Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 02, February -2018

A Survey of Tools for Data Analytics

Sweety Bakyarani.E[1], Dr. Srimathi.H[2]

[1]*Department of Computer Applications, SRM Institute of Science and Technology*.
[2]*Department of Computer Applications, SRM Institute of Science and Technology*.

Abstract- *In recent times Big Data Analytics has garnered a lot of attention from Industry as well as Researchers. Big Data analytics is the process of analyzing,modelling, and inspecting data to discover hidden patterns and come up with some useful information. Big data Analytics helps organizations to harness their data, identify new opportunities andmake timelyas well as critical decisions based on it. There are many powerful analytical tools that are available in the market tthat can help us leverage the volumes of data available. The goal of this paper is to provide a comprehensive view of some such tools that are available. These tools have been chosen because they are open source, easy to use have powerful capabilities and are well documented.*

Keywords: *Big data, Data Analytics, Data Mining, Visualization, Analytical Tools.*

I. INTRODUCTION

Big Data is a term that refers to large volumes of structured, semi structured, or unstructured data that is generated by our day to day business. The amount of digital data that is being generated globally is tremendous and it shows no sign of stopping. And this data that is generated does not fit in to the traditional Data warehouse that is modeled on Relational Database Concepts. Also, a traditional data warehouse cannot handle the processing needs of Big Data as it needsfrequent, continuous updating.

So many organizations that want to harness the power of big data have turned to technologies like Hadoop and related tools like YARN, MapReduce, Spark, Hive , Pig and NoSQL databases. These technologies form the core of an Open Source Frame work that helps us to store and process voluminous data. In Some organizations Hadoop is used as a Data Pool where the organizations raw data is stored.

But the challenge does not end with efficiently storing the data. An organization benefits only when it harnesses the power of the stored data. Big data coupled with Analytics can help take any business to the next level. Data Analytics is the science of examining raw data with the aim of finding hidden patterns and eliciting useful information. Some of the key areas were Business analytics can be used to benefit an organization aredetermining root cause for failures, finding buying patterns of customers, identifying fraudulent transactions, recalculations risk portfolios etc.

Data Analytics can be classified it to

- Descriptive statistics – quantitatively describe features of the collected information
- EDA (Exploratory data analysis) – Find out new features through the data
- CDA (confirmatory data analysis) – verify a hypothesis.

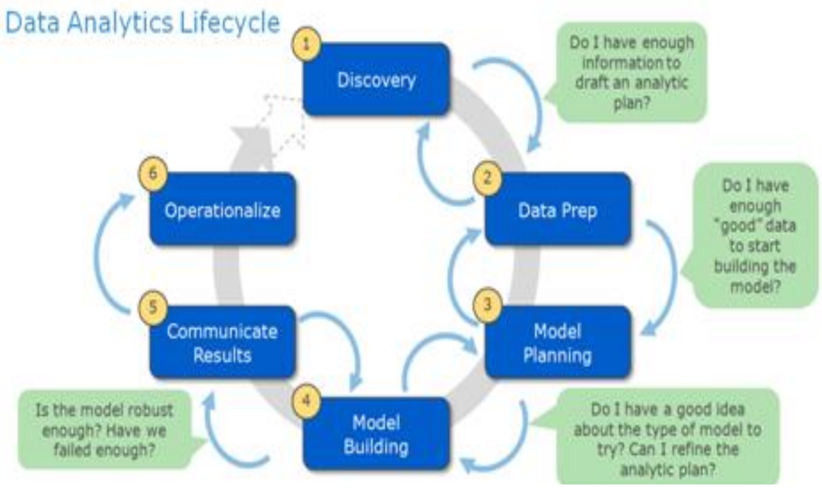

Data Analytics Lifecycle
1 Discovery
Do I have enough information to draft an analytic plan?
2 Data Prep
Do I have enough "good" data to start building the model?
3 Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4 Model Building
Is the model robust enough? Have we failed enough?
5 Communicate Results
6 Operationalize

Figure 1: Data Analytics Lifecycle

There are a host of analytics tools that are freely available that are very helpful to analyze the stored data. These tools help us to get maximum output with minimalistic effort. They use analysis techniques to infer and elicit meaning full information from raw data. The goal of this paper is to survey some of the commonly used analytical tools and understand their unique features.

II. TOOLS FOR ANALYTICS

2.1 ELKI - Environment for DeveLoping KDD-Applications Supported by Index-Structures

ELKI is an open source data mining software developed completely using Java. ELKI's focus is on algorithms but it also emphasis on unsupervised methods in cluster analysis and outlier detection. ELKI is designed in such a way that it is very easy to use for students and researchers. Itprovide researchers with a large collection of parameterizable algorithms, that allows users to evaluate and benchmark their algorithms. Not only can new analytical algorithms be benchmarked ELKI also allows us to use their vast collection of algorithms for analyzing our data. This makes ELKI very unique it also allows us to use arbitrary data types, distance or similarity measures, or file formats.

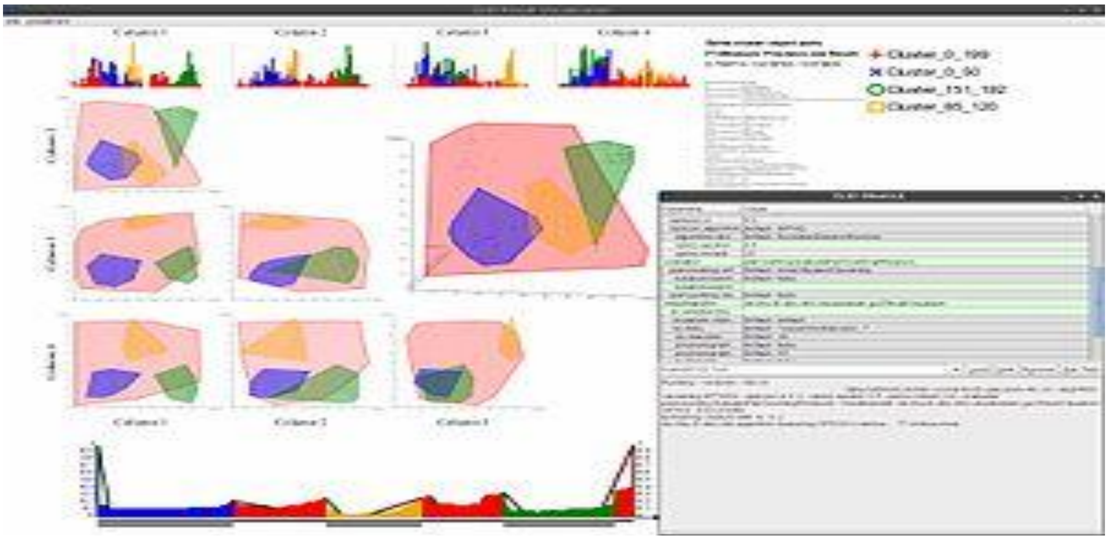


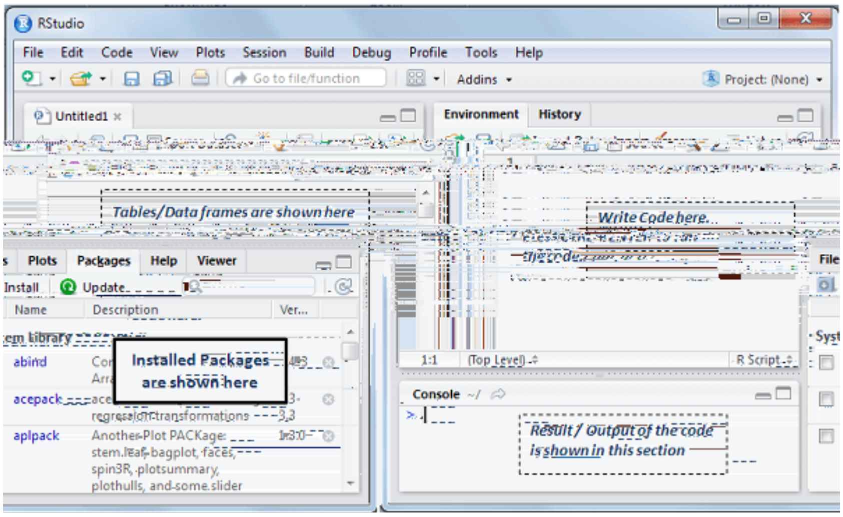
Figure 2:ELKI

2.2. R

R is GNU project, a language and environment for statistical computing and graphics. R provides a wide variety of linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and graphical techniques. It is highly extensible. One of the major strength of R's is the ease with which well-designed quality plots which include mathematical symbols and formulae can be produced. It can compile and runs on a wide variety of UNIX platforms as well as on Windows and MacOS. Some of the key features of R include

- Operators that can perform calculations on arrays and matrices
- A large and integrated collection of intermediate tools for data analysis
- Graphical facilities for data analysis
- Simple language with all the features of any major programming language
- Has its ownLaTex like document format for documentation.

.

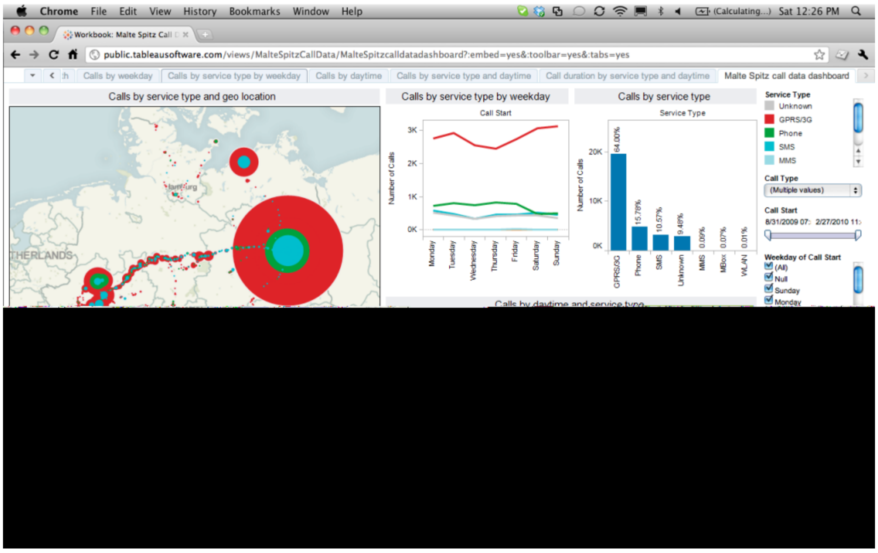


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Environment History
Tables/Data frames are shown here
Write Code here
Plots Packages Help Viewer
Install Update
Name Description Ver...
abind
acepack
aplpack
Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider
Installed Packages are shown here
1:1 (Top Level) R Script
Console ~/
Result / Output of the code is shown in this section

Figure 3: R- Development Environment

2.3.Tableau Public:

It is an elegant simple and intuitive tool. It is exceptionally powerful data visualization It allows us to have a million-row limit on data.It helps us to quicklytest a hypothesis. It requires no programming skill to work with Tableau. Visualizations prepared with Tableau can be easily shared using social media



public.tableausoftware.com/views/MalteSpitzCallData/MalteSpitzcalldatadashboard?:embed=yes&:toolbar=yes&:tabs=yes

Figure 4:Tablaeau Public

Limitaiton:

- All data is public also there is a restriction on the size of the data set
- Cannot be used with R.
- The only way to read is via OData sources, is Excel or txt.

2.4.OpenRefine:

Formerly known as GoogleRefine, this is a software that allows youto clean your data before analysis. OpenRefine contains several clustering algorithms that makes grouping our data very easy. It works with rows of dataorganized as columns, much like relational databases. Apart from Cleaning messy data, it can also be used for:Transformation of data, Parsing data from websites and fetching, adding data from web services

Limitation:

- Does not work well with large data sets

2.5.KNIME:

KNIME allows us to analyze, manipulate and model our data through visual programming. All this can be done without writing a single line of code. It allows us to drop nodes onto a canvas and drag connection points between activities. KNIME can be extended to run R, python, text mining, etc. It can be used to integrate various components for data mining and machine learning through its modular data pipelining concept. It requires no coding required for complex visualizations and supports programming languages.

2.6.RapidMiner:

Like KNIME, RapidMiner extensively supports visual programming and is capable of manipulating, analyzing and modeling data. RapidMiner provides machine learning procedures and data mining including data visualization, processing, statistical modeling, deployment, evaluation, and predictive analytics. It is written in Java. It provides an integrated environment for business analytics, predictive analysis, text mining, data mining, and machine learning.

Limitations

- Size constrain with respect to number of rows.
- Hardware extensive.

2.7. Google Fusion Tables:

It is an incredible tool for data analysis, mapping, and large dataset visualization. It can be used to visualize bigger table data online, filter and summarize across hundreds of thousands of rows and combine tables with other data on web. We can merge two or three tables to generate a single visualization that includes sets of data. We can also can combine public data with your own for a better visualization and create map within minutes

Limitations

- Only the first 100,000 rows of data in a table are included in query results or mapped.
- The total size of the data sent in one API call cannot be more than 1MB.

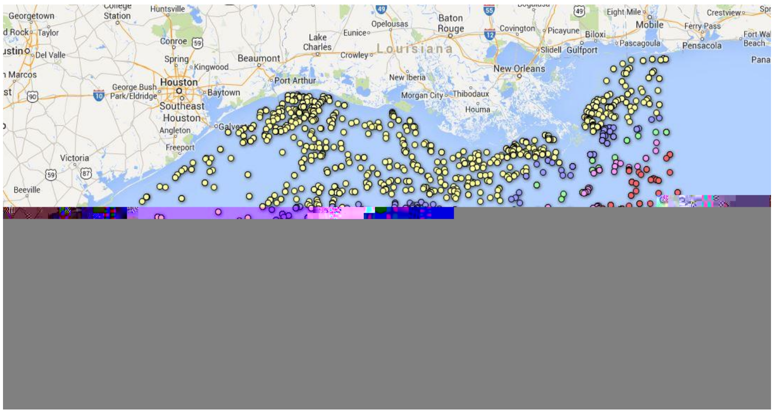
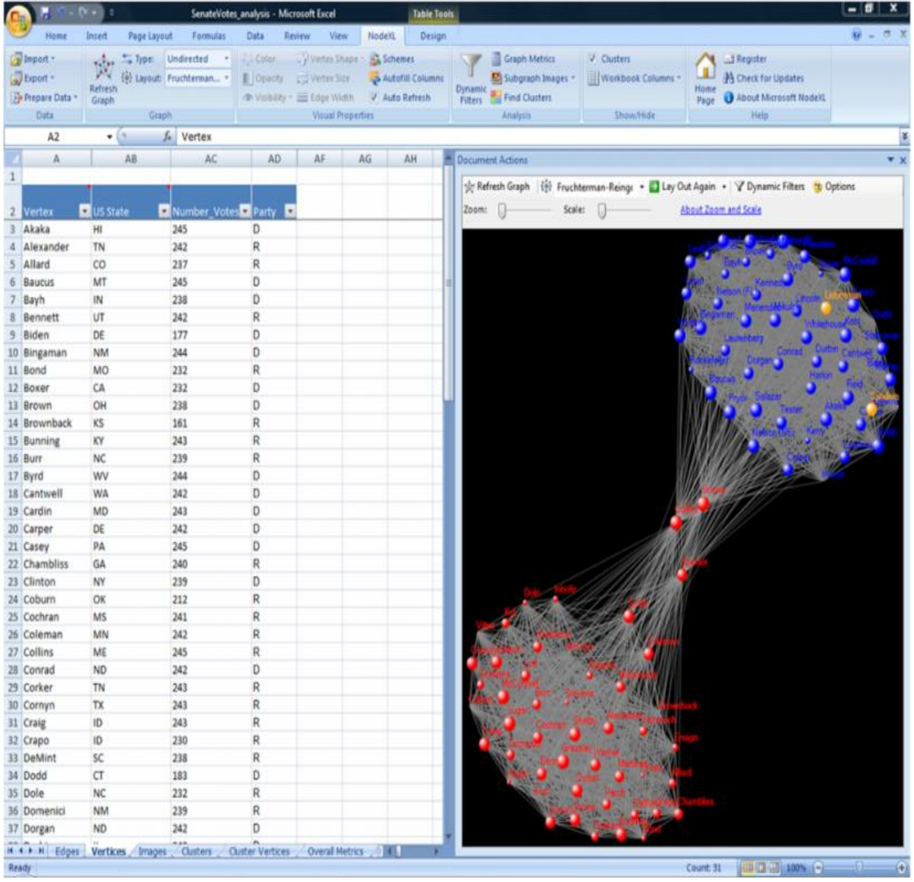


Figure 5: Fusion Table Map Plot

2.8.NodeXL:

NodeXL is a visualization and analysis software of networks and relationships. It is a free software and is one of the best statistical tools for data analysis which includes advanced network metrics, access to social media network data importers, and automation. It can be integrated into Microsoft Excel 2007, 2010, 2013, and 2016. We can import various graph formats like adjacency matrices, Pajek .net, UCINet .dl, GraphML, etc.



Type: Undirected

Layout: Fruchterman...

A2 Vertex

	A	AB	AC	AD
1				
2	Vertex	US State	Number_Votes	Party
3	Akaka	HI	245	D
4	Alexander	TN	242	R
5	Allard	CO	237	R
6	Baucus	MT	245	D
7	Bayh	IN	238	D
8	Bennett	UT	242	R
9	Biden	DE	177	D
10	Bingaman	NM	244	D
11	Bond	MO	232	R
12	Boxer	CA	232	D
13	Brown	OH	238	D
14	Brownback	KS	161	R
15	Bunning	KY	243	R
16	Burr	NC	239	R
17	Byrd	WV	244	D
18	Cantwell	WA	242	D
19	Cardin	MD	243	D
20	Carper	DE	242	D
21	Casey	PA	245	D
22	Chambliss	GA	240	R
23	Clinton	NY	239	D
24	Coburn	OK	212	R
25	Cochran	MS	241	R
26	Coleman	MN	242	R
27	Collins	ME	245	R
28	Conrad	ND	242	D
29	Corker	TN	243	R
30	Cornyn	TX	243	R
31	Craig	ID	243	R
32	Crapo	ID	230	R
33	DeMint	SC	238	R
34	Dodd	CT	183	D
35	Dole	NC	232	R
36	Domenici	NM	239	R
37	Dorgan	ND	242	D

Figure 5: Visualization using NodeXL

2.9. Excel Solver:

The Solver Add-in is a Microsoft Office Excel add-in program that is available when you install Microsoft Excel or Office. It is a linear programming and optimization tool in excel.This allows you to set constraints. It is an advanced optimization tool that helps in quick problem-solving.It uses a variety of methods like nonlinear optimization , linear programming and genetic algorithmsto find solutions.

Limitations

- Poor scaling effects solution time and quality.

2.10. Dataiku DSS

This is a collaborative data science software platform that helps team build, prototype, explore, and deliver their own data products more efficiently. It provides an interactive visual interface where they can build, click, and point or use languages like SQL.This data analytics tool lets you draft data preparation and modulization in seconds.Helps you coordinate development and operations by handling workflow automation, creating predictive web services, model health daily, and monitoring data.

Limitaions

- Limited visualization capabilities
- UI hurdles: Reloading of code/datasets
- Inability to easily compile entire code into a single document/notebook
- Still need to integrate with SPARK

III. CONCLUSION

Data Analytics has applications in all the major fields like Business, Stock Market, Education, Health care to name a few. It helps us to gain insight into numbers and texts. These insights can alter the way we perceive things. In this paper we have surveyed the working of some common analytical tools available in the market. These tools can easily be adapted to both structured and unstructured data. Although these tools make analysis easier, success of it depends on the data we put in and the analysis we make out of it. Also, we identify our needs first before deciding on analytic tool to use.

REFERENCES

[1] D. Magesh Kumar, D. Christy Sujatha, M. Chandra Kumar Peter,A Survey in Data Anaalytic toolsIndian J.Sci.Res. 14 (1): 176-180, 2017.

[2] Mr. Mahesh G Huddar, Manjula M Ramannavar, A Survey on Big Data Analytical Tools, International Journal of Latest Trends in Engineering and Technology (IJLTET), Special Issue - IDEAS-2013, ISSN: 2278-621X.

[3] http://www.digitalvidya.com/blog/data-analytics-tools/

[4]https://www.kdnuggets.com/2014/06/top-10-data-analysis-tools-business.html

[5]Bogdan Batrinca, Philip C. Treleaven, Social media analytics: a survey of techniques, tools and platforms, AI &SOCIETY,February 2015, Volume 30, Issue 1, pp 89–116|.

[6]J. Vijayaraj, R. Saravanan,P. VicterPaul,A comprehensive survey on big data analytics tools, Green Engineering and Technologies (IC-GET), 2016 Online International Conference on19-19 Nov. 2016, DOI: 10.1109/GET.2016.7916733,Publisher: IEEE, Conference location: Coimbatore, India

[7]https://www.datapine.com/data-analysis-tools