



## Analyzing Varieties of Agricultural Data Using Big Data Tools Pig

**BANKIM L. RADADIYA<sup>1</sup> and PARAG SHUKLA<sup>2\*</sup>**

<sup>1</sup>Navsari Agricultural University - Navsari - Gujarat India.

<sup>2</sup>Department of MCA, Atmiya Institute of Technology & Science, Rajkot – 360005, India.

### Abstract

Day by day, data is growing rapidly. Analysis of the data is necessity. As per recent survey data generated in last 2 years is more than the data created in entire previous history of human. Data created in different form and in diversified manner. It can be structured, it can be semi-structured, or it can be unstructured. To analyze diversified by agricultural data we can use the tools of Big Data like Pig. Using Pig, we can analyze varieties of data. Pig is a platform for analysis of data. Biggest advantage of Pig is it can process any diversified data very quickly and allows us to use user defined functions. Use Case of Pig is ETL. It is used to extract the data from sources then after applying transformation we can load it into the data warehouse. We will do analysis of state wise proportion circulation of Numeral of operative properties for all societal collections in 2005-06 and 2010-11 using Pig.



### Article History

Received: 04 December 2017  
Accepted: 11 December 2017

### Keywords

Analysis, Pig, Varieties, Agricultural data, Big Data Tools, Structured, Semi-Structured, Unstructured.

### Introduction


Nowadays, data is growing very speedy. Analysis of the data is necessity for the many organization. As per recent survey data generated in last 2 years is more than the data created in entire previous history of human. Data created in different form and in diversified manner. It can be structured, it can be semi-structured, or it can be unstructured. To analyze diversified by agricultural data we can use the tools of Big Data like Pig. Using Pig, we can analyze varieties of data. Pig is a platform for

analysis of data. Biggest advantage of Pig is it can process any diversified data very quickly and allows us to use user defined functions. Use Case of Pig is ETL. It is used to extract the data from sources then after applying transformation we can load it into the data warehouse.

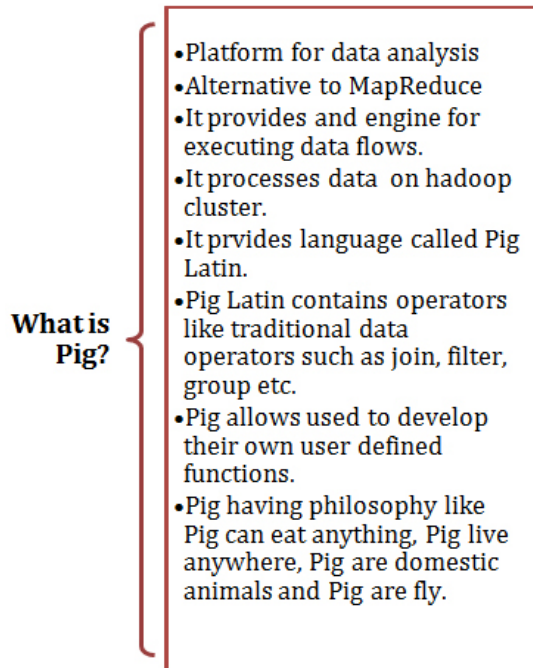
Here, in this study we analyzed varieties of agricultural data using the big data tools Pig.

**CONTACT** Parag C. Shukla  [paragshukla007@gmail.com](mailto:paragshukla007@gmail.com)  Department of MCA, Atmiya Institute of Technology & Science, Rajkot – 360005, India.

© 2017 The Author(s). Published by Techno Research Publishers

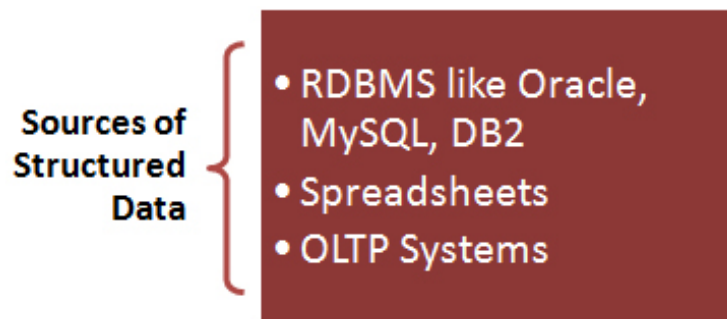
This is an  Open Access article licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits unrestricted NonCommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

To link to this article: <http://dx.doi.org/10.13005/ojcs/10.04.16>

**What is Pig?****Fig.1: What is Pig?****Why Pig? & What Pig Supports?****Fig.2: Why Pig? & What Pig Supports?****Analysis of Structured Agricultural Data Using Pig**

To analyze structured data, first we must identify the source of data. Source of structured data can be

any RDBMS like Oracle, SQL Server, DB2, MySQL, Spreadsheets or OLTP Systems. Following are the source of structured data.

**Fig.3: Sources of Structured Data****Step-1 Load the structured data**

We took the data of state wise proportion circulation of Numeral of operative properties for all societal collections in 2005-06 and 2010-11 from government website<sup>1</sup>.

Once retrieved the comma separated values file from government website, we copied the file on Linux platform. Once we copied on Linux then we moved the same file on HDFS platform. Following is the command to move the file from Linux root directory

to HDFS directory named PARAG. CopyFromLocal command is used to move the file from linux directory to HDFS directory.

```
hadoop fs -copyFromLocal /root/state_data.csv /
PARAG
```

root@sandbox:~

```
[root@sandbox ~]# hadoop fs -copyFromLocal /root/state_data
```

After moving the file from linux root directory to HDFS directory, we can load the data on Pig using Grunt shell

```
1 STATE_DATA = LOAD '/PARAG/state_data.csv' USING PigStorage(',') AS
2   (SR:INT, STATE:CHARARRAY,
3     CENSUS_MARGINAL_05:FLOAT,
4     CENSUS_SMALL_05:FLOAT,
5     CENSUS_SEMI_MEDIUM_05:FLOAT,
6     CENSUS_MEDIUM_05:FLOAT,
7     CENSUS_LARGE_05:FLOAT,
8     CENSUS_MARGINAL_10:FLOAT,
9     CENSUS_SMALL_10:FLOAT,
10    CENSUS_SEMI_MEDIUM_10:FLOAT,
11    CENSUS_MEDIUM_10:FLOAT,
12    CENSUS_LARGE_10:FLOAT )
```

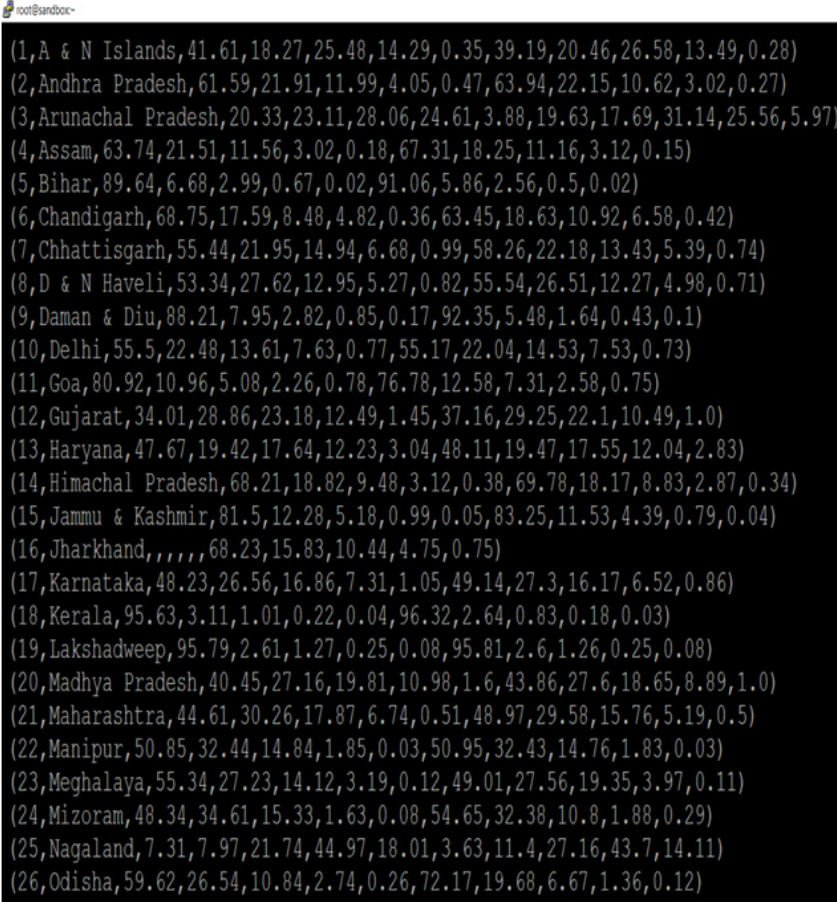
root@sandbox:~

```
grunt> STATE_DATA = LOAD '/PARAG/state_data.csv' USING PigSt
>> AS (SR:INT, STATE:CHARARRAY, CENSUS_MARGINAL_05:FLOAT,
>>   CENSUS_SMALL_05:FLOAT, CENSUS_SEMI_MEDIUM_05:FLOAT,
>>   CENSUS_MEDIUM_05:FLOAT, CENSUS_LARGE_05:FLOAT,
>>   CENSUS_MARGINAL_10:FLOAT, CENSUS_SMALL_10:FLOAT,
>>   CENSUS_SEMI_MEDIUM_10:FLOAT, CENSUS_MEDIUM_10:FLOAT,
>>   CENSUS_LARGE_10:FLOAT );
grunt>
```

**Step-2 Display the loaded data**

We can use dump statement to display the data in Grunt Shell.

```
grunt> DUMP STATE_DATA;
```



```
(1,A & N Islands,41.61,18.27,25.48,14.29,0.35,39.19,20.46,26.58,13.49,0.28)
(2,Andhra Pradesh,61.59,21.91,11.99,4.05,0.47,63.94,22.15,10.62,3.02,0.27)
(3,Arunachal Pradesh,20.33,23.11,28.06,24.61,3.88,19.63,17.69,31.14,25.56,5.97)
(4,Assam,63.74,21.51,11.56,3.02,0.18,67.31,18.25,11.16,3.12,0.15)
(5,Bihar,89.64,6.68,2.99,0.67,0.02,91.06,5.86,2.56,0.5,0.02)
(6,Chandigarh,68.75,17.59,8.48,4.82,0.36,63.45,18.63,10.92,6.58,0.42)
(7,Chhattisgarh,55.44,21.95,14.94,6.68,0.99,58.26,22.18,13.43,5.39,0.74)
(8,D & N Haveli,53.34,27.62,12.95,5.27,0.82,55.54,26.51,12.27,4.98,0.71)
(9,Daman & Diu,88.21,7.95,2.82,0.85,0.17,92.35,5.48,1.64,0.43,0.1)
(10,Delhi,55.5,22.48,13.61,7.63,0.77,55.17,22.04,14.53,7.53,0.73)
(11,Goa,80.92,10.96,5.08,2.26,0.78,76.78,12.58,7.31,2.58,0.75)
(12,Gujarat,34.01,28.86,23.18,12.49,1.45,37.16,29.25,22.1,10.49,1.0)
(13,Haryana,47.67,19.42,17.64,12.23,3.04,48.11,19.47,17.55,12.04,2.83)
(14,Himachal Pradesh,68.21,18.82,9.48,3.12,0.38,69.78,18.17,8.83,2.87,0.34)
(15,Jammu & Kashmir,81.5,12.28,5.18,0.99,0.05,83.25,11.53,4.39,0.79,0.04)
(16,Jharkhand,,,,,68.23,15.83,10.44,4.75,0.75)
(17,Karnataka,48.23,26.56,16.86,7.31,1.05,49.14,27.3,16.17,6.52,0.86)
(18,Kerala,95.63,3.11,1.01,0.22,0.04,96.32,2.64,0.83,0.18,0.03)
(19,Lakshadweep,95.79,2.61,1.27,0.25,0.08,95.81,2.6,1.26,0.25,0.08)
(20,Madhya Pradesh,40.45,27.16,19.81,10.98,1.6,43.86,27.6,18.65,8.89,1.0)
(21,Maharashtra,44.61,30.26,17.87,6.74,0.51,48.97,29.58,15.76,5.19,0.5)
(22,Manipur,50.85,32.44,14.84,1.85,0.03,50.95,32.43,14.76,1.83,0.03)
(23,Meghalaya,55.34,27.23,14.12,3.19,0.12,49.01,27.56,19.35,3.97,0.11)
(24,Mizoram,48.34,34.61,15.33,1.63,0.08,54.65,32.38,10.8,1.88,0.29)
(25,Nagaland,7.31,7.97,21.74,44.97,18.01,3.63,11.4,27.16,43.7,14.11)
(26,Odisha,59.62,26.54,10.84,2.74,0.26,72.17,19.68,6.67,1.36,0.12)
```

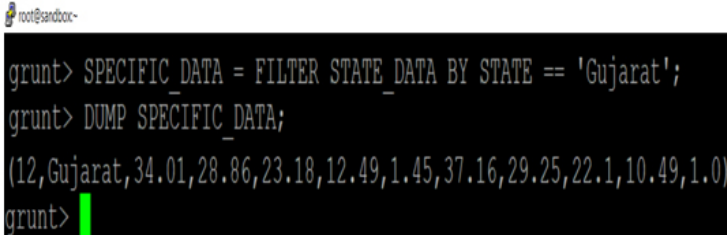
Fig.4: State wise proportion circulation of Numeral of operative properties

**Step-3 Filter Specific Data**

For analysis of any data we can use filter or

aggregate functions. Here, we are filtering the specific data from state Gujarat.

```
SPECIFIC_DATA = FILTER STATE_DATA BY STATE == '
```



```
grunt> SPECIFIC_DATA = FILTER STATE_DATA BY STATE == 'Gujarat';
grunt> DUMP SPECIFIC_DATA;
(12,Gujarat,34.01,28.86,23.18,12.49,1.45,37.16,29.25,22.1,10.49,1.0)
grunt>
```



Finding all state data which census\_marginal of 2005  
is more than 50

```
AGRI_CENSUS_MARGINAL_MORE_50 =
  FILTER STATE_DATA BY CENSUS_MARGINAL_05
DUMP AGRI_CENSUS_MARGINAL_MORE_50;
```

```
grunt> DUMP AGRI_CENSUS_MARGINAL_MORE_50;
Util - Total input paths to process : 1
(2,Andhra Pradesh,61.59,21.91,11.99,4.05,0.47,63.94,22.15,10.62,3.02,0.27)
(4,Assam,63.74,21.51,11.56,3.02,0.18,67.31,18.25,11.16,3.12,0.15)
(5,Bihar,89.64,6.68,2.99,0.67,0.02,91.06,5.86,2.56,0.5,0.02)
(6,Chandigarh,68.75,17.59,8.48,4.82,0.36,63.45,18.63,10.92,6.58,0.42)
(7,Chhattisgarh,55.44,21.95,14.94,6.68,0.99,58.26,22.18,13.43,5.39,0.74)
(8,D & N Haveli,53.34,27.62,12.95,5.27,0.82,55.54,26.51,12.27,4.98,0.71)
(9,Daman & Diu,88.21,7.95,2.82,0.85,0.17,92.35,5.48,1.64,0.43,0.1)
(10,Delhi,55.5,22.48,13.61,7.63,0.77,55.17,22.04,14.53,7.53,0.73)
(11,Goa,80.92,10.96,5.08,2.26,0.78,76.78,12.58,7.31,2.58,0.75)
(14,Himachal Pradesh,68.21,18.82,9.48,3.12,0.38,69.78,18.17,8.83,2.87,0.34)
(15,Jammu & Kashmir,81.5,12.28,5.18,0.99,0.05,83.25,11.53,4.39,0.79,0.04)
(18,Kerala,95.63,3.11,1.01,0.22,0.04,96.32,2.64,0.83,0.18,0.03)
(19,Lakshadweep,95.79,2.61,1.27,0.25,0.08,95.81,2.6,1.26,0.25,0.08)
(22,Manipur,50.85,32.44,14.84,1.85,0.03,50.95,32.43,14.76,1.83,0.03)
(23,Meghalaya,55.34,27.23,14.12,3.19,0.12,49.01,27.56,19.35,3.97,0.11)
(26,Odisha,59.62,26.54,10.84,2.74,0.26,72.17,19.68,6.67,1.36,0.12)
(27,Puducherry,78.95,12.15,6.12,2.43,0.35,85.71,8.36,4.36,1.35,0.21)
(30,Sikkim,54.25,22.53,14.7,7.36,1.16,54.02,22.61,14.43,7.9,1.04)
(31,Tamil Nadu,76.01,15.06,6.62,2.07,0.24,77.19,14.55,6.19,1.86,0.21)
(32,Tripura,86.77,9.63,3.23,0.34,0.03,86.27,9.52,3.72,0.48,0.01)
```

Fig.5: State wise data of 2005 which census marginal is more than 50

Finding all state data which census\_small of 2005  
is more than 30

```
AGRI_CENSUS_SMALL_MORE_30 =
  FILTER STATE_DATA BY CENSUS_SMALL_05 >=30.0;
DUMP AGRI_CENSUS_SMALL_MORE_30;
```

```
grunt> DUMP AGRI_CENSUS_SMALL_MORE_30;
Util - Total input paths to process : 1
(21,Maharashtra,44.61,30.26,17.87,6.74,0.51,48.97,29.58,15.76,5.19,0.5)
(22,Manipur,50.85,32.44,14.84,1.85,0.03,50.95,32.43,14.76,1.83,0.03)
(24,Mizoram,48.34,34.61,15.33,1.63,0.08,54.65,32.38,10.8,1.88,0.29)
grunt> █
```

Finding all state data which census\_marginal of 2010  
is more than 80

```
AGRI_CENSUS_MARGINAL_MORE_80 =
  FILTER STATE_DATA BY CENSUS_MARGINAL_10 >= 80.0;
  DUMP AGRI_CENSUS_MARGINAL_MORE_80;
```

```
(5,Bihar,89.64,6.68,2.99,0.67,0.02,91.06,5.86,2.56,0.5,0.02)
(9,Daman & Diu,88.21,7.95,2.82,0.85,0.17,92.35,5.48,1.64,0.43,0.1)
(15,Jammu & Kashmir,81.5,12.28,5.18,0.99,0.05,83.25,11.53,4.39,0.79,0.04)
(18,Kerala,95.63,3.11,1.01,0.22,0.04,96.32,2.64,0.83,0.18,0.03)
(19,Lakshadweep,95.79,2.61,1.27,0.25,0.08,95.81,2.6,1.26,0.25,0.08)
(27,Puducherry,78.95,12.15,6.12,2.43,0.35,85.71,8.36,4.36,1.35,0.21)
(32,Tripura,86.77,9.63,3.23,0.34,0.03,86.27,9.52,3.72,0.48,0.01)
(35,West Bengal,81.17,14.38,4.04,0.4,0.01,82.16,13.76,3.75,0.32,0.01)
grunt>
```

Fig.6: State wise data of 2010 which census marginal is more than 80

Finding all state data which census\_small of 2010  
is more than 30

```
AGRI_CENSUS_SMALL_MORE_30 =
  FILTER STATE_DATA BY CENSUS_SMALL_10
  DUMP AGRI_CENSUS_SMALL_MORE_30;
```

```
Util - Total input paths to process : 1
(22,Manipur,50.85,32.44,14.84,1.85,0.03,50.95,32.43,14.76,1.83,0.03)
(24,Mizoram,48.34,34.61,15.33,1.63,0.08,54.65,32.38,10.8,1.88,0.29)
grunt>
```

### Analysis of Unstructured Agricultural Data Using Pig

#### Conclusion

We did analysis of agricultural data of state wise proportion circulation of Numeral of operative properties for all societal collections in 2005-06

and 2010-11 using Pig. We analyzed structured agricultural data using Pig. As we know that day by day requirement of analysis of the data is increasing rapidly. To demonstrate the use of analysis using big data tools Pig we used the government agricultural data and did the analysis of data.

Analysis of the data is necessity for the many organization. Data created in different form and in diversified manner. It can be structured, it can be semi-structured, or it can be unstructured. To analyze diversified by agricultural data we can use the tools of Big Data like Pig. Using Pig, we can analyze varieties of data. Pig is a platform for analysis of data. Biggest advantage of Pig is it can process any diversified data very quickly and allows us to use

user defined functions. Use Case of Pig is ETL. It is used to extract the data from sources then after applying transformation we can load it into the data warehouse.

#### Acknowledgment

We wish to thank Open Government Data Platform (OGD) for providing data for analysis & sincere thanks to our mentor.

#### References

- 1 <https://data.gov.in/resources/state-wise-percentage-distribution-number-operational-holdings-all-social-groups-during>
- 2 Apache Pig, <https://pig.apache.org/>
- 3 Apache Pig Architecture and components of Pig [online resource] [https://www.tutorialspoint.com/apache\\_pig/apache\\_pig\\_architecture.htm](https://www.tutorialspoint.com/apache_pig/apache_pig_architecture.htm)
- 4 Pig Philosophy, <https://pig.apache.org/philosophy.html>
- 5 Hive Vs Pig [online resource] <http://www.bigdataanalyst.in/hive-vs-pig/>
- 6 Big Data and Analytics – Wiley Publication, Seema Acharya, Subhashini Chellapan
- 7 Dr. Birendra Goswami, Pradip Kumar Chandra "The Evolution Of Big Data As A Research And Development" *International Journal of Scientific Research and Engineering Studies (IJSRES)* Volume 2 Issue 3, March 2015 ISSN: 2349-8862
- 8 Online Resource <https://data.gov.in/>