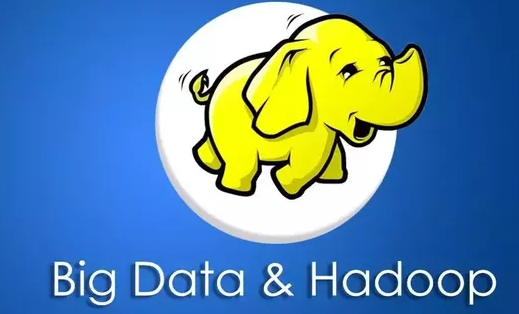
**STATE WISE**

**DEVELOPMENT**

**ANALYSIS**

**ACADGILD**



**acadgild.com**

**LEARN.DO.LEARN**

ABOUT ACADGILD

ACADGILD is a technology education startup that aims to create an ecosystem for skilled development in which people can learn from mentors and from each other. We believe that soft­ware development requires highly specialized skills that are best learned with guidance from experienced practitioners. Online videos or classroom formats are poor substitutes for building real projects with help from a dedicated mentor. Our mission is to teach hands-on, job-ready soft­ware programming skills, globally, in small batches of 8 to 10 students, using industry experts.

**ACADGILD** offers courses in:



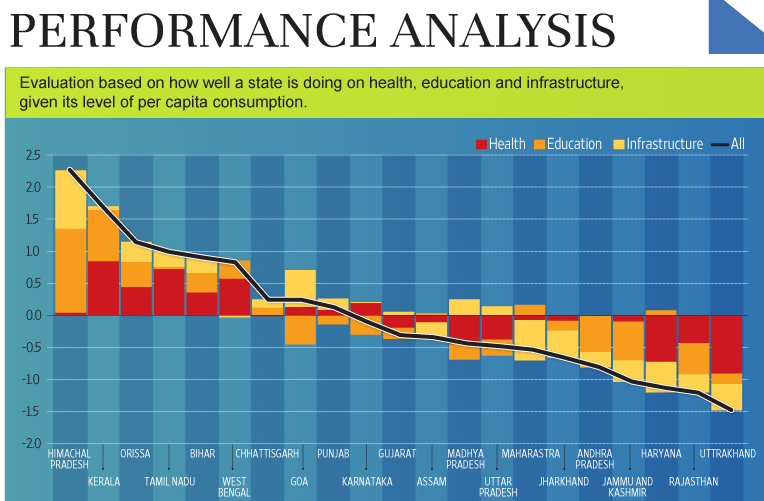
acadgild.com

LEARN.DO.LEARN

|  |
| --- |
| TABLE OF CONTENTS |
| 1.0 Project Description. |
| 2.0 Project Requirements |
| 3.0 Data Files |
| 4.0 Data Ingestion and Validation |
| 5.0 Data Analysis |
| 6.0 Project Implementation and Source Codes |
| 7.0 Conclusions |
| 8.0 Acknowledgement |
| 9.0 Bibliography |
|  |

**PROJECT DESCRIPTION**

The aim of this project is to analyze how various state governments have performed in different developmental schemes. This analysis will be helpful in finding out how successful the government has been in implementing various projects.



**PROJECT REQUIREMENTS**

HARDWARE REQUIREMENTS:

(i) Windows 8.1 with Linux (Intel CentOS) installed in Oracle Virtual Box.

(ii) Ram: 4GB and Linux Virtual Box Ram: 2GB.

(iii) Hard Disk Drive: 1TB.

(iv) Intel Core I3 Processor.

SOFTWARE REQUIREMENTS:

(i) The FLUME job which will format the data and place the data to HDFS

(ii) Pig script in Linux for executing and implementing the data analysis part.

(ii) Microsoft Word 2010 for writing the project report

(i) State\_Name

(ii) District\_Name

(iii) Project\_Objectives\_IHHL\_BPL

(iv) Project\_Objectives\_IHHL\_TOTAL

(v) Project\_Performance-IHHL\_BPL

(vi) Project\_Performance-IHHL\_TOTAL

DATA FILES

DATA INGESTION

(i) Data coming from web applications has xml format.

(ii) If any of the fields from the above fields given in Data Files are absent or null, consider that record as invalid.

**DATA ANALYSIS**

It is not only the data which is important, rather it is the insight it can be used to generate important. Once we have made the data ready for analysis, we have to perform below analysis on a daily basis.

**(i) Find out the districts who achieved 100 percent objective in BPL cards.**

**Export the result to mysql using sqoop.**

**(ii) Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.**

**Export the result to mysql using sqoop.**

**IMPLEMENTATION AND SOURCE CODES**

(i) Copy dataset from local file system to HDFS using flume

Solution:

The dataset link is:

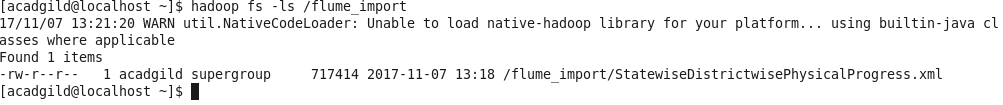
<https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing>

The conf file link is:

<https://drive.google.com/file/d/0B5dejdhAYHztdTdfa0NQNkxvT1U/view>

Firstly we have copied the dataset using flume agent.



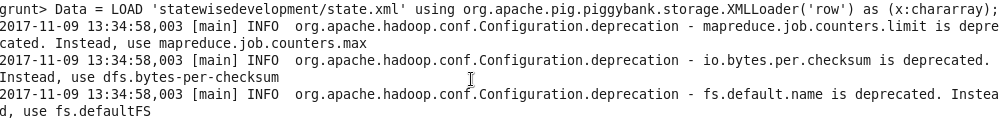


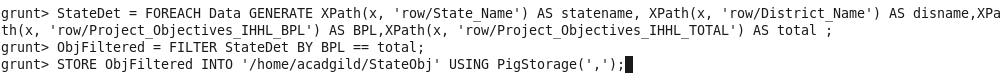
(ii) Find out the districts who achieved 100 percent objective in BPL cards.

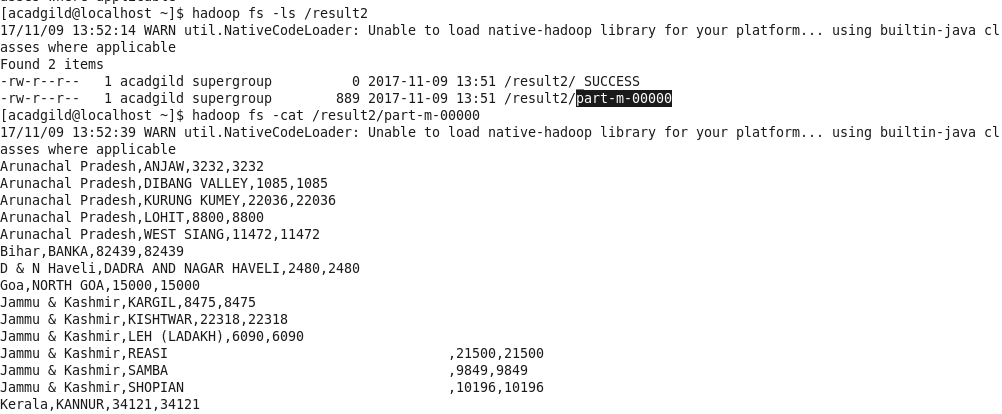
Solutions:

(a)



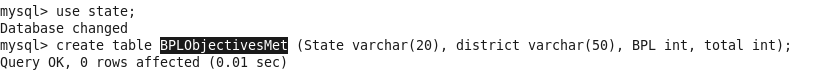




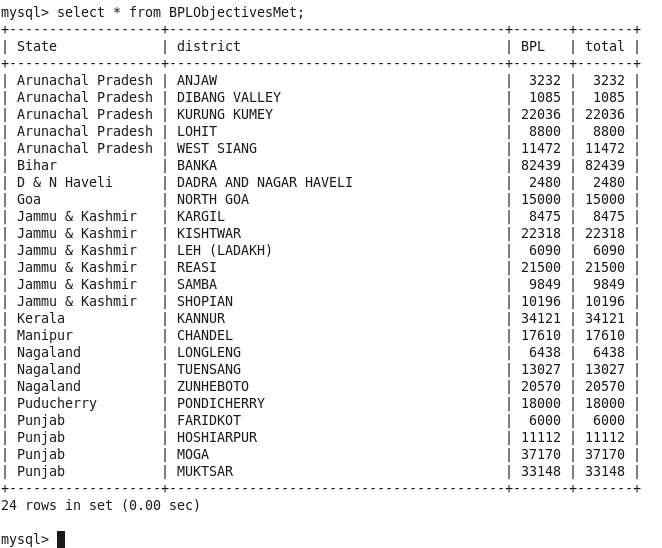


(b) Export the result into mysql using sqoop.

Solutions:



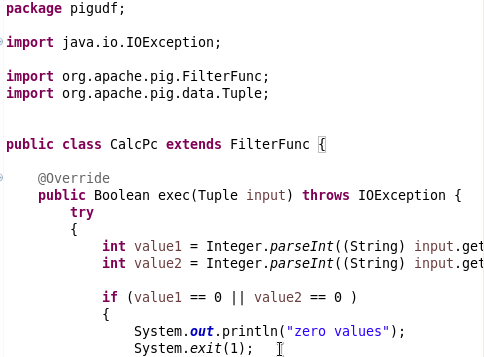




(iii) Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.

Solutions:

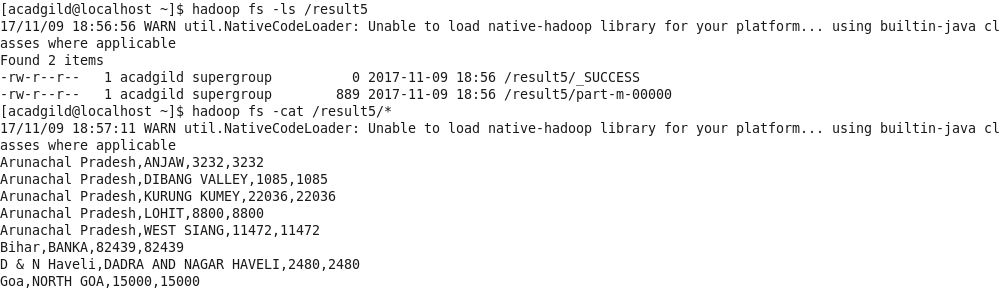
(a) First, the pig udf function is implemented, it needs to be compiled and included in a jar named ‘state.jar’.





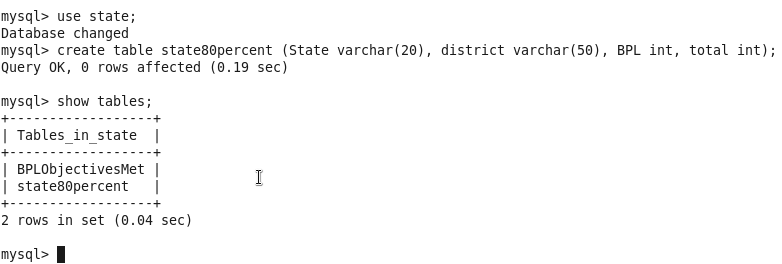




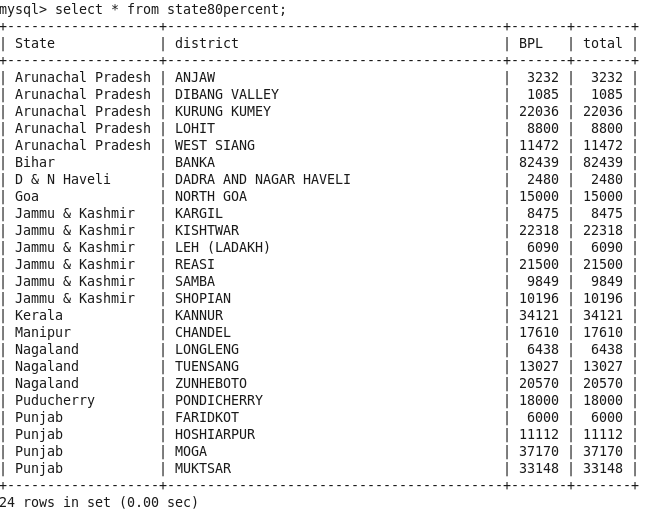


(b) Export the result into mysql using sqoop.

Solution:







**CONCLUSIONS**

A study on Hadoop component apache pig is made in order to write MapReduce code to analyze the Big Data in a minimum of time with simple coding format. An experimental analysis is made using a library data set with 3.40 lakh record made an analysis of frequent library users, authors preferred by the students, date and time accessed by the students frequently are extracted .The above results are extracted within five steps of coding using Pig script and the results are obtained on average of 35 sec of time duration. So from the above analyze it is concluded that using pig script it is possible to handle Big Database in an easy and efficient manner with minimum of time duration with simple coding.

**ACKNOWLEDGEMENT**

I would like to express my special gratitude to all the acadgild technical support team for helping me to complete the project within a span of time and also helping me to clear my doubts and clarification in the project.

Secondly I would like to express my special thanks to my acadgild mentor ‘Mr. Ravi Kshirsagar’ for clearing all the concepts during the entire session of ‘Big Data Hadoop & Spark’ training course.

**BIBLIOGRAPHY**

The references that I have used to complete my project are:

(i) Study material given by acadgild.

(ii) Need the help from internet.



**DONE**

**BY**

**SOHAM NEOGY**