ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

The Rising Flood of Big Data and Its Analysis

Kiran Padwar
Department of Information Technology,
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal, India

Dr. Mahesh Kumar Pawar
Associate Professor
Department of Information Technology,
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal, India

Dr. Ratish Agrawal
Associate professor
Department of Information Technology,
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal, India

Abstract: Today due to digitalization in almost all sectors the data value increases. Many organizations invest millions of rupees in analysis of data. The use of internet by normal citizens enhances data's value. Actually the consumption of internet is regularly increasing almost in every field. As a result a large massive complex data (big data) get produced per day per second. Actually it is not false to say that it is era of big data. Analysis of such huge data is become notable task performed by government and non-government organization. The main reason behind popularity of analysis of big data is extracting useful information .This paper mainly focuses on predictive analytics technologies. The coming time is time of predictive analytics as all sectors organization keenly watches predictive analytics result for business analytics and effective decision making and get tremendous success in their respective businesses.

Keywords: big data, predictive analytics, business analytics, decision making

I. INTRODAUCTION

In almost all field big data is a noteworthy term and is frequently used by all organization either it is government or non-government. Actually, due to digitalization in all sectors, data produced is very large. The data is in terabyte or petabyte or even more and complex too that it is impossible to analyze it by normal database i.e. RDBMS. If saying more precisely many of sites government or non-government organization, all produced exabytes of data or more per day per second. Like FB, Google, yahoo etc. The analysis of such data gives useful information. Such information is highly used in business analytics, decision making, e-commerce, e-governance, health care industry, botanical industry, disease diagnosis, customer choice etc. Actually data produced is static or dynamic. Static data concern to data generally produced by organization, they belong to some kind of report. The other one i.e. Dynamic data which belongs to data that is changing (generally increasing) day by day per second. In this category social networking sites belong.

II. BIG DATA

The term "data" concerns to any type of raw collection of report or information. It may be of many type tabular formed data, normal theoretical data, numerical data, photos, video or audio. In above statement raw means it is not preprocessed data that it needs filtering before analysis. The simple single word "data" is analyzed by RDBMS.As the use of internet increases worldwide in many fields, it produced high volume of data, with high rate. It is so large and complex that it is not able to analyze by normal DBMS. So, it is called "bigdata". Bigdata is very large and complex so, its analysis and management becomes biggest challenge. The analysis of big data is called data mining. It is a data engineering technology, and it means to extract meaningful or desirable information from a large dataset. The only source of production of big data is digitalization in almost all sectors. The only reason behind digitalization is internet. The electronic, gadgets, system, smartphone, sensor devices etc. Are producing massive data every day even every second and continuously increasing. It affects almost every sector like society, governance, business, health, behavior, treatment and many more. So it is not wrong to say that big data analysis and management is leading challenge in almost every organization belongs to analyzing sector. The toughest task is to analyze it i.e. Extracting knowledgeable and useful information from big data. This process of analyzing big data is called data mining. Data mining is emerging technology today and will be highly used technology in future.

A. Static bigdata:

It is that data which may be pre analyzed. This type of data gives high consistency for longer amount of time. For example reports, results, documentation etc. This type of data generally needs storage and analysis both. It is too large data set so it is not stored simply in one system. Here, we use some big data storage technology to store it like hadoop, cloud computing etc. For analysis of big data we use statistical and m/c learning methods to get desired result.

B. Dynamic bigdata:

This is a data which is generated in bulk per day per second. It is totally raw collection of data. The analysis of such data is more difficult than static bigdata as data consistency is changing time to time. It always needs filtering before analysis. Storage and analysis is done

through same way like static analysis. Which technology is used depends on type and need of data.

III. ANALYSIS & ANALYTICS

The term "analysis" refers to extracting meaning information from data. The term "analytics" little bit different from analysis, "analytics" refers to analysis of data with use of technology.

IV. ANALYSIS OF BIG DATA

Analysis of data is done by two ways, they are statistical analysis and machine learning, both have same stages but they are done in different ways. Analysis of big data goes through five stages. We discuss these one by one:-

Fig 1. Process of analysis of big data

A. Data loading:
Data loading is done directly through system or it can be loaded through net directly. The data may be of any type of file like csv, excel or img etc.[1]

B. Filtering:
The data which is imported through internet directly is raw, it needs cleaning, preprocessing before analysis. This process is called filtering.[1]

C. Association:
This stage concern to associate or integrate various data subsets into one set comes after filtering. [1]

D. Analysis:
This stage deals with getting meaningful or desirable result from extraction of dataset. There are various models and technological platform for doing analysis. The analysis models are predictive, perspective or descriptive. As according to data and desired result analysis models are used.[1]

E. Desired result:
In this stage, result is generated as according to need and desire.[1]

V. BIG DATA ANALYTICS PLATFORMS

There are various platforms for analysis some are free, some are business based (commercial) platform which are not free.
Free software platform:-
Dep
knive
octave
orange
R,Rstudio and more.

Commercial software platform:-

MATLAB
Analytica
IBM
SPSS
Model
Rapidminer etc

VI. ANALYSIS OF BIG DATA:- CHALLENGES

A. Volume:-
It concerns about increasing volume of data. Data produced is increasing per second per day, so it is too bulky to analyze by normal DBMS.[2]

B. Velocity:-
The data produced is increasing at very fast rate. So rate of increase in data is very fast.[2]

C. Vareity:-
Data are of many types as it is collected from various sources. It may be structured,, semi-structured or unstructured . Mostly difficulty arises during analysis of semi- structured and unstructured data i.e. Complex data.[2]

D. Variability:-
The data is consistent only for some duration of time, as data regularly changing according to time arise difficulty. As time goes on data gets updated so, this also arise difficulty during analysis.[2]

E. Value:-
Data importance or value changing. It means if some data is important than its value is high at that particular time, as time goes on that data value get changed.[2]

F. Complexity:
Data coming from multiple sources so creates a complex type data which is not easy to analyze. The raw data is not directly carry to link, clean, match and transfer data. So, raw data need filtering before analysis.

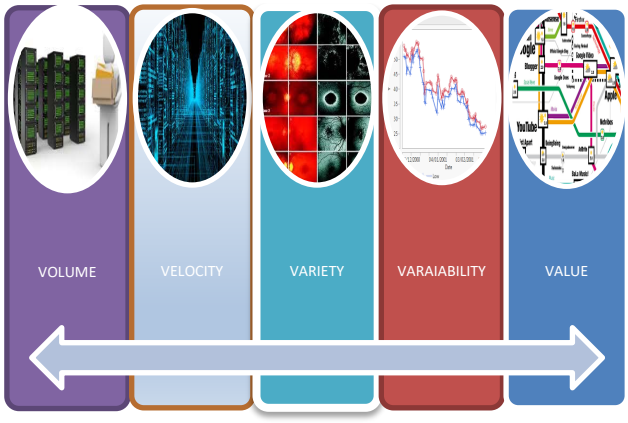

Fig 2. Major challenges of big data

VII. ANALYTICS OF BIG DATA

The term analytics refer to analysis of data by using technology. The analytics of big data divided into three categories based on generating different kind of result.

A. ***Descriptive analytics:-***It concern to presentation of data in meaningful or desirable format or fashion. It gives visualization of data. It represents historical data.[3]

B. ***Predictive analytics:-***It focuses on the predicting future based on the past trends of data. Predicting modeling is done by using statistical trends or machine learning. [3]

C. ***Perspective analytics:-*** It concern to efficiency and decision making. It provides an idea which concern to decision making in business or organization. It deals with business intelligence.[3]

VIII. DESCRIPTIVE ANALYTICS:-

This analytics deals with representation. It describes relation between various parameter. It represent or generate a summary of result. The represented data result may be quantitative or graphical. This concern to descriptive statistics.[4]

A. ***Quantitative :-***

This provides a conclusion either numerically or by plots showing individual parameter value comparing to other and difference between them. For example: summary, mean, median etc. Mean – sum of all values divided by number of values
Median – midpoint of bended set of value
Range – highest longest value
Variance – average squared deviation from mean.

B. ***Visualization****:-*

It deals with graph , which is easily understand by comman man. Example:- pie charts, barplots, scatterplots etc.

- Pie- charts:-representation of value as a slice of circle with same or different colors.
- Bar -plot :-representation of data in rectangular bars with length proportional to value of that variable.
- Box – plots :- it represent the distribution of data in dataset. Minimum, maximum and median, first and third quantile.
- Histogram :- it gives frequencies of value of any variable attached to respective ranges. It applied in continuous varying data.
- R – lineqraphs :- connection of series of points by drawing line segment between them ordered of any one coordinate. It is used to identify trends in data.
- Scatter plots :- it is the combination of many plots in one certain plane. Each point represent values of two variable (one in horizontal and another in vertical)

IX. PREDICTIVE ANALYTICS

It is highly applied technology. It is a technology whose time is come .it is applied by almost all sector of all types. It is highly used by health, social net making sites, online shopping site, manufacturing industry & above all government, stock exchange, behavior recognition. Actually it is a continuous process. It uses various statistical and machine learning techniques for prediction. Basically it uses current historical data to forecast future. It gives approximate result as prediction is based on historical fact which may changed in future.[4]

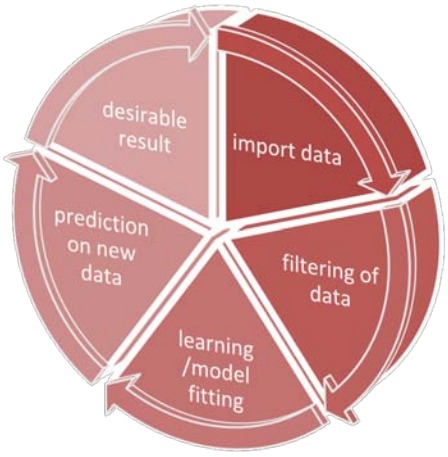

Fig 3. Predictive analytics process

A. Data import

Data input can be done by inputtingfile (data).Thedownloaded data or the data from system both can be imported .there are various types of files , we can import.

From csv
From RODBC
From excel

B. Filtering of data

a) Data treatments:- the goodness of any model depends on the data in which model in formed

b) Missing value :- when association is done result vary with high error due to missing values and also create problem during prediction
Solution:-
i. Case analysis
ii. Estimation method
iii. Synthetic distribution
iv. Replace missing values

b. ***Outlier treatment:*** - from this we have to know about extreme values i.e. low & high values of any variable .for this univariate analysis is used .
c. ***Skewness:-*** it is measure of symmetry i.e. distribution of data in symmetry , if it looks same in left & right part from center point of data , it can be copped according to business need.
d. ***Binning values:-*** sometime binning in variable is necessary to get some sense and easiness for exploring it. It mainly found in numeric data.
e. Breaking data in training & test sample:-
i. Training set :- data used to estimate model parameter.
ii. Test set :- this is validation of data set.
iii. Variable reduction:-
The main motive behind any modeling is to find out parameters that best predict as according to target variable, otherwise having many attributes that not too much contributes for prediction it, will increase implementation cost of model as well as complexity during modeling.
Collinearly measure:-
It explain that variant have similar information for explaining depended attribute

X. PREDICTIVE MODELLING:-

In this step prediction is done .here future is predicted based on past data. It is most important and crucial step during predictive analytics process. Here we create a relationship between various parameters and on basis of this we predict data for other values of variable. It not always give accurate calculation, but it shows approximate result. Here the relationship between parameters used for prediction, these parameters are called training samples or training parameter. The main techniques of prediction analytics are regression and classification. [5]

The main crucial point is to choose model for prediction. Actually every data have parameter relation in different way, so use that model which shows resemblance with data parameter relation. Here attention must be given to data structure, formal information about data and data attribute , what we want to know?

There are some basic steps which must be followed during predictive modeling:-[5]

A. ***Project Definition:*** in this step focus is mainly on what desirable result of any project. What is main aim of doing predictive analytics? So, according to that analysis process various steps schedule. Here, the general information about the analyzing data is also gathered and understand.

B. ***Exploration:*** here data is keenly watched so, that we get idea about which predictive model is suitable for such data.

C. ***Data Preparation***: here arranging and processing data as according to selected model. It is most time consuming step.

D. ***Model Building:*** in this step training datasets are generated and evaluated through other dataset. Here, we get an idea that generated result reaching goal or not?

E. ***Deployment:*** we apply the result in decision making and business intelligence.
F. ***Model Management:*** improve efficiency, power management; model management is done in this step.
There are three types of predictive modeling:-[6]

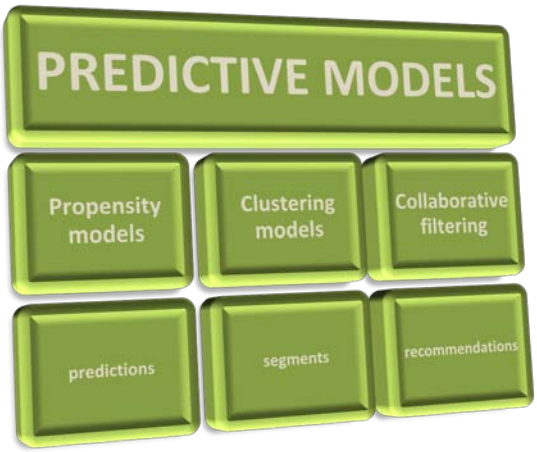

Fig4. Types of predictive modeling

A. ***Propensity models:-*** it deals with what most people think about prediction; it generally helps in prediction which concern to action of customer. Suppose prediction is on lady than what common people think about that lady. examples:-
1) Customer life time
2) Probability to buy
3) Probability to unsubscribe.
B. ***Clustering models:-*** it deals with segment of customer. It is similar to classification but here clustering or segmentation is done on that parameter which is not defined. It generally concerns to behavior prediction .EXAMPLES:-
1) Clustering on brand
2) Clustering on product
3) Clustering in behavior

C. ***Collaborative filtering:-*** it concerns for doing post process filtering. Here rethinking about predictive modeling is done.

XI. MODELLING TECHNIQUES:-

There are many techniques for modeling above models. The two main techniques are statistical modeling and machine learning. Both propensity and clusters models uses either of these two techniques.[6]

A. Statistical modeling:-

This technique concerns to form a relationship between various parameters, and form a mathematical relationship between them. Here mathematical equation is formed .the variable through which predicted variable is calculated called predictor variable (independent variable) and variable which is calculated is called response variable (dependent variable). By filling any value in predictor variable we get its predicted value.

B. Machine learning :-

It is originally developed to enable computer to learn. It is advanced statistical methods for regression & classification. It is applied in many fields like medical diagnosis, face & speed recognition, sometime we are able to predict regression variable without forming any relation between independent variables. Here, we form neural network .they are non linear modeling technique that are able to model complex function. It is generally. It is generally used when we not know about output. Here we give training to input and expected output and match whether calculated output and expected output are same or not if not, we calculate error. There are three types of learning:-

- ***Supervised learning:-*** here we provide desired output with example input , it provide guidance act a teacher. So that coming inputs generate output taking guidance.
- ***Unsupervised learning:-***here no guidance is provided. here data itself generate hidden patterns.
- ***Reinforcement learning:-***here guidance is not given internally rather it is provided externally through external environment.

There are various predictive models:-

- ***Linear regression :-***

  It is highly used statistical tool. Actually it, establish a relation between two attributes (variable. Here the response and predictor variables are fitted in linear line equation. In linear regression these two variable are connected or fitted in a linear line equation which y = ax + b.

  Where y=response
  x=predictor
  a& b =coefficients(constant)
  In R it done by lm ( ) function
  lm (relation of x & y {formula}, data).

- ***Multiple regression:-***

  It is similar to linear regression but it can do some more work. So it is a extension of linear regression. Actually, it can establish relationship between more than two variables in simple linear regression we have one response variable & one predictor, but here we have one response variable & more than one predictor.

  Y = a +b1x1 + b2 x 2 +---------- bn x n

  Y= a + bx + 4 bx+ --------nbnx

  Uses

  Regression method in done, when there exit some relationship between two or more attributes which can result some other senseful attribute or result.

- ***Logistic regression :-***

  In this response variable i.e dependent variable has categorical values like true/false or 0 / 1 etc. It concern to binary response as value of response variable based on relation with predictor.

  $Y = 1 / (+ e^{-(a+b1x12b2x+3b3x+---------)})$
  Y = response x = predictor , a &b = coefficient
  It is done by glm() function in R
  glm (formula , data ,family)
  Uses:-When predictor attribute is categorical ,

- ***Normal distribution:-***

  Generally, when we collect data from various sources, it is observed that distribution of the data is normal when we plot a graph. Here the term “normal” means formulation of bell shape curve. Actually, it means that when we plot a graph values of variables in horizontal axis& respective counts in vertical axis, when we found a bell shaping so, it should that distribution of datais in normal distribution, 50% in left & 50% in right side .
  R has some in built functions to generate normal distribution. They are
  dnorm(x, mean, sd)
  pnorm(x, mean, sd)
  qnorm(p, mean, sd)
  rnorm(n, mean, sd)

  here,
  x is a vector of numbers.
  p is a vector of probabilities.
  n is number of observations
  mean is the mean value of the sample data. Its default value is zero.
  sd is the standard deviation. It's default value is 1.

a) dnorm():- This function provides probability distribution height at every point corresponding to S.D. and mean.
b) pnorm() :_This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".
c) qnorm() takes probability value as input and give cumulative value as output which is corresponding to inputted probability value

d) rnorm():-This function is used to generate random numbers whose distribution is normal.

- ***Binomial distribution:-***

  It deals which the probability of success or failure of any event. Generally, it can to finally probability of any event having only two possible outcomes like tossing of coin etc.

- ***Poission regression:-***

  It deals with the regression in which the response variable is showing number of count. The response variable is not fraction numbers.
  Ex = no. of bias, no of death equation

- ***Analysis of covariance:-***

  It deals with the categorical variable.
  Ex = male / female, yes / no. if we use simple regression, it gives multiple result for each value of categorical variable, so we use analysis of covariance which is also called ANCOVA.

- ***Analysis of variance***

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences between group means and their associated procedures .in the ANOVA, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. ANOVA provides a statistical test of whether or not the means of several groups are equal.

- ***Time series analysis:-***

The term time series refers to the data in which each point in associated with time stamp. Actually each variable value is the value of that variable in a particular time. So, each variable value is different in different time, if there are many time series data then we are multiple time series.

- ***Non – linear least square:-***

Actually, almost real world data analyzed we observe modeling is done through regression we get a linear equation giving linear graph. Generally it involves mathematical function of higher degree. When we plot this we get a curve rather than a line. The main logic behind linear or non – linear regression in to adjust value of models parameter to find linear curve that is used to data that is to be analyzed. So we can calculate data regression variable with good accuracy in least square regression model in which sum of squares of vertical distance of different point from regression curve is minimized. Starting points with defined model & assume some values for coefficient. We apply nls ( ) function of R gets more accurate value along with confidence inequality

- ***Decision – tree:-***

It represents a graph like structure in which there are choices with respective result. So finally it from a tree structure so called decision tree. Here nodes of tree represent conditions. It is generally used to take decision on the basis of condition when some condition is fulfilled we go to that respective choice, so it is not false to say that when we have many choices& sub – choices we use decision tree.

- ***Random – forest :-***

It is a collection of large no. of decision tree. Every observation is analyzed through decision tree & most common choice becomes fined output. So, for conclusion we can say that decision is based on majority of voting between each decision tree in random forest.

- ***Survival analysis:-***

As the same suggest survival analysis concerns prediction of survival i.e. time. So, it is also called failure time analysis or analysis of time to death. It focuses on the remaining time for survival of any parameter chosen from data. It generally used for knowing the remaining life for any patient with danger or death giving decease like cancer ,finding life of any mechanical machine etc.

- ***Chi – square test:-***

It is statistical method to find relation between two categorical variables in data. These two categorical variables belong to same population.

- ***Naive bayes classifier:-***

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes is a popular method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features.

XII. PROPOSED WORK

As the use of predictive analytics in all fields increases, it also enhances interest to dip in predictive analytics. Some think that the process of predictive analytics is so keen process and always will be conducted with high concentration. But, if there is interest than it is totally an artistic kind of work. It need to understanding of parameters and forming a reliable relation between them. [7]

Suppose we have a data i.e. medical data showing liver patients reports. All parameters respective value for respective person defines the patients having liver disorder. The analysis of such data can provide many useful results not only for patients whether can be use for normal fit person.

The datasets defines liver patient laboratory report which is available online in:-
https://archive.ics.uci.edu/ml/datasets.html

The parameters are:-

- Age:- age of patients
- Gender:- gender of patient
- Total biliribin:- a kind of protein in human
- Direct biliribin (DB):- a subtype of total protein
- Alkaphose alkaline phosphotase:- found in high quantity children and pregnant women(very high).
- SGPT alamine aminotransferase:-it found when liver disorder occurs.
- SGOT aspartate amino transferase:- it also found when liver disorder occurs.
- Total protein:- gives total protein value.
- Albumin:- kind of protein , whose normal value gives indication that liver is healthy.
- A/4 ratio:- albumin and globulin ratio. This is high in leukemia and cancer patients.

The above parameters gives indication about human liver is healthy or not. This report can able to generate many kind of useful information through use of predictive and statistical analysis:-

- Which factors are highly responsible for liver disorder, so that awareness related to direct taking of such can be done?
- From which age human have to take care of liver?
- Which gender is highly chronic to liver disorder?
- What are symptoms during liver disorder?
- This data is present report, what will happen in future on the basis of this present data?
- What precautions should be taken?

Some more information can be extracted from above dataset. The analyzed result of such data will be published in next article referencing this article.

XIII. CONCLUSION

As data is increasing day by day because of digitalization increases so data analysis of such huge amount of data becomes biggest challenge. Data mining techniques are highly used technology as it covers all sectors of all kinds of industry and government and non- government organization. It will be highly used technology in future. The predictive analytics is most interesting and crucial to solve. As due to software platform many works can be done in one command. As it is era of big data so analytics of big data may be highly attention paid topic by all business and governance. The predictive analytics will highly used in future .the whole data mining converges in predictive analytics. It can able to generate very useful information from which we will able to take best decision and also able to use it for safety of worldwide .As all humans have high curiosity to know about future. So, predictive analytics is very interesting too.

XIV. ACKNOWLEDGEMENTS

My special thanks to my mentors Dr. Mahesh Kumar Panwar, Associate prof., Dept of IT, RGPV Bhopal and Dr. Ratish Agrawal, Associate prof, Dept of IT, RGPV Bhopal, for their regular support and guidance.

XV. REFERENCES

[1] Thomas A. Runkler "Data Analytics Models and Algorithms for Intelligent Data Analysis"Springer Vieweg, pp 21 august 2012.
[2] RobPeglar :-Introduction Analytics Big Data Hadoop, SNIA education
[3] Hitesh Goyal, Surender Singh " Big Data Analysis Using R (Big Data Analysis Applications, Challenges, Techniques)"Volume 5, Issue 9, September 2015 ,International Journal of Advanced Research in Computer Science and Software Engineering.
[4] Aditi Jain Manju Kaushik "Performance Optimization in Big Data Predictive Analytics"Volume 4, Issue 8, August 2014 , International Journal of Advanced Research in Computer Science and Software Engineering.
[5] Predictive Analytics: A Survey, Trends, Applications, Oppurtunities& Challenges by nishchol Mishra, Dr.SanjaySilakari. International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012,
[6] Predictive Analytics using RbyJeffrey S. StricklandSimulation Educators ,Colorado Springs.
[7] www.sciencedirect.com " Predictive Methodology for Diabetic Data Analysis in Big Data" by drSaravanakumar N M,2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

