The Popular Tools Of Data Sciences: Benefits, Challenges and Applications

Hafiz Burhan Ul Haq[1], Haroon Ur Rashid Kayani[2], Saba Khalil Toor[3], Sadia Zafar[4], Imran Khalid[5]
burhanhashmi64@lgu.edu.pk, hurkayani@gmail.com, sabakhaliltoor@gmail.com, sadiazafar@lgu.edu.pk, imrankhalid@lgu.edu.pk
[1,4,5] Department of Computer Sciences, Lahore Garrison University, Lahore Pakistan.
[2]Data Scientist and Consultant, Lahore Panitan.
[3]Department of Computer Science, Forman Christian College (A Chartered University), Lahore Pakistan.

Abstract

Data Science is a new field and introduced in the United Kingdom (UK), United States of America (USA), European Union, Australia, and Canada, in 2012. The subject such as Statistics, Mathematics, Artificial Intelligence, Machine Learning and Data Mining became an integral part of Data Science. The open-source tools were rejected by International Business Machines Corporation (IBM), Microsoft (MS), Systems Applications and Products (SAP), and Oracle. But open-source tools are essential for all bigger, smaller companies and academic institutions nowadays. This paper discusses the comparative study of the various tools of Data Science. The prime focus of the comparative study is to discuss the benefits, challenges and applications of the Data Science tools for researchers/user to decide which tools are better for their need.

Key words:
Data Science, Big Data, open source tools, Artificially intelligence, Machine Learning, Machine learning and data Visualisation.

1. Introduction

David Donoho described the importance of data science and the improvement of academia Statistics, how it works better than theoretical statistics. He explained that different universities also offering data science in their academia. He also described that statistics encompasses or gaved the sound of all aspects of Data Scientist definition that uses scientific methods to discover new knowledge from raw data [1].

But it has some limitations because it focuses on statistical modelling rather than data preparation and representation. Statistics also deal with data, whether data is small or big. But using the data science deal with traditional data which computing science could never be accommodate and, it also predicts better data analysis across all sciences [2].

Barlas, P et al. described data extraction techniques of data sciences. For this purpose, they proposed the classification schemes to examine 70 open source data science tools, which include different components form different domains like computer learning and pattern recognition, machine learning, programming, and statistical study and data engineering. They have recorded their features and functionalities of different tools. They classified these tools into four categories. The classification of data science tools is based on main features like data mining characteristics, operational characteristics and project activity [3].

Jevin D. West presented the description and emergence of data sciences. He also explains that science cannot examine by studying individual author or single paper. This can be done by studying hundreds of papers, by getting new ideas not only for a single domain but also for multiple domains, for resolving different big problems. In this way scholars will get the new ideas that they don't have listened to yet and able to find something new. They also forced an advanced future study in which the emergence of data science. His aim is to move the graduate, master, etc. to the new discipline. The study of data science is effective for students because it comes from a different discipline like mathematics, statistics, and probability, programming and computer engineering, etc. The field of data science is giving the platform for researchers from different disciplines to work together to solve the world's hard problems. [4].

Rajeswari, C et al., compared the two tools of big data analytics in this paper for getting the better results and performance. He explained in his paper that data is expanding day by day and new required tools that can easily find interesting information. He opted for two tools R and Tableau and discussed three methods for comparing both R and Tableau.

He also analysed three data-sets from the different sources using both tools R and Tableau and compared both tools based on their performance. Moreover, he concluded that Tableau gives a better result and efficient than R in big data analytics, although both have advantages or disadvantages [5].

Longbing Cao described that big data and data economy play the major part in 21 century for understanding the data DNA and its organism He also gives the definition of data science:

"Data science is the science of Data."

Manuscript received May 5, 2020
Manuscript revised May 20, 2020

He also presented a comprehensive survey in his paper and described the key term and journey of data-science that started from “data analysis” in 1962 and a picture of fundamental aspects of data science [7].

2. Methodology

For the deep study, a review approach was adopted. This helped us to integrate the existing work and identify our strategy in broad manner. We made an in-depth study of Data Science tools, revealing their benefits, challenges and applications. The prime focus of our study is to provide comprehensive knowledge to all users to decide which tools are better to meet their need.

Literature Survey

Data science is a multi-disciplinary field that include other scientific fields such as Statistics, Mathematics, Artificial Intelligence, Data Mining and Machine learning to extract knowledge from structured and unstructured data [8-9].

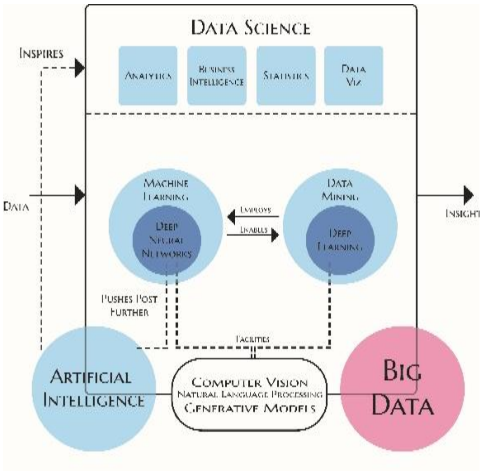


Fig. 1 Flow of Data Sciences in which involve several steps including Analysis, Business understanding, and Data visualisation to deal with Big data by using Machine learning and Data mining techniques.

The term data science is introduced by John Turkey for the reformation of statistics academia, more than 50 years earlier. He pointed to the unrecognised science that is based on data science and its subject is to learn the data. Thus using this concept Bill Cleveland, John Chambers, and Leo Breiman again worked on statistics academia and expanded its boundaries from classical-statistics to theoretical statistics. Chamber focused on data preparation and presentation rather than data modelling. Similarly, the Breiman concept based on prediction instead of inference, and finally, Cleveland suggested the name "Data Science" for fulfilling his vision[2].

Data science is a vast field in which is further divided into different sub-field like artificial intelligence, data mining, deep learning and machine learning that are used to extract significant data from inconsequential data and also helps to create different statistical graphs and patterns easily.

Data mining and Knowledge Data Discovery (KDD) processes are used to analyse the unstructured data and extracting meaningful information and also discovering different patterns by using large and complex datasets[3].

1. Python

Python is developed by Guido van Rossum in the 1980s [10]. It is an open source, interpreted and high-level languages. Specially codes for Python are written in a simple, easy format and extensively used for all kinds of scripts. Over the past years, it becomes popular in the field of data science. It has also many libraries for performing different tasks for different domains like Scipy, Pandas, StatsModels for Statistics. Similarly, Seaborn and Bokeh it is used for visualisation. Python is very popular among the computational scientists, and its use also increases with the passage of time. The major advantage of the python is an object-oriented language, easy to understand, readable and supporting multiple platforms. Besides this, it is slow and difficult to integrate with another language [11].

2. R-Programming

R is the combination of S programming language and lexical scoping [12]. S language was developed by John Chamber in 1976 [13]. R language is most important for performing calculation and visualisation. It is an open source of programming language. Statisticians also using this language for performing different types of analysis, it is also used for parallel processing and handling big datasets, calculation of clusters. R language is also extensible that can easily be integrated with other languages and available as an open-source. Its syntax is very simple and can be easily readable and understandable for users. It is also used in academic institutions and the industries [14].

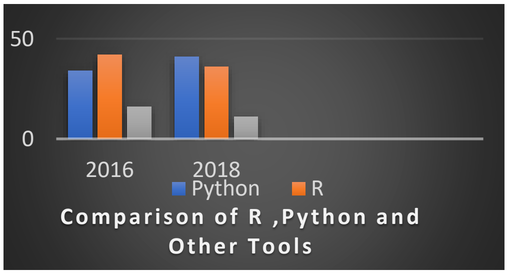
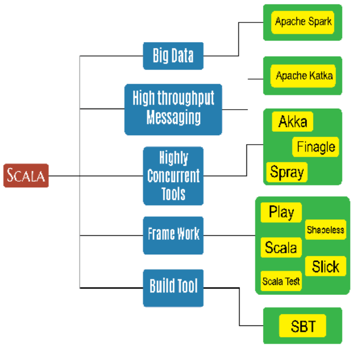


Fig. 7 Comparison of Python, R and Other Languages

3. Scala and Clojure

Both Scala and Clojure are like Julia, the object-Oriented and functional-language that runs on java virtual machine (JVM). JVM is used for saving development time and directly runs on a processor. The popularity of scala increasing in data mining[15]. Clojure was specially designed for concurrent processing and store static datasets. LinkedIn and Netflix are the users of both Scala and Clojure. Scala controls the concurrency of the program and supporting a single object. It also has a possibility of declaring the lazy objects where Clojure also control the concurrency of the program. In Clojure, the program behaves like the data. It also supports java-interoperability[16].



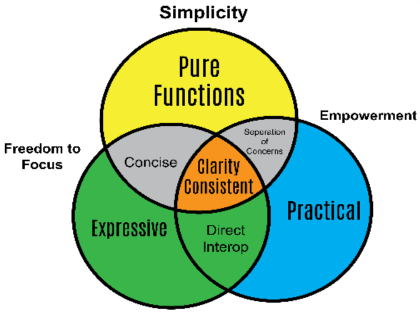


Figure 8: Broad goals of Clojure showing some concepts that underlie the Clojure philosophy and how they intersect.

4. Tibco

Tibco was developed by Vivek Ranadive in 1997 [17-18]. TIBCO StreamBase is used in financial and capital markets. TIBCO (The data transport organisation) items i.e. TIBCO Spotfire is used for information disclosure and visual examination, TIBCO Statistica for information science and machine learning, and TIBCO Stream Base is used for improvement and organisation into a constant foundation. The TIBCO items are open-source inviting and use R/Python scripting for the purpose of life cycle administration. TIBCO has built up its business R motor. TIBCO Enterprise Runtime for R (TERR) that is introduced consequently in TIBCO item appropriations. Note that TIBCO claims the S dialect and the S-PLUS programming item, and a group of designers who have made this conceivable. Tibco provides the facility of lucidity on cloud versus Clarity Enterprise and profiling the data.

5. Excel

Leaner Model was used to develop the Excel software. It is used for analysing unstructured datasets and also used for storing the data in normalised form. It has many processes that include different calculations, conditions, and concatenating the data. It uses the simplest commands that can easily find and replace the word. Similarly, both filtration of data and data sorting can also be possible with this tool. It is also used for password protection, creating different charts and has many built-in formulas[19].

6. SAS

The statistical analysis system (SAS) is the best tool in business intelligence. It deals with unstructured datasets and used for media analytics and data management. It has a procedure that performing various tasks like data-management, analysis, formatting, data-editing, and retrieving. This software is very complicated and costly. The syntax of SAS is easy to understand and has many strong statistical abilities to perform statistical work. It has the ability to analyse the strong data, supporting a different type of Data-format. It also provides Algorithms for the encryption of data[20].

7. MYSQL/Oracle/MongoDB

Structured Query Language is used for storing data in normalised form. These tools are dealing with large datasets and normalise the unstructured data in sequence and categorical. These tools have a different type of keys. MYSQL used RDBMS and the most common softwares, such as XAMPP or WAMP. MongoDB is schema is a schema-less database and dynamically load balance the queries. It is also using aggregation tools for performing data processing in the pipeline. Similarly, MySQL is a relational database and client-server system in which memory allocation system is thread-based and provide the compatibility of multiple commands, but oracle is easy to use and understand as compared to both MongoDB and SQL. It provides a different type of algorithm for data encryption and making highly scalable graphs[21].

8. NoSQL

Johan Oskarsson reintroduced the term NoSQL in 2009 [22]. It is similar to SQL. These are databases used for dealing with structured as well as unstructured data. Because of its high speed and flexibility, NoSQL is the most useful tool in data science. These databases are usually handled through API. No-SQL deals with large datasets like picture video machine to machine communications and convert them in a structured form. This is an advanced tool for dealing with big data. No SQL database supports the non- relational model. It can also support multiple data models, that will help the user to reuse the data within a different type of data model. It uses peer-to-peer architecture, which helps to Handel the complexity of cloud applications. It is highly scalable which helps to improve performance. The no SQL database easily maintains the unstructured, structured, and semi-structured data.

9. Hadoop

Hadoop is open-source software framework that was used for storing and processing the big data. It was developed in 2005 to support for Nutch search engine project. Hadoop is written in java for storing and tackling the big data, this software is specially designed for the data scientist to control the clunky datasets. It provides different frameworks that are used for processing big data. Similarly, Failure detection of hardware against data/applications and handling the issues of the application layer can also be possible with this software. It is an open-source and using hardware for storing large amounts of data. This software can store much data as can be stored without pre-processing. It provides scalability and flexibility and also helpful in Distributed- Processing and provides the facility of data locality. It also handles the repetition[23].

10. RapidMiner

A Data-Science software which is also known as YALE developed in 2001 and its name changed from YALE to the rapid miner in 2007 [24]. Rapid Miner is a very simple and fast software that mine the data perfectly and efficiently. It removes the complexity of data by using the advanced queuing mechanism. It has many features like users can use built-in templates and pre-defined connections. It supports efficiently for model delivery, accesses any type of data, data exploration, data cleaning/preparation.

11. WEKA

Weka is redeveloped in 1997 from java in the university of Waikato [25]. Weka is a mostly use open source machine learning platform. Weka is an arrangement of machine learning that can be connected to an informational collection specifically or called from your own Java code. Weka has different types of instruments for information pre-preparing, bunching relapse, arrangement, affiliation, principles, and perception. Weka provides only the facility of automatic selection of features and the opportunity of data processing. It also gives the facility of the selection and visualisation of features that helps in the mining of data. It is very useful in educational purposes and can also be portable [26].

12. Google Cloud Computing:

The first tool of Google was announced in 2008 and more services were added continuously and called Google Cloud Platform (GCP). It is based on a solid infrastructure and uses a wide range of host services, such as storage, computers, big data, machine learning, networking and internet of things (IoT) through an application programming interface (API). GCP also offers on the same framework end-user products such as YouTube and Google Search. It supports both the window and Linux and also provides the Consistence and Security[27].

13. DataRobot

A robust tool gives the option of machine learning for data scientists and also used to build the predictive model accurately in less time. DataRobot enhancing machine learning with advanced features. It uses text mining for detecting feature engineering and best for data pre-processing. It also uses a distributed system for scaling up to large amounts of data. The Data Robot helps in the preparation of data and data perfection.

Moreover, it supports the distributed architecture, easy to use, understand, and work promptly as compared to the other software[28].

14. Paxata

Paxata is a company that develops software in January 2012. It is self-adaptive data preparation platform that allows the user to get, analyse, re-model and combine the data. Paxata deals with unstructured or raw data and converts into useful information form automatically by using the techniques of machine learning by just on simple clicks without writing the code. It is used for combining data from different sources and checking out the quality of data. Paxata also using the algorithm and machine learning techniques for data preparation and detect the record of any person if it is formatted differently. It can be easily integrable.

15. Trifacta

Trifacta is a company in San Francisco designed software for data preparation and also work on cloud. Different machine learning, Data-visualisation techniques can also be used by Trifacta. This tool is designed for cleaning the raw data [29]. Its application automatically organizes the data is a structured form and do this process very fast. By using this tool, user's can use Pattern Clean, Cluster Clean, and Reference Clean to resolve data quality issues. Transformation of complex data and raw data into clean form that may be possible through Trifacta to enhance the value of an Enterprise's Big Data. Trifacta introduces a new feature that addresses the data quality issues regarding its format and standardisation. It helps in Data Enrichment, Data Preparation and also providing the facility of Profiling Visual Data.

16. Feature Labs

It was developed in 2015, specially designed for feature engineering and machine learning and to build predictive models. Data scientists utilised the techniques of the machine and artificial intelligence using Feature Labs to build the new product that will be useful for an organisation in the future. Some cloud-based tools that are promoting the machine learning techniques such as Alteryx, Qubole, Lumen Data these tools may be used to build predictive models and transform the unstructured data in a structured form. Feature Lab is also a tool which is useful in data transformation. It helps in developing different types of models and improve the flow of the work, also very efficient and scalable[30].

17. Algorithms.io

This tool is used for streaming data from different connected devices and changes the raw data into real-time information. It is also used for the build APIs that connect machine learning with mobile apps and the web. Algorithm.io is a tool which is useful for data streaming, helpful in building APIs, and provides the opportunity of the Data Transformation [31].

18. Apache Hadoop

Apache Hadoop is an open-source software that is used for computation using the network of computers. It was designed for computer cluster and commodity hardware[32]. It is also used for processing large datasets across multiple devices or computers by using different programming techniques. Apache Hadoop handles the issue of application layers instead of hardware level and a useful cluster of higher-end hardware. The Hadoop is maintained by apache Hadoop [33-34].

19. Apache HBase

HBase is based on java and open source programming language. The Hadoop database used to store structured data and also mine the useful information from raw data. Data scientists using this tool to read/write the big data. It has an automation process and creates a different type of tables. Generally, HBase is used when there is a need to access big data continuously and irregularly and for working on the huge level of HDFS. It supports table sharding, automatic fail-over, and supporting APIs [35].

20. MATLAB

MATLAB (matrix laboratory) is multi-paradigm for numerical that is used to integrates programming, computation, visualisation, and makes the environment easier to use, where the problem and their solution are expressed in mathematical form to develop a different type of models as well as algorithms or applications. It is an interactive environment where dimensioning is not required because the element act as an array and many computational problems can be easily solved, especially matrix, vector formulations, and fractions. It has many built-in mathematical functions as well as a vast collection of algorithms to solve different mathematical problems. Similarly, it also allows the user to write a program in C and FORTRAN language and interact with MATLAB. It is also used for used for data visualisation and for building a graphical user interfaces or also allows the customisation of graphics but slow in working [36].

21. Tensor flow

It is a framework of machine learning that is open source created by Google that is used to train, build, and design the deep learning models. Its libraries can also be useful for numerical computations that are performed with a data flow graph. In these graphs, nodes are mathematical operations, and edges are data that is multidimensional form or array. It can also be used to develop solutions using deep learning techniques with deep learning and highly scalable across huge datasets. It can also make the calculation easier [37].

22. Julia

It is an open-source high-performance dynamic programming language that is more convenient for vector and matrice. It is super-fast and more flexible, but the plotting is very difficult. The syntax of Julia is simple and familiar with other languages. It also allows parallel and distributed computing and can easily be integrated, It is easy to understand and read for technical user those are familiar with technical languages. It also provides the best complied, numerical accuracy, and also has a very extensive library for mathematical functions and can also be integrable with libraries of C and Fortran [39].

23. Tableau

Tableau is an application that deals virtually with structured data and generates interactive graphs and reports in less time. It is a business application that allows browsers-based analytics and it is an online software for sharing, distributing the content that is created on the tableau. Similarly, scripting is not required and easy to use and get a solution within minutes. It is also easy to understand due to continuing drag-and-drop features that can be easily translated into queries and also providing the programming-free environment. The speed of tableau is very high, so this software can easily evaluate millions of rows in less moment. Its dashboard is very interactive and data outcomes can also be shared in fewer clicks after examining the data. Besides this, it is very difficult to integrate with other applications [40].

24. KNIME

KNIME (Konstanz Information Miner) is an open-source tool that is used for data modelling, data analysis, and predictive analysis; also it has a graphical interface but not based on the scripting language. It can also integrate with deep learning and data mining through a modular data pipelining concept. KNIME also has nodes repository that allows drag and drop the nodes. It is used in pharmaceutical research [41] and also used for model evaluation as well as feature selection for doing this, it uses a forward feature selection method, this technique can be also used for checking the accuracy of the model [42].

Comparative study of Data Science Tools

We attempted and studied a lot of tools for Data Science. The Open source tools and most demanding tools, that are very important for governments, academic institutions, industries, banking, financial institutions, private companies, researchers and as well for students. Comparative studies of these tools along with their benefits, challenges and applications are discussed in the table [1].

Table [1].Brief Overview of Benefits, Challenges, and Applications/Projects of Data Science tools.

Name	Benefits	Challenges	Applications
Python	Its strong and vigorous toolset with simple syntax helps scientists to generate code that is small and easy to understand. Python provides new facilities with the help of its Object-Oriented and updated structure [43]. The beginner can easily use this language because of its simple syntax which is easy to learn, use, and understand [44].	The real Python program works considerably much slower than the code which is compiled. There is a few scientific libraries related to documentation and Fortran which helps new users [43].	It is helpful in the detection of the face, Machine Learning. It is also helpful in console-based, audio-based enterprise, and video-based applications [45].
R Languages	R language is obtainable for a variety of software and hardware. The R presents a large variety of functions i.e. manipulation of the data, the statistic of modelling, and also helpful in graphics it is also extensible [46]. R gives many facilities regarding operations of the machine learning that includes regression, a classification it also gives facilities regarding the artificial neural network [47].	As compared to python and MATLAB languages the packages of R language are much slower. It is not an easy or simple language. It is not easy to learn because of its steep curve, the user having less experience of programming than is a difficult language for them [47].	It is helpful in the banking sector for the modelling of risk, it's also effective in the discovery of drugs, It is also beneficial in manufacturing companies for evaluating the best opinion[47].
Scala	Scala is a language that is much compatible with Java, its complex coding helps in improving/increasing the performance. With the help of Scala, the developer can generate a code that is concise and functional [48].	There exist a few numbers of developer that use Scala. Its mix/hybrid functionalities sometimes become harder to understand and recognise [47].	Advanced patterns, API [49].

Clojure	Clojure presents various valuable/useful aspects that include the integration of the JVM by assembling to byte code,(STM) Software Transaction Memory. It gives effective tools for the concurrency of data. Clojure has a simple syntax that is easily understandable and it can transform the data through macros. It also has standard libraries [50].	The dense code is difficult to understand [51]. Because it inter-operates uses with Java, it is difficult to use [51].	Clojure use by the Chartbeat, Puppet for trapping blind persons. Clojure also provides its services in the cloud as SAAS [51].
Tibco	TIBCO is a Data Science Software that helps and allows the organisations to produce automation in their workflow. It is also helpful in business to solve real-world problems by using machine learning algorithms, it's also effective in solving a complicated problem and give the solution to produce the best outcomes [52].	Manual training is required for an understanding of the advanced aspects [53]. For prediction and analysis of statistical data, its integration with R language is required [53].	It is helpful in digital factories for intelligent manufacturing purposes [54]. It is helpful in the detection of anomalies and gives the solution for the time-consuming problems [55].
Excel	Prevent from unnecessary coding, make fewer errors, controlling error is easier, assist corporations to implement a consistent style, and give business prospects in selling and buying modules of spreadsheets [56].	The use of Excel is not reliable and helpful in the detection of fraud. Testing is also not possible, not design for the collective work [57].	It gives the facilities to summarise the data by the use of graphs and the charts. Once the data is systematically stored in Excel, then it will be used for multiple purposes. It is helpful in searching, sorting, and analysing the data which makes the work much easier [58].
Statistical analysis system (SAS)	It is helpful in the manipulation of the data for making it effective. They give a better output quality. SAS can manage massive datasets and give opportunities for the documentation [59-60]. It also can handle massive datasets [61].	The representation of graphics is not proper. Problematic Text Mining than R[61].	For reliable clinical research and estimating the intelligence of business SAS is used [62].
MySql	Provide On-Demand Scalability having Complete Control on Workflow with Data Security [63].	Lack of advancement has other relational database management systems [64] and also has stability Issues [65].	Sales Management System, Car Sales and Service System[66].
Hadoop	Hadoop having high Scalability and Resilient to failure [67]. Provide efficient Authentication and Security [68].	Not suitable for Small amount Data and also has some potential Stability Issues [67]	Speech Analysis, Cloud Hosting [69].
Rapid Miner	Having user-friendly interface Robust features that are powerful for applying analytics on real-life data[70]. Rapid Miner also has enormous flexibility[71].	Partitioning ability is Limited for training and testing the dataset[71].	Helpful in data mining and data analysis[72].
Algorithms.io	Provide 99% prediction accuracy on the state of the sensor with a machine learning algorithm and on time series data in Classification & Anomaly Detection.[73] Algorithms.io is beneficial for time-series database and analytics.[73]	-	API for Developers [74]
Apache HBase	Provide modular and Linear scalability with reads and writes Consistent, in strict manner, also provide sharding of table in the automatic	Does not support transaction due to Single point of failure can cause a serious issue as well as supporting issue with	Cryptographic software[75]. storing genome sequences [77]

	form[75].	SQL structure because Apache HBase requires a new design if migrating from RDBMS to HBase servers [76].	
Apache Hadoop	Highly Scalability and Fault Tolerance capabilities [78]	Application only for a small volume of data. [79]	YARN Hortonworks Sqoop [80]
Matlab	Easily processing of images and video and creating simulations and also provide effortless Testing and Implementation of algorithms [81]. External libraries can be utilised easily like OpenCV[81].	Matlab can be easily accessible and efficiently perform the task by using functions but consumes much time to find the answer to my queries[33].	Identification of Vehicle Number plates and face Recognition System[83] Digital Extraction of Brain tumour from MRI using MATLAB[84]
Tensor flow	Provide a better computational graph for visualisations, which areas compared to different libraries like Torch and Theano and very scalable[85]. Having a high level of adoption for finding the resources of TensorFlow[97]	With steep learning curve, Tensor flow having low-level functionality [86] Error detection and debugging are very difficult due to the odd structure. [86].	Speech recognition, Tagging of objects in videos, Summarisation of Text. [87]
tableau	tableau has Remarkable Capabilities for Image Visualisation also having Multiple Connections for Information Supply[88].	Having High Cost with Poor Versioning and with a lot of Security Issues [89].	Helpful to identify the risk of students[90].
Knime	Provide an extension to Big data and suitable for blending of data also having local automation, powerful analytics, and difference of workflow [91].	When loadings looper meta nodes and Old cross-validation, both will be reset [92].	Developing machine learning models and prototypes of workflow [93]
WEKA	Using GNU General Public License Weka software is available in the free version.[94] The Weka is a collection of visualisation tools and GUI for accessing different functionalities Like data analysis and predictive modeling[94]. Weka has various types of machine learning algorithms to deal with real-world data mining issues[95] Compatible and runs on almost any advance computing platform[96]	WEKA is memory intensive due to running on java and required licenses from one or more corporate entities [97].	For performing data-mining tasks, many machine learning algorithms using Weka[98]. Weka is applicable for Data Visualisation and Pre-processing [99].
DataRobot	DataRobot is easy to Integration, also having the better performance of data accuracy [100]. Provide Plugging and playing with Hadoop [101].	Design and suitable for small businesses and supporting web-based applications [100].	Managing AI Cloud [102].
Paxata	Paxata offers various types of deployment such as Paxata-hosted cloud, private cloud, or a hybrid of all of these options and easy to learn [103]. Similarly, by using built-in Paxata	The relational database is not supported as well as not used for complex ETL by Paxata[103]. Paxata deals with a small amount of	Helpful in machine learning, semantic algorithms distributed computing techniques[104].

	function, preparation of projects can be easily done [103.	data and having lacks functionality that makes it smoother [103].	
Trifacta	Visual Data representations are done automatically that depend on the content existed in visual profile which is most compelling also beneficial for predictive Transformation [105].	For performing the work on large files, then be careful about the dependencies on data quality and statistical summaries[106].	Data cleaning and preparation according to Latest platform regarding cloud data lakes and warehouses [107].

3. Conclusion

In this paper, we have discussed the various tools of Data Science. Open source tools and other tools are discussed with their benefits, challenges and applications. But no doubt, the open source tools such as Python, R-Programming and Weka were rejected by big companies such as IBM, MS, SAP and Oracle. Currently aforementioned companies are using open source tools, Python and R-Programming. The crux of our concluding remarks is that Python and R-Programing are free for every one and could be used for data modelling, data visualisation and a lot of other applications. A numerous tools also discussed and provided their useful information for researcher/user to decide which tool is useful for their application such as parallel processing, visualisation, Data Wrangling, Data Management and Big Data as well.

References

[1] Christoph Bernau, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron, and Lorenzo Trippa. Cross-study validation for the assessment of prediction algorithms. Bioinformatics, 30(12):i105–i112, 2014.

[2] Donoho, D. (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745-766.

[3] Barlas, P., Lanning, I., & Heavey, C. (2015). A survey of open source data science tools. International Journal of Intelligent Computing and Cybernetics, 8(3), 232-261.

[4] West, Jevin D. "The Science of Data Science." (2016).

[5] Rajeswari, C & Basu, Dyuti & Maurya, Namita. (2017). Comparative Study of Big data Analytics Tools: R and Tableau. IOP Conference Series: Materials Science and Engineering. 263. 042052. 10.1088/1757-899X/263/4/042052.

[6] Gupta, Vikas. "Prof. Devanand, "A survey on Data Mining: Tools, Techniques, Applications, Trends, and Issues,"." International Journal of Scientific & Engineering Research 4: 20-33.

[7] Longbing Cao. 2017. Data science: A comprehensive overview. ACM Comput. Surv. 50, 3, Article 43 (June 2017), 42 pages. DOI: http://dx.doi.org/10.1145/3076253

[8] Dhar, V. (2013). "Data science and prediction". Communications of the ACM. 56(12): 64–73. DOI:10.1145/2500499.

[9] Leek, J. (2013). The key word in'Data Science'is not Data, it is Science. Simply Statistics, 12.

[10] Hayes B.(January 2, 2018).Data Science, Machine Learning. http://businessoverbroadway.com/2018/01/02/most-used-data-science-tools-and-technologies-in-2017-and-what-to-expect-for-2018/

[11] . Kuhlman, Dave. "A Python Book: Beginning Python, Advanced Python, and Python Exercises". Section 1.1. Archived from the original (PDF) on 23 June 2012.

[12] Cai, Xing & Langtangen, Hans Petter & Moe, Halvard. (2005). On the Performance of the Python Programming Language for Serial and Parallel Scientific Computations. Sci. Program.. 13. 31-56. 10.1155/2005/619804.

[13] Morandat, Frances; Hill, Brandon; Osvald, Leo; Vitek, Jan (2012). "Evaluating the design of the R language: objects and functions for data analysis" (PDF). ECOOP'12 Proceedings of the 26th European Conference on Object-Oriented Programming. Retrieved 17 May 2016.

[14] [Online].Available: https://www.r-project.org/about.html [Accessed 7 July 2019.].

[15] Jessica Davis. 2016. 10 Programming Languages And Tools Data Scientists Used. Retrieved from http://www.informationweek.com/devops/programming-languages/10-programming-languages-and-tool s-data-scientists-use-now/d/d-id/1326034.

[16] [Online].Available:https://www.packtpub.com/big-data-and-business-intelligence/clojure-data-science [Accessed 7 July 2019].

[17] "A Look Back: Vivek Ranadive and TIBCO". TIBCO Software Inc. Retrieved April 26, 2017.

[18] Black, Debra (January 26, 2012). "Davos Elite Get Their Own Facebook – Move Over Twitter and Facebook – There's a New Platform in Town: It's Called TopCom and It's Geared to the World's Leaders". Toronto Star. Retrieved July 19, 2013.

[19] Capterra. 2016. Top Reporting Software Products. Retrieved from http://www.capterra.com/ reporting-software/.

[20] Big Data Analytics: An Assessment of Demand for Labour and Skills, 2012-2017. Retrieved from https://www.thetechpartnership.com/globalassets/pdfs/research-2014/bigdata_report_nov14.pdf Report. SAS/The Tech Partnership.

[21] [Online].Avialable: http://www.computerworld.com [Accessed 7-2-2020].

[22] "NoSQL 2009". Blog.sym-link.com. 12 May 2009. Retrieved 29 March 2010.

[23] [Online].Avialable: https://www.guru99.com/big-data-tools.html [Accessed 7-2-2020].

[24] German Predictive Analytics Startup Rapid-I Rebrands As RapidMiner", TechCrunch, November 4, 2013

[25] Ian H. Witten; Eibe Frank; Len Trigg; Mark Hall; Geoffrey Holmes; Sally Jo Cunningham (1999). "Weka: Practical Machine Learning Tools and Techniques with Java Implementations" (PDF). Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems. pp. 192–196. Retrieved 2007-06-26.

[26] Witten, I., H., Frank, E., Hall, M., A. (2011). Data mining: practical machine learning tools and techniques.

[27] [Online].Avialable : https://cloud.google.com/ [Accessed 7-2-2020].

[28] DataRobot. 2016. DataRobot. Retrieved from https://www.datarobot.com/.

[29] Blattberg, Eric (October 28, 2013). "Paxata grabs $8M to help data scientists skip the dirty work". VentureBeat. Retrieved June 19, 2014

[30] Black, Doug. "Trifactas Data Wrangling Decoder Ring Homogenizes Polygot Data Lakes". Enterprise Tech. Enterprise Tech. Retrieved 11 February 2016.

[31] [Online] Available.: https://www.owler.com/company/featurelabs [Accessed 7-2-2020].

[32] Stringfellow, A.(2017,August 21) Top Tools for Data Scientists: Analytics Tools, Data Visualization Tools, Database Tools, and More. Retrieved from https://www.ngdata.com/top-tools-for-data-scientists.

[33] Judge, Peter (2012-10-22). "Doug Cutting: Big Data Is No Bubble". silicon.co.uk. Retrieved 2018-03-11

[34] Woodie, Alex (2014-05-12). "Why Hadoop on IBM Power". datanami.com. Datanami. Retrieved 2018-03-11.

[35] Hemsoth, Nicole (2014-10-15). "Cray Launches Hadoop into HPC Airspace". hpcwire.com. Retrieved 2018-03-11.

[36] "Apache HBase," Apache, 22 December 2014. [Online].Available: http://hbase.apache.org/. [Accessed 06 January 2020].

[37] [Online].Available: https://www.trustradius.com/reviews/matlab-2018-01-19-10-57-25 [Accessed 25-Dec-2019]

[38] [Online].Availablehttps://www.datacamp.com/community/tutorials/tensorflow tutorial?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=m&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=1t1&utm_creative=278443377095&utm_targetid=aud-392016246653:dsa-498578051924&utm_loc_interest_ms=&utm_loc_physical_ms=1011082&gclid=EAIaIQobChMI6oTi9Ymc5AIVmK3tCh229gciEAAYASAAEgIpM_D_BwE[accessed 2-6-2019].

[39] [Online].Available: https://medium.com/@jayeshbahire/introduction-to-julia-d3341b9cd24c [Accessed 7-7-2019]

[40] [Online].Available: https://intellipaat.com/blog/tutorial/tableau-tutorial/introduction-tableau/ [accessed 5-6-2019]

[41] [Online].Available:https://towardsdatascience.com/guided-analytics-using-knime-analytics-platform-b6543ebab7e2 [Accessed 5-Jun-2019]

[42] Tiwari, Abhishek; Sekhar, Arvind K.T. (October 2007). "Workflow based framework for life science informatics". Computational Biology and Chemistry. 31 (5–6): 305–319. doi:10.1016/j.compbiolchem.2007.08.009.

[43] Lin, J. W. B. (2012). Why Python is the next wave in earth sciences computing. Bulletin of the American Meteorological Society, 93(12), 1823-1824.

[44] [Online].Available:https://www.invensis.net/blog/it/benefits-of-python-over-other-programming-languages/ [Accessed 5-July-2019].

[45] [Online].Available https://www.educba.com/uses-of-python/[Accessed 5-July-2019].

[46] [Online].Available:https://www.dummies.com/programming/r/the-benefits-of-using-r/ [Accessed 5-July-2019].

[47] [Online].Available:https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/[Accessed 5-July-2019].

[48] [Online].Available:https://dzone.com/articles/advantages-of-scala [Accessed 5-July-2019].

[49] [Online].Available:https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/https://github.com/gothinkster/scala-play-realworld-example-app[Accessed 19-Aug-2019]

[50] [Online].Available:https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/http://www.informit.com/articles/article.aspx?p=2464444 [Accessed 5-July-2019].

[51] [Online].Available: https://www.quora.com/What-are-the-downsides-of-Clojure [Accessed 5-July-2019].

[52] [Online].Available:https://www.tibco.com/products/data-science Accessed 5-July-2019].

[53] [Online].Available:https://www.getapp.com/business-intelligence-analytics-software/a/tibco-spotfire/reviews/ [Accessed 5-July-2019]

[54] [Online]. Available: https://www.tibco.com/solutions/business-activity-monitoring [Accessed 5-July-2019].

[55] [Online].Available:https://www.tibco.com/solutions/anomaly-detection [Accessed 5-July-2019].

[56] Paine, J. (2008). Excelsior: Bringing the benefits of modularisation to Excel. arXiv preprint arXiv:0803.2027.

[57] [Online]. Available:https://blog.blackcurve.com/11-disadvantages-of-using-excel-to-manage-your-pricing [Accessed 19-Aug-2019].

[58] [Online]. Available:https://magoosh.com/excel/10-best-uses-microsoft-excel/ [Accessed 19-Aug-2019]

[59] Fernandez, G. (2010). Statistical data mining using SAS applications. CRC press.

[60] [Online]. Available: https://www.researchgate.net/post/Advantages_of_using_SAS_software [Accessed 19-Aug-2019]

[61] [Online].Available:https://www.newgenapps.com/blog/sas-review-what-is-it-pros-cons-suitability [Accessed 19-Aug-2019]

[62] [Online].Available:https://data-flair.training/blogs/sas-application/ [Accessed 19-Aug-2019]

[63] [Online].Available:https://www.datamation.com/storage/8-major-advantages-of-using-mysql.html [Accessed 7-Sep-2019].

[64] [Online].Available:https://www.smartfile.com/blog/the-pros-and-cons-of-mysql/ [Accessed 7-Sep-2019].

[65] [Online].Available:https://www.datarealm.com/blog/five-advantages-disadvantages-of-mysql/[Accessed 7-Sep-2019]

[66] [Online]. Available:https://www.freeprojectz.com/mysql-projects [Accessed 7-Sep-2019]

[67] [Online].Available:https://www.mindsmapped.com/hadoop-advantages-and-disadvantages/ [Accessed 7-Sep--2019]

[68] [Online].Available:https://www.quickstart.com/blog/5-business-benefits-of-hadoop [Accessed 7-Sep-2019]

[69] [Online].Available:https://www.dezyre.com/article/8-common-hadoop-projects-and-spark-projects/182 [Accessed7-Sep-2019]

[70] [Online].Accessible:http://comparecamp.com/rapidminer-review-pricing-pros-cons-features/ [Accessed 7-Sep-2019]

[71] Gulia, P. Comprehensive Study of Open-Source Big Data Mining Tools.

[72] [Online].Available:http://mtechproject.com/computer-science/rapid-miner-projects/ [Accessed 7-Sep-2019]

[73] [Online].Available: https://stackshare.io/stackups/algorithms-io-vs-nanonets [Accessed 7-Sep-2019]

[74] [Online].Available: http://blog.algorithms.io/ [Accessed 5-Jun-2019]

[75] [Online].Available:http://hbase.apache.org/ [Accessed 7-Sep-2019]

[76] [Online].Available:https://data-flair.training/blogs/hbase-pros-and-cons/ [Accessed 7-Sep-2019]

[77] [Online]. Available: https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781783985944/1/ch01lvl1sec15/applications-of-hbase [Accessed 5-Nov-2019]

[78] [Online]. Available: https://www.guru99.com/learn-hadoop-in-10-minutes.html [Accessed 5-Nov-2019]

[79] [Online].Available:https://www.knowledgehut.com/blog/big-data/top-pros-and-cons-of-hadoop [Accessed 25-Dec-2019]

[80] [Online].Available:https://www.quora.com/Where-can-I-learn-Hadoop-real-time-projects-for-free [Accessed 25-Dec-2019]

[81] [Online].Available: https://guides.libraries.uc.edu/c.php?g=461109&p=3152738 [Accessed 25-Dec-2019]

[82] [Online].Available:https://www.mathworks.com/matlabcentral/answers/82408-the-advantages-of-matlab-over-other-programing-languges-for-image-processing [Accessed 25-Dec-2019].

[83] [Online].Availablehttps://www.skyfilabs.com/project-ideas/latest-projects-based-on-Matlab [Accessed 25-Dec-2019]

[84] [Online].Available:https://www.skyfilabs.com/project-ideas/digital-extraction-of-brain-tumor-from-mri [Accessed 25-Dec-2019]

[85] [Online].Available:https://data-flair.training/blogs/tensorflow-pros-and-cons/ [Accessed 25-Dec-2019]

[86] [Online].Available:https://medium.com/swlh/googles-artificial-intelligence-system-tensorflow-pros-and-cons-464c4107a6fc [Accessed 25-Dec-2019].

[87] [Online].Available:https://dzone.com/articles/tensorflow-for-real-world-applications [Accessed 25-Dec-2019].

[88] [Online].Available:https://data-flair.training/blogs/tableau-pros-and-cons/ [Accessed 25-Dec—2019]

[89] [Online].Available:https://www.sam-solutions.com/blog/tableau-software-review-pros-and-cons-of-a-bi-solution-for-data-visualization/ [Accessed 25-Dec-2019]

[90] [Online].Available:https://www.tableau.com/learn/articles/business-intelligence-examples [Accessed 25-Dec-2019].

[91] [Online].Available:http://comparecamp.com/knime-analytics-platform-review-pricing-pros-cons-features/ [Accessed 25-Dec-2019]

[92] [Online]. Available:https://www.knime.com/known-issues-0 [Accessed 5-Jun-2019]

[93] [Online].Available:https://www.knime.com/knime-analytics-platform. [Accessed 3-Jan-2020]

[94] [Online].Available:http://en.wikipedia.org/wiki/Weka_(machine_learning). [Accessed 3-Jan-2020]

[95] [Online].Available: http://www.cs.waikato.ac.nz/ml/weka/[Accessed 3-Jan-2020]

[96] Jagtap, S. B. (2013). Census data mining and data analysis using WEKA. arXiv preprint arXiv:1310.4647.

[97] [Online].Accessible: https://alternative.me/weka[Accessed 3-Jan-2020]

[98] [Online].Available:https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/ [Accessed 25-Dec-2019]

[99] [Online].Available: http://mtechproject.com/computer-science/weka-projects/[Accessed25-Dec-2019].

[100] [Online].Available:http://comparecamp.com/datarobot-review-pricing-pros-cons-features/

[101] [Accessed 25- Dec-2019].

[102] [Online].Available: https://stackshare.io/datarobot [Accessed 4-Jan-2020].

[103] [Online].Available: https://www.datarobot.com/platform/managed-cloud/ Accessed 4-Jan-2020].

[104] [Online].Available: https://blog.aptitive.com/pros-and-cons-to-using-paxata-53af380ef836 [Accessed 4-Jan-2020].

[105] [Online].Available:https://www.sdxcentral.com/listings/paxata/ [Accessed 4-Jan-2020]

[106] Online].Available:https://stackshare.io/trifact a [Accessed 4-Jan-2020].

[107] [Online].Available: https://www.computerworld.com/article/3104769/data-wrangling-tool-trifacta-aims-to- ase-analysis-pain.html [Accessed 4-Jan-2020].

[108] [Online].Available:https://www.trifacta.com[Accessed 4-Jan-2020].

Dr Haroon Ur Rashid Kayani is Data Scientist and Consultant for promoting and training the Data Science Education in Pakistan. He was granted dual research award from the Engineering and Physical Sciences Research Council, UK and National Physical Laboratory, UK for PhD and completed his Ph.D. from the University of Warwick, UK. He also completed his M.Phil. & Master degrees from the University of

Aberystwyth, Wales, UK by obtaining scholarship award from the Ministry of Education, Islamabad, Pakistan. He has been teaching graduate, undergraduate students at national and international universities since 1987. Kayani's teaching and research interests primarily include Big Data, Data Science, Data Security, Data Privacy, Cyber Security, Data Mining, Machine Learning, AI, IoT, Modelling & Simulation and Data Analysis.

Dr Saba Khalil Toor has more than 21 years of working experience, with 6 year in software development industry and 15 years in academia. In Software industry, she worked at leading software houses of the country such as CresSoft (Pvt.) and Vroom (Pvt). In academia, Dr Saba worked at Virtual University of Pakistan for more than 14 years as assistant professor and Head of Computer Science department. There she was also a Co-director for Centre of Education and Technology and conducted research on several educational tools/technologies. She has also been Director, Centre for Learning and Software Development at Imperial College of Business studies and worked to establish this center. Saba's research interests primarily include Educational Technology, eLearning and Software Engineering. She has written several papers and presented them in Pakistan as well as abroad.

Hafiz Burhan Ul Haq received the BSc degree from Punjab University in 2015 and MSc degree from Lahore Garrison University in 2017. He is doing MSCS from Lahore Garrison University. Burhan's research interests primarily include Web Development, Deep Learning, Artificial intelligence and Networking.

Sadia received the B.com-IT degree from Punjab University in 2015 and Mcs degree from Lahore Garrison University in 2017.He is doing MSCS from Lahore Garrison University. Burhan's research interests primarily include Web Development, Deep Learning, Artificial intelligence and Networking.

Imran Khalid received the BSCS degree from Lahore Garrison University in 2017. He is doing MSCS from Lahore Garrison University. Imran's research interests primarily include Apps &amp; Web Development, Particle Swarm Optimisation, Deep Learning and Image Processing.