iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** X **Month of publication:** October 2016

DOI:

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Big Data Analytics on Unemployment & Health Records

Kamparapu. Bhanu Rajesh Naidu[1], Battula. Lakshmana Rao[2]
[1]PG Scholar, Department of CSE, [2]Assistant Professor, Head of the Department CSE
Kakinada Institute of Technology & Science, Divili (Tirupathi), India

Abstract: ***Now a day's huge amount of data is produced from the social web sites, government industries, public sectors etc., Storing and Maintaining the data is big challenge which is solved by special databases and programming languages. Traditional data bases and programming languages are in sufficient to handle big data. Hadoop is a tool which is designed for storing and processing data in distributed way. In this paper, analysis of big data is done by using hive, R programming. The graphical representation of big data is done by suing Tableau software. In this, unemployment dataset analyzed by hive with built functions and health records data set is analyzed by statistical chi-square and fisher test .This analyzing makes a better understanding of data and which improves the prediction of diseases according to their habits***
Keywords: Bigdata, Hive, R programming, Chi-square, Unemployment Dataset, Health records

I. INTRODUCTION

Bigdata[1-2] is huge amount of data which is not stored in the single and traditional data base. Bigdata is analyzed by special processing techniques to predict decisions for business applications. Today, the world is surrounded by the data. People uploading videos, audios, images, text messages to the social web sites and machines also generating more and more data. The exponential growth of data which brings challenges to the users and business people in terms of storage and processing. Bigdata is mainly characterized by 3V's. Volume, velocity and variety. Volume refers to huge amount of data. Variety refers to different types of data. Velocity refers to the speed of processing data. The following are properties of big data.

A. Volume

The main characteristic of Bigdata[3] is volume. Currently the data size is in peta bytes and it is supposed to increase to zeta bytes in nearby future. The social websites like face book, twitter etc producing huge amount of data every day which difficult to be handled and processed by traditional data based and programming languages.

B. Variety

The data produced from the sources like websites, sensors, machines etc are not in the single format like structured it may be different formats like semi structured and structured data. Examples of semi structured are xml files, JSON objects. Examples of unstructured are video, audio etc.

C. Velocity

Velocity refers to the speed of data generated from various sources. This velocity is not being limited to the speed of incoming data but also speed at which the data flows. For example the data from the sensor devices would be constantly moving to the database store and this amount won't be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion. Bigdata is generated by various industries [4] like health care, sensors, machines etc.

II.HADOOP

A. Building Bloacks Of Hadoop

Hadoop[5] is an open source frame work for writing and executing applications in distributed approach. Hadoop have two important modules. Those are 1Hadoop distributed file system (HDFS) 2.Map Reduce(MR).HDFS is used for storing data in distributed

systems and map reduce is processing of big data. Hadoop cluster is a group of machines which are connected through the network in one location for storing data and processing data in distributed way. The hadoop cluster contains different machines comes under the following categories.

1) Name Node
2) Data Node
3) Secondary Name Node
4) Job Tracker
5) Task Tracker.

B. Hive

Hive is a data warehouse tool under Hadoop which handles huge amount of structured data by Hive query language. The Hive QL supports commands create, alter, drop, select, load data to manage data similar to sql.

III. R PROGRAMMING

R [6] is a programming language for a statistics and representation of graphics for data in easy way. R programming has number of features and built in functions for analysis purpose.

A. Features of R

R contains programming features similar to other languages like c, java etc. R contains conditional statements, loops, functions etc.,

R has efficient data handling and storage facility.
R contains different data types like arrays, vectors etc.
R provides different graphical representation functions.
R provides the following graphical representations of data
Pie charts
Bar charts

IV. TABLEAU

Tableau [7] is a data image tool with a primary focus on business intelligence. You can create maps, bar charts, scatter plots and more without the need for programming. They recently released a web connector that allows you to connect to a database or API thus giving you the ability to get live data in visualization. The following figure shows graphical representation of data in Tableau.

Fig 1: Graphical representation of data in Tableau

V. EXISTING SYSTEM

In existed system [8] wireless data set is analyzed using my hadoop and performed queries using hive joins. In existed only wireless data set is analyzed and reports are generated based on the cpu time and no of attributes. The sql queries written by using hive query language. In this only query analyzing is applied and does not apply any data mining and statistical algorithms.

VI. PROPOSED METHOD

In proposed, we analyzed the data by using Hive query language, R programming and Tableau software. Hive is a query language which is used for analyzing structured data in Hadoop environment. R is a programming language for statistics and graphical representation. Tableau is software which represents data in graphical format in easy way. The proposed scheme has considered two datasets. One is unemployment data set. From this data set, identifying which country has more unemployment [9], minimum employment, which year having more unemployment and what is the average etc. These details are identified by representing the data set in the graphic format by using Tableau and analyzed by hive. Another data set is cancer patient details. We represented these data by using statistical programming language called R programming which is having number of built in function to perform statistical tests on data set.chisq.test() function is used to perform the chi-square test on dataset and we evaluated the results in R Studio.

Chi-Square test is a statistical method to determine if two categorical variables have a significant correlation between them. Both those variables should be from same population and they should be categorical like − Yes/No, Male/Female, Red/Green etc.

For example, we can build a data set with observations on people's ice-cream buying pattern and try to correlate the gender of a person with the flavor of the ice-cream they prefer. If a correlation is found we can plan for appropriate stock of flavors by knowing the number of gender of people visiting.

A. Syntax

The function used for performing chi-Square test is chisq.test ().

The basic syntax for creating a chi-square test in R is −

chisq.test (data)

VII. EXPERIMENTAL RESULTS

The implementation of proposed scheme is done in three ways. We have applied hive query language; Hive is installed in Ubuntu virtual machine. Initially the virtual machine contains Hadoop and java then we installed hive. The second way of analysing data is done by using Tableau software. We installed software and imported data set from local file system to Tableau which represents data in various graphical formats. Third way is using R studio and programming language for analysing cancer patient data using chi square test. The following are evaluated results.

A. Queries Performed Using Hive Quer Language

Q1: hive>CREATE TABLE Unempset(Year BigInt,Jan Double,Feb Double,Mar Double,Apr Double,May Double,June Double ,July Double ,Aug Double ,Sep Double ,Oct Double ,Nov Double ,Dec Double)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

Q2: hive> LOAD DATA LOCAL
> INPATH '/home/lalitha2/Desktop/dd.txt'
> OVERWRITE INTO TABLE unempset;

Q3: SELECT * FROM unempset;

Q4: SELECT Year, Jan FROM unempset

Q5: SELECT max (jan) from unempset ;

The following screens shows creation of tables in hive graphical representation of unemployed dataset using Tableau and R programming.

Volume 4 Issue X, October 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

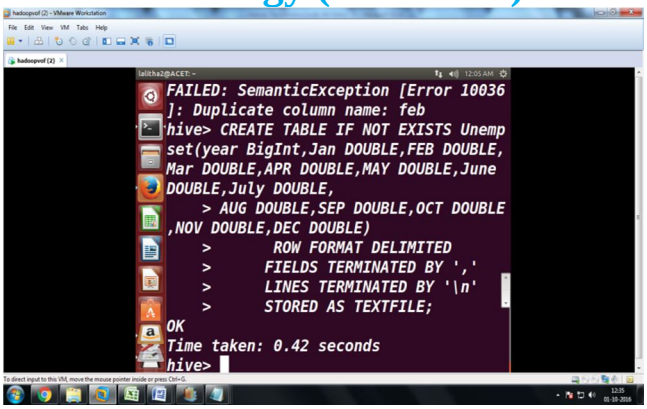


Fig 2:Creation of table in Hive

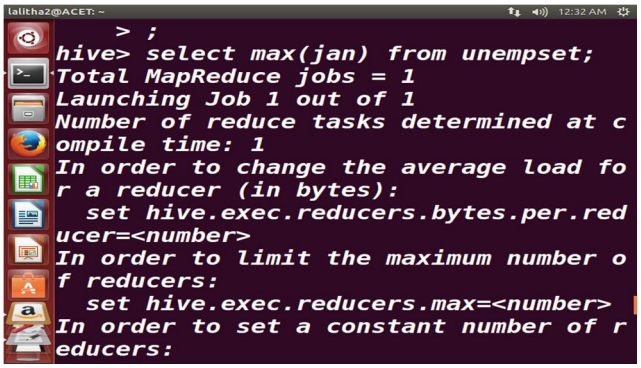


Fig 3:Finding max percentage of unemployment in the month Janauary

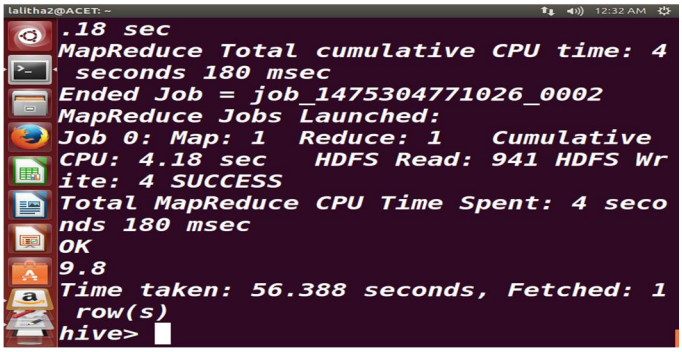


Fig 4:Displaying max percentage of unemployment in the month Janauary

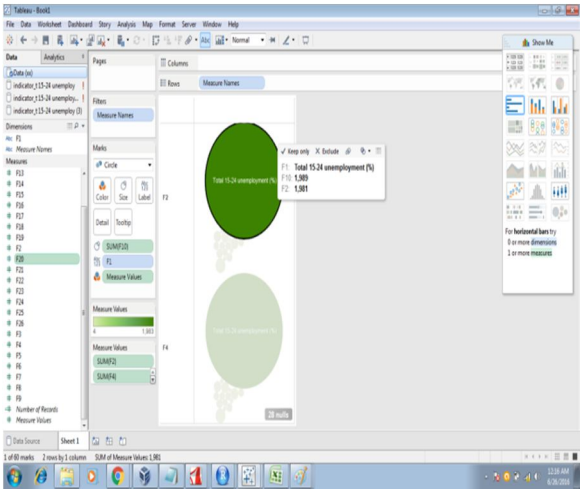


Fig 5: Representing unemployed data set in Bubble Format

Volume 4 Issue X, October 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

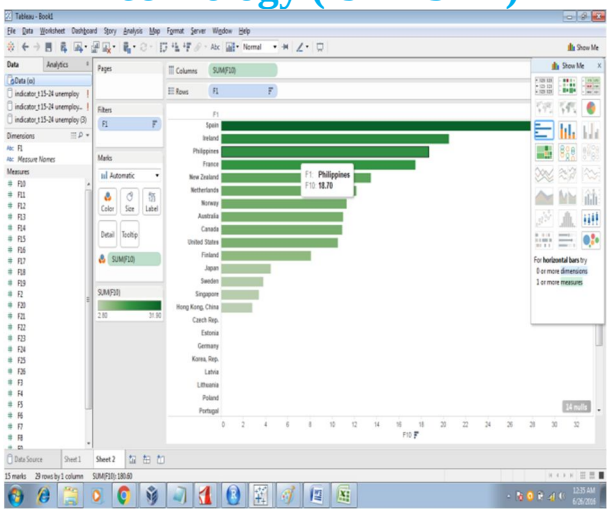


Fig 6: Representing highest unemployment in the year 1989 in the descending Order

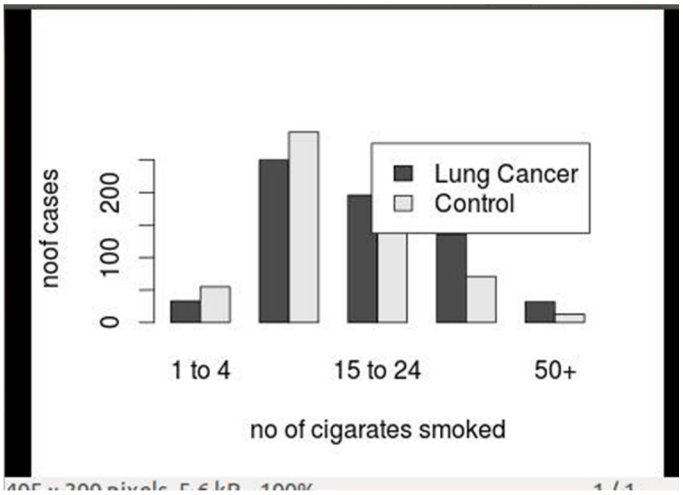


Fig 7: Graphical representation of cancer patient data in R programming

VIII. CONCLUSION

In this paper, presented basic building blocks of hadoop, R programming and Tableau software. In this unemployment data set is down loaded from the web and aggregation function is applied on unemployment data set by using Hive Query language. The graphical representation of data set is presented by using tableau software. Also analyzed cancer patient data set in R-studio with the help of built in functions which improves the prediction of diseases according to their habits. In this, we analyzed data by query approach, statistical approach and graphical view which make more easy understanding and analyzing of big data.

REFERENCES

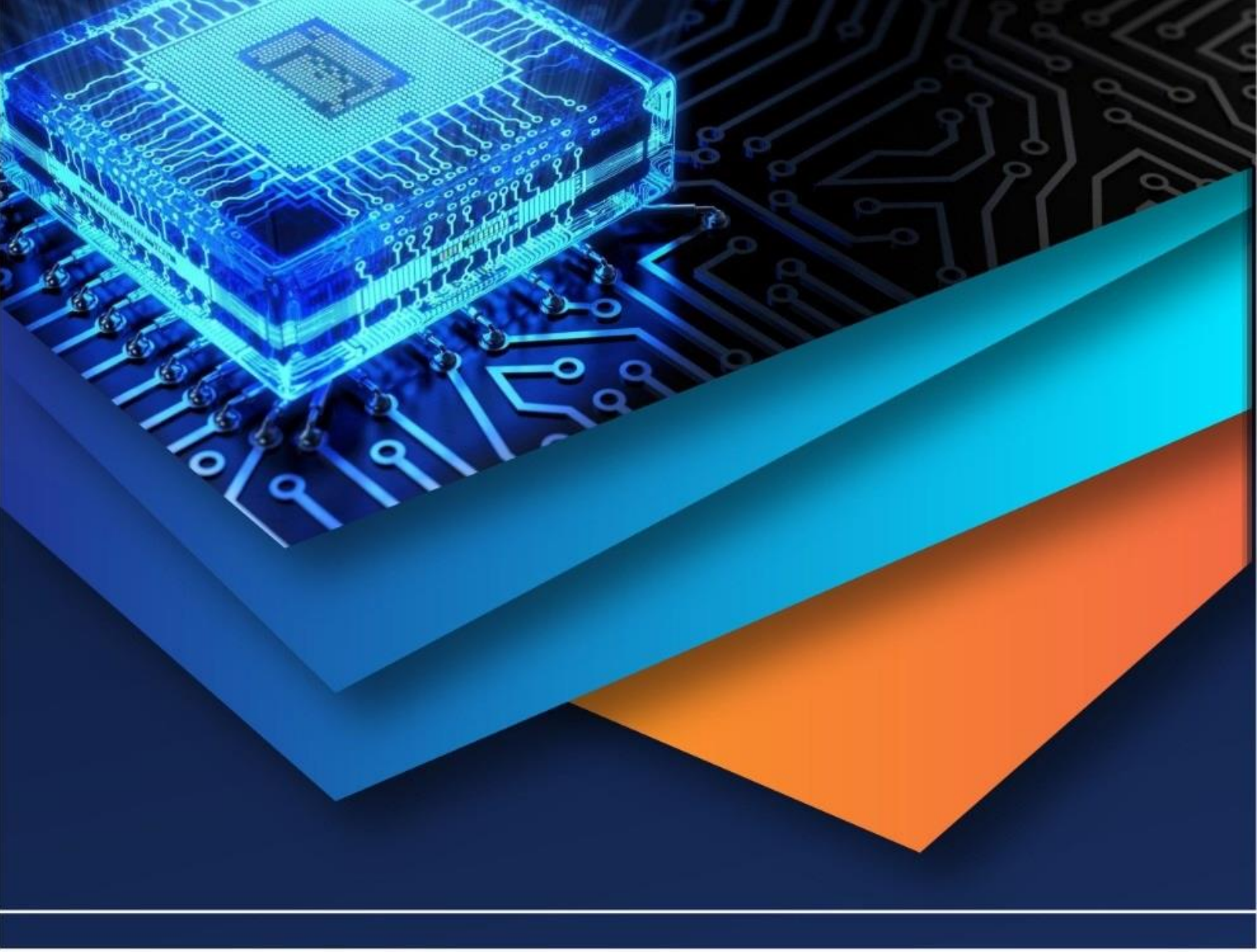
[1] Joonas Tuhkuri Etla, The Research Institute Of The Finnish Economy And The University Of Helsinki. Available: http://sites.uclouvain.be/aiece/password/BIG-DATA-ETLA-11-2015-PRESENTATION.pdf]
[2] Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang, Fellow, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4
[3] Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges Jasleen Kaur Bains Department of Computer Science and Applications, Panjab University(www.ijarcsse.com ISSN: 2277 128X)
[4] Chandan K. Reddy Department of Computer Science Wayne State University" Big Data Analytics for Healthcare."
[5] February 2015 Oracle Enterprise Architecture White Paper – Improving Healthcare Payer Performance with Big Data Author: Venu Mantha, Robert Stackowiak, and Art Licht.
[6] www.tableau.com
[7] Vibha Bhardwaj1, Rahul Johari2, Priti Bhardwaj Query Execution Evaluation in Wireless Network Using MyHadoopieeexplore.ieee.org/iel7/7321997/7359191/07359278.pdf

[8] Kim Seefeld, MS, M.Ed.* Ernst Linder, Ph.D., "Statistics Using R with Biological Examples," University of New Hampshire, Durham, NH Department of Mathematics & Statistics.
[9] Makoto Nakajima Federal Reserve Bank of Philadelphia "WORKING PAPER NO. 11-8 A Quantitative Analysis Of Unemployment Benefit Extensions", February 8, 2011.

AUTHOR'S BIOGRAPHIES

Mr. k. Bhanu Rajesh Naidu is a M.Tech student of Kakinada Institute of Technology & Science, Divili (Tirupathi). Presently pursuing M.Tech CSE from this college he is received his B.Tech degree from Kakinada Institute of Technology & Science, Divili (Tirupathi), Affiliated to JNTU Kakinada University in the year 2014. His area of interest includes Big Data Analytics, Web Technologies and Database Management systems, all current trends and technologies in Computer Science.

Mr. LakshmanaRao Battula, is received B.Tech CSE from Loyola engineering college Guntur and M.Tech CSE from Gayathri Vidya Parishad College of Engineering, Visakhapatnam, and working as Assistant Professor and Head of the CSE department, Kakinada Institute of Technology & Science, Divili (Tirupathi), Andhra Pradesh, India. His area of Interest includes Big Data Analytics, Computer Networks, Network Security and Cryptography.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY