UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

Design of Mechanism for Enhancement the Security of Hadoop Processing Tool "Hive" With VMWARE Platform

Anjali Devi

Master of Technology, Department of Computer Science & Engineering, Bhagat Phool Singh Mahila Vishvavidyalya, Khanpur Kalan, Sonipat, Haryana, India

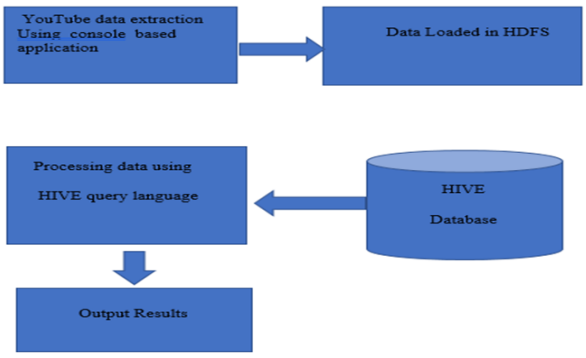
Abstract: Today's we are active in the age of big information with complex hazards. In the ancient time, the volume of the data will be less the well managed structure, form data provide a greater output using RDBMS concept. The current theme defines the unstructured and semi structure data define the complexity as a challenging area So on this day we reduce the problem with the Hadoop concept with fetching huge information, manage, provide security. RDBMS doesn't handle this problem so the technology moves for the new platform i.e. called Hadoop. To write this paper I have Study about improving data security by means of Hadoop as well as Crypto technique with HIVE ecosystem. Next the detailed study of Hadoop using hive w.r.t cryptography we define Kerberos algorithm for the security purpose, I am contribution my work. This paper is distributing in four sectors. In sector-I, I am donating just basic introduction about Hadoop with hive algorithm for improving data Security. In sector-II, I am present detailed description, in sector-III, I am presenting working procedure with results, and in sector IV, I am offering conclusion and references where I have completed my research.

Keywords: RDBMS Solutions, Big Data Analytics, Hadoop, Big Data Platform, GFS, HDFS, MapReduce, Real-Time Processing, Hive, Pig, VMware, Cloudera.

I. INTRODUCTION

In the Hadoop, we have HDFS to stores filesystem metadata and request data distinctly. As in other distributed filesystems, like GFS stands for google file system, then HDFS notation defined as a Hadoop distributed file system to supply metadata on a devoted server, as a NameNode. Hadoop MapReduce is a database design model and software context for writing submissions that quickly procedure vast bulks of data into equivalent on big clusters of computer arrangements, as a Data Node. In this paper, I have provided an overview, architecture and its various property and implementations with security of data in Cloudera Environments using VMware.

II. DETAILED DESCRIPTION



High Level Flow Diagram of HIVE Using YouTube

HADOOP

Hadoop is an open-source framework to store w.r.t HDFS and process w.r.t MapReduce, Big Data in a distributed environment. Hadoop having two modules Hadoop Distributed File System (HDFS) and MapReduce

MapReduce: MR is a parallel programming model for processing bulky amounts of structured, semi-structured, and unstructured information on large clusters of cheap hardware.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

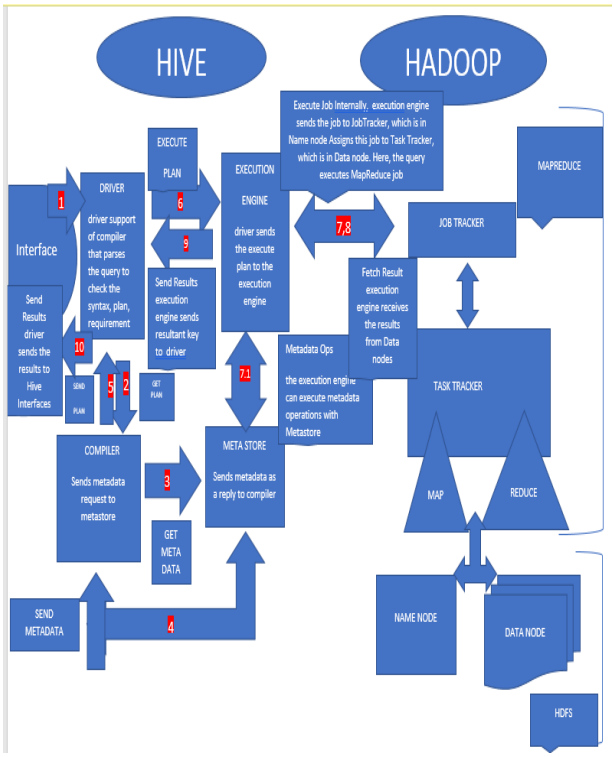
ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

HDFS: Hadoop Distributed File System is a part of Hadoop framework, used to store the datasets. It provides a fault-tolerant file system to run on commodity hardware.

HIVE-

Hadoop ecosystem Hive is used to develop Structure Query Language type scripts to do MapReduce fundamentals.[1]



Solution Extraction Using HIVE

- **Step 1** Use the following command to 'Create a Table' in HIVE
- This command will create a Hive table named 'YouTube_data_table' in which rows will be delimited and rows fields will be terminated by commas.
- **Step 2** Load YouTube data into the Hive Table Use the command given below to load YouTube data into the Hive table created in
- Hive> load data local inpath'/home/MyYouTubeProject/YouTubeDataset.txt' overwrite into table YouTube_data_table;
- This command will load the YouTube dataset from the given path to the table (YouTube_data_table) created in Hive so that meaningful processing of the dataset can be done using the Hive queries
- Calculate the result using map reduce ecosystem give top channels 5 categories with maximum number of videos uploaded, the top 10 rated videos in YouTube, Security with Authentication.

CLOUDERA

Cloudera Inc. is a United States-based software company that provides Apache Hadoop-based software, support and services, and training to business customers.

Cloudera's open-source Apache Hadoop distribution, Cloudera Distribution Including Apache Hadoop targets enterprise-class deployments of that technology.

Cloudera speaks that more than 50% of its engineering output is donated upstream to the various Apache-licensed open source projects that combine to form the Hadoop platform.

Cloudera is also a sponsor of the Apache Software Foundation. Three engineers from Google, Yahoo and Facebook (Christophe Bisciglia, Amr Awadallah and Jeff Hammerbacher, respectively) joined with a former Oracle executive (Mike Olson) to form Cloudera in 2008

Architect Doug Cutting, also a former chairman of the Apache Software Foundation, authored the open-source Lucene and Nutch search technologies before he wrote the initial Hadoop software in 2004.

He designed and managed a Hadoop storage and analysis cluster at Yahoo! before joining Cloudera in 2009. Service Apache Hadoop distribution with support, professional services and training. [2]

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

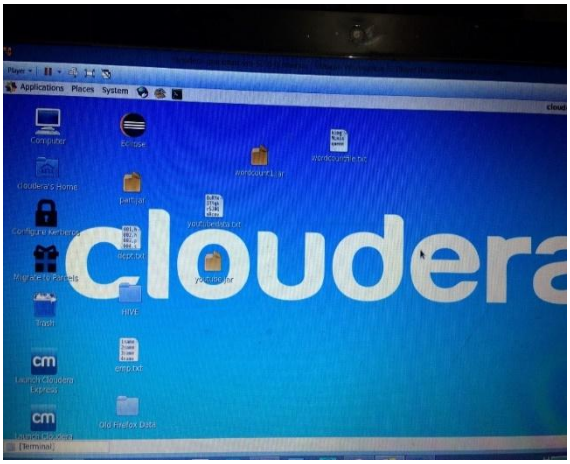
Vol. 4, Issue 7, July 2017

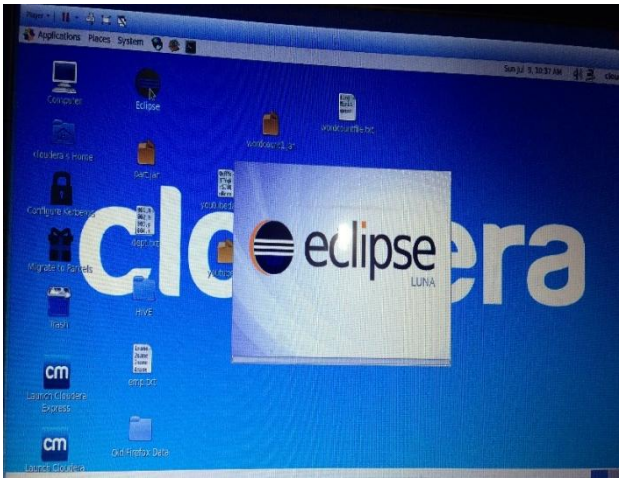
VMWARE –

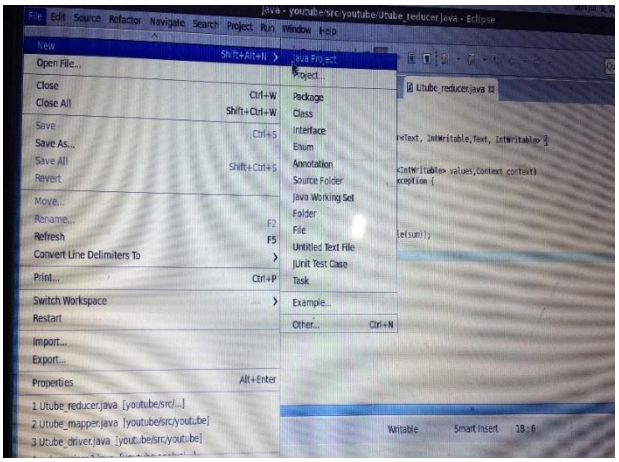
VMware Workstation includes the ability to group multiple virtual machines in an inventory folder. The machines in such a folder can then be powered on and powered off as a single object, useful for testing complex client-server environments Development status is Active. Written in F#, C, C++ Operating system Windows Linux, Platform x86-64 only. Type Hypervisor License Freeware (Workstation Player) Trialware and commercial (Workstation Pro) [3]

III. IMPLEMENTATIONS ALGORITHM WITH RESULT

Cloudera QuickStart Virtual Machines single-node cluster style it easy to quickly become hands-on with Cloudera Distribution Including Apache Hadoop for testing, demo, and self-learning purposes, and include Cloudera Manager for managing your cluster. Cloudera QuickStart VM platform also includes a sample data, and scripts for getting started. In the eclipse, we create java project YouTube name and classes like driver, mapper, reducer.







UGC Approved Journal

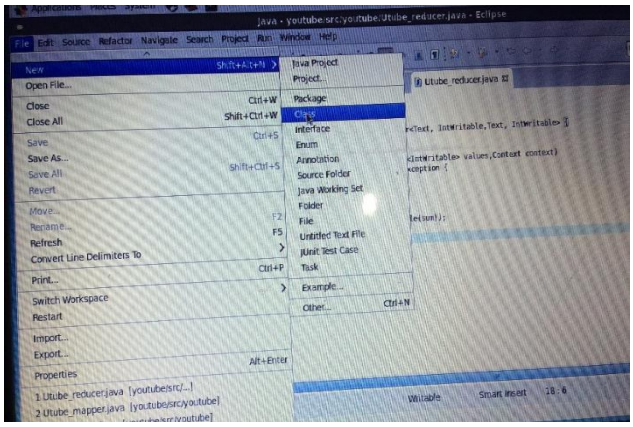
IARJSET

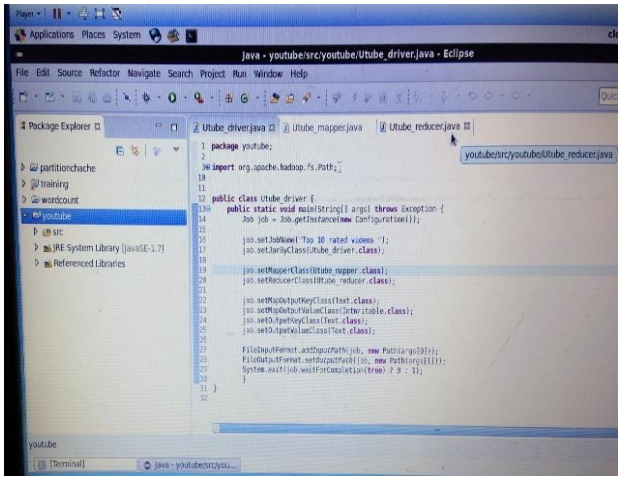
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

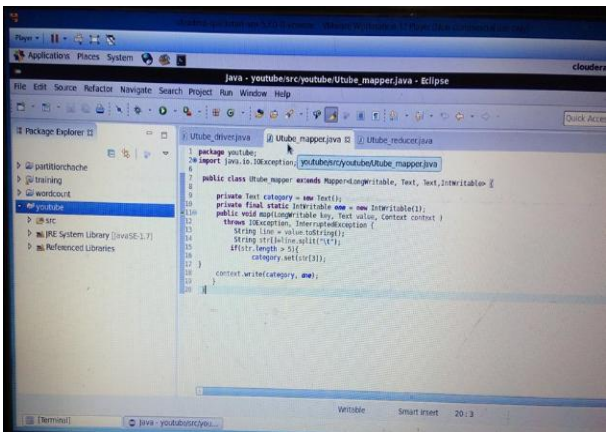
International Advanced Research Journal in Science, Engineering and Technology

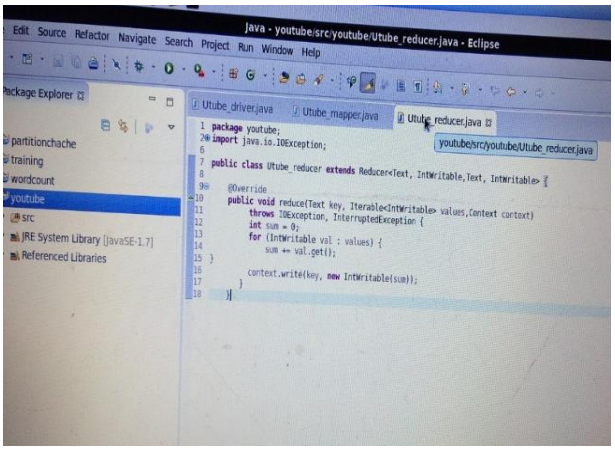
ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017









UGC Approved Journal

IARJSET

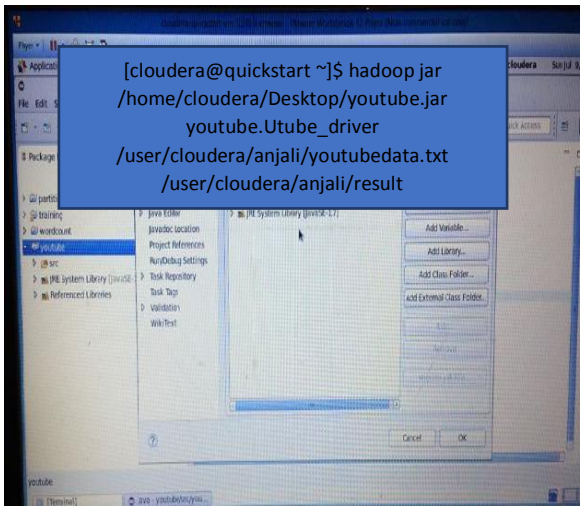
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

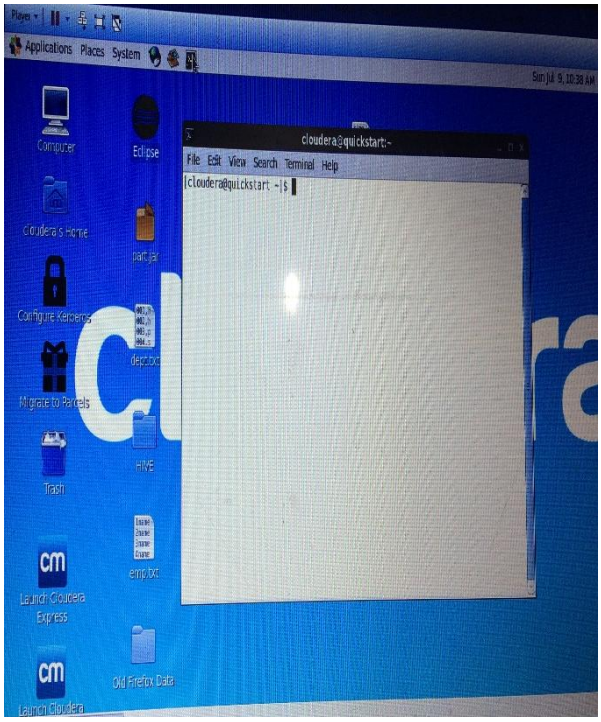
Vol. 4, Issue 7, July 2017

After creating project, classes. Now add jar files which is as follow: -



Check the cloudera desktop there by default created jar file folder with the given name by user. Now, move to terminal mode so this is basic command for the cloudera platform.

```
[cloudera@quickstart ~]$
```



After adding jar files, we have defined the location with creating class name as driver with database location, at last we define result location

If the command is correctly working they move automatically for the YARN resource manager connectivity with the address location

```
17/06/17 08:23:49 INFO client.RMProxy: Connecting to
ResourceManager at /0.0.0.0:8032
```

UGC Approved Journal

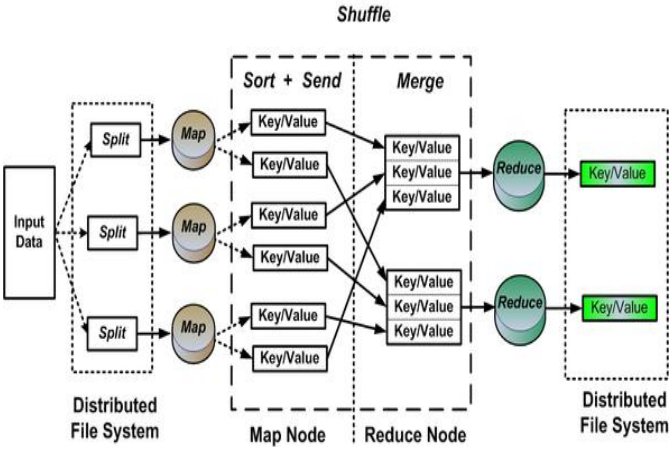
IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

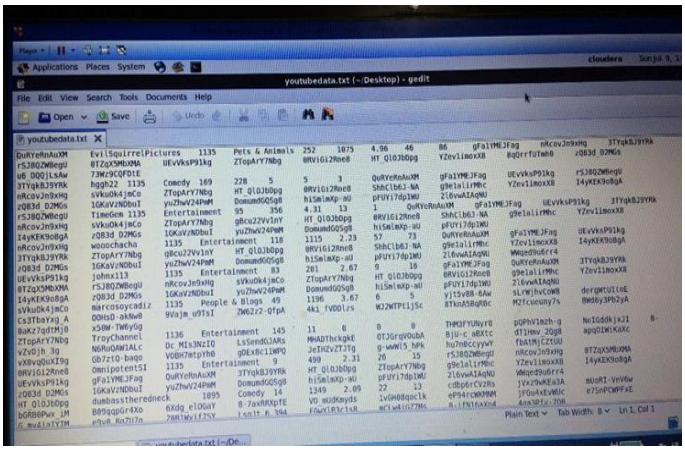
International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017



Block Diagram For The MapReduce [4]



This is database for the working area as a implementation zone According to, block diagram of MapReduce we input, MapReduce job submitter, shuffle, reduce, output as a key value format.

17/06/17 08:23:52 INFO input.FileInputFormat: Total input paths to process : 1

17/06/17 08:23:52 INFO mapreduce.JobSubmitter: number of splits:1

17/06/17 08:25:16 INFO mapreduce.Job: Counters: 49

File System Counters

FILE: Number of bytes read=73143

FILE: Number of bytes written=373779

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=969522

HDFS: Number of bytes written=257

HDFS: Number of read operations=6

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

Job Counters

Launched map tasks=1

Launched reduce tasks=1

Data-local map tasks=1

Total time spent by all maps in occupied slots (ms)=29250

Total time spent by all reduces in occupied slots (ms)=12970

Total time spent by all map tasks (ms)=29250

Total time spent by all reduce tasks (ms)=12970

Total vcore-seconds taken by all map tasks=29250

Total vcore-seconds taken by all reduce tasks=12970

Total megabyte-seconds taken by all map tasks=29952000

Total megabyte-seconds taken by all reduce tasks=13281280

Map-Reduce Framework

Map input records=4100

Map output records=4100

Map output bytes=64937

Map output materialized bytes=73143

Input split bytes=133

Combine input records=0

Combine output records=0

Reduce input groups=15

Reduce shuffle bytes=73143

Reduce input records=4100

Reduce output records=15

Spilled Records=8200

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

GC time elapsed (ms)=730

CPU time spent (ms)=8560

Physical memory (bytes) snapshot=347336704

Virtual memory (bytes) snapshot=3007991808

Total committed heap usage (bytes)=226562048

Shuffle ErrorsBAD_ID=0

CONNECTION=0

IO_ERROR=0

WRONG_LENGTH=0

WRONG_MAP=0

WRONG_REDUCE=0

File Input Format Counters

Bytes Read=969389

File Output Format Counters

Bytes Written=257

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

The working of commands is correctly, so using this location, which path is given during processing is display the result with the file name.

```
[cloudera@quickstart ~]$ hadoop fs  -ls /user/cloudera/anjali/result
Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2017-06-17 08:25
/user/cloudera/anjali/result/_SUCCESS
-rw-r--r--   1 cloudera cloudera        257 2017-06-17 08:25
/user/cloudera/anjali/result/part-r-00000
```

Final output, shown w.r.t this cat command and result location so in the anjali name folder having file with the output as a new database according to user requirements

```
[cloudera@quickstart ~]$ hadoop fs -cat
/user/cloudera/anjali/result/part-r-00000
```

```
UNA        32
Autos & Vehicles 77
Comedy    420
Education  65
Entertainment 911
Film & Animation 261
Howto & Style 138
Music 870
News & Politics 343
Nonprofits & Activism              43
People & Blogs 399
Pets & Animals 95
Science & Technology               80
Sports 253
Travel & Events 113
[cloudera@quickstart ~]$
```

IV. CONCLUSION

The mission of big data analysis is not only significant but also a requirement. In fact, various administrations that have applied Big Data are understanding significant inexpensive benefit linked to further establishments with no Big Data efforts. In this planned to analyse the YouTube Big Data and come up with important insights which cannot be strongminded otherwise. The output results of YouTube data security analysis project show key insights that can be induced to other use cases as well. One of the output results labels that for a specific uploader of the video, Interval between the day of establishment of YouTube and the date of uploading of the video, Category of the video, Length of the video, Number of views for the video, Rating on the video, Number of ratings given for the video, Number of comments done on the videos, Related video ids with the uploaded video. Most of the companies are uploading their product launch on YouTube and them excitedly await their subscribers' reviews. Handle data sets that do not have proper structure and how to sort the output of reducer.

REFERENCES

[1] https://www.tutorialspoint.com/hive/hive_quick_guide.htm
[2] https://en.wikipedia.org/wiki/Cloudera
[3] https://en.wikipedia.org/wiki/VMware_Workstation
[4] https://www.google.co.in/search?q=mapreduce+architecture&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjUvZSh2PzUAhWMP48KHQ-DAZoQ_AUICigB&biw=1336&bih=611#imgdii=smJGzaHE14WD7M:&imgrc=VxfMxnsADcHsEM: