International Journal of Engineering & Technology, 7 (3.31) (2017) 10-15

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Review of Leading Data Analytics Tools

Sridevi Bonthu*, K Hima Bindu

Computer Science and Engineering, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202
**Corresponding author E-mail:sridevi.db@gmail.com ; Tel +919885880382*

Abstract

Data Analytics has become increasingly popular in uncovering hidden patterns, correlations, and other insights by examining large amounts of data. This led to the emergence of a variety of software tools to analyze data. Before adopting the tool, organizations need to know how they will fit into their larger business goals. Due to ever changing requirements from people practicing Data Analytics, many new tools are entering into the market and few tools are losing importance. A review of current popular tools is provided in this paper to help the analytics practitioners to choose the appropriate tool for the requirement at hand. This paper provides a review of seven popular tools viz., R, Python, RapidMiner, Hadoop, Spark, Tableau, and KNIME.

Keywords: *Data Analytics; Hadoop; KNIME; Python; RapidMiner; R language; Spark; Tableau; Tool;*

1. Introduction

Data is the basic building block upon which any organization runs and flourishes. It is not possible to imagine the world without data [1]. "Data is increasing with a faster pace than ever before and by 2020, nearly 1.7MB of new data will be formed each second for each human being on the planet"[2]. Through the advancements in technologies and therefore the Internet, the data and the information are increasing every second. New business models based on data have emerged during past few decades like Facebook, Yahoo, Microsoft, Google, Linkedln, youtube, Twitter etc. We all are aware how the technical giants like Google and Facebook are using "data as new oil".

In recent times, the term "Big Data" has been applied to data that grow up so huge and become uncomfortable to work and manage with using traditional age-old database management systems. Size of these datasets is beyond the capability of frequently used tools and storage systems to retrieve, store, and administer, as well as process the data within an average elapsed time [3].

We all are starving for information, even though the information is drowning [4]. Data Analytics is the process of analyzing data, which converts information into useful knowledge [4]. This knowledge helps to understand the world better, and in many contexts enable us to make better decisions. Big data analytics is applying advanced analytic techniques on big datasets. Big data and advanced analytics are now actively busy in transforming the enterprise. The marketplace for data analytics is growing at faster pace. According to the World Economic Forum forecasts, by 2020, the most in-demand job of the future years or present is Data Analyst [5]. The data analytics tools are the next big things, as organizations are treating data as an asset today. Hence, it is essential to know about data analytics and the tools that match our purpose the best. Data analytics based on huge datasets reveals hidden patterns and helping for a business change. However, it is more difficult to manage the larger set of data [6]. There were nearly more than hundred tools available to perform steps of analytics. Selecting a tool from the vast amount of existing tools is a challenging task. When choosing analytical tools to use, it is useful to grasp the dimensions of the software's market share and whether or not it's growing or shrinking.

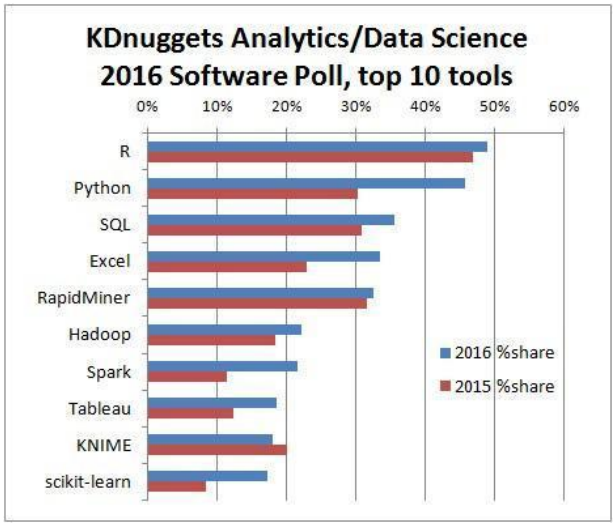

Fig. 1: Adopted from KDnuggets Software poll on top 10 popular Analytics/Data Science tools in 2016 [10]

This paper describes the characteristics of seven most popular software tools for data analytics. Results of a software poll, articles from AnalyticsVidya[7], Kaggle[8] and report of Gartner Magic Quadrant for Data Analytics platforms[9] motivated to pick these tools. One of the primary sources was provided by the 17th annual Software Poll of KD nuggets. It conducted a survey on the software adopted for Data Analytics, Mining, Science, and Machine Learning projects in the year 2016[10]. The poll has enormous participation from 2895 voters and the obtained results were shown in **Figure 1**. All the voters are chosen from analytics, data science community and vendors chose from 102 different tools. **Figure 1** represents an adopted poll result on data science/analytics tools. From a study on digitalvidya[11], The

income from the data and business analytics services sales will lift more than 50% to $187 billion by the year 2019. Companies that do not seem to make powerful utilization of data analytic tools and techniques are diminishing.

Motivating from these results, this paper reviews R[12], Python[13], RapidMiner[14], Hadoop[15], spark[16], Tableau[17], KNIME[18].

The life cycle of Data Analytics is discussed in section 2, Overview of every tool is presented in section 3, Discussion on all tools is taken care in section 4, and section 5 presents the conclusion.

2. Analytics Techniques

To give clear insights to customers from data, a framework which enables to think of it as a cycle with different stages is needed. This framework involves various actions to be carried out in analyzing the data. **Figure 2** depicts different phases of Data Analytics life cycle along with the flow of data in between [19], they are identifying the problem, preparing data, model planning, and building, communicating the obtained results with an operationalization of the product.

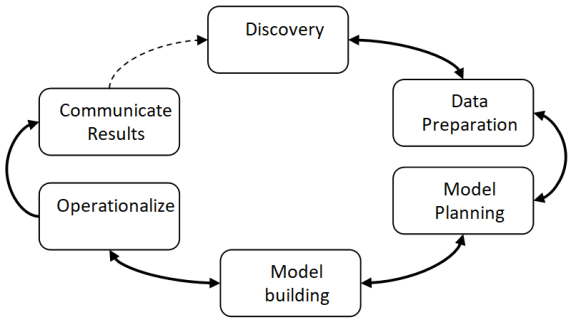

Fig. 2: Data Analytics Life Cycle

i) Discovery - Understand the problem and identifying whether enough information is available or not to prepare an analytic plan and share for another study is very important. In this phase, the business organization wants to make predictions over the data to make required decisions. Because of this reason, analytics are carried over the datasets. The team assesses the resources like people, technology, time, and information. Other activities in this phase comprise framing the problem and formulating early hypotheses.
ii) Data preparation - Checking whether the data available is of good quality or not to start building the model is also important. This phase includes steps to explore, pre-process and order the data. This phase requires execution of extract, load, and transform (ELT). In this stage, the team also needs to be acquainted or got familiarize themselves with the data systematically to put data in well-ordered format.
iii) Model planning - Finding whether an idea which is available is good to try for a model is crucial. In this phase, the team finds out the process, methods, and order it needs to follow for the model building. The group determines the information to know about the variables, how they are related and consequently select key variables and suggests suitable models.
iv) Model building - To continue further, the model which is planned is robust or not has to be checked. In this stage, the team members put together datasets for training, production and testing functions. In this phase, the model built is executed to test work done in the planning phase is supported or not. The team also tests for suitability of the existing tools to run the prepared models.
v) Operationalize - In this phase, the team delivers a final documentation, summaries, project code, and technical documentation.
vi) Communicate results - This part of the life cycle determine whether the results are a success or a failure based on the analytic plan made at the discovery stage.

3. Overview of the Tools

To analyze data and to extract commercially relevant actionable information, one should depend on good software. There are an ample number of tools; both commercial and open-source are available for data analytics in the market today. One has to invest in the suitable tools and skills to discover new opportunities. Software tools of data analytics employ several types of investigation methods to store, manipulate and find suggestive implication from the given datasets. Some of the tools are even doing well in producing summarization reports and better visualization, thereby helping us in getting accurate results with negligible effort.

3.1. R

R is an Open Source software program, and a service developed by volunteers to the community of scientists, researchers, and data analysts and it is maintained by R foundation for statistical computing [12]. R is available freely under GNU General Public License. It is widely used by statisticians and lot of guidance is available online. Interested readers can refer to R-bloggers, ucanalytics [20], introductoryr [21], r4stats.com [22] websites for further exploration. KDnuggests poll[5] conveyed that the usage of R is more among data analytics tools. It says that 49% of the voters are using R and there is an increase of 4.5% from 2015 to 2016. The poll conveyed that the usage of R is skyrocketing. Surveys of data-miners, polls and studies of literature on databases are showing that the popularity of R has increased to a large extent in recent years. The development and usage of R has challenged 40 years of monopoly by SAS language. The software environment of R is developed using C, Fortran and R itself. R has a good command line interface and all the commands of R are easy to apply and understand. Several graphical front-ends like R studio, IntelliJ, visual studio etc. are available nowadays.

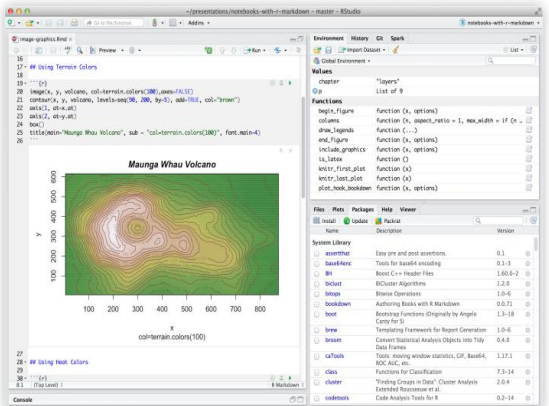

Fig. 3: R Studio IDE [12]

Figure 3 shows a glimpse of popular IDE, R Studio. **Figure 3** is a glimpse of R Studio IDE which integrates workspace, results, variables etc. R language is an interpreted one. Users normally access it through a command-line interpreter. RMD abbreviated as "R Mark Down" allows easy exchange of code along with output. RMD files can be exported as pdf, word, HTML and latex files using *knitr*.
It has a vast number of, nearly 12000 statistical, graphical, and analytical packages as part of it [23]. Function of R can be enhanced through functions easily, and the R community is renowned for its contributions in terms of packages. Some libraries which proved to be handy to do analytics are dplyr, tidyr,

stringr, lubridate, ggplot2, ggvis, rgl, shiny, rMarkdown, RMySQL, RSQLite etc.
Using R markdown, one can make their data to tell a story by turning their analyses into documents, presentations, reports and dashboards. Strengths of R include static graphics to produce publication-quality graphs. Dynamic and interactive graphics are available through extra packages [24]. Few more strengths of this tool are, promotion of reproducible research, and record of how the analysis was done. Commands can be modified, re-run, forwarded, commented etc. It is more of user run software, which means that anyone is allowed to provide code enhancements and add new packages. The limitation of R is that, if it is not taking the advantage of parallel processing, the data it can process is equal to the memory of system on which it is running. Few more limitations of R are, it not so easy to use for the novice, cannot scale properly with large sets of data, some procedures could take days to run.

3.2. Python

Guido van Rossum developed Python language is a majorly followed general purpose programming language, released in 1991[13]. Python is friendly, easy to learn, powerful and fast programming language. This language is developed under an Open Source license which is OSI approved, thus made it freely usable and distributable, even for the commercial user base. Python Software Foundation administrates python's license. Lots of third-party modules for Python are hosted by Python Package Index, simply PyPI. The standard library of Python and the modules contributed by its community provide endless possibilities. The poll[10] found that the usage of python 45.8% and there is a slight decrease. Python is highly used to perform data analytics with the help of rich set of libraries it has. Python is vastly attracting Data Analysts. Interested readers can refer to datacamp [26], datasciencecentral [27], byteacademy [25] for online reference. Some of the best python IDEs that are best for Data Analytics are Spyder, PyCharm, Rodeo, Atom, Jupyter notebook [26]. Jupyter notebook shown in **Figure 4** is not only an easy-to-use, interactive environment, but also a presentation or education tool. It includes code along with output in the same notebook which can be shared as pdf and many other file types including latex.

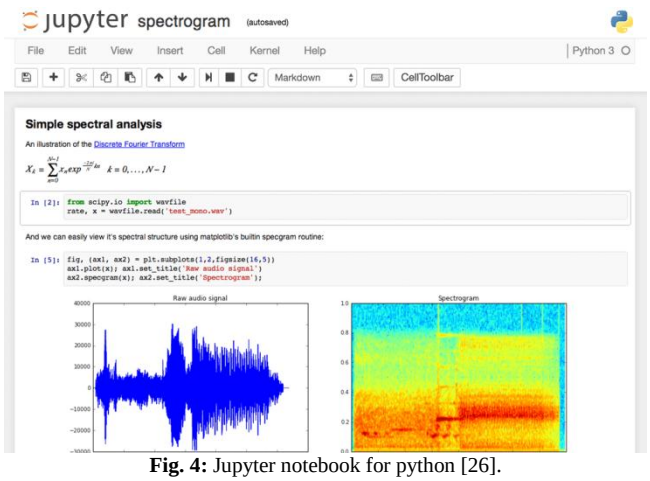

Fig. 4: Jupyter notebook for python [26].

The data analytics libraries are pandas, statsmodel, scikit-learn, Mlpy, NumPy, SciPy, matplotlib, NLTK, Theano[27]. Pandas library, written in python itself is used for data manipulation and data analysis. To manipulate numerical and time series data, python offers data structures and number of operations. Statsmodels module helps users to explore data, estimate the statistical models, and to perform statistical tests. Various features of scikit-learn include classification, regression and clustering algorithms consisting of support vector machines (SVM), k-means and DBSCAN, gradient boosting. It can also interoperate with Bayes, random forests. Mlpy is built on top of NumPy/SciPy, which provides machine learning methods for the supervised and unsupervised. Methods which run faster for numerical routines, to support matrices are provided by the library Numpy. SciPy provides modules to perform functions of linear algebra, interpolation, integration, optimization, signal, FFT and image processing. matplotlib is an excellent plotting library for NumPy. Its main objective is to enable the user to generate plots into application with the help of GUI toolkits. The Natural Language Toolkit (NLTK), is a group of excellent libraries, which include graphical demonstrations and sample data. Theano is a useful Python library that enables us to play with mathematical expressions like defining, optimizing, evaluating which involves arrays of any dimensions efficiently [27].
The limitation of python as a data analytics tool is, it is more programmer friendly, where as R is statisticians friendly. Limitations of python include lack of commercial support, does not have proper multiprocessor support, lack of UI development framework.

3.3. RapidMiner

RapidMiner, one of the commercial data analytic tool today, which offers machine learning measures and data mining measures including statistical modeling, processing, visualization, deployment of the product, evaluation, and predictive analytics[14]. It is developed by RapidMiner. This software is coded by using Java programming language. It supports most of the steps of the machine learning process. It is available under the AGPL license. Though it is new software when compared with leads like R and Python, it is managed to capture the attention of users. According to poll [10] it is used by 32.5% of voters and its usage is raised from last year. RapidMiner offers a GUI-based data science platform, which best suits beginner and expert data analysts. It is leading in the new 2017 Gartner Magic Quadrant for Data Science Platforms [9].
A suite of products like RapidMiner Studio, Server, Radoop, and Streams are offered by RapidMiner. All these allow data analyst to put up new data mining processes, set up predictive analysis, and many more. Its key characteristics include Graphical user interface, taking data from the files, database, web, and through cloud services, in-memory, database and Hadoop analytics, interactive, easy to use, shareable dashboards, predictive analytics, analysis of remote data, processing, filtering, merging, joining and aggregating, build, train and validate predictive models, and can run more than 1500 operations etc. The typical customers of RapidMiner are large enterprises, mid-size business. Hardware resources requirement is more for this software.

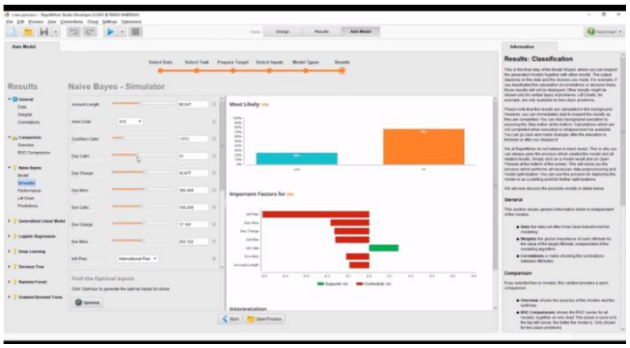

Fig. 5: Rapidminer working environment [14]

Strengths of RapidMiner are, it has a very nice working environment as shown in **Figure 5**, it provides wizards for data management, it supports and can accept new data drivers, over 1500 methods for data integration, transformation, analysis and modelling, easy to debug the errors, integration of maximum algorithm of such tools. Limitations of RapidMiner includes more

suitable for the people who are accustomed to working with databases, limited partitioning abilities for dataset to training and testing sets.

3.4. Hadoop

Hadoop is a top-level Apache Foundation project [15] which is a distributed file system, works on any platform. Hadoop has become a de facto standard and the companies that utilize big data sets and analytics use this. It is an elastic architecture for gigantic scale computation and processing of data on a network of service hardware. It can scale from one server to any number of machines; every machine has their own storage and processing. Hadoop can be used for so many purposes, one of them is analytics. It has huge, energetic user base, mailing lists, user groups. Hadoop is outperforming the newcomer Spark in some of the business applications. Adoption of Hadoop is slowed down; it is presently used by 22.1% voters.

Figure 6 presents the architecture of Hadoop. Main components of Hadoop are NameNode, DataNode, secondaryNameNode, Job Tracker, Task Tracker, Yarn, Client Application and Application Master.

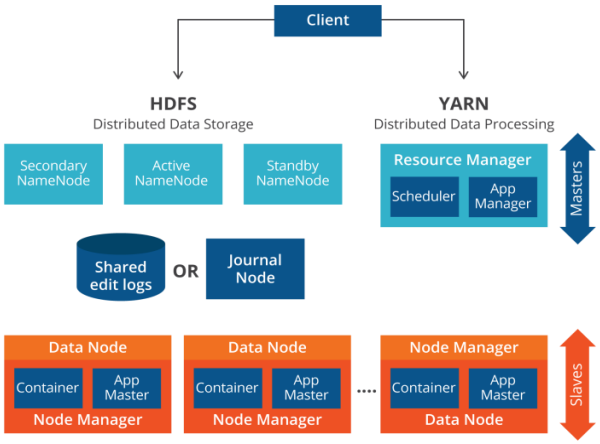

Fig. 6: Hadoop Architecture [28].

One can do data analytics on top of Hadoop by using MapReduce, Kudu, Pentatho, IBM Biginsights, Apache Spark tools. Strengths of Hadoop include scalability, cost effectiveness, flexibility and resilience to failure. Few limitations of this include concerns regarding security, vulnerability by nature, unfit for little datasets, potential stability issues.

3.5. Spark

Apache Spark™ is a rapid and general mechanism for large-scale data processing [16]. It was developed to answer the limitations of the MapReduce computing paradigm. It can run applications 100 times faster than Hadoop in-memory and 10 times faster on disk [16]. Spark analyzes data in real time and proved as an excellent tool for in-memory computations. Spark is completely built around fast computing, sophisticated analytics, ease-of-use there by making it as a very dominant open source processing engine. It was developed in 2009 at UC Berkeley [29]. Now-a-days many companies are adopting Spark quickly and providing community support. It is currently the fastest growing big data technology and is being used at several leading companies in production. Though it is a recent one, it slowly captured the attention of the developer community. Spark is becoming next big thing in data analytics due to the reasons 1) it makes advanced analytics a reality 2) makes everything easier 3) can speak more than one language 4) doesn't care which Hadoop vender is being used 5) it accelerates results. Spark has its own place in data space because of its speed and its ease of use nature. As shown in **Figure 7**, spark powers a set of libraries SQL, MLib, GraphX and Streaming. ASingle application can use all these libraries.

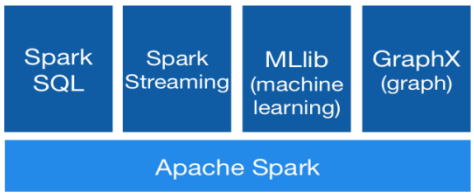

Fig. 7: Apache Spark Generality [11]

The Spark ecosystem comprises multiple components that complement each other are designed to work with each other seamlessly. The versatile and general structure of Spark allowed for specialized libraries geared towards specific workloads to be built on top of it like SparkSQL for querying structured data through an SQL interface, MLib for machine learning, Spark Streaming for processing data streams in motion and GraphX for graph computations. Underpinning any of these components is the Spark core engine that defines the basic structure of Spark including its core abstraction, the Resilient Distributed Dataset (or RDD). Spark ships with user-friendly APIs for Scala, Java, Python, Spark R and Spark SQL. Spark SQL.

3.6. Tableau

Tableau made visualization available to everyone as an elegantly simple and perceptive tool. It has become powerful in business as it communicates the insights of data through visualization [17]. Despite of hundreds of alternatives, Tableau is providing a great playground for individuals because of its million row limit. Its visuals permit you to investigate the hypothesis quickly, exploration the data before embarking on a treacherous statistical journey in the analytics process. This elegant drag-and-drop analytics solution is completely free for students and it can be downloaded from "https://www.tableau.com/academic/students". A sample screen of this easy-to-use tool is shown in **Figure 8**.

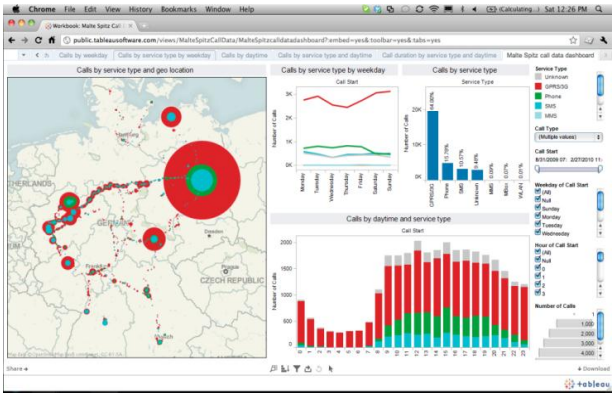

Fig. 8: Tableau working environment [17]

Uses of Tableau include free service through which everyone can publish data visualizations to the web, there is no requirement of coding skills, easy upgradation, integrates fine with third party big data platforms, offers support for Google BigQuery API. Storytelling from the data, report generation, and dashboard creation for at-a-glance view can be easily created by using this tool. The excellent user interface is carried further onto mobile devices and the dashboard reports it generates are also mobile optimized. Limitations of Tableau Public include limitation on data size, cannot connect with R, all data is public, data can be read only through OData Sources, Excel or txt. Gives the ability to set up "row level" security at the data level which is a bit risky, Businesses and IT sectors are struggled to understand needs, heavy losses due to Shrink, complex issues are difficult to address, static nature.

3.6. KNIME

KNIME is an Open Source platform which enables us to do data analytics, reporting and integration [18]. This tool is recommended especially to those who are beginner to data analytics and also to those who are highly skilled [30]. This tool is an integration of various components of machine learning and of data mining. An easy to understand GUI lets assembly of nodes for data pre-processing, modeling, data analysis and visualization. KNIME.com AG is its developer and it works well on various operating systems like Windows, Linux, OS X etc.

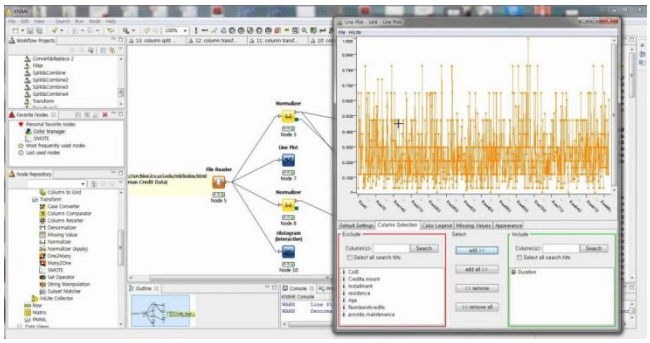

Fig. 9: KNIME interface [18]

KNIME has now become one of the leading analytics platforms for innovations on data which discovers the potentiallly hidden patterns, mining for insights which are new, or predict new features [31]. It has thousands of modules, and every module has hundreds of ready-to-run sample codes, ample number of integrated tools, and the vast choice of algorithms. This tool uses modular pipelining concept for the integration of components. To use KNIME, one has to simply drag and drop connection points between activities thereby enabling a novice user to use it. It even supports programming languages. **Figure 9** shows the working environment of this tool and how a user can drag to drop the activities..

KNIME tool is more suitable for Molecular Analysis. Limitations of KNIME are poor visualization, not suitable from large and difficult workflows, cannot partition more than one data set, has limited error measurement methods, no wrapper methods, preliminary results are not available.

4. Discussion

The overview of tools provided an insight into how all these data analytics tools can be differentiated. Table 1 gives a clear picture about the developers, and its home-page. One can move to the downloads section with the help of the specified website of **Table 1** to download and install it.

Table 1: Developers and website names of opted tools

Tool	Developer	website
R	R Core Team	https://www.r-project.org/
Python	Python Software Foundation	https://www.python.org/
RapidMiner	Rapid Miner, Germany	https://rapidminer.com/
Hadoop	Apache Software Foundation	http://hadoop.apache.org/
Spark	Apache Software Foundation	http://sparkapache.org/
Tableau	Tableau Software	http://www.tableau.com/
KNIME	KNIME.com AG	http://www.knime.org/

Most of the programmers are comfortable with Command Line Interface, which is provided by most of the leading tools. The general characteristics like present version, support to command line, licence, and language they were written and the functionality of selected software tools is presented in the **Table 2**.

Table 2: General characteristics of the tools

Tool	GUI/ CLI	Current version	License	Language used for development	Functionality
R	Both	3.3.3	GNU GPL v2	C, Fortran, R	Statistical Techniques
Python	Both	3.6.1.rc1	Python software foundation license	C, Python	General purpose programming
RapidMiner	GUI	7.4	Profession edition is proprietary, basic edition is AGPL	Java	Business Intelligence, Machine Learning, predictive analytics
Hadoop	Both	3.0.0 alpha2	Apache License 2.0	Java	Distributed environment
Spark	Both	2.1.0	Apache License 2.0	Scala, Java, Python, R	Data Analytics, Machine Learning
Tableau	GUI	10.1	Commercial	C++	Data Visualization
KNIME	GUI	3.3.1	GNU general public	Java	Data Analysis, Text Mining

Data analytics can be performed on various platforms like web, Iphones, through the mobile applications of android and windows phone. **Table 3** gives a clear picture on which tool can be used on which platform. Python can be used with all type of platforms where as R, Spark, Tableau and KNIME cannot be utilised in windows phone app. As Hadoop is a distributed environment, it cannot be used with all platforms.

Table 3: supported Platforms

Tool	web	I-phone	android	Windows phone
R	✓	✓	✓	
Python	✓	✓	✓	✓
RapidMiner	✓			
Hadoop	✓			
Spark	✓	✓	✓	
Tableau	✓	✓	✓	
KNIME	✓	✓	✓	

Data analytics can be performed by self-employed or by the businesses which are small, medium and large scale. **Table 4** illustrates which software tool can be utilized by which type of customers. It is clear that all the tools suit the needs of mid-size and large enterprises.

Table 4: Typical Customers

Tool	Free-lancers	Small business	Mid-size business	Large enterprise
R	✓	✓	✓	✓
Python	✓	✓	✓	✓
RapidMiner		✓	✓	✓
Hadoop			✓	✓
Spark			✓	✓
Tableau			✓	✓
KNIME		✓	✓	✓

The languages R and Python can be used by any kind of customers due to their ease of use. As Hadoop and Spark are distributed work environments, therefore suitable only for medium and Large scale businesses. As Tableau supports visualizations more, it is commercial software and a recent addition to the analytics race, it is not yet adopted by freelancers and small-size businesses. Among all the studied tools R language is purely statistical and suits well for exploratory data analysis and all data analytical task. It also has good packages for visualization. All these features made users to adopt R and stood R in top position in the software tools race. The discussion also made clear that RapidMiner and Spark are also capturing the market in a rapid pace. According to

AnalyticsVidya [32] few more competing tools were added recently like DataRobot, BigML, Google Cloud Prediction API, Paxata, Trifacta, Narrative Science, ML Base, Automatic Statistician. Most of these tools were developed for nonprogrammers.

5. Conclusion

This paper presented popular data analytics tools and their characteristics. Strengths and limitations of every tool are discussed. This paper clearly highlighted which tool is more suitable for which Data Analytics task. From the discussions, researchers/data analysts can easily check suitability of tool for their requirements.

References

[1] Bonthu, Sridevi, Y S S R Murthy, M. Srilakshmi. "Building An Object Cloud Storage Service System Using Openstack Swift." International Journal of Computer Applications102.10 (2014): 39-42. available online: http://dx.doi.org/10.5120/17854-8827

[2] "Forbes Welcome." Forbes.com. N.p., 2018. Web. 9 Sept. 2017.

[3] Kubick, Wayne. "Big Data, Information And Meaning." Appliedclinicaltrialsonline.com. N.p., 2018. Web. 28 March 2018.

[4] Introduction to Data Analytics - Course. (2017). Onlinecourses.nptel.ac.in. Retrieved 28 March 2017, from https://onlinecourses.nptel.ac.in/noc15_mg05

[5] The most in-demand job of the coming years. from https://www.morningfuture.com/en/article/2018/02/21/data-analyst-data-scientist-big-data-work/235/

[6] Russom, P. Big data analytics. TDWI best practices report, fourth quarter, (2011). 1-35

[7] Analytics Vidhya, https://www.analyticsvidhya.com/

[8] Kaggle: Your Home for Data Science, https://www.kaggle.com/

[9] Linden, A., Krensky, P., Hare, J., Idoine, C., Sicular, S., & Vashisth, S. "Magic Quadrant for Data Science Platforms. Gartner.com." (2017). Retrieved 28 March 2017, from https://www.gartner.com/doc/3606026/magic-quadrant-data-science-platforms

[10] Piatetsky, Gregory. "R, Python Duel As Top Analytics, Data Science Software – Kdnuggets 2016 Software Poll Results." Kdnuggets.com. N.p., 2018. Web. 28 March 2018.

[11] Behl, Sakshi. "Top 10 Data Analytics Tools | Tools Used For Data Analysis." Digital Vidya. N.p., 2018. Web. 28 March 2018.

[12] "R: The R Project For Statistical Computing." http://R-project.org. N.p., 2018. Web. 28 March 2018.

[13] "Welcome To Python.Org." Python.org. N.p., 2018. Web. 28 March 2018.

[14] "Lightning Fast Data Science Platform | Rapidminer." RapidMiner. N.p., 2018. Web .28 March 2018.

[15] "Welcome To Apache™ Hadoop®!." Hadoop.apache.org. N.p., 2018. Web. 9 May 2018.

[16] "Apache Spark™ - Unified Analytics Engine For Big Data." Spark.apache.org. N.p., 2018. Web. 28 March 2018.

[17] "Tableau Software." Tableau Software. N.p., 2018. Web. 28 March 2018.

[18] "KNIME - Open For Innovation." Knime.org. N.p., 2018. Web. 26 March 2018.

[19] Services, EMC Education. Data Science And Big Data Analytics. Somerset: Wiley, 2015. Print.

[20] Explore the power of predictive analytics http://ucanalytics.com/

[21] R resources for beginners http://www.introductoryr.co.uk

[22] R for Statistics from http://r4stats.com/

[23] Behl, S. How to Choose Data Analytics Specialization: Python, R, SAS, Excel or SQL?. (2017). Digital Vidya. Retrieved 28 March 2017, from http://www.digitalvidya.com/blog/choose-data-analytics-specialization/#

[24] Lewin-Koh, Nicholas. "CRAN Task View: Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization." Cran.r-project.org. N.p., 2018. Web. 9 May 2018.

[25] Byte academy http://byteacademy.co/

[26] Top 5 python IDEs for Data Science from https://www.datacamp.com/community/tutorials/data-science-python-ide

[27] 9 Python Analytics Libraries. (2017). Datasciencecentral.com. Retrieved 28 March 2017, from http://www.datasciencecentral.com/profiles/blogs/9-python-analytics-libraries-1

[28] An introduction to hadoop architecture from http://www.bmc.com/guides/hadoop-architecture.html

[29] What is Apache Spark?. (2017). Databricks. Retrieved 28 March 2017, from https://databricks.com/spark/

[30] Dwivedi, S., Kasliwal, P., & Soni, S. "Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime)". In Colossal Data Analysis and Networking (CDAN), (2016, March). Symposium on (pp. 1-8). IEEE.

[31] KNIME | KNIME Analytics Platform. (2017). Knime.org. Retrieved 28 March 2017, from https://www.knime.org/knime-analytics-platform

[32] 19 Data Science Tools for people who aren't so good at Programming from https://www.analyticsvidhya.com/blog/2016/05/19-data-science-tools-for-people-dont-understand-coding/