IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-13, May 2019 E-ISSN: 2347-2693

Novel Approach to Seat Matrix Prediction using Hadoop

Ravikiran M D[1*], Poonam Gouli[2]

[12]Department of Computer Science & Engineering, RVCE, Bengaluru

Corresponding Author: ravikirandevadiga@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si13.2532 | Available online at: www.ijcseonline.org

Abstract— Big data is a term that is used to describe huge data sets having large, varied and complex structure with the hardness of analysing, storing and visualizing for further analysis or results. The analysis into large amounts of data to expose secret correlations and hidden patterns can be termed as big data analytics. [1] Big Data has proved to be beneficial as it helps to gain richer and deeper insights into the underlying mass of data. Common Entrance Test is a flat form for the students to opt for colleges to pursue Under graduation. Every year approximately 150000 students take up CET. Thus, these students will compete for seat among 220 engineering colleges across Karnataka, that are enrolled to the CET cell. Student can opt for a College based on availability of seat or branch for the rank they obtain in CET. Predictive analytics is the basic enabler for big data. On a day to day basis, Businesses collect large quantity of customer data which is used by predictive analytics along with historical data, coupled with customer insight, to forecast future events. For predicting a college several tools are taken into consideration they are: HBase for database, MapReduce for data processing mahout's distributed naive Bayes classification for classifying and training data.

Keywords—CET, HADOOP, PIG, MAHOUT, MAPREDUCE

I. INTRODUCTION

Allocation of seats to a college as per KEA norms is an elaborate process. The patterns and the schemes that are followed for the allotment each year are subjected to timely reforms that are imposed by the governing organization. Two of the recently followed strategies are Offline seat matrix display and Online seat allotment in terms of mock and actual placement.

This paper comprises of an overview of big data's scope, methods, samples, challenges associated with it. The project also aims at putting an end to the eagerness of the parents along with the forthcoming candidates who are waiting for their turn to get into a reputed college to pursue their graduation / post-graduation. This tool will help the students/parents to understand the probability of getting a seat in a college. Section I contains the Introduction of "*Novel approach to seat matrix prediction using Hadoop*", Section II contains the Related work done for the paper, Section III contains the Experiment and Data , Section IV contains the Methodology, Section V contains Results, Section VI contains conclusion and future Scope, Section VII contains References.

II. RELATED WORK

Literature survey describes the existing and established theory and research thus simplifying the process of achieving the end result of the project. It also helps in identifying the deviations in the analysis that are done over the same concept. Here, a list of all the papers that were referred in the due course of the project as been listed.

In the distributed environment [2] mahout works well. Mahout effectively scales in the cloud using the Apache Hadoop library and provides the developer with a ready-to- use framework for data mining tasks on large volumes of data. The two basic deciding factors considered from the candidates end in the admission process are: AIEEE rank and family pressure. This makes it difficult for them to decide of another branch in case they don't get a seat in the desired branch.

Regular evaluation of the teaching patterns of universities [3] plays an important role in improving management levels. The methodology used in college teaching has the characteristics of the Markov chain. Using the Markov chain, it selects four ranks of comprehensive evaluation results in different specialties in order to carry out an analysis in a state shift; it then carries out a gradual assessment of specialized teaching through the progressive matrix of the transfer matrix and its degrees of efficiency;

finally, according to the vector of a steady state, it forecasts the long-term effect of specialized teaching.

Seat matrix prediction is a non-linear classification problem. This paper [4] proposes a new approach for Prediction of college entrance examination (CEE) based on the learning algorithm of vector machines. Data to be collected include ranking in all subjects, CEE score, the number of college admission plan and relevant data of the past two years. With all the data gathered a training set is formed.

III. EXPERIMENTS AND DATA

Relevant details should be given including experimental design and the technique (s) used along with appropriate statistical methods used clearly along with the year of experimentation (field and laboratory).

3.1 Data Description

Data set was downloaded from the kea website. Dataset used is in PDF format. Inside the PDF format the data was there in xml format. The dataset contains Rank, CET NO, Name of the student, Category, College/Course etc. from KEA Website.

3.2 Research Challenge

One of the main challenges encountered while beginning the project was that of Data Collection, as it wasn't available in the desired format. It was placed in such a manner that the main PDF document housed has an XML document inside it, which can be referred to as semi structured format. This cannot be converted directly into csv format. Data wrangling method from Visual Basic script was used to convert the original document into csv format. Next challenge that came in was that that algorithm selection. Since a large collection of libraries were available in Mahout such as distributed naive based classification method was chosen as the base algorithm.

3.3 Motivation

Big Data is the most important asset of any organization which is further processed to produce useful information [5]. Big Data techniques are widely used in industrial sectors to generate patterns that are helpful for earning more profits and expansion of business. One such tool can be to make appropriate predictions in the education sector where in decisions pertaining to college allotment can be made. It can serve to be an effective means of information to both institutions and aspiring students to represent and understand
the statistics.

3.4 Methodology

3.4.1 Extraction transformation load:

HBase is another example of an environment for non-
relational data management that distributes massive data sets across the Hadoop framework. It has a column- based data layout that can be used to store and manipulate large data tables. It is a reasonable alternative as a persistent storage
platform when running map reduce application. Well-timed updating, analysis and storage of the unstructured data will be conducted on a regular basis by utilizing HBase.

3.4.2 Training Phase:

Direct querying of unstructured data is not feasible in terms of big data. Hence a combined approach involving map reduce application along with mahout library for machine learning algorithms, text mining and binary search tree algorithms is made use of for achieving the task of querying into the big data.

3.4.3 Execution phase:

The two earlier stages are collaborated with each other.
A good walk-through of the previously stored data values is made at appropriate locations to determine what the future trend might be, for any given scenario

3.4.4 Results:

For example, if a candidate has secured 150th rank and belongs to the GM category. Then this software will make a thorough analysis of the college allotment data pertaining to the earlier years. i.e. the allotment strategy followed for the 150th and its neighbouring rank holders to arrive at a predicted conclusion that the current candidate may or may not be allotted to the same college as that of the earlier pattern that was followed.

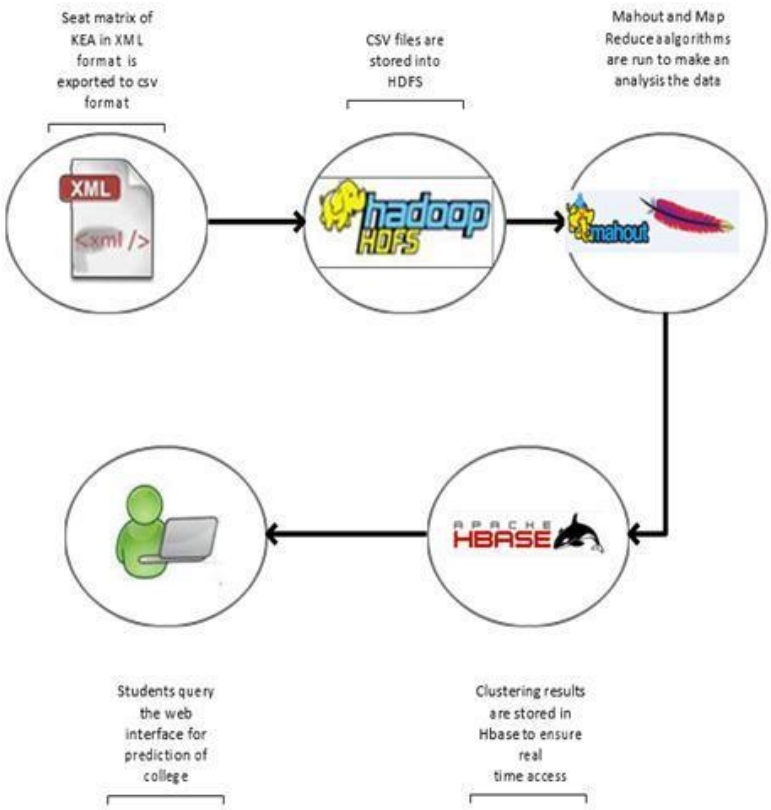

Fig 1 Methodology of seat matrix prediction

IV. ANALYSIS AND DESIGN

Steps followed to Prediction of Rank

1. Data Preprocessing
2. Store the KEA DATA to HDFS.
3. Mapper Phase.
4. Reducer Phase
5. Prediction phase with mahout algorithm,
6.Final Predicted Rank Stored in HBase.

1) Data Preprocessing: The Data was in PDF Format with unformatted Data Cells. PDF data is Converted to CSV with removing of unnecessary column values using data wrangling method.
2)Store the KEA DATA to HDFS: Converted CSV files from PDF will be saved to the HDFS file systems for Data post processing. The files will be divided into chunks of data in HDFS.
3)Mapper Phase: In the Mapper phase all CSV files from the HDFS will be collected together and key value pairs will be produced as a intermediate output
4)Reducer Phase: In the reducer phase intermediate output from mapper phase will be taken as input. And intermediate results will be sorted and results of reducer phase will be stored to HDFS file system.
5)Prediction phase with mahout algorithm: In the prediction phase previous results from the Reducer phase will be considered as input. we will user mahout machine learning module from apache to predict the rank.
6)Final Predicted Rank Stored in HBase: At last we will find out the predicted rank from the prediction algorithm we will store this results our own database in Hbase to make sure that remaining candidates will not get the same college as previous candidates predicted result.

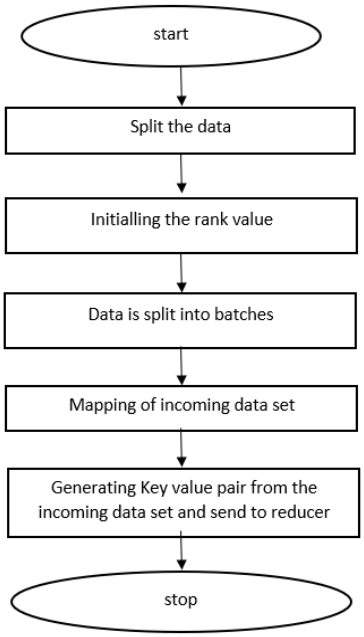

FIG 2: Flowchart of Mapper Phase

In mapper module, the important concept is generating <key,value> pairs. These pairs are stored in an intermediate file which will be accessed by reducer. Figure 3 shows flowchart for mapper Phase.

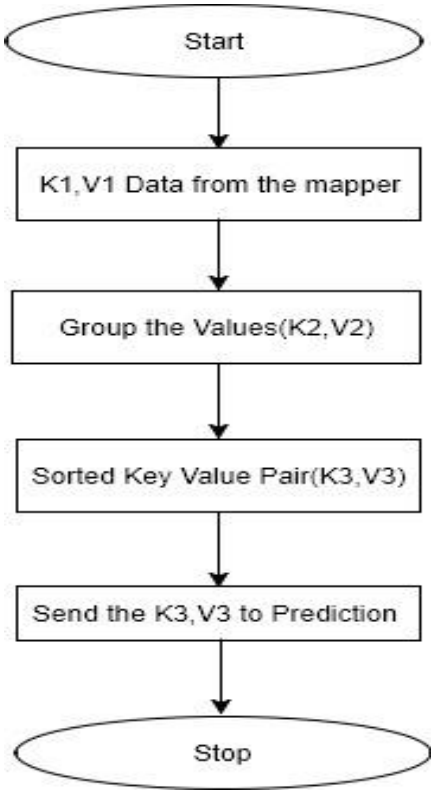

FIG 3: Flowchart of Reducer Phase

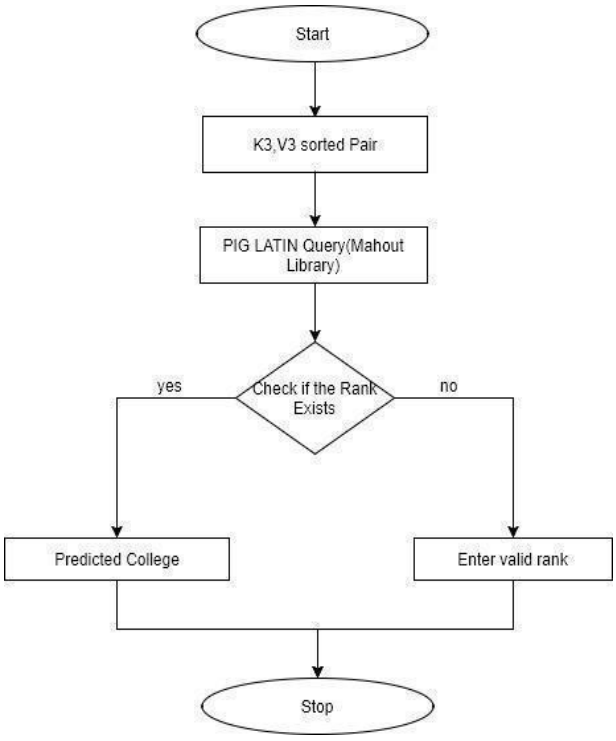

Fig 4 Flowchart of Prediction phase

In mapper module, the important concept is generating <ke y, value> pairs. These pairs are stored in an intermediate file which will be accessed by reducer. The major work of reducer is to collect data from mapper and then sort the desired result. Figure 5.4 shows flowchart for reducer. The reducer first collects the intermediate data from all mappers [6]. It calculates the incoming data from these intermediate data. This will be the updated or create the new row in the database. The data of whole process will be in sorted order and these values will be the input for prediction. The final output from the reducer Figure 5.4 shows flowchart for predicton i.e. the (key,value) pair (K3, V3) is sent to the PIG Latin query which makes use of the mahout library. This query performs the basic check that searchers whether the rank is present in the CSV file. If it is present, then the probable/predicted college will be displayed as output. If it's not present, then an alert pops up saying "Please enter a valid rank".

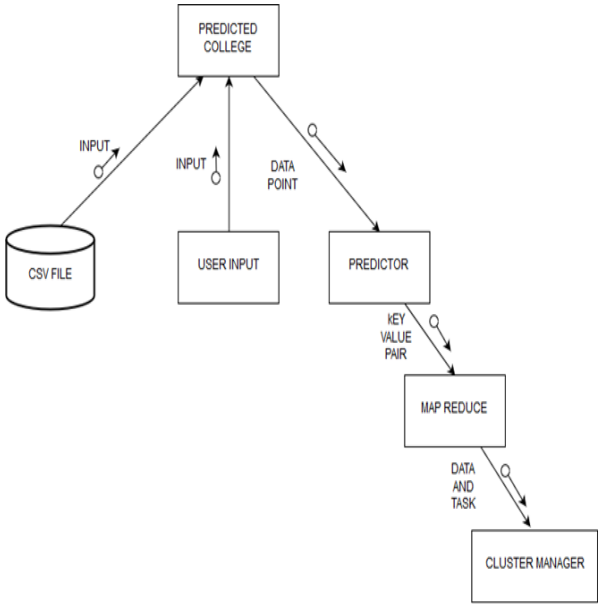

Fig 5 Structural Chart of Seat Matrix Prediction

V. RSULTS AND DISCUSSION

Sample of CSV file stored in HDFS: The contents of the CSV file with multiple parameters that will be available in HDFS is shown in fig 6.

Sl.No	CET NO	CANDIDAT	ENGG. RAI	Category(	Category/	Course	Alloted Date	Adm. Ord.	Fee Collec	Qual_Exan	Qual_Exan	Qual_Exan	Qual_Exan	College Name
1	NJ599	SYED ABU	3733	GM	GM	Civil	09-Jul-11	43095	Rs.18,090	96/100	85/100	97/100	90/100	College :1 - University Visveswariah College of Engineering,Bangalore.
2	LQ064	RANJITH J	4745	GM	GM	Civil	09-Jul-11	43902	Rs.18,090	98/100	96/100	94/100	75/100	College :1 - University Visveswariah College of Engineering,Bangalore.
3	LE128	HAREESHA	6530	3BR	SNQ	Civil	11-Jul-11	45442	Rs. 3,090	90/100	88/100	95/100	72/100	College :1 - University Visveswariah College of Engineering,Bangalore.
4	YA174	DHARANE!	7236	3BG	SNQ	Civil	11-Jul-11	46006	Rs. 3,090	93/100	98/100	94/100	89/100	College :1 - University Visveswariah College of Engineering,Bangalore.
5	CG550	SUSHMA P	7712	GM	GM	Civil	11-Jul-11	46498	Rs.18,090	89/100	93/100	78/100	74/100	College :1 - University Visveswariah College of Engineering,Bangalore.
6	HN561	SALMAN	7792G55	2BG	GM	Civil	11-Jul-11	46511	Rs.18,090	87/100	83/100	88/100	70/100	College :1 - University Visveswariah College of Engineering,Bangalore.
7	AL524	CHANDRA	7826	GM	GM	Civil	11-Jul-11	46608	Rs.18,090	93/100	82/100	96/100	81/100	College :1 - University Visveswariah College of Engineering,Bangalore.
8	CU155	MEGHANA	8000	GM	GM	Civil	11-Jul-11	46739	Rs.18,090	91/100	97/100	87/100	94/100	College :1 - University Visveswariah College of Engineering,Bangalore.
9	AY367	VIDHYA G	8037	GM	GM	Civil	11-Jul-11	46766	Rs.18,090	79/100	84/100	95/100	/100	College :1 - University Visveswariah College of Engineering,Bangalore.
10	YZ092	HONNESH	8241	3AG	GM	Civil	12-Jul-11	46925	Rs.18,090	88/100	89/100	93/100	91/100	College :1 - University Visveswariah College of Engineering,Bangalore.
11	XB118	AJITH SHA	8283	GM	GM	Civil	03-Sep-11	46958	Rs.18,090	85/100	72/100	71/100	/100	College :1 - University Visveswariah College of Engineering,Bangalore.
12	CF433	PRASAD S.	8521	GM	GM	Civil	12-Jul-11	47171	Rs.18,090	78/100	82/100	95/100	79/100	College :1 - University Visveswariah College of Engineering,Bangalore.
13	GB145	VEERESHB	8543	3AG	SNQ	Civil	03-Sep-11	47177	Rs. 3,090	79/100	74/100	95/100	86/100	College :1 - University Visveswariah College of Engineering,Bangalore.
14	LQ265	DEVEGOW	9097	3AR	GM	Civil	12-Jul-11	47666	Rs.18,090	92/100	88/100	93/100	96/100	College :1 - University Visveswariah College of Engineering,Bangalore.
15	PP189	NAGESH	9149	GM	GM	Civil	12-Jul-11	47647	Rs.18,090	88/100	88/100	95/100	90/100	College :1 - University Visveswariah College of Engineering,Bangalore.

Fig 6 Contents of CSV file

Mapper reducer execution displaying the status as 100% complete as as shown as fig 7 and fig 8

```
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local852967345_0005
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases ket,ket_short3,newone3
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: ket[1,6],ket[-1,-1],newone3[6,10],ket_short3[5,13] C:  R:
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
ner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner -
ner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task:attempt_local852967345_0005_m_000000_0 is done. And is in the process of committing
ner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner -
ner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task attempt_local852967345_0005_m_000000_0 is allowed to commit now
ner Map Task Executor #0] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_local852967345_0005_m_000000_0' to file:/home/cloudera/Desktop/ketoutput2/
000000
ner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - map
ner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task 'attempt_local852967345_0005_m_000000_0' done.
ner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local852967345_0005_m_000000_0
INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local852967345_0005
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
```

Fig 7 Execution of MapReduce

```
24081,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
25515,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
25926,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
27425,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
27972,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
28507,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
29711,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
29782,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
30923,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
31048,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
31536,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
31683,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
32024,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
32752,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
33519,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
33575,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
33617,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
34379,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
34564,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
34578,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
35072,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
35374,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
35429,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
35529,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
35604,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
36658,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
36831,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
36936,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
37271,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
37720,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
37868,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
38378,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
44711,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
56613,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
58081,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
58591,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
80593,"College : 508 - Jain College of Engineering,Belgaum (IInd Shift).")
```

Fig 8 output of MapReduce

Search string that is supplied to the Map Reduce output is as follows: newone3 = filter ket_short3 by eng_rank == 3490; where newone3 is the resultant search from the PIG query and ket_short3 and eng_rank are variables. The result / predicted college based on the above input string is shown as in fig 9

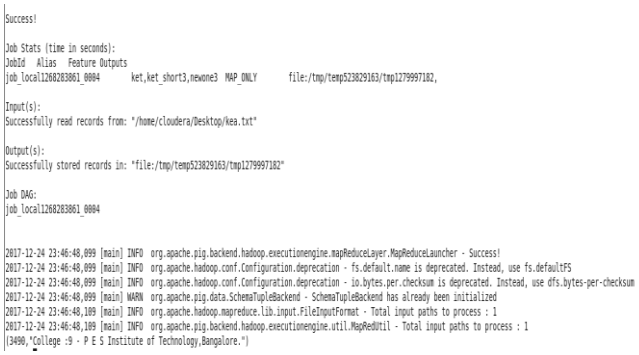

Fig 9 Predicted College in Command Prompt

The below command shows how to store the above result into a file in HDFS:

store newone3 into '/user/cloudera/ketoutput' using PigStorage('\t');

All the results that are stored in this manner can be used for comparisons whenever required. The contents of the resultant file will be stored with the name "part-m- 00000". The below Fig. 10 shows a view when the file is opened.

Metrics used to verify the results are Rank and college. Its approximately accurate for the year 2014 after verifying the 2014 CET [7] allotment sheet. But for the year 2016 it was not accurate. Need to include Category and branch to get the best accurate result.

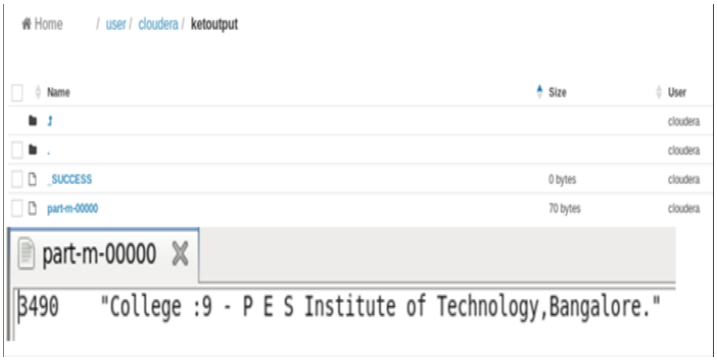

Fig 10 Predicted College stored in HBase and HDFS

Conclusion and Future Scope

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. Predictive analytics is an enabler of big data: Businesses collect vast amounts of real-time customer data and predictive analytics uses this historical data, combined with customer insight, to predict future events. Predictive analytics enable organizations to use big data (both stored and real-time) to move from a historical view to a forward-looking perspective of the customer. Advantages 1) Predicting the college based on rank is made easier.2)Time consumed in prediction is drastically reduced.3) Predicted results can be stored for future use3)Result matrix for the years to come can be paced along with the existing data in HDFS Limitations and future work 1)As of now Web UI is not made available to the user to key in the input. Prediction of college can be made by taking candidates belonging to reserved categories into consideration. The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

REFERENCES

[1] Sanjay Kapoor , Deepak Kumar , Sandeep Jain , Brijesh Khandelwal , Anil Mittal, Manoj Kumar "An optimized e- Allotment algorithm for admissions in educational institutes," International Conference on Technology Enhanced Education (ICTEE),IEEE, 2012, pp.17 25.

[2] Rotsnarani Sethy, "Big Data Analysis using Hadoop: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, July 2015, pp 1153- 1157

[3] SHIVAM AGARWAL,"DATA MINING CONCEPTS AND TECHNIQUES", International Conference on Machine Intelligence and Research Advancement, 2013, pp 203-207
[4] B.N. Lakshmi, G.H. Raghunandhan, "A Conceptual Overview of Data Mining," National Conference on Innovations in Emerging Technology-2011, pp 27-33
[5] V.K.Deepa, J. Rexy R. Geetha," Rapid Development of Applications in Data Mining", International Conference on Green High Performance Computing March 14-15, 2013 pp 13-16
[6] Jinlong Wang ,Jing Liu , Russell Higgs , Li Zhou , "The Application of Data Mining Technology to Big Data", IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) 21-24 July 2017 pp 284-289
[7] http://kea.kar.nic.in/

AUTHORS PROFILE

Mr.Ravikiran Devadiga pursed Bachelor of Engineering from University of VTU, Belgaum in 2012 and M.Tech from VTU University in year 2019. He is currently working as Programmer in Department of Computer Sciences, RNSIT since 2012.

Dr Poonam G is Currently Working as Associate Prof in RVCE Bangalore. Her area of interest in Research are Distributed Clustering, MapReduce And Machine Learning.

