iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Analysis of Library Data Using Pig Latin

Sonali G Dalasania [1], Prof. Krunal N Vaghela[2]
[1]M. Tech, Department of Computer Engineering, School of Engineering RK University, Rajkot-India.
[2]Head of Department (CE, IT, BCA, MCA), School of Engineering RK University, Rajkot-India.

Abstract: ***As per the advancement in organizations across the world in public as well as private sectors have made unprecedented growth in data. The massive growth in data is increased day by day. This massive growth of data or high volume of data with high velocity and different variety is known as Big Data. Analyzing, Processing and Storing of this data we require some efficient technique because the traditional techniques are not able to deal with this data. Apache PIG is the platform on which we can process this large amount of data with minimum time and minimum lines of coding. Apache PIG uses HDFS(Hadoop Distributed File System) to store data and Map Reduce to processing the data. The purpose of this research is to find 1)The increment or decrement in the requirement of the book. 2)How many books are available for the specific subject by various authors. 3)Different Edition of the specific book used by student in one week. 4) Use of Specific book by students of various departments. In this research dataset format will be .csv format .***
Keywords: HDFS (Hadoop Distributed File System), Apache Hadoop, Apache Sqoop, Map Reduce, Pig Latin.

I. INTRODUCTION

Collection, Storage and Processing of Big Data with Traditional Technique is not so easy as we think.. As people are Accepting the digitalization ,The growth of the Digital Data is increasing rapidly. This Big Data may have Different formats like Structure ,Unstructured and Semi structure Data. Pig Latin can be used for storing, processing and analyzing of Big Data. Pig Latin is query language for different formats of Big Data. Specific Requirement for enhancing the library stuff can be found using the Pig Latin. Increment and Decrement of the book requirement can be found using following questions.1)The increment or Decrement in the requirement of the book . 2)How many books are available for the specific subject by various authors. 3)Different Edition of the specific book used by student in one week 4) Use of Specific book by students of various departments.

A. Introduction to Hadoop:[1]

Apache Hadoop is an open source Framework Which is used for Distributed Storage and processing of extremely large amount of the data sets at high velocity. Apache Hadoop consists of a storage part known as Hadoop Distributed file System(HDFS), and a processing part known as Hadoop Map Reduce. Hadoop splits files into large blocks and distributes them across the nodes.

1) Apache Hadoop framework is composed of the following :

a) Hadoop Common: It contain the basic library

b) Hadoop Distributed File System (HDFS): a distributed file-system that stores data in distributed manner. .

c) Hadoop YARN: It is a resource manager which is responsible for the managing and computing the resources.

d) Hadoop Map Reduce : It performs processing of the Data using Map and Reduce function

2) Hadoop Distributed File System(HDFS):

HDFS is designed for storing very large amount of files in Distributed Manner. It splits all the data into blocks and these Blocks are placed in to Data Nodes. Every blocks are replicated many times and replicas are stored in different nodes for fault tolerance. this provides detection of fault and fast, automatic recovery.

HDFS is represented as master slave architecture which has single Name Node as master and one or more Data Node as slave. Name Node manages the file system and keep all the information about the Data Node.

The Name Node knows in which block files are located in the Data Node. A file is splited into one or more blocks (default 64MB or 128MB) and these blocks are stored in Data Node Secondary Name Node communicates with the Name Node to take details of the HDFS metadata periodically. It is not a Backup of Name Node but in case of failure of Name Node It Works as Name Node.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

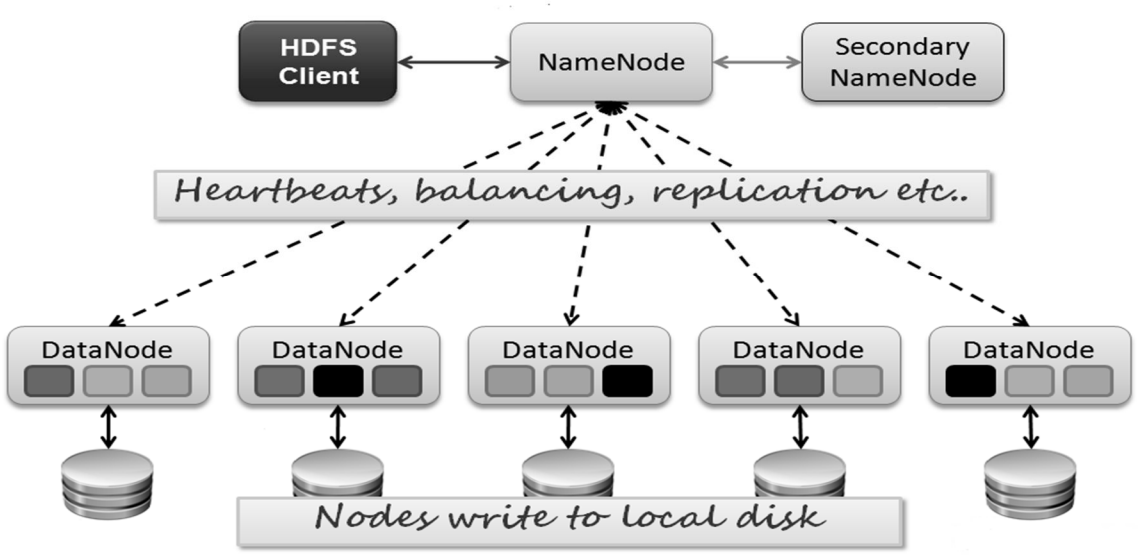


Figure 1 Architecture of HDFS

B. Introduction to Map Reduce: [1][2]

Map Reduce is a framework for processing the data which are extremely in large amount with various type at high velocity. Map Reduce can process the data either it stored in structured manner or Unstructured manner.

The processing of Map and Reduce function can be understood using Following figure.

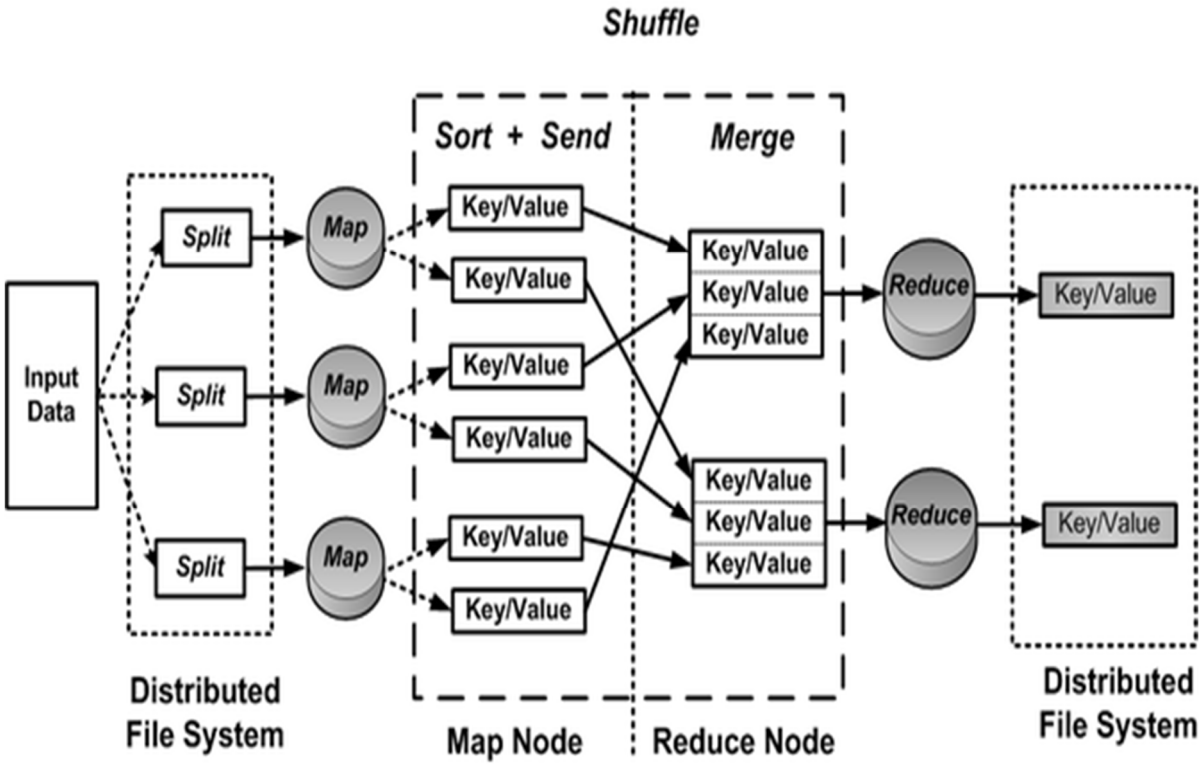


Figure 2 Map Reduce Framework

C. Introduction to Pig Latin.[1]

Apache Pig Components:

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

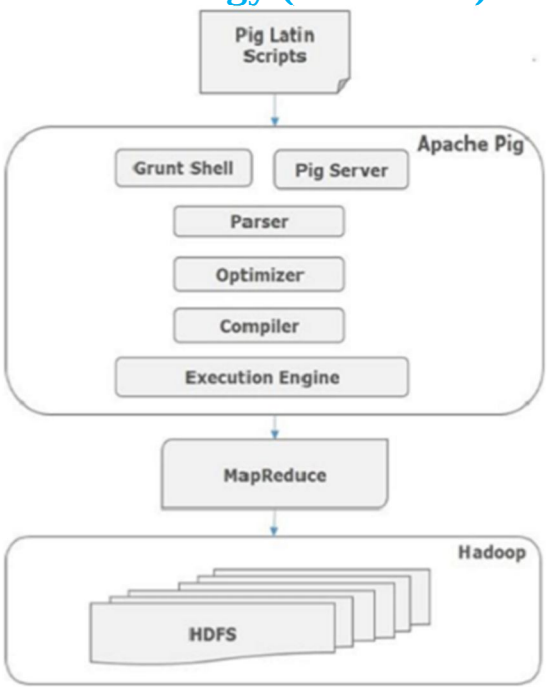


Figure 3 Apache Pig Component

Parser: Pig script is initially handled by the parser in the Pig Latin environment. after this it will checks the syntax of the script and other checks performed by the parser. Output of the parser is represented in the directed acyclic graph(DAG).
Optimizer: output of the parser is passed to the Optimizer ,then logical optimization is carried out.
Compiler: Optimized logical plan is complied into the series of the map reduce jobs by the compiler.
Execution engine: sorted order of Map reduce jobs are submitted to the Hadoop.

1) Apache Pig Execution Mechanisms
Three Ways to execute the Apache Pig
a) Interactive Mode (Grunt shell) : enter the Pig Latin statements and get the output.
b) Batch Mode (Script) : writing the Pig Latin script in a single file with .pig extension
c) Embedded Mode (UDF) : Provide ability to create our own function as in java.

II. RELATED WORK

A. An Insight on Big Data Analytics Using Pig Script: [3]
Apache Pig is extremely important Hadoop component which is use to process the Big Data with the minimum time and with less technical knowledge. From the analysis the regular library users and books accessed regularly by users and others mostly preferred by students at most regular time and date are analyzed.

1) Hadoop: It is a framework consisting Hadoop components like HDFS,YARN, Map reduce, Pig, Hive, HBASE etc.
2) HDFS and Map reduce are two major part of Hadoop.
3) YARN : It is a technology for the cluster management which is known as yet another resources negotiation. It also provides a central platform for consistent operation.
4) Here are some Hadoop components with their Functionalities.
a) Hadoop Common: Contains library files.
b) Map Reduce: Distributed processing fault tolerance.
c) HDFS: Storage and replication.
d) YARN: Cluster resources management.
e) HBASE: fast read and Wright access management system.
f) PIG: Scripting.

©IJRASET: All Rights are Reserved

g) HIVE: HQL
h) OOZIE: workflow and scheduling.
i) Mahout: machine learning..

Time Optimization in Pig :Execution time in Pig Latin is directly proportional to the size of input data.
T Proportional N
Execution time Proportional input size
So, f(n)= n.
So, time complexity for Pig Latin execution time is O(n).

B. Indian Health Care Analysis using Big Data Programming Tool.[4]

Author of this paper have analyzed the health care dataset for different research queries by using Pig Latin query language. Continuous growth in population and pollution increasing day by day so health problem may also increasing. In this situation Government have to provide better health care facility to the people.

1) Research methodology:
a) First of all data related to the analysis was collected then file will be located into the HDFS using copy command.
b) After this clearing of data extracting information, removing duplicated and noisy data , conversion of data into the standard format is done.
c) Storing of data will be done using HDFS, and data are duplicated over the multiple data nodes. So, fault tolerance can be handle, better quality of services provided.
d) Processing of data will be done using Pig Latin scripting language in this research.
e) After getting the result , the results are analyzed to get the meaning full details and accurate decision making.Query Formulation:
2) Here are the queries for the research:
a) Aggregate number of hospitals in between 1950-2015.
b) Aggregate number of physician in between 2005-2015.
c) Male-Female life expectancy in various states.
d) Percentage of people happy with the healthcare standards in different state.

Over the few time the health care services are provided in better way with high quality because doctors have higher qualification and increase number of hospitals.

C. Pig Latin: A Not So Foreign Language for Data Processing:[5]

In this paper basic introduction of the Big Data and its Hadoop ecosystem is explained as in the introduction and part of this paper. Features and motivation behind using pig Latin is the dataflow language and can process unstructured data also other feature is quick start and interoperability. Pig Latin data types are explained in the next part which also explained before in this Paper.

1) Some specific commands of Pig Latin:
a) Specifying input data: LOAD
b) Per-tuple processing: FOREACH
c) Discarding unwanted data: FILTER
d) Getting related data together: COGROUP
e) Join for joining two records: JOIN
f) Union, Cross, Order and Distinct also available.
2) Implementation part contain 3 steps:
a) Building the logical plans.
b) Compilation of map reduce plan.
c) Efficiency with nested bags.

D. Hadoop Ecosystem and Its Analysis on Tweets:[6]

Hadoop is a programming framework for distributed storage and processing of high volume of data with bulk of low information at a time. Hadoop consist mainly two components one is HDFS and other is Hadoop map reduce. HDFS is used for distributed storage

©IJRASET: All Rights are Reserved

of large amount of files and map reduce for processing high volume of data. Tweets are collected and analyzed using Hadoop ecosystem.

1) Hadoop and it's Architectural environment:

a) Map reduce: It is mainly designed for processing the distributed data at top on Hadoop. It has mainly two process: Job Tracker and task tracker, there is only on job tracker and one or more task tracker.

Job tracker schedules the task and manage the job where task tracker execute the task and perform map and reduce function.

b) HDFS: HDFS splits the data into the blocks and distributes to data node in the server .Each blocks are replicated in different node to manage fault tolerance and fact execution .It also has name node and database as in master-slave architecture. Name node contains all the info about the data node and manages the file system .It is almost act like file manager and it has all the details about on which it is stored .Data nodes are as same as the worker nodes. Secondary name node communicates with the name node and in the failure of name node it will work like name node.

2) Analysis of tweets with Hadoop:

Steps of analysis of tweets:

a) Tweet are collected using twitter 4j API (twitter 4j application programming interface) which is unofficial java library.
b) Searching word is also used for collecting tweet.
c) For sequential data collection from the twitter the thread mechanism is used.
d) After all the thread cycles the tweets are stored in the local disk then sended to the HDFS by java job.
e) All these tweets are analyzed using map and reduce algorithm.
f) After this analysis the result is transferred to the local disk.
g) After this result will shown in the form of chart or graph.

III.PROPOSED FRAMEWORK

A. Basic Method for Analysis

1) Loading of Data: In this research library data ,Which is structured data so for loading this structured data in HDFS Apache Sqoop is required.
2) Storing of Data: HDFS is used to store big size file across multiple nodes in a cluster.
3) Processing of Data: Map Reduce is used to process Big Size data.
4) Analysis: Collection of Data is not only requirement of the modern digital environment but to analysis of data and mine some meaningful information from this is more required. Analysis will be done using Pig Latin query Language.

B. Basic Requirement

1) Before installation of Hadoop, The System must have Ubuntu Operating System with Java and Open ssh server installed.
2) Mode of Hadoop must be Pseudo-Distributed Mode (Hadoop 2.7.3)
3) Apache Pig Version 0.16.0 is used here.
4) Apache sqoop-1.4.6 for loading structured data into HDFS.
5) Data Set used for this research is in .csv format.

IV.EXPERIMENTAL EVALUATION

There are 3 Different files for Books list, Member list, Issue-return book data list. (before analysis of these data joining of these 3 files required).

Books list have various fields like title (varchar), edition(varchar), year(int), author name(varchar).

Member List file have various fields like Barcodeno (UID), Fname (varchar), Mname (varchar), Lname (varchar), dept (varchar), degree (varchar).

Issue-return book data file have various fields like mem_cd (UID), accn_no (UID), iss_dt(date), recv_dt(date).

Books list file have list of 4595 books.

Member list file have list of 7819 members.

Issue Return book data file have 395536 records in the file

A. The increment or decrement in the requirement of the book.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

1) Query: Find the increment or decrement in the requirement of the book Computer Networks: A pragmatic approach
2) Input: Library Data Se
3) Output: Decrement in requirement of the book Computer Networks: A pragmatic approach.

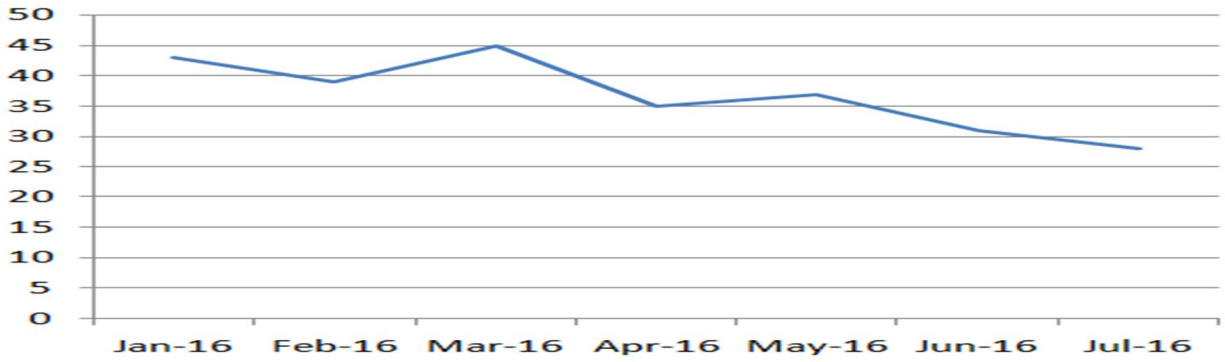


Figure 4 Decrement in Requirement in the book

4) Query: Find the increment or decrement in the requirement of the book Computer Big Data Analytic
5) Input: Library Data S
6) Output: Increment in requirement of the book Big Data Analytics

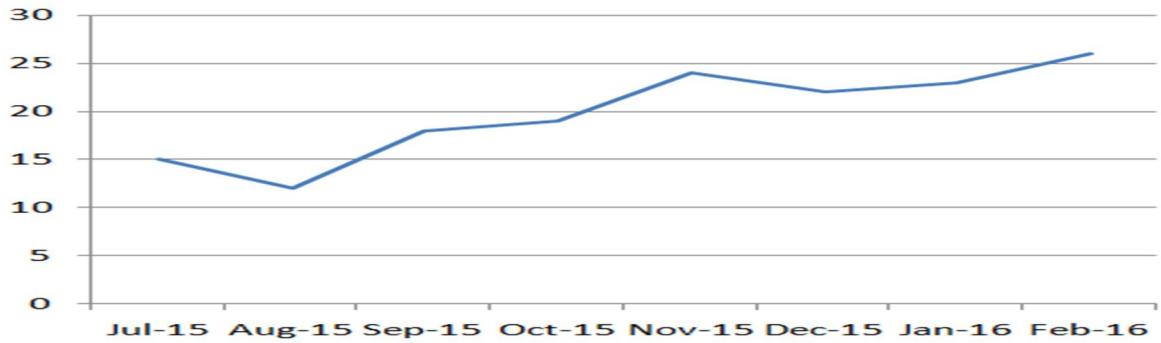
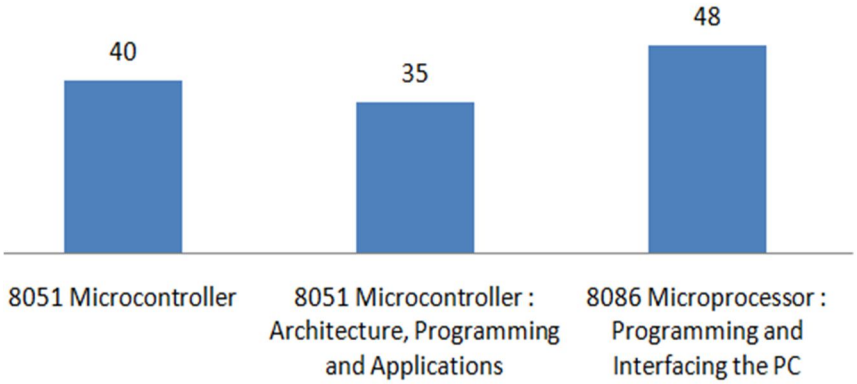


Figure 5 Increment in requirement of the book



B. How many books are available for the specific subject by various authors.

1) Query: Find Number of Books of basic environmental studies available by various autho
2) Input: Library Data Se
3) Output: Books of Basic of Environmental Studies available by various authors

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

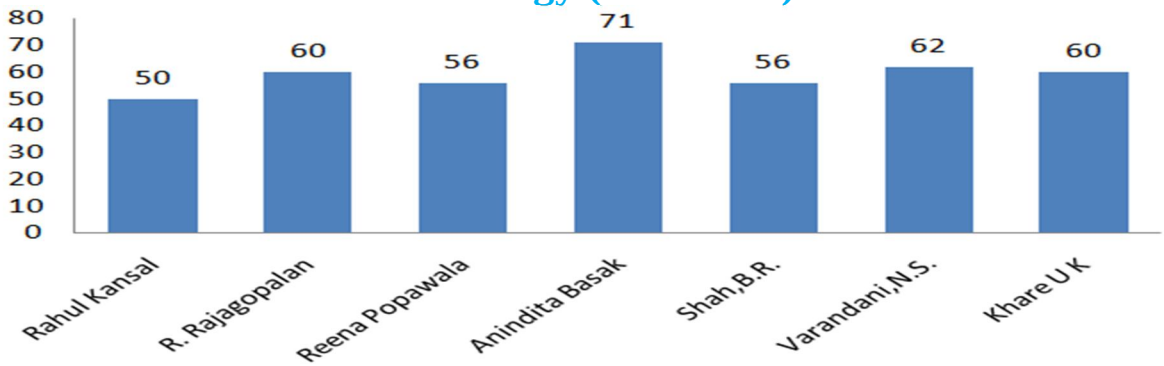
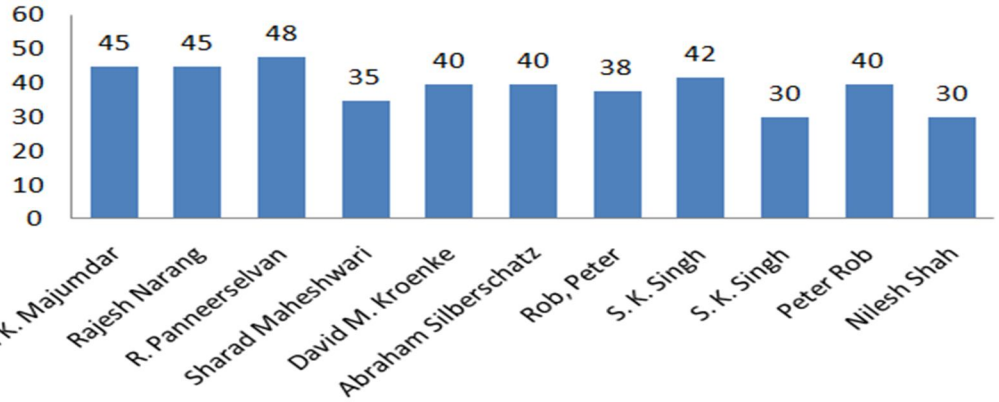


Figure 6 : Books of Basic of Environmental Studies available by various authors

4) Input: Library Data Se

5) Output**:** Books of Database management system by various author



6)

Figure 7 Books of Database management system by various author.

7) Input: Library Data Set

8) Output: Books of Mechanics of solid by various author

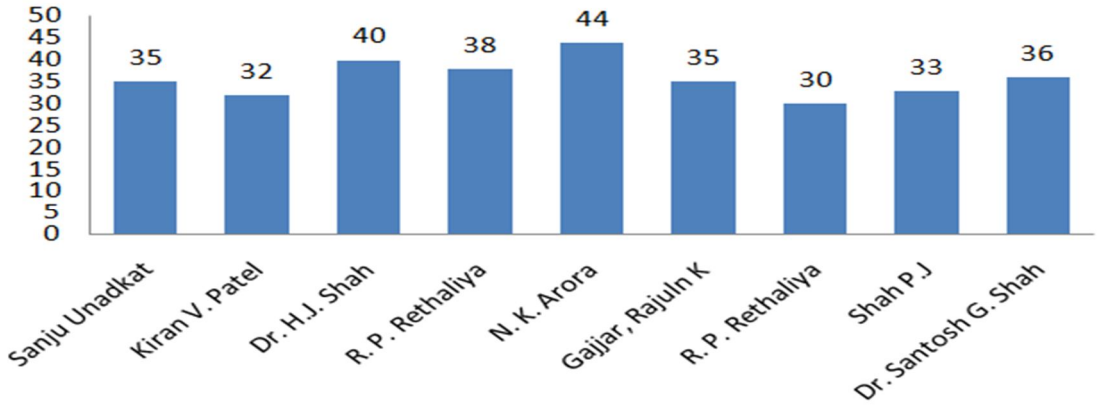


Figure 8 Books of Mechanics of solid by various author

C. Different Edition of the specific book used by student in one week.

1) Query: Different edition of the book of 8051 microprocessor

2) Input: Library Data Se

3) Output: Various edition of 8051 microprocessor book.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

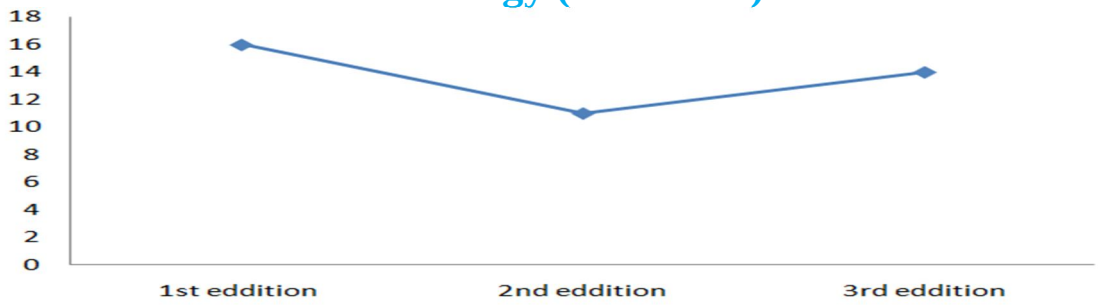


Figure 9 Use of Different edition of same subject book

D. Use of Specific book by students of various departments.

1) Query: Use of the book C language and numerical methods by student of various department

2) Input: Library Data Set

3) Output: Various users of book C language and numerical methods by student of various department.

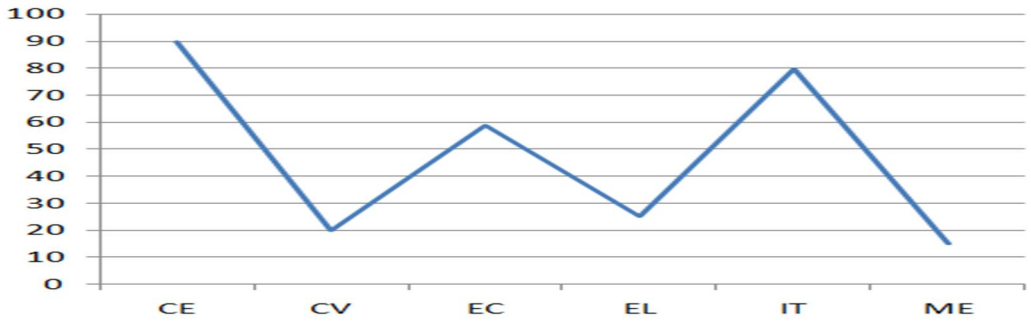


Figure 10 Various users by department.

4) Query: Use of the book Mechanics of solid by student of various department.

5) Input: Library Data Set.

6) Output: Various users of book Mechanics of solid by student of various department.

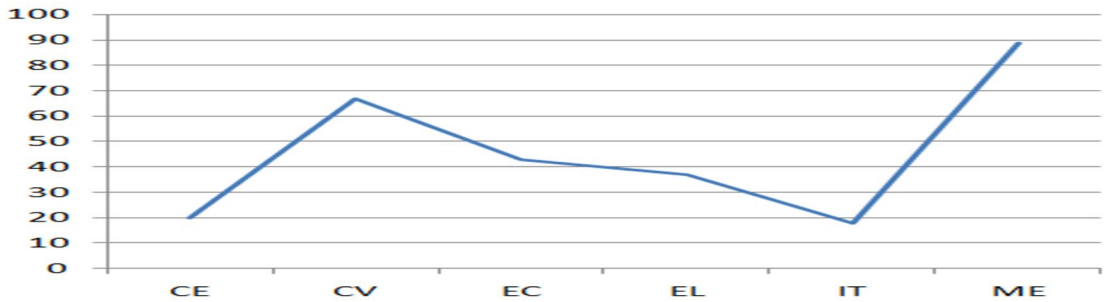


Figure 11 Various users by department.

V. ISSUES AND LIMITATION

This Research requires data in .csv format only. so if data is in other format then we need to convert it in .csv format. This conversion takes time and it is difficult when it is in high volume. Difficult to Provide High Quality of service when Volume of data is increased.

VI. CONCLUSION

Big Data can be analyzed using pig latin. But It requires Map Reduce for processing and HDFS for file storage
.when processing with Big Data, technical knowledge about Map Reduce and Hadoop is required .Data which is analyzed by pig latin is only in .csv format. any other format of is not considered in this analysis. execution speed of the Pig Latin is much faster than the Java language. Apache pig is executed in three ways interactive mode, batch mode, embedded mode. Most of the authors

©IJRASET: All Rights are Reserved

have used interactive mode for the execution.

VII. ACKNOWLEDGEMENT

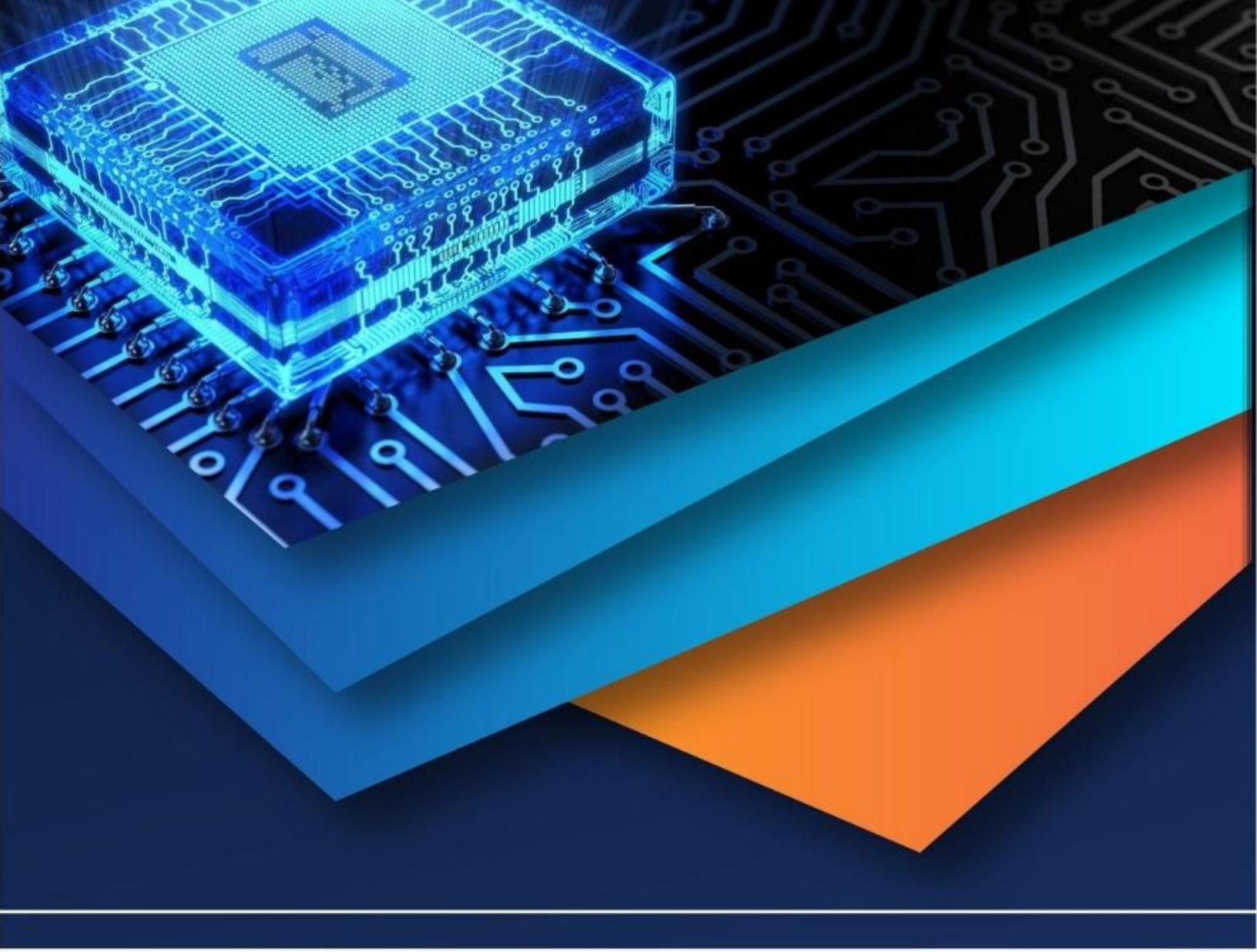
We would like to thank our institute School of Engineering RK University for providing Required information and Library Data for this research. We would also like to thank Vice Chancellor of RK University , Director of School of Engineering RK University ,Our Family ,Friends and all the people who helped us to make this possible.

REFERENCES

[1] Y. Demchenko, C. De Laat, and P. Membrey, Defining architecture components of the Big Data Ecosystem, 2014 Int. Conf. Collab. Technol. Syst. CTS 2014, pp. 104112,
[2] H. Hu, Y. Wen, T. S. Chua, and X. Li, Toward scalable systems for big data analytics: A technology tutorial, IEEE Access, vol. 2, pp. 652687, 2014
[3] R. Sparks, A. Ickowicz, and H. J. Lenz, An Insight on Big Data Analytics, vol. 4, no. 6, pp. 3348, 2016.
[4] M. Dayal and N. Singh, Indian Health Care Analysis using Big Data Program-ming Tool, Procedia Comput. Sci., vol. 89, pp. 521527, 2016
[5] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, Pig latin, Proc. 2008 ACM SIGMOD Int. Conf. Manag. data - SIGMOD 08, p. 1099, 2008
[6] C. Uzunkaya, T. Ensari, and Y. Kavurucu, Hadoop Ecosystem and Its Analysis on Tweets, Procedia - Soc. Behav. Sci., vol. 195, pp. 18901897, 2015
[7] A. Z. Bhat and I. Ahmed, Big data for institutional planning, decision support and academic excellence, 2016 3rd MEC Int. Conf. Big Data Smart City, ICBDSC 2016, pp. 118122, 2016
[8] . 2014U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, Critical analysis of Big Data challenges and analytical methods, J. Bus. Res., 2016
[9] N. Chakraborty and S. Gonnade, Big Data and Big Data Mining: Study of Approaches , Issues and Future scope, vol. 18, no. 5, pp. 221223, 2014
[10] A. Jain and V. Bhatnagar, Crime Data Analysis Using Pig with Hadoop, Procedia Comput. Sci., vol. 78, no. December 2015, pp. 571578, 2016
[11] M. Srivathsan and K. Y. Arjun, Health Monitoring System by Prognotive Com-puting Using Big Data Analytics, Procedia Comput. Sci., vol. 50, pp. 602609, 2015
[12] M. Jadhav, Big Data: The New Challenges in Data Mining, Ijircst.Org, no. 2, 2013
[13] M. Dayal and N. Singh, An Anatomization of Aadhaar Card Data Set A Big Data Challenge, Procedia Comput. Sci., vol. 85, no. Cms, pp. 733739, 2016.
[14] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, Trends in big data analytics, J. Parallel Distrib. Comput., vol. 74, no. 7, pp. 25612573, 2014
[15] K. Sin and L. Muthu, Application of big data in education data mining and learning analytics-A literature review, Ictact J. Soft Comput. Spec. Issue Soft Comput. Model. Big Data, vol. 5, no. 4, pp. 10351049, 2015.

.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)