of other systems.

A reliable **YOUTUBE DATA ANALYSIS SYSTEM** can be developed in the considerable amount of time.

**Chapter 2**

**Analysis:**

The analysis of YouTube Datasets using Hadoop Map Reduce is as follows:

1. The top categories in which the most number of videos are uploaded.

2. To identify the top above 4.90 rated videos.

3. To identify the most liked videos are uploaded.

4. To identify the most disliked videos are uploaded.

5. To identify the top views of uploaded videos.

6. To identify the Average ratings of uploaded videos.

**Problem Statement:**

Big Data is a term used to suggest huge data sets (several gigabytes / terabytes /petabytes) of data. The data is so large and complex that it would become difficult to process using traditional data processing applications.

For providing Solution for this problem HADOOP is introduced

**YouTube Data Analysis Using Hadoop,** Here we will find out what are the Top rated videos, most liked videos, and most disliked videos, Top views of videos, Average ratings of videos which are uploaded.

**Evolution of Data/Big Data**

In today’s world, due to advancements of technology there is a huge (several terabytes /petabytes) amount of data that’s being constantly captured. Natural curiosity about truly important things, like whether more teenagers like Justin Bieber than millennials, demand processing Twitter data, which is huge.

**Characteristics of Big Data**



**Characteristics of big data**

The characteristics of Big Data are popularly known as the three V’s of Big Data

**Introduction To Hadoop**

Hadoop is an open source framework. It is capable of processing large amounts of data sets in a distributed fashion across clusters using a simplified programming model.

Hadoop provides a reliable way to store, process and analyze the data

Hadoop runs on clusters of commodity servers and can scale up to support thousands of hardware [nodes](https://searchnetworking.techtarget.com/definition/node) and massive amounts of data. It uses a namesake distributed file system that's designed to provide rapid data access across the nodes in a cluster, plus fault-tolerant capabilities so applications can continue to run if individual nodes fail. Consequently,

Hadoop was created by computer scientists Doug Cutting and Mike Cafarella, initially to support processing in the Nutch open source search engine and web crawler

**Hadoop Architecture**

The Hadoop architecture is illustrated in Figure:



**Hadoop architechture**

Hadoop works in a master-slave fashion. Hadoop has two core components-HDFS(Hadoop Distributed File System) and MapRededuce.

**Hadoop Components**

1. **HDFS (HADOOP DISTRIBUTED FILE SYSTEM)**

HDFS offers a reliable and distributed storage. It replicates the data across multiple nodes on clouds or commodity computers. Unlike a regular file system, when data is pushed

into HDFS, it internally splits into multiple data blocks (configurable parameter with default size of 64Mb). Each incoming file is broken into 64Mb data block by default and all blocks which make up the file are of the same size 64Mb except the last block which could be less than 64Mb depending upon the size of the incoming file.

**HDFS works in a master/slave fashion.**

**NameNode:** NameNode is a master component and holds the information about all other nodes in the Hadoop cluster, files present and their locations in the cluster .There is only one NameNode per cluster.

**DataNode:** DataNode is a slave node and holds the user data in the form of data blocks. There can be a lot of DataNodes in a Hadoop cluster.

1. **MAPREDUCE:**

MapReduce offers a framework/analysis system which performs complex computations on large datasets in a parallelized fashion. This system breaks down the complex computations into multiple smaller tasks and assigns those to individual slave nodes and takes care of the co-ordination and consolidation of the results. These tasks run independently on various nodes across the cluster. There are primarily two types of tasks: Map tasks and Reduce tasks.

As in HDFS, MapReduce (computation part) also works in master/slave fashion.

* **JobTracker:** Keeps track of the tasks assigned and co-ordinates the exchange of information and the results with the slave nodes. Its responsibility also includes rescheduling of failed tasks and monitoring the overall progress of the job. There is only one JobTracker per cluster.
* **TaskTracker:** Acts as a slave and is responsible for running the tasks assigned by the JobTracker and providing the results back to the JobTracker. There can be multiple Task Tracker nodes that can exist in a cluster. Data processing using the MapReduce frame work is highlighted in Figure



**Figure . Data processing using MapReduce framework.**

**FileInputFormat:** This is the input file/data which need to be processed.

**Split :**Hadoop splits the incoming data into several blocks.

**RecordReader:** RecordReader helps to read the data line by line and converts into key/value pairs to be passed as the input to the Mapper.

**Mapper:** Mapper contains the logic to process input data. The Map function transforms the input records to intermediate records.

**Combiner:** This is an optional step often used to improve the performance by reducing the data to be transferred across the network .

**Shuffle:** Output of all the mappers is collected, shuffled and sorted, to be sent to the Reducer.

**Reducer:** Reducer applies logic to aggregate the data and provide it to an FileOutputFormat class.

**FileOutputFormat:** It is a pre-defined class provided by the MapReducer framework through which final output can be written to HDFS.

**Hadoop Characteristics**

* Hadoop provides a reliable shared storage system (HDFS) and data analysis system (Map Reduce).
* Cost effective, as it can work with commodity hardware and doesn’t need expensive hardware.
* Flexible and can process both structured as well as un-structured data sets.
* Is optimized for large and very large data sets. It takes a lot less data processing time due to parallel processing, when compared with traditional data base management systems.
* Very scalable. As a result, the Hadoop cluster can contain hundreds or thousands of servers.
* Provides a very reliable system as data is replicated across multiple nodes (replication factor is configurable).

**Objective of Project:**

The main objective of this project is to focus on how data generated from YouTube can be mined and utilized by different companies to make targeted, real time and informed decisions about their product that can increase their market share. This can be done by using Hadoop concepts.

The given project will focus on how data generated from YouTube can be mined and utilized. There are multiple applications of this project. Companies can use this project to understand how effective and penetrative their marketing programs are. In addition to the view counts, subscribers and shares, audience retention count, companies can also evaluate views according to date range. This can tell the companies when the slow period or spike in viewership is and attribute the same to certain marketing campaign.

Applications for YouTube data can be endless. For example, Companies can analyse how much a product is liked by people. This project can also help in analysing new emerging trends and knowing about people's changing behaviour with time. Also people in different countries have different preferences. By analysing the comments/feedbacks/likes/view counts etc. of the videos uploaded, companies can understand what are the likes/dislikes of people around the world and work on their preferences accordingly.

This project uses following concepts and tools throughout its lifecycle.

1. Java API

2. Hadoop

3. Linux (Ubantu)

**Software & Hardware Requirements:**

|  |  |  |
| --- | --- | --- |
| **Hardware Requirements** | | |
| Processor | RAM | Disk Space |
| Dual-core Processor | 4 GB | 1 TB |

|  |
| --- |
| **Software Requirements** |
| * OPERATING SYSTEM*:* Install Hadoop on Linux based operating systems (Ubantu). * JAVA: You need to install the Java 8 package on your system. |

**Steps for installing Hadoop:**

### ****Step 1:**** download the Java 8 Package. Save this file in your home directory.

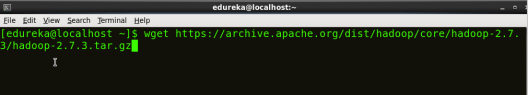
### ****Step 2:**** Extract the Java Tar File.

Command**:** tar -xvf jdk-8u101-linux-i586.tar.gz

https://d1jnx9ba8s6j9r.cloudfront.net/blog/wp-content/uploads/2016/11/untar-java-install-hadoop-edureka-528x60.png

**Step 3:**Download the Hadoop 2.7.3 Package.

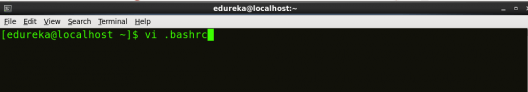
Command**:** wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz



### ****Step 4:****Extract the Hadoop tar File.

Command: tar -xvf hadoop-2.7.3.tar.gz

C:\Users\SAI\Desktop\folder\Extract-Hadoop-Package-Install-Hadoop-Edureka-528x60.png

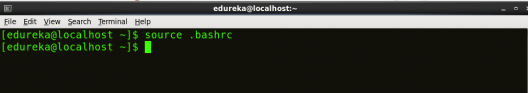
**Step 5:**Add the Hadoop and Java paths in the bash file (.bashrc).Open**.** **bashrc** file. Now, add Hadoop and Java Path as shown below. Command**:**  vi .ba****shrc

## C:\Users\SAI\Desktop\folder\Add-Java-and-Hadoop-variable-in-BASH-Install-Hadoop-Edureka-528x177.png

Then, save the bash file and close it.

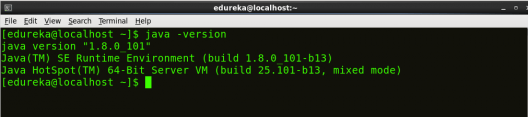
For applying all these changes to the current Terminal, execute the source command.

Command**:** source .bashrc

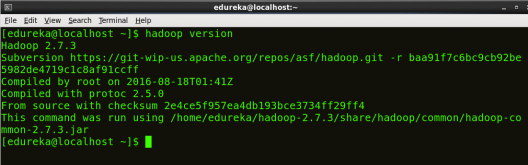


To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

Command**:**java -version

****

Command**:**hadoop version

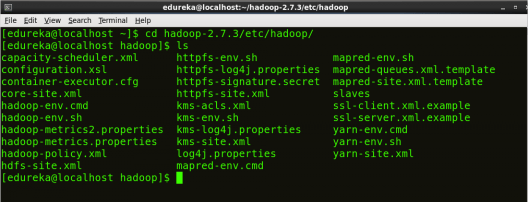
****

### ****Step 6:****Edit the [****Hadoop Configuration files****](https://www.edureka.co/blog/explaining-hadoop-configuration/).

***Command:*** cd hadoop-2.7.3/etc/hadoop/

***Command:*** ls

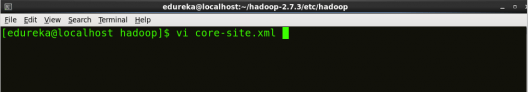
All the Hadoop configuration files are located in **hadoop-2.7.3/etc/hadoop** directory.

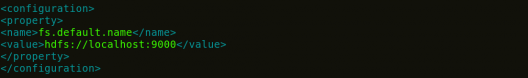


### ****Step 7:****Open core-site.xml and edit the property mentioned below inside configuration tag:

### core-site.xml

Command**:** vi core-site.xml

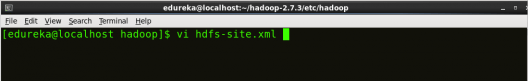


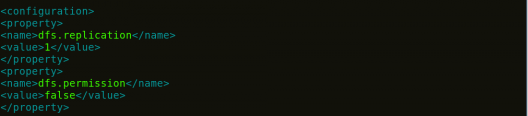


### ****Step 8:**** Edit hdfs-site.xml and edit the property mentioned below inside configuration tag:

hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.

Command**:** vi hdfs-site.xml



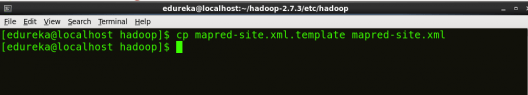


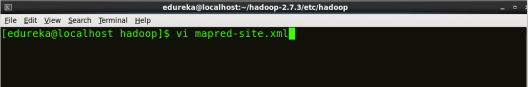
### ****Step 9:****Edit the mapred-site.xml file and edit the property mentioned below inside configuration tag:

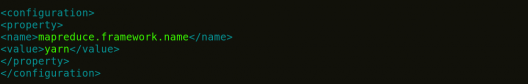
In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template.

Command**:** cp mapred-site.xml.template mapred-site.xml

Command**:**vi mapred-site.xml.

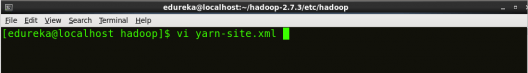


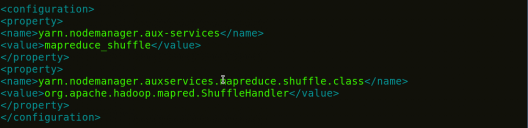




### ****Step 10:**** Edit yarn-site.xml and edit the property mentioned below inside configuration tag:

Command**:** vi yarn-site.xml

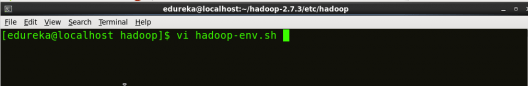
****

****

### ****Step 11:****Edit hadoop-env.sh and add the Java Path as mentioned below:

hadoop-env.sh contains the environment variables that are used in the script to run Hadoop like Java home path, etc.

Command**:** vi hadoop–env.sh



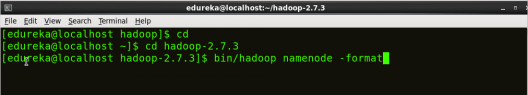
**C:\Users\SAI\Desktop\folder\Property-of-Hadoop-env-Install-Hadoop-Edureka-528x48.png**

### ****Step 12:**** Go to Hadoop home directory and format the NameNode.

Command**:** cd

Command**:** cd hadoop-2.7.3

Command**:** bin/hadoop namenode –format



Formatting the file system means initializing the directory specified by the dfs.name.dir variable.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

### ****Step 13:**** Once the NameNode is formatted, go to hadoop-2.7.3/sbin directory and start all the daemons.

Command: cd hadoop-2.7.3/sbin

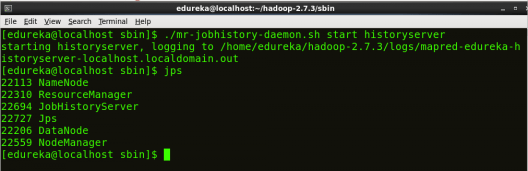
Either you can start all daemons with a single command or do it individually.

***Command:*** ./start-all.sh

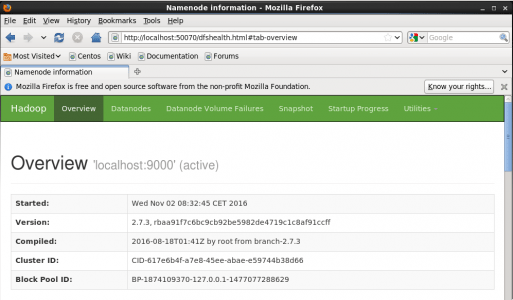
The above command is a combination of start-dfs.sh, start-yarn.sh & mr-jobhistory-daemon.sh

### ****Step 14:**** To check that all the Hadoop services are up and running, run the below command.

Command: jps



### ****Step 15:**** Now open the Mozilla browser and go to ****localhost****:****50070/dfshealth.html**** to check the NameNode interface.



**Acceptance criteria:**

A study of resource availability that may affect the ability to achieve an acceptable system. This evaluation determines whether the technology needed for the proposed system is available or not

* Can the work for the project be done with current equipment existing software technology & available personal?
* Can the system be upgraded if developed?
* If new technology is needed then what can be developed?

This is concerned with specifying equipment and software that will successfully satisfy the user requirement.

The technical needs of the system may include:

**Front-end and back-end selection**

An important issue for the development of a project is the selection of suitable front-end and back-end. When we decided to develop the project we went through an extensive study to determine the most suitable platform that suits the needs of the organization as well as helps in development of the project.

The aspects of our study included the following factors.

**Front-end selection:**

1. It must have a graphical user interface that assists employees that are not from IT background.

2. Scalability and extensibility.

3. Flexibility.

4. Robustness.

5. According to the organization requirement and the culture.

6. Must provide excellent reporting features with good printing support.

7. Platform independent.

8. Easy to debug and maintain.

9. Event driven programming facility.

10. Front end must support backend

According to the above stated features we selected **Pycharm** as the front-end for developing our project.

**Back-end Selection:**

1. Multiple user support.

2. Efficient data handling.

3. Provide inherent features for security.

4. Efficient data retrieval and maintenance.

5. Stored procedures.

6. Popularity.

7. Operating System compatible.

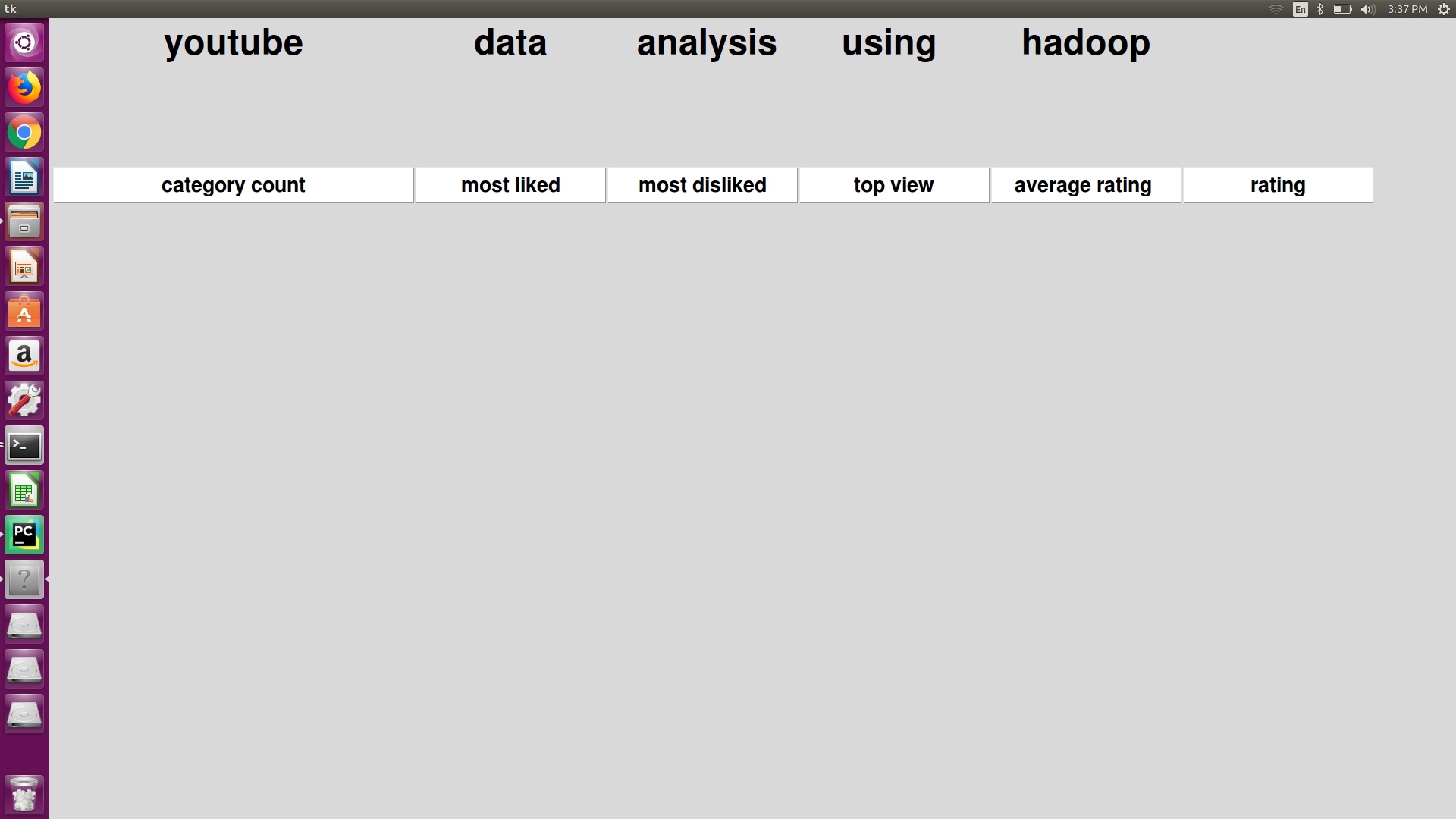
8. Easy to install.

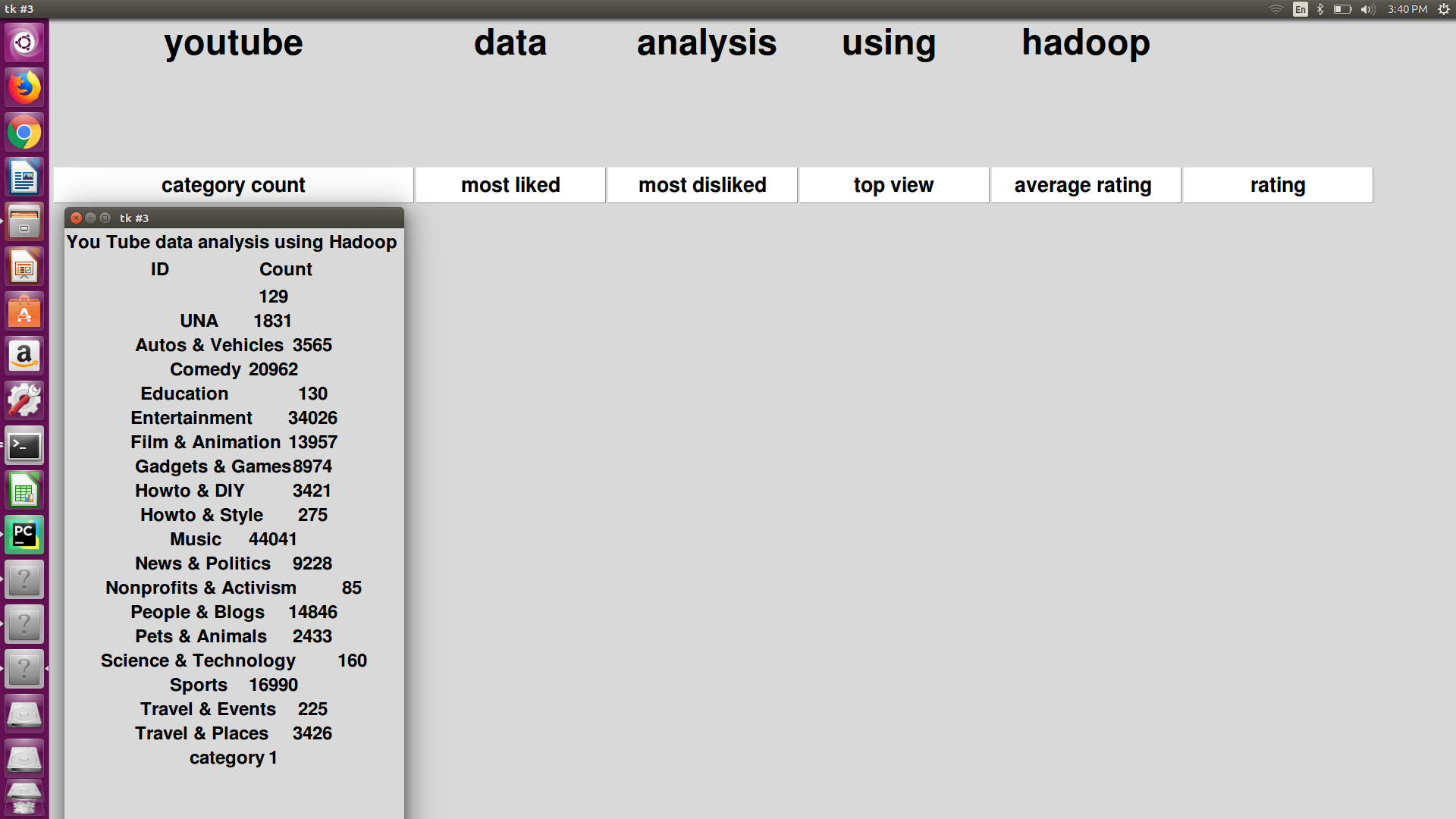
9. Various drivers must be available.

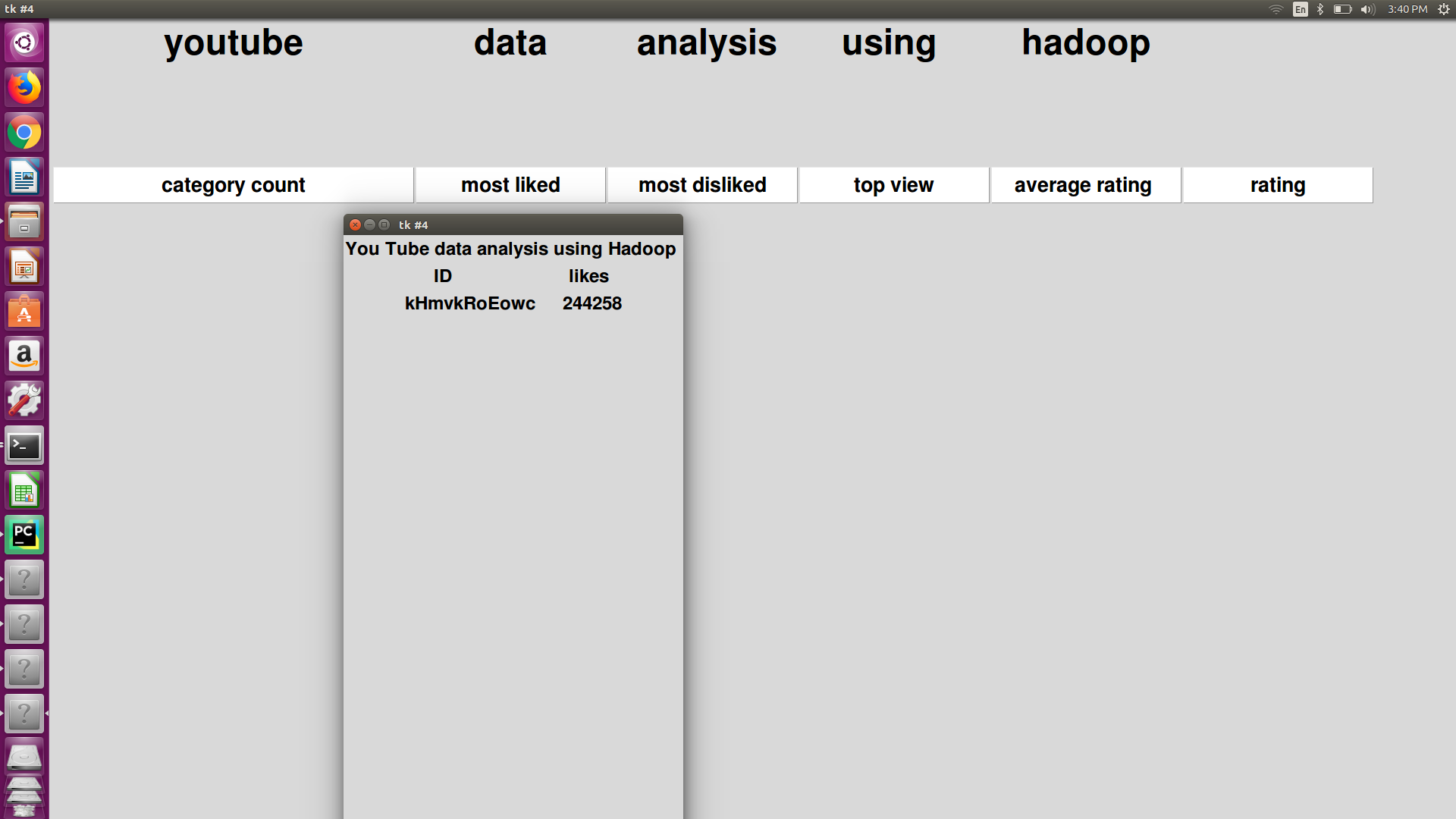
10. Easy to implant with the Front-end.

**Chapter 3**

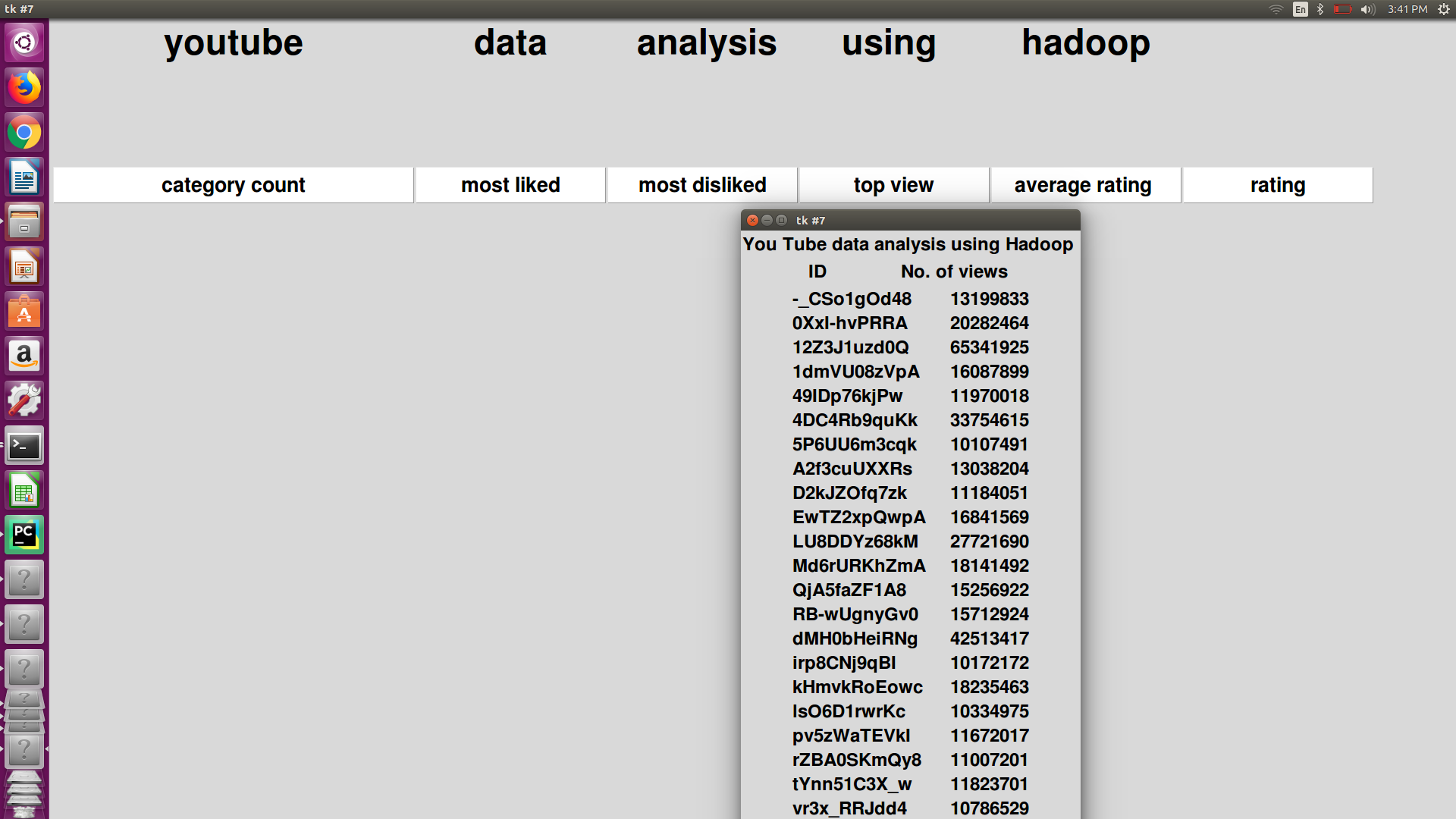
**UI design:**

****

**Output of category count: **

**Output of Most liked videos:** ****

## Output of Most Disliked videos: F:\SCREENSHOT\dislike.png

**Output of Most Top view:** ****

**Chapter 4**

**Coding:**

## DATA SET DESCRIPTION

|  |  |
| --- | --- |
| video ID | an 11-digit string, which is unique |
| uploader | a string of the video uploader's username |
| Age | an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment) |
| category | a string of the video category chosen by the uploader |
| Length | an integer number of the video length |
| Views | an integer number of the views |
| Rate | a float number of the video rate |
| Ratings | an integer number of the ratings |
| comments | an integer number of the comments |
| related Ids | up to 20 strings of the related video Ids |

### SOURCE CODE

Now from the mapper, we want to get the video category as key and final int value ‘1’ as values which will be passed to the shuffle and sortphase and are further sent to the reducer phase where the aggregation of the values is performed.

### MAPPER CODE

public class Top5\_categories {

public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>{

private Text category = new Text();

private final static IntWritable one = new IntWritable(1);

public void map(LongWritable key, Text value, Context context )

throws IOException, InterruptedException {

String line = value.toString();

String str[]=line.split("\t");

if(str.length > 5){

category.set (str[3]);

}

context.write(category, one);

}

}

### REDUCER CODE

public static class Reduce extends Reducer<Text, IntWritable,Text,IntWritable>{

public void reduce(Text key, Iterable<IntWritable> values,Context context throws IOException, InterruptedException {

int sum = 0;

for (IntWritable val : values) {

sum += val.get();

}

context.write(key, new IntWritable(sum));

}

}

.

### CONF CODE

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(IntWritable.class);

This two configuration classes are included in the main class whereas to clarify the Output key type of mapper and the output value type of the Mapper.

### ****HOW TO EXECUTE****

**hadoop jar top5.jar /youtubedata.txt /top5\_out**

### Output

Entertainment 911

Music 870

Comedy 420

Sports 253

Education 6

**Languages / tools used:**

**Language:** java with Hadoop packages, python

**Tools:** Hadoop setup is required, Pycharm to create a GUI

**Chapter 5**

**Test Plan:**

In this GUI we performing the Manual Testing for testing this GUI following testing types are used

* GUI testing
* Performance Testing
* Usability testing

**Testing:**

|  |  |  |
| --- | --- | --- |
| sr no | Test ID | Test case |
| 1 | TC-01 | verify that the heading is written in center |
| 2 | TC-02 | verify that the buttons are correctly placed one after another |
| 3 | TC-03 | verify that the buttons are not overlapped |
| 4 | TC-04 | check that the buttons are working properly |
| 5 | TC-05 | check that buttons are performing the right functions |
| 6 | TC-06 | check that it is easy to use for User |

**References:**

[**https://hadoop.apache.org**](https://hadoop.apache.org)

[**https://acadgild.com/**](https://acadgild.com/)

[**https://www.edureka.co/blog/install-hadoop-single-node-hadoop-cluster**](https://www.edureka.co/blog/install-hadoop-single-node-hadoop-cluster)