International Journal on Applications in Information and Communication Engineering
Volume 2: Issue 5: May 2016, pp 9-13. www.aetsjournal.com ISSN (Online) : 2394 -6237

Data Visualization & Applications of Big Data in various sectors

Soumya Jha, Nikhil Kumar Agrawal

Abstract— Big data is a term for massive data sets having large, varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. Analysis of massive amount of raw data by visualization of the data, for decisions to be made on the basis of data management. These useful information for companies or organization with the help of gaining richer and deeper insights and getting an advantage over huge data sets. The paper presents an overview of big data's management and visualization, its scope in various sectors, samples, methods, advantages and challenges according to pilot study.

Index Terms— Data Munging, Data Visualization, Variety, Velocity, Veracity, Volume, Statistical Analysis System, Machine learning.

I. INTRODUCTION

Big data generates value from very large data set which can't be analyzed with traditional computing techniques. Quantity of computed data generated on planet Earth is growing exponentially for many reasons. This digital trace (data) we can use and analyze. Big data refers to our ability to make use of ever increasing data.

Characteristics of Big Data can be: Volume, Velocity, Variety, & Veracity. Volume refers scale of data. Velocity determines analysis of stream of data. Variety is different forms of data. Veracity being uncertainty of data.

Data Munging or data wrangling is loosely the process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data with the help of semi-automated tools.

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

II. VISUALIZATION OF DATA

Improvements in data visualization: just because you have a great data set with an awesome DB architecture doesn't mean your users are going to benefit from your application unless they understand the significance of what they are seeing. With the advent of touch sensitive navigation, interactive visualization technologies and these are being taken to another level. Whether the user is a data analyst or a stay at home mom, the ability to understand and act on data is going to be democratized with these new visualization tools.

The various programming languages serving data visualization are *R programming, Python, SAS and also other Hadoop based languages.* Data needs to be visualized from various raw datasets. Data set described here includes variables like income per person, oil consumption, co2 emission, life expectancy, employment rate etc. of 214 countries.

Countries	incomepe	co2emissi	lifeexpect	employrt
CHINA	2425.471	1.01E+11	73.45	72.8
GERMANY	25306.19	4.12E+10	80.414	53.5
INDIA	786.7	3.04E+10	65.438	55.4
JAPAN	39309.48	4.61E+10	83.394	57.3
USA	37491.18	3.34E+11	78.531	62.3
ZIMBABWE	320.7719	5.90E+08	51.384	66.8
UK	28033.49	7.25E+10	80.17	59.3

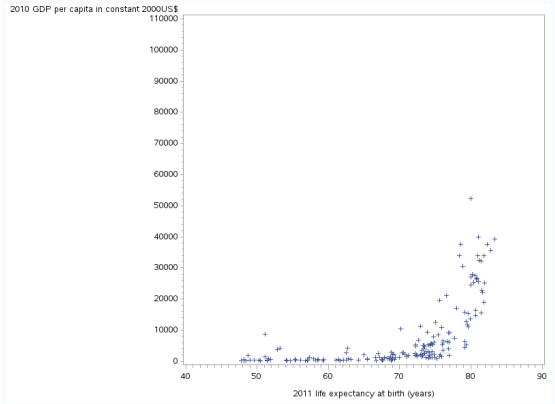
This is the abstract raw data of few countries. [1]

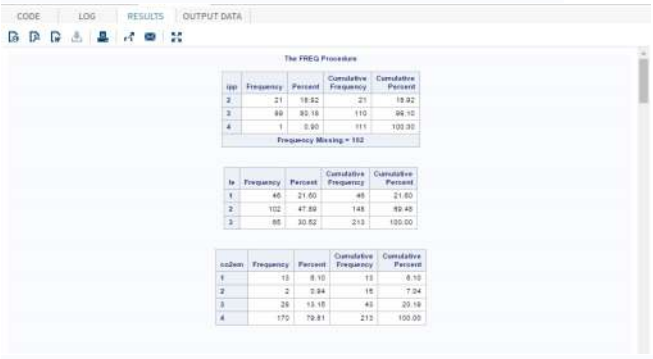
Data Munging: Out of R, Python, SAS, etc., SAS (Statistical Analysis System) is used here in accordance with the chosen data set.

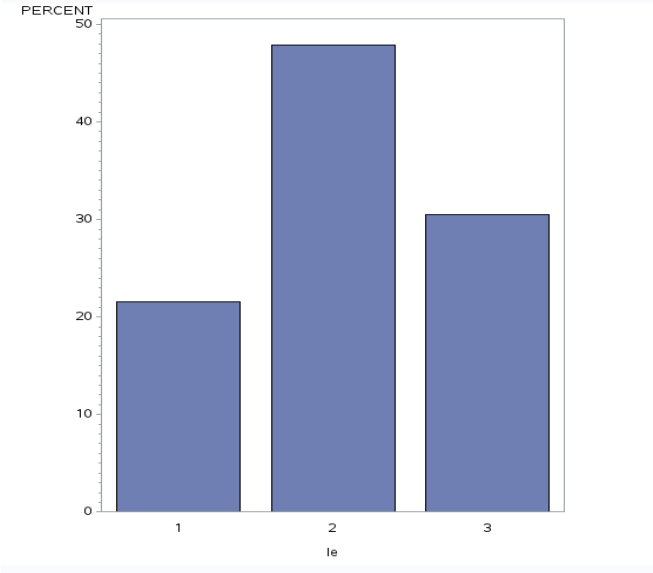
```
LIBNAME mydata "/courses/d1406aa6ba17fe300" access=readonly;
DATA NEW; SET mydata.gapminder;
label incomeperperson = "2010 GDP per capita in constant 2000US$"
      co2emissions = "2006 cumulative CO2 emission (metric tons)"
      lifeexpectancy = "2011 life expectancy at birth (years)";
keep country incomeperperson lifeexpectancy co2emissions ipp le co2em;

if incomeperperson < 500 then le = "1";
if incomeperperson >= 500 and incomeperperson < 1000 then ipp = "2";
if incomeperperson >= 1000 and incomeperperson < 10000 then ipp = "3";
if incomeperperson >= 100000 then ipp = "4";

if lifeexpectancy < 55 then le = "1";
if lifeexpectancy >= 55 and lifeexpectancy < 75 then le = "2";
if lifeexpectancy >= 75 and lifeexpectancy < 90 then le = "3";
if lifeexpectancy >= 90 then le = "4";


if co2emissions < 100000 then co2em = "1";
if co2emissions >= 100000 and co2emissions < 1000000 then co2em = "2";
if co2emissions >= 1000000 and co2emissions < 10000000 then co2em = "3";
if co2emissions >= 10000000 then co2em = "4";
proc freq; tables ipp le co2em;
RUN;

/*UNIVARIATE GRAPH*/
proc gchart; vbar ipp/ Discrete type = PCT WIDTH=15;
RUN;

proc gchart; vbar le/ Discrete type = PCT WIDTH=15;
RUN;


proc gchart; vbar co2em/ Discrete type = PCT WIDTH=15;
RUN;

/*BIVARIATE GRAPH */
PROC GPLOT; PLOT incomeperperson*lifeexpectancy;
run;
```

Soumya Jha, (Email: 55soumya.j@gmail.com) Nikhil Kumar Agrawal . (Email: nikhilanant21@gmail.com)

PROC FREQ gives frequency distribution table with respect to various constraints. PROC GCHART, for graphical construction for certain number of variables. The independent, or explanatory variable, is plotted on the X axis. The dependent, or response variable, is plotted on the Y axis.







III. BIG DATA APPLICATIONS

Big Data is an emerging field of technology and is growing by leaps and bounds. There are numerous fields in which research can be done when it comes to Big Data. [4] Here is a list:

1. Retail Sector
2. Banking
3. Health Sector
4. Finance
5. Telecom
6. Digital Media Solutions
7. Machine Learning
8. Sentiment Analysis etc.

A. Big Data for Agriculture

To meet the demand of growing population it is estimated that the food production should increase by 60% according to Food and Agriculture Organization United Kingdom. Changes need to be implied in not just the way of farming but also understanding farming. A large and complex set of data which needs to be analyzed through applications is called Big Data. It is believed that 90% of the world's data came into existence in the past 2 years. The principle behind big data is that more and better information enhances competitiveness and better decision making. Big data can help farmers decide which crops to plant where and when.

Consider the example of farm equipment that can take soil samples in real time, directly perform the relevant analysis and feed the results to a large database stored in the cloud. Combined with weather predictions, these results can be used to make precise adjustments to nitrogen applications.

If more farmers contribute to more production data about crops it will help them better understand which crop to plant, when to sell, available pesticides, weather conditions etc. In every supply chain food is wasted till the point of consumption. Let us assume this wastage to be around 30%. If we aim at getting this number down even by a smaller value, we can improve the quality in food supply chains. In 2013 a social media study was performed and with the help of Big Data the food inflation prices could easily by estimated which is nearly impossible with traditional ways of data collection.

Big Data Farming is also known as Precision Farming and is believed to play an important role in the near future helping farmers to produce more supply of food. Farmers can get easy access to data with the help of mobile phones. They would have automated irrigation, real-time optimization of farming machinery, monitor gain prices in the market etc. United States already operate cloud-based farming o has the potential to increase the yield production. Farmers can easily predict the real time cost of grains in the local markets around them and automatically compute the transportation costs. Precision farming has already been started in USA and is gaining is foothold in other countries.

B. Big Data for Medical Science

Now that data scientists are supplementing doctors, it's not so far that they might be put in prime positions. Time and again, when doctors mandate a treatment, whether its surgery or an over-the-counter medication, they are using a typical treatment or some variation that is centered on their own intuition, hoping for the best. The sad reality of medicine is that we don't really understand the relationship between treatments and outcomes.

We have studies to show that various treatments will work more often than try-ons; but, we know that much of our medicine doesn't work for half or our patients, we just don't know which half. At least, not in advance. One of data science's many promises is that, if we can collect data about medical treatments and use that data effectively, we'll be able to predict more accurately which treatments will be effective for which patient, and which treatments won't.

We believe that data science has the potential to revolutionize health care. There are big changes happening in healthcare right now, and the implementation of EHR (electronic health records) in particular is a great example of how data scientists will be working with doctors in the future. All of these electronic patient records spell out Big Data for the healthcare fields, and data scientists — like all quantitative folks – love data. These medical data could not only offer tremendous insights that change the face of modern medicine, but also offer rewarding opportunities to the data scientists who must decipher the data. Patient care also stands to receive enormous benefits from data science. While a doctor may be trained to look for many factors when diagnosing an ailment, some of these diseases are impossibly complex, and patients could stand to gain faster, safer treatment if left in the hands of a well-developed machine, or even a physician aided by one.

Two factors lie behind this new approach to medicine: a different way of using data, and the availability of new kinds of data. It's not just stating that the drug is effective on most patients, based on trials (indeed, 80% is an enviable success rate); it's using artificial intelligence techniques to divide the patients into groups and then decide the difference between those groups. We're not asking whether the drug is effective; we're asking a fundamentally different question: "for which patients is this drug effective?" We're asking about the patients, not just the treatments. A drug that's only effective on 1% of patients might be very valuable if we can tell who that 1% is, though it would indeed be rejected by any traditional clinical trial.

McKinsey & Company compiled a report for the Center for US Health System reform which identified four main sources of big data in the healthcare industry.

Activity (claims) and cost data: These are the basic figures showing the amount of care which has been supplied by providers in the system, and the cost of paying for that care. Analysis of this tells us about the spread of diseases, and the priority that should be given to dealing with specific health threats. The most cost-effective treatments for specific ailments can be identified and the number of duplicate or unnecessary treatments can be significantly reduced.

Clinical data: These include patient medical records and images gathered during examinations or procedures, as well as doctors' notes.

Pharmaceutical R&D data: Over the last few years a large number of partnerships have sprung up between pharmaceutical companies – becoming aware of the huge benefits of pooling their knowledge.

Patient behavior and sentiment data : This is data from over-the-counter drug sales combined with the latest "wearables" which monitor your activity and heart rates, patient experience and customer satisfaction surveys as well as the vast amount of unstructured information about our lifestyles broadcast every day over social media.*What*

Personalized Medicine: One of the top goals is to create a personalized treatment plan based on individual biology. Instead of treating your patient with a drug that works 80% of the time (e.g., the breast cancer drug, Tamoxifen), you can employ data science to custom-tailor a regimen just for her.

Genomics: Inexpensive DNA sequencing and next-generation genomic technologies are changing the way health care providers do business.

Self-Motivated Care: It's a "patient heals thyself" world, now. Developments like personal genetic testing (e.g., 23andMe.com), online patient networks and behavioral apps like Be Well are allowing individuals to take control of their own health.

Disease Modeling and Mapping: One of the flashiest uses of data science in the past few years has been in tracking (and finding ways to halt or prevent) diseases.

"*Prevention is better than cure*" has led to a focus on predicting problems in the early stages when they are easier to treat, and outbreaks can be more easily contained.

In the future we are likely to recover more quickly from illness and injury, and we will live longer. New drugs will come into existence and our hospitals and surgeries will operate more efficiently – all thanks to **BIG data**.

C. Big Data's Impact on Finance Industries

Banks have been storing enormous amounts of data since ages and this valuable data has helped understand customer

behavior and helped prevent thefts. Banks in the United States have already been using this enormous amount of data in various spheres such as sentiment analysis, risk management, product cross selling etc. Big Data can be applied to banking firms in areas of such as analyzing the spending patterns of the users, the possible loans a user would take in the future, fraud analysis, customer segmentation, product cross selling, customer segmentation, sentiment analysis etc.

The major areas of work in a Banking firm where Big Data can have a drastic impact are:

1. Customer Centric
2. Risk Management
3. Transactions

Customer Centric: Big Data can be applied to measure the client feedback, strategy planning, decide the next best offer , sentiment analysis, analyze the customer life events,measuring the quality of leads,etc.

Risk Management: The ways in which data analysis is being used to find out and evaluate financial crime management (FCM) solution rules, by early detection of the correlation between financial crime and attributes of the transaction, or series of transactions are MIS reporting, real time keyboard conversations etc.

Transaction Analysis: Transactions tend to reveal a lot about the nature of trade, log analysis etc. The ways in which this can be done are log analytics, B2B merchant insights etc.

With the help of Big Data analytics banks can perform better Risk Management, Transactional Analysis and also get Customer Centric data faster.

D. Big Data in Aviation

Advances in technology have increased our ability to collect, store and analyze data. Big data emerged due to the following three major trends. First, it has become easier to generate data due to smart devices, Internet of things, sensors etc and all this being stored at a minimal cost. Second, it is easier to process this huge amount of data due to cloud computing, multi core CPUs etc. Thirdly, many people have ways and access to this data and use it for valuable decision making. The most popular definition of Big Data can be defined as volume, velocity and variety of data. Volume refers to the size of data sets and storage, velocity is the speed of incoming data and variety is the data types. Business has always wanted to derive insights from Big Data for faster and better decision making. The aviation industry deals with huge amounts of data and many airports cannot manage the amount of data they receive. Data from various sources such as passenger flow, weather conditions, sensors, cost reduction, departure and arrival timings, services and feedback, revenue enhancement etc.From a recent study conducted by a leading software company, big data analytics has become the highest priority for aviation (61%) followed by wind (45%) and manufacturing (42%) companies.

There are about 35 million flight departures per year and it is very important for decision making by airports and airlines. Data collection could be effectively used by the airline websites with respect to transactional data sets and booking. A study shows that there are greater than $140 million ticket transactions made through airline ticketing websites. The current look to book ratio stands to 10:1 which says that a person would look up to 10 websites before booking an airline ticket. However, airlines do not track this data. They record only their transactional data which leads to them missing out on possible marketing strategies. Tracking the IP address, time and fares offered by other websites would allow the airlines a greater insight into its customer needs. The simplest example can be route development where airlines can see if a customer searched for a route it does not offer.

Over the past two decades the rise of a new industry too place whose main asset is data. The use of this vast amount of data occurs in Internet-based industries. A study by FAA states that during a year a jet engine generates data equivalent to 20TB. Most of this data is not used for any analytics purpose since this data is unstructured .Big data analytics can be used to predict the fault in the component by analysing data

obtained from various sensors. Big data can help operations for airline companies and airports to reduce redundant variability. Using the data, airlines can offer personalized incentive for every type of customer resulting in more auxiliary sales and greater percentages of repeat business.

REFERENCES

[1] www.gapminder.org
[2] www.computerworld.com
[3] www.csc.com
[4] www.forbes.com