iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

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

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

Health Care Analysis for Cost Reduction using Hadoop

Ahmed Sayed [1], Akshay Wadkar[2], Rohan Bendre[3], Akshay Adagale[4], Anantha Subramaniam[5]
[1,2,3,4,5] *Computer Department, Savitribai Phule Pune University*

Abstract: ***The world we live in is changing drastically day by day and as we know that the world population is exceeding over 7Billions in coming ages. So there is also a rapid increase in data provided by these people. So if we see in past few ages there has been 2.5 quintillion bytes of data generated per day. So the issues with storage of this data. Our paper is responsible for storing this data and manipulating it overall known as Big Data.***
So if see this data is coming from various sources daily used so we can't afford to lose the data. But the most important data is the Health Data there are numerous patients overall in the world so there should be machines generated to store the data of these patients. So it will be benifitial to them when further accessing for any reason. This paper focuses on accessing data from a data source that is un-structured and finding data-sets on which analysis can be done to reduce the cost.
Keywords: Big Data, Medical Analysis, Data Analytics , HIVE, BIRT Reporting, Sqoop.

I. INTRODUCTION

Big data analytic is a growth area with the potential to provide useful insight in health care[1]. While many dimensions of big data still present issues in its use and adoption, such as managing the volume, variety, velocity, veracity and value. Volume is nothing but the quantity of generated and stored data, Variety is the type and nature of the data and Velocity is refers to the speed at which the data is generated and processed to meet the demands and challenges It has provided tools to accumulate, manage, analyse , and assimilate large volumes of disparate, structured, and unstructured data produced by current health care systems[1].
Big data what does this word means? Big Data in simple words can be defined as data that cannot be processed by normal computing machines, which are used in day to day life. For this we don't need to buy separate machines costing 100K(Lakhs)
we are provided with a technology Hadoop which will help us in solving this problem.
The problem with Big Data is solved but there is another problem caused by various types of data. There are various types of data such as:

A. Structured Data.
B. Un- Structured Data.
C. Semi-Structured Data.

So basically Structured Data is something which comes in the form of Rows & Columns which we generally see in tables where we use Structured Query Language(SQL) to manipulate this data. But Unstructured Data is a type of data which in un-defined type of format the reason for this is data comes from various resources and it cannot be manipulated using SQL.
The third type is Semi- Structured Data where data is represented in the form of tags using Extended Mark-up Language (XML) we can manipulate the data.
So the problem is with un-structured data because we don't have any technology yet which can solve this problem but we can achieve this by using Map-Reduce concept provided by Hadoop. So Map-Reduce helps in converting the un-structured data into structured data using key- value pairs. Once this is done we can then fire queries an access data from data resources.
Big data analytic is a growth area with the potential to provide useful insight in health care[1]. While many dimensions of big data still present issues in its use and adoption, such as managing the volume, variety, velocity, veracity and value. Volume is nothing but the quantity of generated and stored data, Variety is the type and nature of the data and Velocity is refers to the speed at which the data is generated and processed to meet the demands and challenges It has provided tools to accumulate, manage, analyse , and assimilate large volumes of disparate, structured, and unstructured data produced by current health care systems[1].

II. EXISTING SYSTEM

SQL as we know is the Structured Query Language where we fire the predefined format queries to achieve data sets or data in the

form of rows and columns that is in table format. SQL provided easy format or syntaxes for queries there are three types:
DDL(Data Definition Language)
DML(Data Manipulation Language)
DCL(Data Control Language)
Where, Data Definition Language (DDL) is a standard for commands that define the different structures in a database. DDL statements create, modify, and remove database objects such as tables, indexes, and users. While DML is used to retrieve, store, modify, delete, insert and update data in database. While a data control language (DCL) is a syntax similar to a computer programming language used to control access to data stored in a database (Authorization). In particular, it is a component of Structured Query Language (SQL).
But the only problem with SQL is its Structured in nature which means it only works with Structured Data. When other types of data are passed such as Un-Structured or Semi-Structured then SQL does not work. So SQL fails in this case.
As discussed earlier we cannot loose data as data is very important and it comes from various sources so predicting its type is not possible so we need a system which will accept all the data as it comes irrespective of its format and then convert it into a specific structured format so that accessing data sets from it become easier for us. So this part is done by our system.
As we know we are working on Big Data there are 3 aspects of Big-Data :
Volume: Big Data doesn't sample, it just observes and tracks what happens.
Velocity: Big Data is often available in real-time.
Variety: Big Data draws from text, images, audio, video.
So these above aspects states that Big Data comes in greater volume with a high velocity and also comes in variety of data. So machines should be capable of handling all these aspects. So to overcome this problem in existing Systems Hadoop was introduced which solves this problem.
In an existing system when un-structured data is passes we can use hadoop. So what basically hadoop will do is that it will accept the data as it is irrespective of its type and manner. It will first store the data and the using it Map-Reduce it will convert it into key value pairings and then this data will be converted into a Structured Format .
Another concept provided by Hadoop i.e Hive will be used to fire queries on this Big Data to generate data sets.
As we know storing was a problem for this BIG DATA, so then on further survey we found that using Hadoop can solve this problem. Big data analytics provides great potential for health care operations to increase patient safety and quality health care while reducing medical errors and costs. This report gives an insight of how we can uncover additional value from the data generated by health care. Large amount of heterogeneous data is generated by hospitals. But without proper data analytic methods these data became useless. With the development of smart devices more and more public health data can be collected from various sources and can be analysed in an unprecedented way.

III. PROPOSED SYSTEM

After understanding with the problem and the drawbacks lets discuss about the solution . The solution for handling Big Data is Hadoop, which is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

A. Now the Over-All Structure for Solution IDs Divided into Following 4 Steps

1) *Understanding the Un-structured Data:* Understanding here means to analyse the un-structured data i.e. the key elements the no of columns. Counting the columns and thereby dividing them into appropriate manner.
2) *Converting the Un-Structure Data into Structured Data using Map-Reduce:* Passing the un-structured file as an input to the Map-Reduce while on generating the output in the form of key- value pairs which will give us the Structured Data.
3) *Accessing Data-Sets in the Form of Rows & Columns (i.e. Tables):* So using HIVE which acts like the SQL of Hadoop we can fire queries and generate data accordingly.
4) *Using BIRT (Business Intelligence Reporting Tool) to generate the Reports:* The BIRT is a report generation tool which is given as a plug-in and HIVE data source is accessed by BIRT and then reports are generated accordingly.

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

B. The following Diagrams will give Some Examples about the Entire Process:

```
1245012188863032   130110410    2 211111
2
2
52
1001011              1            1            3
010112101002400101012000196            01 (
0631531555320302103217244912188842062
2
1121772
62
1001011              1            1            2
010111101002480101012000196            01 (
101203201201201101212243712588861042   1
2
1112
31
1001011              1            1            1
```

Fig 1:Un-Structured Data.

NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
5160	1	4	9999	243	1	1	1
1							
1	1	1	1	1	3003	42	1
1	2	3	3	3	1	2	2
NULL	NULL	2	NULL	NULL	NULL	NULL	NULL
1	7	2	NULL	NULL	210	140	505
3	1	2	2	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
1	1	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
000	1	4	9999	350	3	2	4
1							
1	1	1	2	1	3003	49	1
2	3	5	15	3	2	NULL	NULL

Fig 1: Structured Data

c1	impage
2,123	18
1,852	19
2,115	20
2,444	21

Fig 3: Data Set Accessed.

C. Final Outputs will Generated in the Following Forms

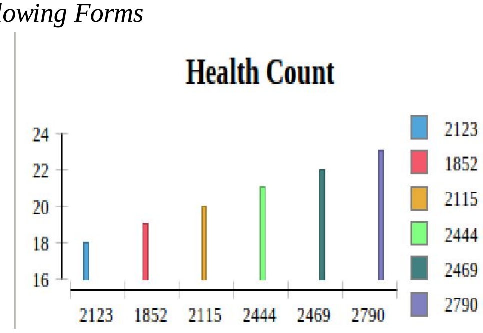


Fig 4: Report 1

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

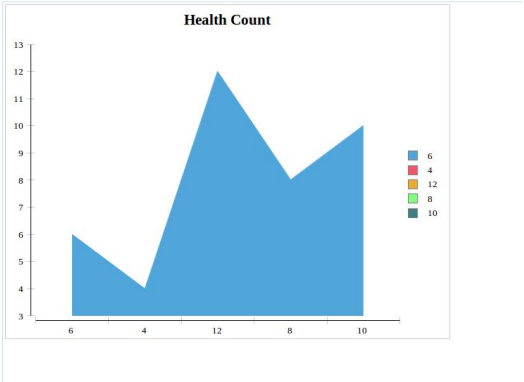


Fig 5: Report 2

So we can see that the reports are generated in this format. Now the further work is done by the Analyst. Decisions are taken by them by analysing the report.

IV. CONCLUSIONS

In this paper Hadoop was used to differentiate the structured data and un-structured data.. This model produces the good conversion using Map Reduce. Technique such as Hive, Map-Reduce ,Flume & Scoop of Hadoop was used .The existing model problems were overcome by this model by generating structured data sets. Various types of reports were generated and based on the Analyst decisions are taken.

REFERENCES

[1] Hao Zhang, Gang Chen, Member, IEEE, Beng Chin Ooi, Fellow, IEEE, Kian-Lee Tan, Member, IEEE, and Meihui Zhang, Member, IEEE ,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 7, JULY 2015, In-Memory Big Data Management and Processing: A Survey.

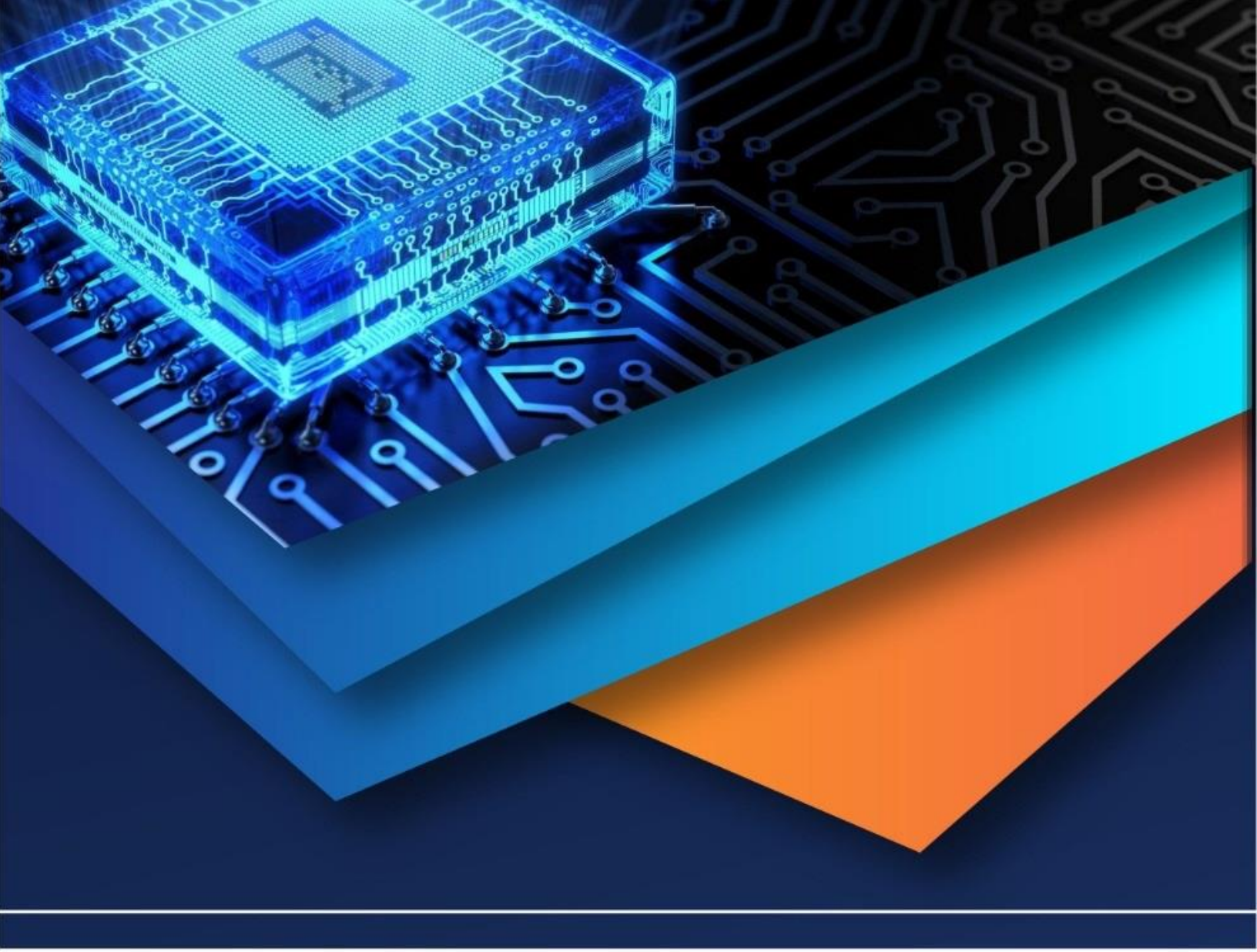
[2] Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrield, Stephen T. C. Wong, and Guang-Zhong Yang, Fellow, IEEEIEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4, JULY 2015, Big Data for Health.

[3] Marco Viceconti, Peter Hunter, and Rod Hose 10.1109/JBHI.2015.2406883, IEEE Journal of Biomedical and Health Informatics, Big data, big knowledge: big data for personalised healthcare.

[4] TECHNOLOGY AND TRENDS TO HANDLE BIG DATA: SURVEY Sunaina Sharma ME IT, Department of Information technology UIET, Panjab University Chandigarh, India E , .Veenu Mangat Professor, Department of Information technology UIET, Panjab University Chandigarh, India .

[5] Efficient Heart Disease Prediction System using Decision Tree Purushottam Research Scholar (PhD.), R.G.T.U. Bhopal (M.P), India. puru.mit2002@gmail.com Prof. (Dr.) Kanak Saxena, Prof & Head, Department of Computer Application, S.A.T.I. Vidisha (M.P), India. kanak.saxena@gmail.com Richa Sharma Assistant professor Amity University Uttar Pradesh Noida,India s.richa.sharma@gmail.com.

[6] Using Data Science & Big Data Analytics to Make Healthcare Green Nina S. Godbole, Green IT Professional, Certified Information Privacy Professional and Healthcare Researcher, Pune, India ninagodbole@yahoo.com John Lamb, PhD, Adjunct Faculty Mathematics Department, Pace University, Pleasantville, NY, USA jlamb@pace.edu.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY