IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Big data analytics using Apache Hadoop: A case study on different fertilizers requirement and availability in different states of India from 2012-2013 to 2014-2015

Avinash Pratap Budaragade
avinashpb1008@gmail.com
Visveswaraya Technological University, Belgaum, Karnataka

Jones Temitope Mary
topsonwise86@gmail.com
The Federal University of Technology Akure, Ondo, Nigeria

ABSTRACT

In today's digital world large volume of data is being generated from various sources including social media, healthcare, transportation, industries, sensors, etc. This data includes structured, semi-structured and unstructured data. This huge volume of data cannot be stored and processed using traditional systems thus it is termed as big data. To store and analyze this type of data parallel storage and analysis is required. This can be achieved by using big data analytics. Using Apache Hadoop such huge volume data can be analyzed efficiently. In this paper, a case study is performed on different fertilizers requirement and availability in different states of India in three years from 2012-2013 to 2014-2015 using Apache Hadoop.

Keywords*— Big data, Apache hadoop, Apache pig, Pig latin, HDFS*

1. INTRODUCTION

The digital universe is flooded with a large amount of data generated by the number of users worldwide. These data are of diverse in nature, come from various sources and in many forms. Every time we use the Internet, send an email, make a phone call, or pay a bill, we create data. All this data needs to be stored in huge data chunks. These data chunks are stored in thousands of disks or hard drives. It consists of structured data as relational data, semi-structured data as XML data & unstructured data as Word, PDF, Text, and Media logs. Combination of all these contains a huge amount of information. Big data is a collection of complex and large data sets, which include information, may be produced by multiple services such as Black Box Data, Social media, Stock exchange, Search engine, sensors used for climate information, digital pictures, traffic, software logs etc.

Big Data is not just about being big in size. The definition is broadened using five characteristics or "V's". These are:

- **Volume:** This characteristic signifies huge voluminous data; it is in orders of terabytes and even pet bytes.
- **Velocity:** This characteristic signifies the high velocity with which the data is generated.
- **Variety:** This characteristic refers to the huge variety in the big data.
- **Value:** This characteristic refers to the intrinsic value contained in big data.
- **Veracity:** This characteristic refers to uncertainties in big data such as missing, duplicate and incomplete entries.

This huge volume of data cannot be stored and processed using traditional systems thus it is termed as big data. To store and analyze this type of data parallel storage and analysis is required. This can be achieved by using big data analytics. Using apache Hadoop such huge volume data can be analyzed efficiently. In this paper data on fertilizers requirement and availability in different states of India in three years from 2012-2013 to 2014-2015 is collected and can be analyzed using big data analytics technology.

Hadoop is an Apache open source framework written in Java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from a single server to thousands of machines, each offering local computation and storage.

Along with the storage of data parallel, we need to process that data. This can be done by using an open source data processing tool called apache pig. Pig is an open-source high-level data flow system. It provides a simple language called Pig Latin, for queries and data manipulation, which are then compiled into MapReduce jobs that run on Hadoop. Pig is important as companies like Yahoo, Google and Microsoft are collecting huge amounts of data sets in the form of click streams, search logs and web crawls. Pig is also used in some form of ad-hoc processing and analysis of all the information.

To write data analysis programs, Pig provides a high-level language known as Pig Latin. This language provides various operators using which programmers can develop their own functions for reading, writing, and processing data. To analyze data using Apache Pig, programmers need to write scripts using Pig Latin language. All these scripts are internally converted to Map and Reduce tasks. Apache Pig has a component known as Pig Engine that accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs.

2. RELATED WORK

Various Big Data analytics tools in terms of Big Data Process in [1]. In this paper, authors proposed a framework for the selection of tool as each stage in the data process should use the appropriate tool for that stage for optimum utilization of CPU time, cost and accuracy. Authors conducted a case study that proved the efficiency in tool selection and utilization will lead to the efficient management of data and decision making.

Bigdata analytics is performed in [2], using pig and hive on significant issues faced by consumers which helps the institutions or corporations to rectify these issues, provide proper satisfaction to the consumers, improvement in services, to keep a check on issues and to build up goodwill in the market. This provides consumers to distinguish properly among the institutions and make the service provider selection vigorously

Storage, processing and analyzing big data by using apache Hadoop and apache pig by taking an example of crime datasets from the year 2000 to 2014 is done in [3]. Authors showed visually how crimes against women are becoming an increasingly worrying and disturbing problem for the government. They found a number of such crimes must be found, especially the ones against young women (age between 18-30 years) [3].

3. OBJECTIVES AND SOFTWARE REQUIREMENT

This subsection describes the objectives of the proposed work and also provides information about tools used for data analysis on fertilizers requirement and availability.

3.1 Objectives

- To load the datasets of different chemical fertilizers requirement and availability in different states of India on apache HDFS.
- To integrate the different datasets and process those datasets to analyze the following queries using apache pig.

Query 1: What is the overall requirement and availability of different fertilizers?
Query 2: Check the requirement and availability of different fertilizers in different states.
Query 3: Check the shortage of different fertilizers in Karnataka state.
Query 4: What is the requirement and availability UREA in Karnataka state in the year 2012-2013.
Query 5: Which is the most demandable fertilizer from 2012-13 to 2014-15.

3.2 Software requirement

Linux operating system is used for the implementation of this research. Software requirement for implementation is listed below:

- Apache Hadoop-3.0.3
- Apache Pig-0.17.0
- JDK-11.0.1
- Visualization Engine v3.0

4. SYSTEM ARCHITECTURE

The scalability of increase in volume, velocity and variety of data in an organization will be benefited by selecting appropriate big data technologies. Appropriate selection of tool will be the basis of global competition result in optimum investment in big data analytics, production growth and strengthening consumer surplus. Big Data Analytics tool made the entire data management cycle technically and economically feasible from the collection and storing of larger datasets to analyze the data in order to provide new and valuable insights.

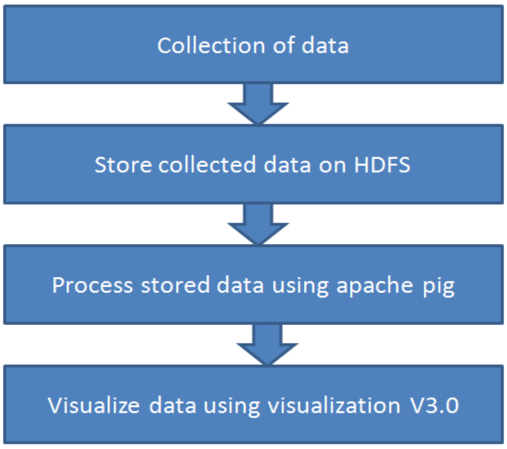

Collection of data
Store collected data on HDFS
Process stored data using apache pig
Visualize data using visualization V3.0

Fig. 1: Big data analytics tool process

- **Collection of data:** Data is collected from the government database on fertilizer requirement and availability.
- **Storage of data:** Collected data is stored on the Hadoop Distributed file system (HDFS).
- **Processing:** Data stored on HDFS is processed by apache pig using apache pig Latin.
- **Visualizing:** After processing data, there is a need for visualization. Visualization V3.0 is used for visualization.

5. IMPLEMENTATION AND RESULTS

This subsection describes the implementation and results of the proposed work. All the implantation are done by using Ubuntu 18.04-LTS.

To check the versions of Apache Hadoop, apache pig and java following commands can be executed.

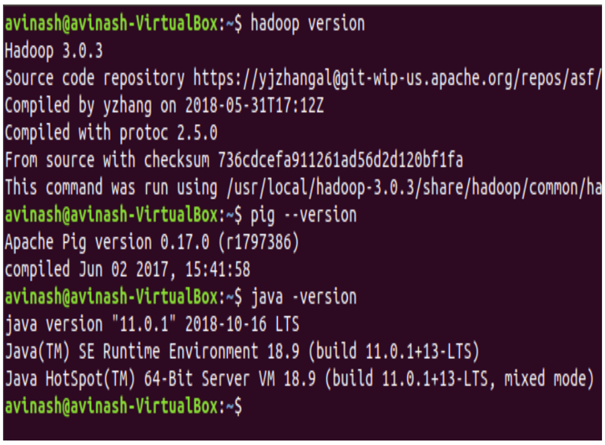
```
$ Hadoop version
$ pig -- version
$ java –version
```

Implementation of objectives

Objective 1: To load the datasets of different chemical fertilizers requirement and availability in different states of India on apache HDFS.

Following commands are executed to start HDFS, apache pig and load datasets.

```
$ service sshd restart
$ cd /usr/local/hadoop-3.0.3
$ ./start-all.sh
$ pig –h;
$ pig –x local;
```

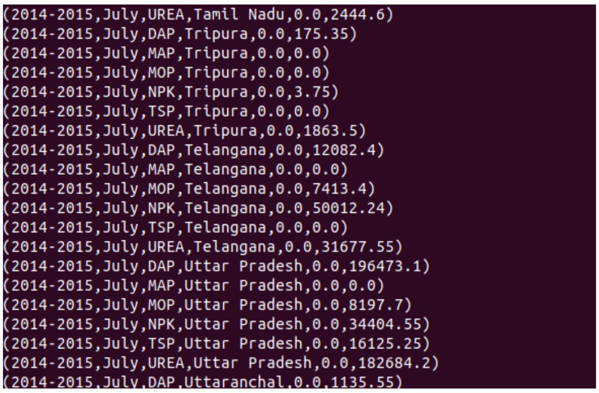


```
avinash@avinash-VirtualBox:~$ hadoop version
Hadoop 3.0.3
Source code repository https://yjzhangal@git-wip-us.apache.org/repos/asf/
Compiled by yzhang on 2018-05-31T17:12Z
Compiled with protoc 2.5.0
From source with checksum 736cdcefa911261ad56d2d120bf1fa
This command was run using /usr/local/hadoop-3.0.3/share/hadoop/common/ha
avinash@avinash-VirtualBox:~$ pig --version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
avinash@avinash-VirtualBox:~$ java -version
java version "11.0.1" 2018-10-16 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.1+13-LTS)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.1+13-LTS, mixed mode)
avinash@avinash-VirtualBox:~$
```

Fig. 2: Apache hadoop, pig and Java version

To load the dataset and show the loaded dataset following commands are executed

```
grunt > A = LOAD ‘all_data.txt’ AS (year:chararray,
month:chararray, product:chararray, state:chararray,
requirement:double, availability:double);
grunt >DUMP A;
```



```
(2014-2015,July,UREA,Tamil Nadu,0.0,2444.6)
(2014-2015,July,DAP,Tripura,0.0,175.35)
(2014-2015,July,MAP,Tripura,0.0,0.0)
(2014-2015,July,MOP,Tripura,0.0,0.0)
(2014-2015,July,NPK,Tripura,0.0,3.75)
(2014-2015,July,TSP,Tripura,0.0,0.0)
(2014-2015,July,UREA,Tripura,0.0,1863.5)
(2014-2015,July,DAP,Telangana,0.0,12082.4)
(2014-2015,July,MAP,Telangana,0.0,0.0)
(2014-2015,July,MOP,Telangana,0.0,7413.4)
(2014-2015,July,NPK,Telangana,0.0,50012.24)
(2014-2015,July,TSP,Telangana,0.0,0.0)
(2014-2015,July,UREA,Telangana,0.0,31677.55)
(2014-2015,July,DAP,Uttar Pradesh,0.0,196473.1)
(2014-2015,July,MAP,Uttar Pradesh,0.0,0.0)
(2014-2015,July,MOP,Uttar Pradesh,0.0,8197.7)
(2014-2015,July,NPK,Uttar Pradesh,0.0,34404.55)
(2014-2015,July,TSP,Uttar Pradesh,0.0,16125.25)
(2014-2015,July,UREA,Uttar Pradesh,0.0,182684.2)
(2014-2015,July,DAP,Uttaranchal,0.0,1135.55)
```

Fig. 3: Dataset loaded into apache HDFS

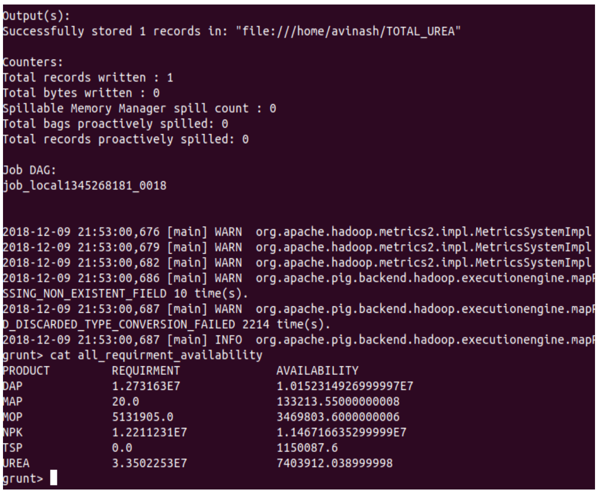
Objective 2: To integrate the different datasets and process those datasets to analyze the following queries using apache pig.

Query 1: What is the overall requirement and availability of different fertilizers?

Total requirement and availability of all fertilizers from all states from 2012-13 to 2014-15 can be calculated by executing the following commands.

```
grunt > FILTER_DAP = FILTER A BY PRODUCT == ‘DAP’;
grunt > GROUP_DAP = GROUP FILTER_DAP ALL;
grunt > RESULT_DAP = FOREACH GROUP_DAP
GENERATE SUM(FILTER_DAP.requirement),
(FILTER_DAP.availability);
grunt > STORE RESULT_DAP INTO ‘FINAL_DAP’ USING
PigStorage(‘,’);

grunt > FILTER_MAP = FILTER A BY PRODUCT ==
‘MAP’;
grunt > GROUP_MAP = GROUP FILTER_MAP ALL;
grunt > RESULT_MAP = FOREACH GROUP_MAP
GENERATE SUM(FILTER_MAP.requirement),
(FILTER_MAP.availability);
grunt > STORE RESULT_MAP INTO ‘FINAL_MAP’ USING
PigStorage(‘,’);
grunt > FILTER_MOP = FILTER A BY PRODUCT ==
‘MOP’;
grunt > GROUP_MOP = GROUP FILTER_MOP ALL;
grunt > RESULT_MOP = FOREACH GROUP_MOP
GENERATE
SUM(FILTER_MOP.requirement),
(FILTER_MOP.availability);
grunt > STORE RESULT_MOP INTO ‘FINAL_MOP’ USING
PigStorage(‘,’);
grunt > FILTER_NPK = FILTER A BY PRODUCT == ‘NPK’;
grunt > GROUP_NPK = GROUP FILTER_NPK ALL;
grunt > RESULT_NPK = FOREACH GROUP_NPK
GENERATE
SUM(FILTER_NPK.requirement), (FILTER_NPK.availability);
grunt > STORE RESULT_NPK INTO ‘FINAL_NPK’ USING
PigStorage(‘,’);
grunt > FILTER_TSP = FILTER A BY PRODUCT == ‘TSP’;
grunt > GROUP_TSP = GROUP FILTER_TSP ALL;
grunt > RESULT_TSP = FOREACH GROUP_TSP
GENERATE
SUM(FILTER_TSP.requirement), (FILTER_TSP.availability);
grunt > STORE RESULT_TSP INTO ‘FINAL_TSP’ USING
PigStorage(‘,’);
grunt > FILTER_UREA = FILTER A BY PRODUCT ==
‘UREA’;
grunt > GROUP_UREA = GROUP FILTER_UREA ALL;
grunt > RESULT_UREA = FOREACH GROUP_UREA
GENERATE
SUM(FILTER_UREA.requirement),
(FILTER_UREA.availability);
grunt > STORE RESULT_UREA INTO ‘FINAL_UREA’
USING PigStorage(‘,’);
grunt > ALL_DATA = JOIN FILTER_DAP BY requirement,
FILTER_MAP BY requirement, FILTER_MOP BY requirement,
FILTER_NPK BY requirement, FILTER_TSP BY requirement,
FILTER_UREA BY requirement
grunt > ALL_DATA = JOIN FILTER_DAP BY availability,
FILTER_MAP BY availability, FILTER_MOP BY availability,
FILTER_NPK BY availability, FILTER_TSP BY availability,
FILTER_UREA BY availability
```



```
Output(s):
Successfully stored 1 records in: "file:///home/avinash/TOTAL_UREA"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1345268181_0018


2018-12-09 21:53:00,676 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl
2018-12-09 21:53:00,679 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl
2018-12-09 21:53:00,682 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl
2018-12-09 21:53:00,686 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapR
SSING_NON_EXISTENT_FIELD 10 time(s).
2018-12-09 21:53:00,687 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapR
D_DISCARDED_TYPE_CONVERSION_FAILED 2214 time(s).
2018-12-09 21:53:00,687 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapR
grunt> cat all_requirment_availability
PRODUCT         REQUIRMENT              AVAILABILITY
DAP             1.273163E7              1.0152314926999997E7
MAP             20.0                    133213.55000000008
MOP             5131905.0               3469803.6000000006
NPK             1.2211231E7             1.1467166352999999E7
TSP             0.0                     1150087.6
UREA            3.3502253E7             7403912.038999998
grunt> █
```

Fig. 4: Total requirement and availability of different fertilizers

Query 2: Check the requirement and availability of different fertilizers in different states.

Requirement and availability of different fertilizers in all states from 2012-13 to 2014-2015 are shown figure xxx is done from data got from query 1.

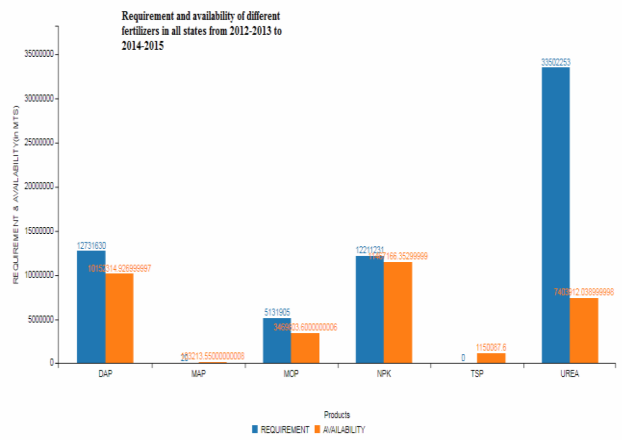
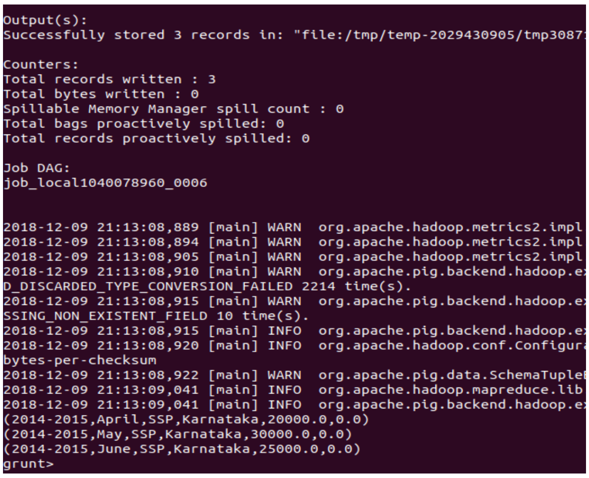


Fig. 5: Requirement and availability of different fertilizers in all states from 2012-13 to 2014-2015

Query 3: Check the shortage of different fertilizers in Karnataka state.

To find the shortage of all fertilizers in Karnataka state from 2012-13 to 2014-15 following command is executed.

grunt > C = FILTER A BY state == 'Karnataka' AND requirement != 0 AND availability == 0;
grunt > DUMP C;



```
Output(s):
Successfully stored 3 records in: "file:/tmp/temp-2029430905/tmp3087

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1040078960_0006


2018-12-09 21:13:08,889 [main] WARN  org.apache.hadoop.metrics2.impl
2018-12-09 21:13:08,894 [main] WARN  org.apache.hadoop.metrics2.impl
2018-12-09 21:13:08,905 [main] WARN  org.apache.hadoop.metrics2.impl
2018-12-09 21:13:08,910 [main] WARN  org.apache.pig.backend.hadoop.e
D_DISCARDED_TYPE_CONVERSION_FAILED 2214 time(s).
2018-12-09 21:13:08,915 [main] WARN  org.apache.pig.backend.hadoop.e
SSING_NON_EXISTENT_FIELD 10 time(s).
2018-12-09 21:13:08,915 [main] INFO  org.apache.pig.backend.hadoop.e
2018-12-09 21:13:08,920 [main] INFO  org.apache.hadoop.conf.Configur
bytes-per-checksum
2018-12-09 21:13:08,922 [main] WARN  org.apache.pig.data.SchemaTuple
2018-12-09 21:13:09,041 [main] INFO  org.apache.hadoop.mapreduce.lib
2018-12-09 21:13:09,041 [main] INFO  org.apache.pig.backend.hadoop.e
(2014-2015,April,SSP,Karnataka,20000.0,0.0)
(2014-2015,May,SSP,Karnataka,30000.0,0.0)
(2014-2015,June,SSP,Karnataka,25000.0,0.0)
grunt>
```

Fig. 6: Shortage of fertilizers in Karnataka state

Query 4: What is the requirement and availability UREA in Karnataka state in the year 2012-2013?

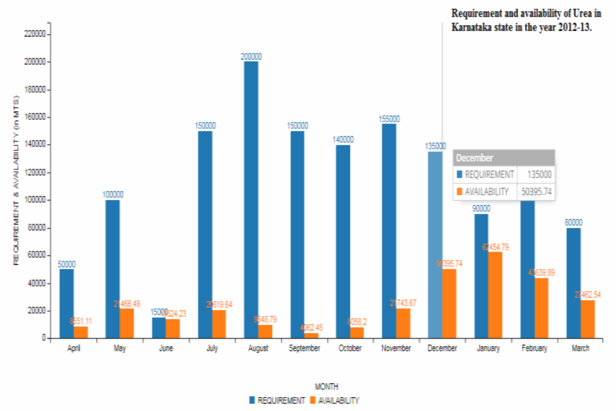


Fig. 7: Requirement and availability of UREA in Karnataka state

Query 5: Which is the most demandable fertilizer from 2012-13 to 2014-15.

Below figure 8 shows the requirement of different fertilizers from all states of India from 2012-13 to 2014-15. The pie chart indicates Urea is the most demandable fertilizer followed by DAP.

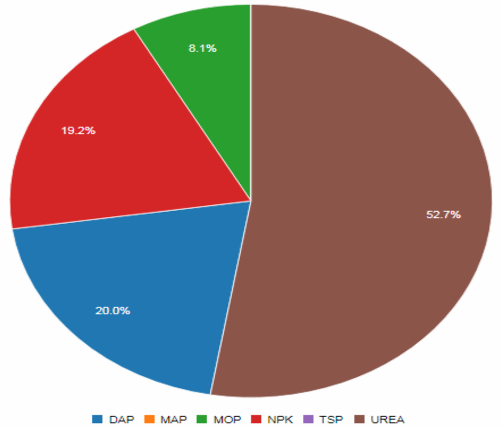


Fig. 8: Demands for different fertilizers from 2012-13 to 2014-2015

6. CONCLUSION

Big data refers to the huge volume of data that cannot be stored and processed using the traditional system. This data includes structured, semi-structured and unstructured data. To store and analyze this type of data parallel storage and analysis is required. This can be achieved by using big data analytics. Using apache Hadoop such huge volume data can be analyzed efficiently. Datasets on fertilizer requirement and availability provided by the government are processed and analyzed to get various results, which helps to understand the requirement, availability, shortage of different fertilizers, and most demandable fertilizer in different states of India in different years.

7. REFERENCES

[1] Sanjib Kumar Sahu, M Mary Jacintha, Amit Prakash Singh (2017). "Comparative study of tools for Big Data Analytics: An Analytical Study". International Conference on Computing, Communication and Automation (ICCCA2017).

[2] Pooja Jain, Prof. Jay Prakash Maurya (2017). "Comparative Analysis Using Hive and Pig on Consumers Data". (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 8 (2), 2017, 285-291.

[3] Arushi Jaina, Vishal Bhatnagar(2015). "Crime Data Analysis Using Pig with Hadoop". International Conference on Information Security & Privacy, Procedia Computer Science 78 (2016) 571 – 578.

[4] Ms Sarika Rathi (2017). "A Brief Study of Big Data Analytics using Apache Pig and Hadoop Distributed File System". International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 6, Issue 1.

[5] www.data.gov.in