ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Data Analytics Types, Tools and their Comparison

Sofiya Mujawar[1], Aishwarya Joshi[2]

Final year student, Computer Engineering, MKSSS' Cummins college of Engineering for Women, Pune, India[1,2]

Abstract: In this paper, we have presented an overview of the concepts of big data and big data analytics. Furthermore the various types of data analytics have been discussed and popular tools being currently used for data analytics have been comprehensively explained. Lastly we have presented a tabular view of the comparison of the tools on numerous parameters.

Keywords: Big data, Data analytics, Analytics tools.

I. INTRODUCTION

Data is expanding exponentially day by day. By 2020, experts say that there will be 4300% of annual data increase. Hence it's not only the size of big data that makes it unique but also its unstructured form that can cause serious issues for handling it. With the way data has been expanding every minute, new technique and analysis tools have been made to handle them. These tools analyse large data sets simultaneously and storage on cloud on secure data centres has made their analysis easy and on the go. Hence, big data is not only unique in its size and form but also in its processing and knowledge discovery. Big data in petabytes are analysed quickly and give more accurate interpretation of respective queries than ever before. Most raw data, especially big data, are not suitable for human consumption, but the information we derived from the data is. Big data is one of the misunderstood (and misused) terms in today's market. But with a clearer understanding the how to apply big data to business intelligence (BI), you can help customers navigate the ins and outs of big data, including how to get the most from their big data analytics. Hence efficient tools for analysing data has become the need of the hour.

II. WHAT IS DATA ANALYTICS?

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories. [2] In any big data setup, the first step is to capture lots of digital information analytic excellence leads to better decisions. The majority of raw data, particularly big data, doesn't offer a lot of value in its unprocessed state. Thus, by applying the right set of tools, we can pull powerful insights from this stockpile of bits. It is not about more data. It is about deeper look.

III. TYPES OF DATA ANALYTICS

There are four types of big data BI that really aid business:

A. Prescriptive:

Prescriptive analytics is a type of predictive analytics, it's basically when we need to prescribe an action, so the business decision-maker can take this information and act. Also it does not predict one possible future, but rather multiple futures. The various types of methods that analyse current and historical facts to make predictions about future events. In essence, to use the data on some objects to predict values for another object. The models predicts, but it does not mean that the independent variables cause. Accurate prediction depends heavily on measuring the right variables. Although there are better and worse prediction models, more data and a simple model works really well. Prediction is very hard, especially about the future references. Predictive analytics is the next step up in data reduction. It utilizes a variety of statistical, modelling, data mining, and machine learning techniques to study recent and historical data, thereby allowing analysts to make predictions about the future. The purpose of predictive analytics is not to tell you what will happen in the future, it cannot do that. It can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature. The input to the model is plain text and the output of that model is a sentiment score, whether it's positive, negative, or something between +1 or -1. The emerging technology of prescriptive analytics goes beyond descriptive and predictive models by recommending one or more courses of action and showing the likely outcome of each decision.

In addition, prescriptive analytics requires a predictive model with two additional components: actionable data and a feedback system that tracks the outcome produced by the action taken. Since a prescriptive model is able to predict the possible consequences based on different choice of action, it can also recommend the best course of action for any pre-specified outcome.

B. Predictive:

The extensive use of data and mathematical techniques to uncover explanatory and predictive models of business performance representing the inherit relationship between data inputs and outputs. Predictive analytics uses the understanding of the past to make "predictions" about the future. Predictive analytics is applied both in real-time to affect the operational process (ex: real-time retention actions via chat messages or real-time identification of

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

suspicious transactions) or in batch (target new customers on Web site or direct mail to drive cross-sell/up-sell, predict churn etc.). These predictions are made by examining data about the past, detecting patterns or relationships in this data and then extrapolating these relationships forward in time. For example, a particular type of insurance claim that falls into a category (pattern) that has proven troublesome in the past might be flagged for closer investigation.

Predictive modeling techniques can also be used to examine data to evaluate hypotheses. If each data point (or observation) is comprised of multiple attributes, then it may be useful to understand whether some combinations of a subset of attributes are predictive of a combination of other attributes. For example, one may examine insurance claims in order to validate the hypothesis that age, gender and zip code can predict the likelihood of an auto insurance claim. Predictive modeling tools can aid in both validating and generating hypotheses. This is particularly useful when some of the attributes are actions determined by the business decision-makers.

C. Descriptive:

It is the simplest class of analytics, one that allows you to condense big data into smaller, more useful nuggets of information. Most raw data, especially big data, are not suitable for human consumption, but the information we derived from the data is. Estimated that more than 80% of business analytics most notably social analytics are descriptive. Descriptive. The discipline of quantitatively describing the main features of a collection of data. In essence, it describes a set of data. Typically the first kind of data analysis performed on a data set. Commonly applied to large volumes of data, such as census data. The description and interpretation processes are different steps. Univariate and Bivariate are two types of statistical descriptive analyses.

Descriptive analytics may begin by providing a static view of the past, but as more instances are accumulated in the data sources that document past experience, the steps of evaluation, classification and categorization can be performed repetitively by fast algorithms, endowing the overall work process with a measure of adaptability. As descriptive analytics reach the stage where they support anticipatory action, a threshold is passed into the domain of predictive analytics. Predictive analysis applies advanced techniques to examine scenarios and helps to detect hidden patterns in large quantities of data in order to project future events. It uses techniques that segment and group data (transaction, individuals, events, etc.) into coherent sets in order to predict behavior and detect trends. It utilizes techniques such as clustering, expert rules, decision trees and neural networks.

IV. DATA ANALYTICS TOOLS

Data analysis is used in different domains like science, business, and social science. With the increasing need of data analysis some tools that directly analyse the data and derive conclusions are in demand in the market. Data analysis tools use many types of analysis techniques to store, manipulate and find meaningful inference from provided data sets. Some tools also generate reports to summarize the conclusion and provide better visualization. Data analysis tools help in deriving accurate results with minimum efforts. Now we are going to see some of the top tools used for data analysis in different business domains. These tools can be used right from a beginner to an expert who may or may not be from a technical background. We will consider six tools which make analysing data sets, visualization and presentation of data easy and accurate.

D. Data Wrangler

Wrangler is an interactive tool for data cleaning and transformation. [4] Wrangler is designed to quicken the process of data manipulation, helps u spend less time transforming your data and more time learning from it. Wrangler allows interactive transformation of messy, real-world data into the data tables which analysis tools expect which in turns allows spending less time formatting and more time analysing your data. It is a Web-based service from Stanford University's Visualization Group which is designed for cleaning and rearranging data so it's in a form that other tools such as a spreadsheet app can use. Click on a row or column, and Data Wrangler will suggest changes.

For example, if you click on a blank row, several suggestions pop up such as "delete row" or "delete empty rows." Text editing is especially easy. For example, when I selected "Alabama" in one row of sample data headlined "Reported crime in Alabama" and then selected "Alaska" in the next group of data, it led to a suggestion to extract every state name. Hover your mouse over a suggestion, and you can see affected rows highlighted in red.

Limitations: Not all suggestions are useful**.** And while the fact that Data Wrangler is a Web-based service makes it convenient to use, don't forget that it sends your data off to an external site -- which means it isn't an option for sensitive internal information. [4]

Skill level: Advanced beginner.

Runs on: Any Web browser.

E. The R Project:

R is a free software environment for statistical computing and graphics. R is a general statistical analysis platform that runs on the command line.[4] R can find means, medians, standard deviations, correlations and much more, including linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing.

R also graphs, charts and plots results. There are numerous add-ons to this open-source project that significantly extend functionality. [4]

There is excellent functionality in R, including quite a number of visualization options as well as numerical and spatial analysis.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Limitations: The fact that R runs on the command line means that users will have to take the time to learn which commands do what, and not all users will be comfortable with a text-only interface. Those dealing with large data sets may hit a memory barrier [4]

Skill level: Intermediate to advance. Comfort with command-line prompts and a knowledge of statistics are a musts for the core application.

Runs on: Linux, Mac OS X, Unix, Windows XP or later.

F. Time Flow:

This is desktop software for analyzing the time attribute. TimeFlow can generate visual timelines from text files, with entries color- and size-coded for easy pattern spotting. It also allows the information to be sorted and filtered, and it gives some statistical summaries of the data. [4]

TimeFlow makes it incredibly easy to interact with data in various ways, such as switching views or filtering by criteria such as date ranges or earthquakes of magnitude 8 or more. The timeline view offers a slider so you can zero in on a time period. While many applications can plot bar graphs, fewer also offer calendar views. TimeFlow is a desktop application that makes it quick and painless to edit individual entries.

Limitations: There are no facilities for publishing or sharing results other than taking a screen snapshot, and additional development appears unlikely in the near future. [4]

Skill level: Beginner.

Runs on: Desktop systems running Java 1.6, including Windows and Mac OS X.

G. NodeXL:

NodeXL is a visualization and analysis software of networks and relationships. It uses a technology referring to the discipline of finding connections between people based on various data sets. [5]

This Excel plug-in displays network graphs from a given list of connections, helping you analyze and see patterns and relationships in the data.
NodeXL merges the older and current definitions of Software Network Analysis. Its "optimized for analyzing online social media -- it includes built-in connections to query the APIs of Twitter, Flickr and YouTube, allowing you to draw networks of users and their activity. It also handles e-mail and conventional network analysis files. [4]
Runs on: Excel 2007 and 2010 on Windows.

H. Tableau Plateau:

Data visualization tools allow anyone to organize and present information intuitively. It is exceptionally powerful in business because it communicates insights through data visualization. [5] This tool can turn data into any number of visualizations, from simple to complex. You can drag and drop fields onto the work area and ask the software to suggest a visualization type, then customize everything from labels and tool tips to size, interactive filters and legend display. [4]
Tableau Public offers a variety of ways to display interactive data. You can combine multiple connected visualizations onto a single dashboard, where one search filter can act on numerous charts, graphs and maps; underlying data tables can also be joined. And once you get the hang of how the software works, its drag-and-drop interface is considerably quicker than manually coding in JavaScript or R for most users, making it more likely that you'll try additional scenarios with your data set. [4]

Limitations: In the free version of Tableau's business intelligence software, your visualization and data must reside on Tableau's site. Whenever you save your work, it gets sent up to the public website -- which means you can't save work in progress without running the risk that it will be seen before it's ready . And once it's saved, viewers are invited to download your entire workbook with data. All that functionality comes at a cost: Even with the drag-and-drop interface, it'll take more than an hour or two to learn how to use the software's true capabilities, although you can get up and running doing simple charts and maps before too long. [6]

Skill level: Advanced beginner to intermediate.

Runs on: Windows; native OS X version planned for 2014.

I. CSV Kit:

CSVKit contains tools for importing, analyzing and reformatting comma-separated data files. CSVKit makes it quick and easy to preview, slice and summarize your file to examine it. [1]

For example, you can see all your column headers in a list -- which is handy for super-wide, many-column files -- and then just pull data from a few of those columns. In addition to inputting CSV files, it can import several fixed-width file formats -- for example, there are libraries available for the specific fixed-width formats used by the Census Bureau and Federal Elections Commission.
Two simple commands will generate a data structure that can, in turn, be used by several SQL database formats. The SQL code will create a table, inferring the proper data type for each field as well as the insert commands for adding data to the table. [4]
Limitations: Working on a command line means learning new text commands (not to mention the likely risk of typing errors), which might not be worthwhile unless you work with CSV files fairly often. Also, be advised that this tool suite is written in Python, so Windows users will need that installed on their system as well.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Skill level: Expert

Runs on: Any Windows, Mac or Linux system with Python installed.

V. COMPARISON

Tool	Category	Multi-purpose visualization	Mapping
DataWrangler	Data Cleaning	No	No
R Project	Statistical Analysis	Yes	With plugin
TimeFlow	Temporal data analysis	No	No
NodeXL	Network analysis	No	No
CSVKit	CSV file analysis	No	No
Tableau	Visualization app/service	Yes	Yes

Tool	Skill Level	Data Stored or Processed?	Designed for Web Publishing	Platform
Data Wrangler	2	External Server	No	Browser
R Project	4	Local	No	Linux, Mac OS X, Unix, Windows XP or later
TimeFlow	1	Local	No	Desktops running Java
NodeXL	4	Local	As image	Excel 2007 and 2010 on Windows
CSVKit	3	Local or external server	Yes	Linux, Mac OS X or Linux with Python installed
Tableau	3	Public external server	Yes	Windows

VI. CONCLUSION

Thus we discussed concepts like big data, big data analytics and some varied tools that perform data analysis, cleaning and presentation. These tools save the time spent on coding and testing by giving customized and accurate results. These tools can be used in various fields where data analytics is required. Data analysis tools play a vital role in all business domains. Many more tools have been introduced in the market and the existing products are also under constant improvement. The demand for better analytics tools is increasing constantly which is only going to increase further in future.

REFERENCES

[1] http://www.dataversity.net/3-types-data-analytics-descriptive predictive-prescriptive
[2] https://www.google.co.in/webhp?sourceid=chrome-instant&rlz=1C1CHWA_enIN623IN624&ion=1&espv=2&ie=UTF-8#q=data%20analytics%
[3] http://www.im-techsolutions.com/big-data/four-types-of-big-data-analytics-and-examples-of-their-use
[4] http://www.computerworld.com
[5] http://www.kdnuggets.com/2014/06/top-10-data-analysis-tools-business.html
[6] http://www.tableau.com
[7] Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.
[8] Blog post: Mark van Rijmenam titled "Why the 3V's Are Not Sufficient to Describe Big Data".
[9] Jean Yan, April 9, 2013 "Big Data, Bigger Opportunities Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.
[10] "Research in Big Data and Analytics: An Overview" International Journal of Computer Applications (0975 – 8887) Volume 108 – No 14, December 2014
[11] Blog post: Thoran Rodrigues in Big Data Analytics, titled "10 emerging technologies for Big Data", December 4, 2012