Empirical Aspects to Analyze Population of India using Apache Pig in Evolutionary of Big Data Environment

Yogesh Kumar Gupta, Tanusha Mittal

Abstract: Big data contains great variety of data arrives in incrementing volume and with high velocity. The data sets are so voluminous that the conventional data processing software just can't able to manage them. Hence, big data tools i.e. Hadoop came into the glare due to its high scalability, availability and the cluster environment mechanism which provides the facility to work in the distributed manner. One of the important components of Hadoop is MapReduce which is able to handle the unstructured data but to use this, high programming skills are needed. Therefore, due to the reason of high programming skill, users are now a days moving towards the tool i.e. Apache Pig, as we can analyze the data simply by executing the queries. In this paper, we analyze the gender ratio of India according to the age group of 0 to 24 from the year of 2001-2018 that is further analyzed through Pig Latin scripts and results are represented in the pictorial form. The government of India introduces a policy i.e. Two-Child Policy. The policies are implemented by disallowing the people with more than two children from serving the government. Firstly, the policy was implemented by Assam in 2017. The motive of this paper is to analyze whether the introduced policy of government is fulfilling or not.

Index Terms : Big data, Hadoop, Apache Pig, and Gender Ratio

INDRODUCTION

The term “Big Data” is utilized to label the data that is generated from different origins such as social media sites, medical data, stock market data, etc. on daily basis. The data in the form of structured files can be handle with the traditional contrivances. Semi structured data can be in the form of CSV or XML files while unstructured data contains audio, video, pictures. The Big Data cannot be handle using the traditional database management system. Therefore, we have to use some highly parallel software to acquire, organize and analyze these types of data. To handle this large volume of data with high efficiency, here is mainly accepted tools is Apache Hadoop. Apache Hadoop- Hadoop is the release structure written in java. The Hadoop platform is highly scalable for processing the large volume of structured and unstructured data. In Hadoop framework, data is stored in the form of block of size i.e. 128MB or 256MB. Hadoop architecture elaborates the nodes i.e. Name node or master node which is used to maintain, control and manage the data nodes. Data nodes (slave nodes) send the heartbeat to the name node on the regular basis for the surety of its aliveness. Name node stores all the records of Data nodes in HDFS. If any node is crash or fails, then name node provides the facility to replicate the data. MapReduce is a programming paradigm that gives huge scalability crosswise over hundreds or thousands of servers in a Hadoop clusters. As the processing component, MapReduce is the center of Apache Hadoop. The term “MapReduce” refers to two unique and different tasks that Hadoop programs perform. The first is the map job, which accepts a set of data as input and converts it into another set of data, where individual data are separated into tuples (key/value pairs). The reduce job takes the yield from a map job as input of it and joins those data tuples into a smaller set of tuples.Apache Pig- Pig is one of the important and core components of Hadoop which works on the Latin Scripts. It is the data flow language which fills the gap between the high-level declarative language (SQL) and low-level procedural language (MapReduce). Now the question arises why Apache Pig is better than MapReduce. To deal with the big data through MapReduce, developer has to write several lines of code which require more efforts and time. Also, a user who has less knowledge of java programming language cannot handle the Big Data using MapReduce. That's the main reason behind using Apache Pig instead of MapReduce Programming. There are some features of Apache Pig which are described below –

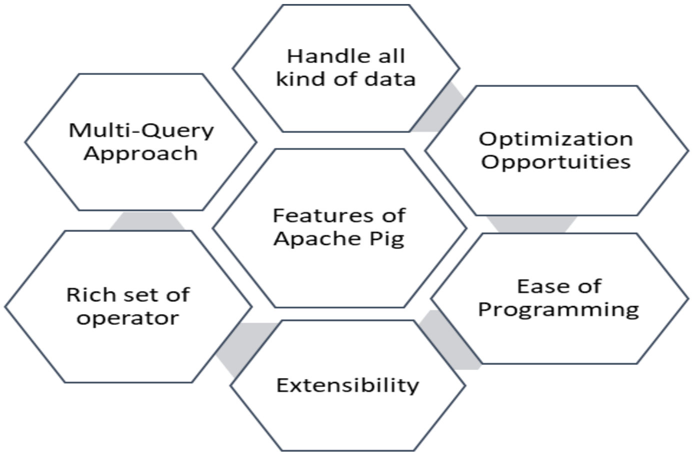


Figure 1: *Features of Apache Pig*

LITERATURE REVIEW

Zhang, Cherkasova et al. defines the performance valuation model with the pre-defined resources or a certain deadline with the Apache Pig approach. This model also uses PigMix benchmark for evaluation. To overcome the problem of performing the work in particular deadline performance model is generated. This benchmark and cluster are used to evaluate the performance of model. This paper is used to help the Apache Pig to utilize the resources in efficient manner with particular time limit with the help of performance model.[1] Sanjeev Dhawan et al. defines that how data is stored on

- *Yogesh Kumar Gupta*, Computer Science Department, Banasthali Vidyapith, Niwai, India Email:gyogesh@banasthali.in*
- *Tanusha Mittal, Computer Science Department, Banasthali Vidyapith, Niwai, India. Email: tanushamittal5544@gmail.com*

Hadoop distributed file system using the tool pig and hive. In this paper, it compares and concludes both the techniques. It is concluded that the dataflow is specified in PigLatin. Hive is a technology which is used for turning the Hadoop into a data warehouse. In this paper, initially a MapReduce job is successfully created using Apache Pig & then it is used to analyze a big database to get the required and efficient results. Finally, a MapReduce job is created using Hive & then it is used to analyze a big database to get results. The final result shows that the analysis done by both of the MapReduce machine is successful. The performance or results of both i.e. pig and hive was nearly same.[2] Munesh Kataria et al. explains that big data can't be handle by using conventional database management system. So, we have to use some highly parallel software to acquire, organize and analyze these types of data. Firstly, we acquire the data from different sources. Then, we can organize the acquired data using HDFS that means a distributed file system. At last, we can analyze them for faster access and query response. The components like Pig, Hive and Jaql takes very less time for performing the analysis. Hive can analyze a database of more than 5 lakh records in just 34 seconds. So, all these components make it possible to handle the database in an easy and very efficient way. [3] Swarna C. et. al, introduce the concept of Pig and its associated language i.e. Pig Latin which provides a new data processing environment deployed at Yahoo. Pig Latin is the parallel dataflow language which is designed in such a way to fit between SQL and MapReduce. Pig Latin script explicates a directed acyclic graph, where data flows are represented as edges and operators are represented as nodes. Pig Latin expressions are compiled using Pig into a sequence of MapReduce jobs and therefore, starts the execution of these jobs on Hadoop environment. Pig structure is susceptible to substantial parallelization. [4] Arushi Jain, Vishal Bhatnagar explains that the crimes and crime rate is increasing with the increased population. This crime related data is the huge issue for governments to make strategic decisions so as to maintain law and order. If from the particular state, the number of complaints is very high, then the extra security must be provided to the residents that means increasing the police presence, quick redressal of complaints and strict vigilance. Now a days, crimes against women are becoming a worrying and distributing problem for the government of India. This concludes that Big Data Analytics helps to analyze certain trends so that the laws and orders can be maintain properly. [5] Preethi Elavarasi et. al. defines big data analytics using some of its core components, Apache Pig using Pig Latin and Apache Hive using Hive QL like languages and also explores the architecture of both. Apache Hive defines as data warehouse systems and also used for ad-hoc query analysis. Apache Pig defines as a data flow system. This concludes the performance of Pig and Hive. [6] After goes during the comprehensive revision of research papers on Pig, we conclude that nearly each author illustrates Pig Latin to remove the complications of MapReduce programming and defines the proficiency of Pig Latin over MapReduce. Many authors work on various kind of data using Apache Pig, Like Arushi Jain, Vishal Bhatnagar analyze huge data related to crime; no one has worked on Gender Ratio. Therefore, in this paper, we analyze the gender ratio data according to different age group by using Apache Pig. Gender Ratio-Gender Ratio can be defined as the ratio between the males and the female population. Across the world, there are differences in the ratio of males and females with their different life styles and stages. The inequality between the population of males and females in some cases can be traced back to birth: in many countries the number of boys and girls born every year is significantly bias. The sex ratio or in other words the share of population of female- varies across the world. The share of women in the world was 49.6% in the year of 2017. Paper Organization: - We categorize our paper into different sections. The first section depicts the perception of big data and emphasizes on the analysis of population data of India corresponding to the categories of males and females with the help of one of the famous big data tools “Apache Pig”. The section 2 defines the literature review related to Apache Pig. Section 3 illuminates the requirements for the proposed algorithm and process model to analyze the population data. The Section 4 makes us understand how to analyze the data in a very efficient manner. Last section, concludes the impact of big data to analyze the population data of India using tools and techniques.

RESEARCH METHODOLOGY

In this research, we acquire data from the secondary source i.e. internet. In this study, we focused on the population of India according to different age groups. To analyze the age population data, we categorize the data into the 3 different groups corresponding to years i.e. 2001-2006, 2007-2012 and 2013-2018. In each group, data categories corresponding to age with gap of 5 years. In this research we took the data from the age of 0 to 24 in both categories of males and females.

3.1 Experimental Setup

For the Hadoop environments we used software's such as VMware has been elected as graphical user interface with defined version is 5.2.16r123759 (Qt5.6.2) to run the Hadoop (2.6.0-cdh5.13.0) on Cloudera environment with CentOS 6.7 non-windows operating system. To execute the queries of pig we have used Pig genre 0.12.0-cdh5.13.0. and hardware Intel(R) Core (TM) I3-3210 CPU@3.20GHz Processor, 4GB RAM and 32-bit Operating System.

3.2 Proposed Model to analyze the population of India using Apache Pig

In this paper, we have contrivance the model to process the population data of India of males and females in three different groups corresponding to years i.e. 2001-2006, 2007-2012 and 2013-2018. In each group, there is the age of females and males in groups of 5 from 0 to 24. The process model to analyze the population data using pig step by step is shown in to the following figure.

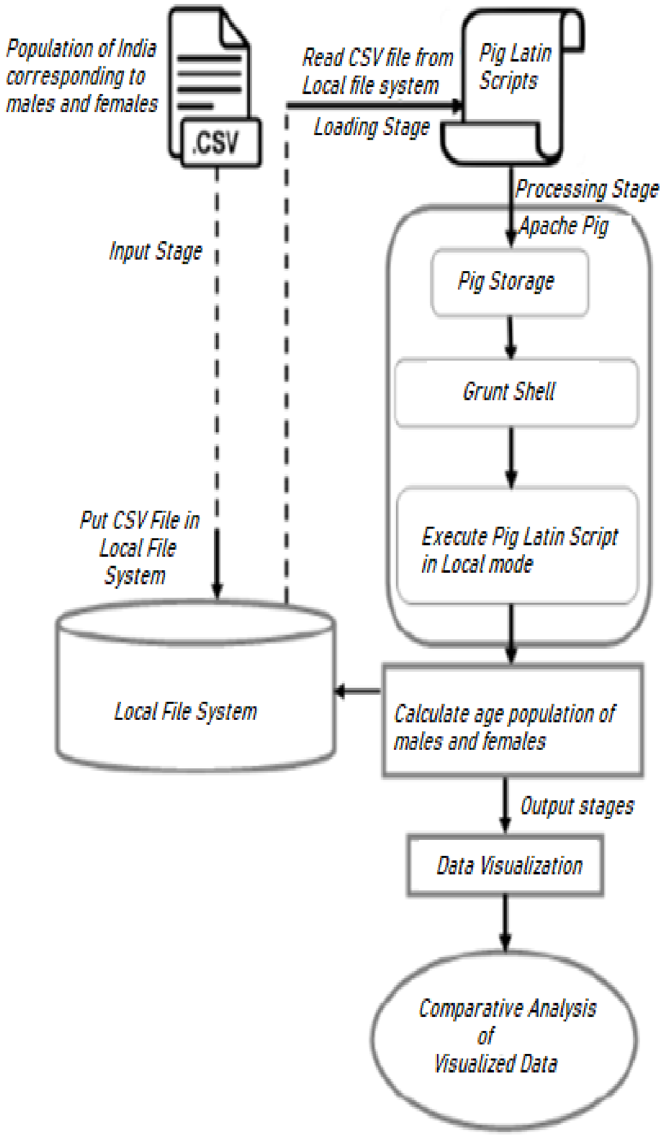
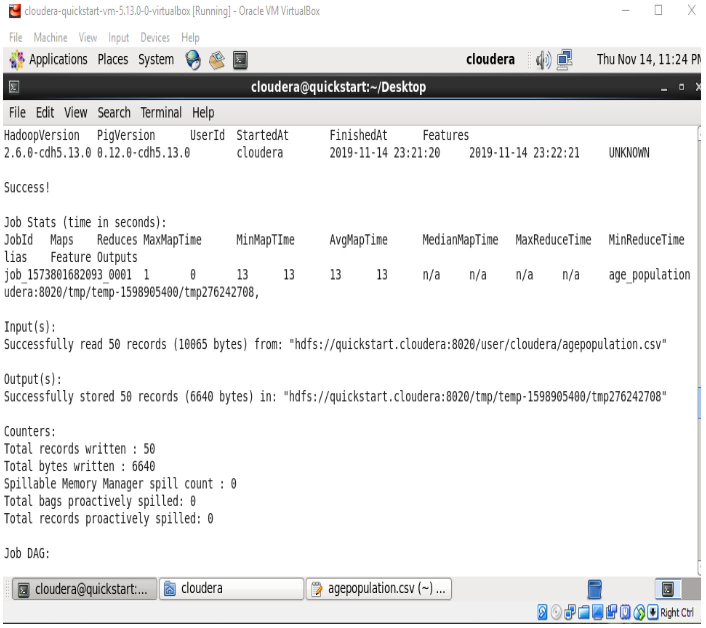

Population of India corresponding to males and females
.CSV
Read CSV file from Local file system
Pig Latin Scripts
Loading Stage
Input Stage
Processing Stage
Apache Pig
Pig Storage
Grunt Shell
Execute Pig Latin Script in Local mode
Put CSV File in Local File System
Local File System
Calculate age population of males and females
Output stages
Data Visualization
Comparative Analysis of Visualized Data

Figure 2: *Model to Analyse Population of India*

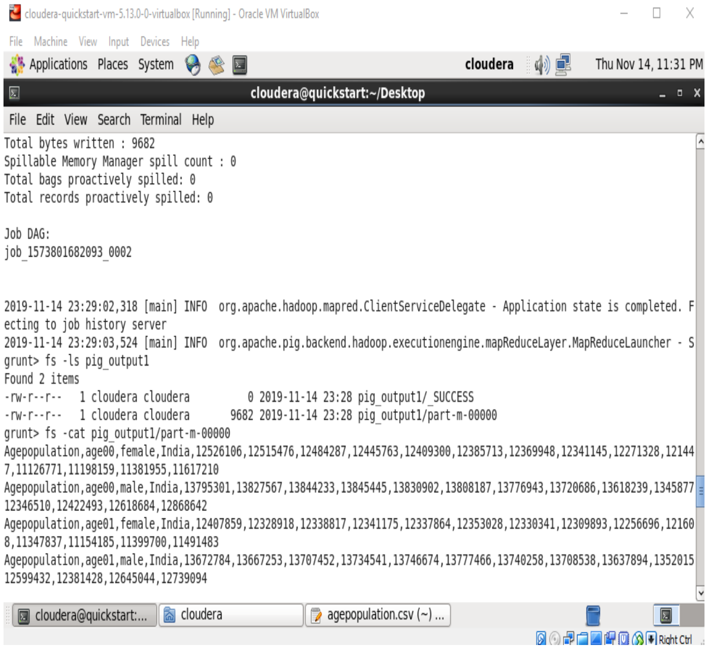
First of all, we acquired data of population corresponding to males and females of India. We have to load the data from HDFS to pig. While loading the population data, total number of records are 50, total bytes are 6640, the maximum, minimum and average map time is 13 as shown in the below figure. The total time taken while loading the data is 00:01:01 or 1 minute and 1 second. After loading the data into Pig, we analyse the data as discussed in the above figure.


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera
Thu Nov 14, 11:24 PM
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
cloudera@quickstart:...
cloudera
agepopulation.csv (~) ...
Right Ctrl

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0        cloudera        2019-11-14 23:21:20     2019-11-14 23:22:21     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime
lias    Feature Outputs
job_1573801682093_0001  1       0       13      13      13      13      n/a     n/a     n/a     n/a     age_population
udera:8020/tmp/temp-1598905400/tmp276242708,

Input(s):
Successfully read 50 records (10065 bytes) from: "hdfs://quickstart.cloudera:8020/user/cloudera/agepopulation.csv"

Output(s):
Successfully stored 50 records (6640 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1598905400/tmp276242708"

Counters:
Total records written : 50
Total bytes written : 6640
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
```

Figure 3: *Loading Population data into Pig*

After loading and analyzing the data in pig, the output will again have to store in HDFS. The time taken to store the data from Pig to HDFS is less than the loading time i.e. 00:00:50 or 50 seconds. The other related information regarding storing is shown below in the figure.


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera
Thu Nov 14, 11:31 PM
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
cloudera@quickstart:...
cloudera
agepopulation.csv (~) ...
Right Ctrl

```
Total bytes written : 9682
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1573801682093_0002



2019-11-14 23:29:02,318 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. F
ecting to job history server
2019-11-14 23:29:03,524 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - S
grunt> fs -ls pig_output1
Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2019-11-14 23:28 pig_output1/_SUCCESS
-rw-r--r--   1 cloudera cloudera       9682 2019-11-14 23:28 pig_output1/part-m-00000
grunt> fs -cat pig_output1/part-m-00000
Agepopulation,age00,female,India,12526106,12515476,12484287,12445763,12409300,12385713,12369948,12341145,12271328,12144
7,11126771,11198159,11381955,11617210
Agepopulation,age00,male,India,13795301,13827567,13844233,13845445,13830902,13808187,13776943,13720686,13618239,1345877
12346510,12422493,12618684,12868642
Agepopulation,age01,female,India,12407859,12328918,12338817,12341175,12337864,12353028,12330341,12309893,12256696,12160
8,11347837,11154185,11399700,11491483
Agepopulation,age01,male,India,13672784,13667253,13707452,13734541,13746674,13777466,13740258,13708538,13637894,1352015
12599432,12381428,12645044,12739094
```

Figure 4: *Storing resultant data from Pig*

RESULTS AND DISCUSSION

At this time, we put the population dataset according to the different age group i.e. 0 to 24 with different years i.e. from 2001 to 2018. After processing this stored data, the related output is generated as the population of males and females in different years with different age groups. Following results are depicted by analyzing the data in Apache Pig that shown in the following figure.

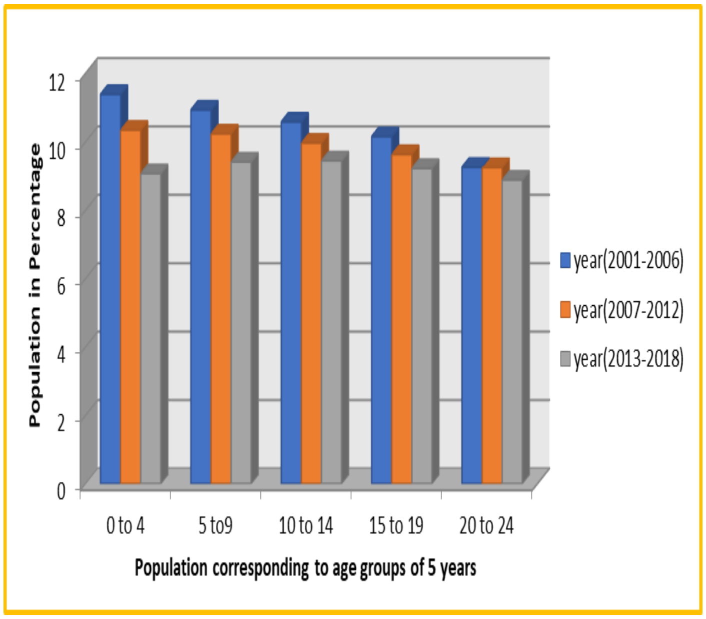


*Figure 5:*Females Population in Age group 0 to 24

Above figure describes the female population in three different groups corresponding to the years i.e. 2001-2006, 2007-2012 and 2013-2018. The age of females is categorized into groups of 5 from 0 to 24. In 2001-2006, the female population was higher and in 2013-2018 the population was lower in each age group. That means the female population of India in age group 0 to 24 rapidly decreases yearly from 2001 to till 2018.

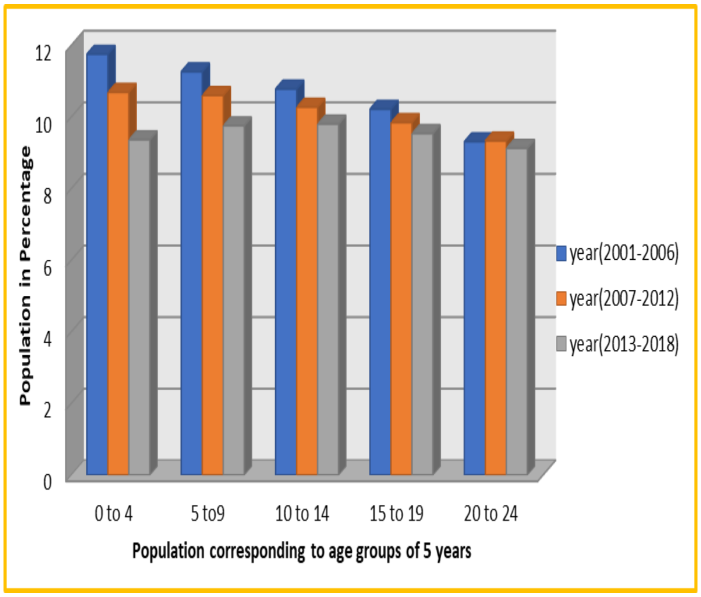


Figure 6: Males Population in Age group 0 to 24

Above figure describes the male population in three different groups corresponding to the years i.e. 2001-2006, 2007-2012 and 2013-2018. The age of males is categorized into groups of 5 from 0 to 24. In 2001-2006, the male population was higher and in 2013-2018 the population was lower in each age group. That means the male population of India in age group 0 to 24 rapidly decreases yearly from 2001 to till 2018.

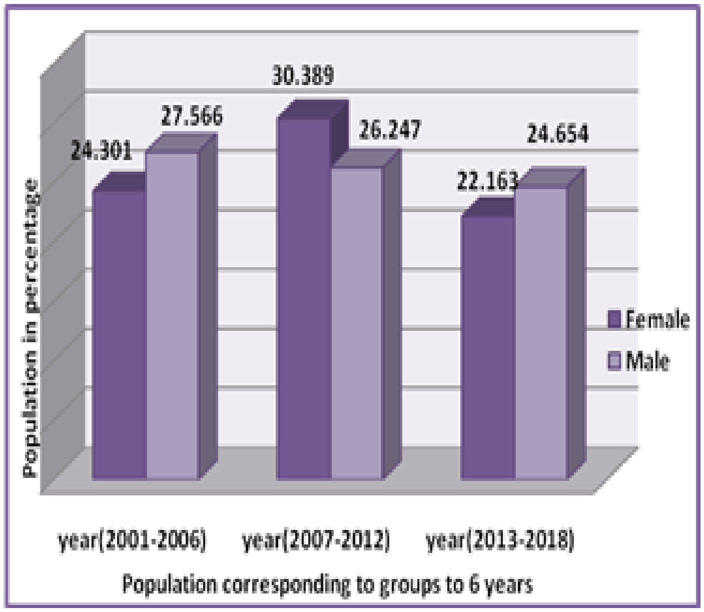


Figure 7: Gender ratio in the age group of 0 to 24

Above figure depicted, in the year of 2001-2006, the female population was lower than the male population and in 2007-2012, the female population were increased than male population but again in next six years the female population were decreased than the male population in the age group of 0 to 24 only. The gender ratio in the year of 2001-2006 is around 3.27%, in the year of 2007-2012 is around 4.14%and in the next six year i.e. 2013-2018 the gender ratio was decreased that is around 2.49%.

CONCLUSION

The huge formation of big data creates difficulties for users to extract the meaningful information i.e. the highly efficient tools needs to analyze the data in the limited number of times. The purpose to choose the Hadoop platform is to speed up the task since it provides the facility of distributed cluster environment to perform the work and reduce the load of single machine by doing the work on multiple nodes simultaneously. There are many tools of Hadoop are available to analyze the big data and we used Apache Pig. The key point of this paper is to analyze the gender ratio of India with the age group of 0 to 24 yearly from 2001 to 2018 using Pig. The results are represented in the graphical format. The government of India introduces a policy i.e. Two-Child Policy. The policies are implemented by disallowing the people with more than two children from serving the government. Firstly, the policy was implemented by Assam in 2017. The motive of this paper is to analyze whether the introduced policy of government is fulfilling or not. Therefore, by means of the above analysis depict that the population of India is decreasing rapidly year by year from 2017.

REFERENCES

[1] Z. Zhuoyao, C. Ludmila, V. Abhishek and L.T. Boon, “Meeting service level objective of Pig programs”. Proceedings of the 2nd International Workshop on Cloud Computing Platforms. ACM, 2012.

[2] Dhawan, S. and Rathee, S. “Big Data Analytics using Hadoop Components like pig and hive”. American International Journal of Research in Science, Technology, Engineering &

Mathematics, pp.88-93, March-May, 2013.

[3] Kataria, M. And Mittal, P. “Big Data and Hadoop with components like Flume, Pig, Hive and Jaql”, International Journal of Computer Science and Mobile Computing, Vol. 3 Issue 7, pp. 759-765, July 2014.

[4] Swarna, C. and Ansari, Z. “Apache Pig- A Data Flow Framework, Based on Hadoop MapReduce”, International Journal of Engineering Trends and Technology (IJETT)- Vol. 50, Number 5, pp. 271-275, August 2017.

[5] Jain, A. and Bhatnagar, V. “Crime Data Analysis Using Pig with Hadoop”, International Conference on Information Security & Privacy (ICISP2015), 11-12 December, 2015, Procedia Computer Science 78, Pp. 571-578 (2016).

[6] R. A. Preethi and J. Elavarasi. “Big data analytics using Hadoop tools – Apache Hive vs Apache Pig.” Int. J. Emerg. Technol. Computer. Sci. Electron, Vol. 24, 2017.

[7] F.G. Alan, D. Jianyong and N. Thejas “Apache Pig’s Optimizer.” IEEE Data Eng. Bull. pp. 34-45, 2013.

[8] Gupta, Y.K. and Sharma, S. “Impact of Big Data to Analyze Stock Exchange Data Using Apache Pig”. International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8, Issue-7. pp. 1428-1433, May 2019

[9] Gema Bello-Orgaz, Jason J. Jung, David Camacho, “Social big data: Recent achievements and new challenges”. Information Fusion, Volume 28. Pp. 45-59, March 2016.

[10] E. Nada and E. Ahmed “Big data analytics: a literature review paper.” Industrial Conference on Data Mining. Springer, Cham, pp. 214-227, 2014.

[11] Ouakine, Keren, Michael Carey, and Scott Kirkpatrick. “The PigMix Benchmark on Pig, MapReduce and HPCC systems.” Big Data (Big Data Congress), International Congress on. IEEE, 2015.

[12] Shvachko, Konstantin, “The Hadoop distributed file system.” Mass storage systems and technologies (MSTT), IEEE 26th symposium on. IEEE, 2010.

[13] Agarwal, S. and Khanam, Z. “Map Reduce: A Survey Paper on Recent Expansion.” International Journal of Advanced Computer Science and Applications 6.8, pp.209-215 (2015).

[14] Olshannikova, Ekaterina. “Conceptualizing Big Social Data.” Journal of Big Data 4.1 (2017).

[15] Gupta, Y.K. and Barhaiya, G. “Analysis of Crime Rates of Different States in India Using Apache Pig in HDFS Environment”, Recent Patents on Engineering 13: 1. https://doi.org/10.2174/1872212113666190227162314, ISSN 2212-4047 (2019).