A Review on Data Science Technology and Big Data Analytics with Python

Anamika Pandey
Assistant Professor
Department of Computer Science & Engineering
IIMT College of Engineering,
Greater Noida, U.P., India

Sovers Singh Bisht
Assistant Professor
Department of Computer Science & Engineering
IIMT College of Engineering,
Greater Noida, U.P., India

***Abstract*— Data Science is an area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data. A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms.**

Keywords— Analysis, Bigdata, Data Science, Data Set, Python, Structured, Unstructured.

I. Introduction

Data Science is the extraction of learning from substantial volumes of information that are unorganized or unstructured, which is a continuation of the field of information mining and perceptive investigation, otherwise called information disclosure and information mining. In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. It provides evolutionary breakthroughs in many fields with collection of large datasets. In general, it refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications. These are available in structured, semi structured, and unstructured format in petabytes and beyond.

It is expected that the growth of big data is estimated to reach 25 billion by 2015 [1]. From the perspective of the information and communication technology, big data is a robust impetus to the next generation of information technology industries [2], which are broadly built on the third platform, mainly referring to big data, cloud computing, internet of things, and social business. Generally, Data warehouses have been used to manage the large dataset. In this case extracting the precise knowledge from the available big data is a foremost issue. Most of the presented approaches in data mining are not usually able to handle the large datasets successfully. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications. A fundamental problem is how to quantitatively describe the essential characteristics of big data. There is a need for epistemological implications in describing data revolution [3]. Additionally, the study on complexity theory of big data will help understand essential characteristics and formation of complex patterns in big data, simplify its representation, gets better knowledge abstraction, and guide the design of computing models and algorithms on big data [2]. Much research was carried out by various researchers on big data and its trends [4], [5], [6].

II. BASIC STEPS OF DATA SCIENCE

The three segments included in data science are arranging, bundling and conveying information (the ABC of information). However bundling is an integral part of data wrangling, which includes collection and sorting of data. However what isolates data science from other existing disciplines is that they additionally need to have a nonstop consciousness of What, How, Who and Why. A data science researcher needs to realize what will be the yield of the data science transform and have an unmistakable vision of this yield. A data science researcher needs to have a plainly characterized arrangement on in what manner this yield will be accomplished inside of the limitations of accessible assets and time. A data scientist needs to profoundly comprehend who the individuals are that will be included in making the yield. The steps of data science are mainly: collection and *preparation* of the data, alternating between running the *analysis* and *reflection* to interpret the outputs, and finally *dissemination* of results in the form of written reports and/or executable code. The following are the basic steps involved in data science.

(a) Data wrangling and munging

Collecting data from relevant areas and the process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption and manipulation of the data with the help of semi-automated tools is referred to as data wrangling[7] or munging[8].Sorting out data includes the physical stockpiling

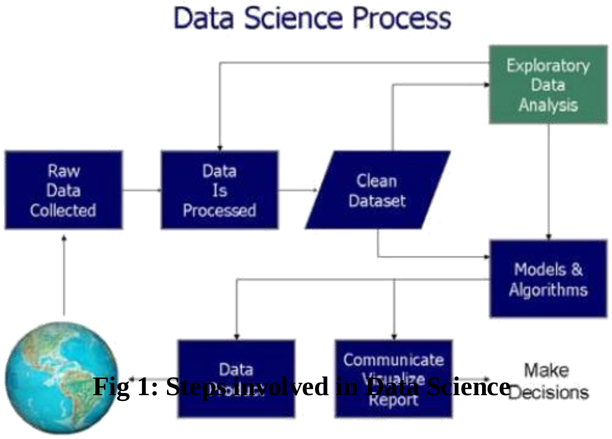

Fig 1: Steps involved in Data Science

and arrangement of information and joined best practices in information administration. It basically includes moving individuals and frameworks from current to new (left to right) and from learner to master (start to finish). Propelling advances and abilities is the pith of development.

Bundling data is the next step that follows arranging data. Bundling data includes consistently controlling and joining the fundamental crude information into another representation and bundle. Bundling data is actually the opposite of sorting out data and includes moving individuals and frameworks from new to current (right to left) and from master to apprentice (base to beat). This is the specialty of making things basic yet not less complex.

(b)Data Analysis

Analysis or investigation of data is a procedure of assessing, changing, and demonstrating information with the objective of finding helpful data, recommending conclusions, and supporting decision-making. The data is processed using various algorithms of statistics and machine learning to extract meaning and useful conclusions from the large volumes of data.

(c)Convey Data

Conveying data includes methods to transform the mathematical or statistical conclusions drawn from the data into a form that can be easily understood and interpreted by those in need of it. Conveying data is empowering the development starting with one perspective then onto the next, empowering a beginner to turn into an expert, current technology to appear to be new and allowing the modeled information to be seen by apprentices and making new technology to appear like it was an integral part of the system.

III. CHALLENGES IN BIG DATA ANALYTICS

Recent years big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, Thomson Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges.

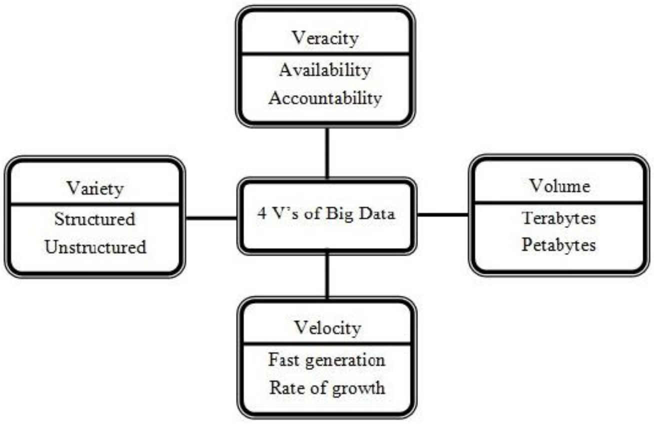

Fig 2: Characteristics for Big Data

To handle the challenges we need to know various computational complexities, information security, and computational method, to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data. Various challenges that the health sector face were being researched by much researchers [9], [10]. Here the challenges of big data analytics are classified into four broad categories namely data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and information security. We discuss these issues briefly in the following subsections.

A. Data Storage and Analysis

In recent years the size of data has grown exponentially by various means such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers etc. These data are stored on spending much cost whereas they ignored or deleted finally because there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on the top priority for the knowledge discovery and representation. The prime reason is being that, it must be accessed easily and promptly for further analysis. In past decades, analyst use hard disk drives to store data but, it slower random input/output performance than sequential input/output. To overcome this limitation, the concept of solid state drive (SSD) and phrase change memory (PCM) was introduced. However the available storage technologies cannot possess the required performance for processing big data.

Another challenge with Big Data analysis is attributed to diversity of data. With the ever growing of datasets, data mining tasks has significantly increased. Additionally data reduction, data selection, feature selection is an essential task especially when dealing with large datasets. This presents an unprecedented challenge for researchers. It is because existing algorithms may not always respond in an adequate time when dealing with these high dimensional data. Automation of this process and developing new machine learning algorithms to ensure consistency is a major challenge in recent years. In addition to all these Clustering of large datasets that help in analyzing the big data is of prime concern [11]. Recent technologies such as hadoop and mapReduce make it possible to collect large amount of semi structured and unstructured data in a reasonable amount of time. The key engineering challenge is how to effectively analyze these data for obtaining better knowledge. A standard process to this end is to transform the semi structured or unstructured data into structured data, and then apply data mining algorithms to extract knowledge. A framework to analyze data was discussed by Das and Kumar [12]. Similarly detail explanation of data analysis for public tweets was also discussed by Das et al in their paper [13].

The major challenge in this case is to pay more attention for designing storage systems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources. Furthermore, design of machine learning algorithms to analyze data is essential for improving efficiency and scalability.

B. Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is a prime issue in big data. It includes a number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation. There are several tools for knowledge discovery and representation such as fuzzy set [14], rough set [15], soft set [16], near set [17], formal concept analysis [18], principal component analysis [19] etc to name a few. Additionally many hybridized techniques are also developed to process real life problems. All these techniques are problem dependent. Further some of these techniques may not be suitable for large datasets in a sequential computer. At the same time some of the techniques has good characteristics of scalability over parallel computer. Since the size of big data keeps increasing exponentially, the available tools may not be efficient to process these data for obtaining meaningful information. The most popular approach in case of large dataset management is data warehouses and data marts. Data warehouse is mainly responsible to store data that are sourced from operational systems whereas data mart is based on a data warehouse and facilitates analysis.

Analysis of large dataset requires more computational complexities. The major issue is to handle inconsistencies and uncertainty present in the datasets. In general, systematic modeling of the computational complexity is used. It may be difficult to establish a comprehensive mathematical system that is broadly applicable to Big Data. But a domain specific data analytics can be done easily by understanding the particular complexities. A series of such development could simulate big data analytics for different areas. Much research and survey has been carried out in this direction using machine learning techniques with the least memory requirements. The basic objective in these research is to minimize computational cost processing and complexities [20], [21], [22].

However, current big data analysis tools have poor performance in handling computational complexities, uncertainty, and inconsistencies. It leads to a great challenge to develop techniques and technologies that can deal computational complexity, uncertainty, and inconsistencies in a effective manner.

C. Scalability and Visualization of Data

The most important challenge for big data analysis techniques is its scalability and security. In the last decades researchers have paid attentions to accelerate data analysis and its speed up processors followed by Moore’s Law. For the former, it is necessary to develop sampling, on-line, and multi resolution analysis techniques. Incremental techniques have good scalability property in the aspect of big data analysis. As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift in processor technology being embedded with increasing number of cores [23]. This shift in processors leads to the development of parallel computing. Real time applications like navigation, social networks, finance, internet search, timeliness etc. requires parallel computing.

The objective of visualizing data is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper interpretation. However, online marketplace like flipkart, amazon, e-bay have millions of users and billions of goods to sold each month. This generates a lot of data. To this end, some company uses a tool Tableau for big data visualization. It has capability to transform large and complex data into intuitive pictures. These help employees of a company to visualize search relevance, monitor latest customer feedback, and their sentiment analysis. However, current big data visualization tools mostly have poor performances in functionalities, scalability, and response in time.

We can observe that big data have produced many challenges for the developments of the hardware and software which leads to parallel computing, cloud computing, distributed computing, visualization process, scalability. To overcome this issue, we need to correlate more mathematical models to computer science.

D. Information Security

In big data analysis massive amount of data are correlated, analyzed, and mined for meaningful patterns. All organizations have different policies to safe guard their sensitive information. Preserving sensitive information is a major issue in big data analysis. There is a huge security risk associated with big data [24]. Therefore, information security is becoming a big data analytics problem. Security of big data can be enhanced by using the techniques of authentication, authorization, and encryption. Various security measures that big data applications face are scale of network, variety of different devices, real time security monitoring, and lack of intrusion system [25], [26]. The security challenge caused by big data has attracted the attention of information security. Therefore, attention has to be given to develop a multi level security policy model and prevention system.

Although much research has been carried out to secure big data [25] but it requires lot of improvement. The major challenge is to develop a multi-level security, privacy preserved data model for big data.

IV. Python

In this paper, we are going to introduce the characteristics of Python. Python is a general-purpose, high-level programming language which is widely used in the recent times [27][28][29].

The most important feature in Python being it supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. Python supports a dynamic type system and automatic memory management and has a large and comprehensive standard library. Python interpreters are available for many operating systems.

Characteristics of python:

Python is a well designed language that can be used for real world programming. Python is a very high-level, dynamic, object-oriented, general purpose programming language that uses interpreter and can be used in a vast domain of applications. Python was designed to be easy to understand and use. Python is termed as a very user-friendly and beginner-friendly language in the recent times. Python has gained popularity for being a beginner-friendly language, and it has replaced Java as the most popular introductory language. As a dynamically typed language, Python is really flexible. Furthermore, Python is also more forgiving of errors, so you'll still be able to compile and run your program until you hit the problematic part. Python is a flexible, simple coding programming language. This language can support different styles of programming including structural and object-oriented. Other styles can be used, too. Python is very flexible, because of its ability to use modular components that were designed in other programming languages. For example, you can write a program in C++ and import it to python as a module. Then add something else to it (for example design a GUI for it).

Features of Python:

Python is simple and lovely

It is a very high-level language that has many sources for learning. Python supports a wide variety of third party tools which makes it much easier to use and motivates the users to continue with. Python has a very simple and elegant syntax. It's much easier to read and write Python programs compared to other languages like: C++, Java, C#. Python makes programming fun and allows you to focus on the solution rather than syntax. If you are a newbie, it's a great choice to start your journey with Python.

Python is portable

Python scripts can be used on different operating systems such as: Windows, Linux, UNIX, Amigo, Mac OS, etc. You can move Python programs from one platform to another, and run it without any changes. It runs seamlessly on almost all platforms including Windows, Mac OS X and Linux.

Python is open source

Even though all rights of this program are reserved for the Python institute, but it is open source and there is no limitation in using, changing and distributing. You can freely use and distribute Python, even for commercial use. Not only can you use and distribute softwares written in it, you can even make changes to the Python's source code. Python has a large community constantly improving it in each iteration.

Python supports other technologies

It can support COM, .Net, etc objects.

Extensible and Embeddable

Suppose an application requires high performance. You can easily combine pieces of C/C++ or other languages with Python code. This will give your application high performance as well as scripting capabilities which other languages may not provide out of the box.

A high-level, interpreted language

Unlike C/C++, you don't have to worry about daunting tasks like memory management, garbage collection and so on. Likewise, when you run Python code, it automatically converts your code to the language your computer understands. You don't need to worry about any lower-level operations.

Large standard libraries to solve common tasks

Python has a number of standard libraries which makes life of a programmer much easier since you don't have to write all the code yourself. For example: Need to connect MySQL database on a Web server? You can use MySQLdb library using import MySQLdb . Standard libraries in Python are well tested and used by hundreds of people. So you can be sure that it won't break your application.

Object-oriented

Everything in Python is an object. Object oriented programming (OOP) helps you solve a complex problem intuitively. With OOP, you are able to divide these complex problems into smaller sets by creating objects. Python is a multi-paradigm programming language: object-oriented programming and structured programming are fully supported. Python uses dynamic typing and a combination of reference counting and a cycle-detecting garbage collector for memory management. An important feature of Python is dynamic name resolution (late binding), which binds method and variable names during program execution. Python was designed to be highly extensible. Python can also be embedded in existing applications that need a programmable interface. Python has a large standard library, commonly cited as one of Python's greatest strengths, providing tools suited to many tasks. For Internet based applications, a large number of standard formats and protocols (such as MIME and HTTP) are supported. Modules for creating graphical user interfaces, connecting to relational databases, pseudorandom number generators, arithmetic with arbitrary precision decimals, manipulating regular expressions, and doing unit testing are also included.

Python can be used to write a wide variety of programs:

Python is a well designed language that can be used for real world programming. The most common program types that can be written by Python are categorized below:

System programming

Pythons Internal interfaces support working with services of operating system and hence makes it a suitable language for system programming. The standard library of Python can support the different types of platforms and operating systems. It contains some tools for working with system resources such as environmental variables, files, sockets, pipe, processes, multiple treats, command line, standard stream interfaces, shell programming, etc.

Graphical User Interface (GUI)

Tkinter and wxPython are the basic interfaces for designing GUIs in Python. Tkinter is a standard object-oriented interface that is distributed with Python interpreter. It provides the essential tools for designing GUI.

Network and internet programming

Various modules are embedded in Python standard library that provide many tools for network programmers, such as: client-server connection, socket programming, FTP, Telnet, email functions, RPC, SOAP, etc.

Components integrity

Python is able to make an integrated connection between its codes and other components. Tools such as Swing and SIP can import the compiled codes of other languages for using in Python.

Database programming

Python supports most of the common databases like Sybase, Oracle, Informix, MySQL, PostgreSQL, SQLite, etc. Pickle is a standard module that can store and recover objects in files. Also, ZODB is a pure object-oriented tool for working with databases. From Python 2.5 on, SQLite was considered as a standard part of Python.

Other programming applications

Python dominates a wide extent of programming areas. For example, PyGame is a tool for game programming and PIL is used for image processing. For robotic programming, PyRo exists. A complete package for artificial intelligence, network simulation, and shell programming was published under the title NLTK. Almost in all area you can find sufficient modules that can help you to get to your goals. There are different tools for Python users with different needs. This good feature makes Python suitable for any kind of programming. Large amount of using Python by popular websites and applications is the best evidence for this matter.

V. CONCLUSION

For sure the future will be crowded with people trying to applying data science in all problems, kind of overusing it. But it can be sensed that we are going to see some real amazing applications of DS for a normal user apart from online applications (recommendations, ad targeting, etc). In recent years data are generated at a dramatic pace. Analyzing these data is challenging for a general man. In this paper, we introduced the Python programming language as a suitable choice for learning and real world programming. The paper has discussed the basic steps of data science, challenges in big data analytics, and the characteristics, features, types of programming support offered by python.

REFERENCES

[1] C. Lynch, Big data: How do your data grow?, Nature, 455 (2008), pp.28-29.

[2] X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2(2) (2015), pp.59-64.

[3] R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) (2014), pp.1-12.

[4] C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275 (2014), pp.314-347.

[5] K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.

[6] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285 (2014), pp.112-137.

[7] Parsons, MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science.HYDROL PROCESS. 18:3637-653. http://www3.interscience.wiley.com/cgi-bin/jissue/109856902

[8] Data Munging with Perl. DAVID CROSS. MANNING. Chapter 1 Page 4.

[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.

[10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, 2013, pp.17-22.

[11] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[12] T. K. Das and P. M. Kumar, Big data analytics: A framework for unstructured data analysis, International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.

[13] T. K. Das, D. P. Acharjya and M. R. Patra, Opinion mining about a product by analyzing public tweets in twitter, International Conference on Computer Communication and Informatics, 2014.

[14] L. A. Zadeh, Fuzzy sets, Information and Control, 8 (1965), pp.338 353.

[15] Z. Pawlak, Rough sets, International Journal of Computer Information Science, 11 (1982), pp.341-356.

[16] D. Molodtsov, Soft set theory first results, Computers and Mathematics with Aplications, 37(4/5) (1999), pp.19-31.

[17] J. F.Peters, Near sets. General theory about nearness of objects, Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.

[18] R. Wille, Formal concept analysis as mathematical theory of concept and concept hierarchies, Lecture Notes in Artificial Intelligence, 3626 (2005), pp.1-33.

[19] I. T.Jolliffe, Principal Component Analysis, Springer, New York, 2002.

[20] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, Efficient machine learning for big data: A review, Big Data Research, 2(3) (2015), pp.87-93.

[21] Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, International Neurourology Journal, 18 (2014), pp.50-57.

[22] P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C.Jain, H. S.Behera, J. K.Mandal and D. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2 (2014), pp. 89-97.

[23] A. Jacobs, The pathologies of big data, Communications of the ACM, 52(8) (2009), pp.36-44.

[24] H. Zhu, Z. Xu and Y. Huang, Research on the security technology of big data information, International Conference on Information Technology and Management Innovation, 2015, pp.1041-1044.

[25] Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, Congresso da sociedada Brasileira de Computacao, 2014, pp.1-6.

[26] I. Merelli, H. Perez-sanchez, S. Gesing and D. D.Agostino, Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives, BioMed Research International, 2014, (2014), pp.1-13.

[27] TIOBE Software Index (2011). "TIOBE Programming Community Index Python". 1

[28] "Programming Language Trends - O'Reilly Radar". Radar.oreilly.com. 2 August 2006.

[29] "The RedMonk Programming Language Rankings: January 2011 – tecosystems". Redmonk.com.