A Study on Importance of Python in Static and Real Time Big Data Analysis

Lekha C. Warrier
Assistant Professor, ISBR COLLEGE, Electronic City, Bangalore

Abstract

The era of data-driven society has aroused. The data emanating due to advanced technologies are increasing exponentially. The massive, fast-growing diverse data, which cannot be processed by the traditional methods, are called Big Data. Huge junk of raw data is available which needs to be consumed for various business needs and it requires advanced programming techniques to break down into meaningful insights. More sophisticated systems, as well as advanced visualization techniques, is necessary for big data analysis. Python, an open source general purpose programming language is considered as a leader in data science programming. Its rich set of utilities & libraries for data processing & analytics tasks keeping it in the prominent place. In this study paper, we explain the significance of Python for Big Data analysis and the framework used for big data in Python platform for both static big data and real time big data.

Keywords— **Big Data, Python, Analytics, Real time data, Importance of Python, Static data.**

I. INTRODUCTION

The era of artificial intelligence and machine learning had come. The technology is developing in each second and people should grow accordingly. The huge data coming out from diverse sources needs to be stored and processed in a very fast manner. The ecosystem of our life has given more importance to the internet. The basic necessity of a human being itself includes the internet. Instrumentation makes us to sense everything around us. Each activity of people in devices making billions of data at a time. Traditional technology and database cannot handle this huge amount of fast moving, diverse data. Thus, the term big data has aroused.

The data produced by Internet of Things, social media, sensors etc contain lot of noisy contents. Filtering of this data are essential in all the matters. Cleansing of data requires some technology that provides the valuable insights from the pool of data. The database that using for normal data was mainly designed for structured data. Earlier the main form of data was document file. As the technology the variety of data have produced from various resources.

Image, audio, video etc. are the different forms of data that comes from plenty of devices while using internet or the latest technologies. This type of data neither has a format nor have a particular style. So, this data is difficult to store and process in a stipulated time. The emergency of efficient big data technology has aroused due to the rapid growth of internet usage.

Organisations now utilizing the opportunities to capture all the data that streams into their business. On this captured data they can apply analytics and get significant value from it. This led them to various business benefits, effective marketing, better customer service etc.

Python has a unique capacity of being general purpose programming language, which is easy to learn and easy to use in analytical and quantitative computing. In this paper we are recognising the impact of python in big data analysis. It also explains the frame works of Hadoop and Apache Spark interfacing with Python language in the analysis of static and real time big data.

II. BIG DATA

Large volume of datasets that are growing in very rapid manner and contains different forms of data with or without any proper format can be called as big data. Big data is so called because the traditional data processing application and technology are inadequate to handle this type of data.

Big data is not just big. It is described by mainly three characteristics- Volume, Velocity and the Variety [4]. As the magnification of data is increasing exponentially and the different varieties of data are generated in each fraction of second. The data can be useful for some purpose for better market analysis or customer service. But the extraction of such meaningful data is a challenge. For the storage and processing of this huge amount of data big data technologies have come to the screen [6].

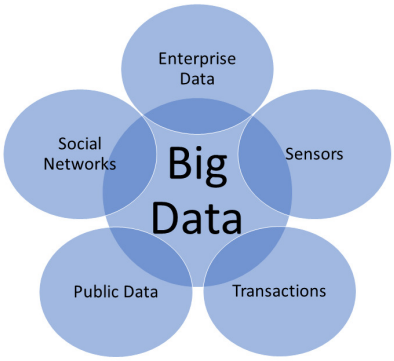

Fig. 1 Big Data Sources

III. BIG DATA TOOLS

The goal of organisation is to uncover never- before seen business insights in less time span in an economic way. Big data technology contains a lot of process including extraction, storage, cleaning, mining, visualizing analysing and integrating. This paper explains the processes comprises of big data analysis and the tools used in each process.

A. Data storage and Management

An efficient storage provider should give good and effective infrastructure facilities to store the data and queries. The important tools that is used for big data storage and management are

1) *Hadoop:* Hadoop is an open source framework for distributed storage of very large datasets on computer clusters. The repository of data in Hadoop frame work have the advantage that we can scale up your data [3][1].

2) *Mongo DB:* It is the modern database, which is good for managing the data that changes frequently or data that is unstructured or semi structured data [8].

B. Data Cleaning

The purpose of the task data cleaning is refining and reshaping the data into a usable data set.

The major tools that is used for cleaning the huge amount of data are

1) Open Refine: It is the open source tool that is dedicated to clean messy data. The noise contents would be filtered out as the outcome of the process.

2) *Data Cleaner:* This tool is mainly for transforming messy semi structured data sets into clean readable data sets.

C. Data Mining

It is the process of discovering insights with in a data base. The extraction of meaningful data from the pool of data is called as data mining process. In big data processing the main task is data mining. The main tools used for mining of big data are

1*) Rapid Miner:* It is the tool for predictive analysis. It is powerful and has an opens source community behind it.

2) *IBM SPSS Modeller:* This tool is mainly designed for text analysis, decision management and for optimization.

D. Data Analysis

The data analysis is the process of breaking the data down and accessing the impact of those patterns over time. The main objective of analysis is to support decision making. In the next section we will look into more detail about big data analytics. The generally used tools for data analysis are

1) BigML: The tool is using the concept of machine learning. They offer machine learning service with a user-friendly interface for you to import your data and get predictions out of it
2) Quoble: It speeds and scales analysis of big data in any level. This is an enterprise level solution.

E. Visualization

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects contained in graphics. The goal is to communicate with users efficiently. The technology used in big data visualization is

1) Tableau: It is a tool for visualization base on business intelligence.

2) CartoDB: This is the tool that specialises in making maps. It helps user to visualize location data.

F. Data Integration

It is the operation of combining data from several diverse sources to accomplish task. The main tool for integration is

1) Pentaho: This tool offers big data integration with zero coding. It also provides embedded analytics and business analytics services too.

In the previous section we have seen the exercises in handling big data. The process which undergoes in big data technology is very much significant in this data driven world. The concept behind the analysis of big data had been explained in detail in this section.

IV. BIG DATA ANALYTICS

Big data analytics is the process of examining large and varied datasets to uncover hidden patterns or unknown correlations, market trends and other useful information that can help organizations to make better business decisions. To get significant value from the pool of data the analytics in big data is essential.

Organizations need to follow some steps in the analysis of big data for the better outcome. The steps are described here in a brief manner [2]

1) Find the relevant data which need to be analyzed: It is not necessary to process a company's whole data. The company can decide what data to include and what

data to leave out. The data which will lead to valuable insights only need to be analyzed

2) Build effective business rules and translate into analytics pattern: To get the meaningful information from the data, it is necessary to create some business methodology. In order to filter the data, the rules could be used as proper benchmarks.
3) Make effective business plan: To provide support for iterative development process in dynamic business environments proper maintenance plan is required.

The analytics of big data can be done using some programming languages like Python, R, Julia etc. In this paper we are focusing on python language and its benefits for big data analytics.

V. PYTHON IN BIG DATA ANALYTICS

Python is a general purpose, open source programming language. The most important feature of python is its rich set of utilities and libraries for data processing and analytics task. It also has easy to use functions which supports big data processing. It is a preferable language for making scalable applications. To generate quick and valuable insights organization use python in any platform. Another success of python is that it is growing every day. According to the technical news python is the most important language to learn in the coming years. To make effective and productive programs is preferred.

In this section we will go through the python libraries which are more popular for data analysis.

Python Libraries

1) Numpy: It is the fundamental package for scientific computing. Mainly it deals with N Dimensional arrays, universal functions etc. [8].
2) Pandas: Pandas contain high level data structures and manipulation tools to make data analysis fast and easy in python.
3) Matplotlib: It controls the style and color of a figure, graph coordinates and complex figures.
4) Pydoop: This library provides a simple API for Hadoop.
5) Scipy: This library widely used in scientific and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering [8].

VI. BIG DATA FRAMEWORKS IN PYTHON

The main advantage of python is that it has the fundamental building blocks necessary for doing data analysis. Now Python is the center of an emerging trend that is unifying traditional high-performance computing with big data applications [9]. In Big data technology, Hadoop and Map Reduce have prominent place. The data sets are stored in HDFS cluster files. The processing of these data sets is accomplished by Map Reduce function using parallelization technique. The computational speed is less in Hadoop Map reduce.

TABLE 1
COMPARISON OF FRAMEWORKS

PySpark	Pydoop
Interface python with Apache Spark	Interface python with Hadoop framework
Speed is high	Less speed
Easy to program with RDD	Difficult to code
Real time data analysis	Not preferable for real time data
Less Secure	More secure
Cost is high	Less Cost

Apache Spark [9] is an open source big data framework. It is fast, easy to use general engine for big data processing. It is a general-purpose cluster computing framework. It distributes data and computation across multiple computers.

The interfacing of python with Spark is named as PySpark. PySpark provides an easy to use programming abstraction and parallel run time.

VII. REAL TIME BIG DATA ANALYSIS IN PYTHON FRAME WORK.

Real time data analytics is the method of analyzing the data as soon as it is produced. Without any latency the data must be processed and analyzed. In fraud detection, user sentimental analysis of social media and live traffic management produce real time data. When big data is effectively collected and efficiently analyzed, Companies can gain a more comprehensive understanding of their business, products, customers and competitors. The extraction of valuable insights from this is a challenge when the huge, unstructured and very rapid data is generating.

Real time big data processing system must have strong timeliness which means it must quickly respond to the request from system terminals in a very short time delay. So, at first, real-time big data processing system must have powerful computing ability for big data.

The framework to analyze this type of data should be effective and productive. Apache Spark in Python interface (PySpark) can provide an efficient ecosystem for the real-time data analysis. Python can be easily integrated into web applications to carry out tasks requiring machine learning.

Resilient Distributed Dataset, RDD, is an abstract use of distributed memory as well as the most fundamental abstract of Spark, achieving operating the local collection to operate the abstract of a distributed data set [5].

Data coming from the different sources are collected by PySpark and separating the data into corresponding data sets. Then PySpark analyze each data set as soon as it is full. Python interface helps in mathematical computing and statistical modelling in an efficient manner. The timeliness in Computing and analyzing the data can be more acquired when

we are using PySpark frame work. The filtering of data and the removal of redundant data can be done efficiently with python framework.

VIII. ADVANTAGES OF PYTHON FRAME WORK IN BIG DATA ANALYTICS

The study of python framework in big data analytics details these advantages.

1. For better decision making in business, the proper analysis of various data have to be done. With python interface the fast analysis happens for real time data also
2. The cost of Python interface for data analysis is comparatively less as it is open source.
3. The support and maintenance of big data tools in python analytics is easy as very big online community is there in whole over the world.
4. The portability and scalability of python helps to make analysis in an effective manner.
5. The protection of data also assured in python framework
6. The modules and extensions in python language interfaces with Twitter, LinkedIn, Sensex data, social media etc.

IX. CONCLUSION

Python is the outstanding and efficient programming language in data analysis, cyber forensics etc. The power of communication with live and real-time servers makes python frame work in a top position in big data analysis. The protection of data is another major advantage of python in data analysis. We understood that analysis of big data in python platform is the deep investigation of intelligent and valuable insights from the heterogeneous data. The big data is the buzz word in the technical ecosystem. The future of technology and business depends on the big data analysis. Interfacing with Python makes big data analysis better in all the aspects. This paper reviewed the significance of python in the analysis of big data. More contribution is required in this field especially in the security area as big data and python are the future.

X. References

[1] Munesh Kataria, Pooja Mittal, Big Data: A Review, IJCSMC, Volume 3, Issue-7, pp. 106-110, , July 2014 ,

[2] Prashant Kumar, Kushboo Pandey,,Big Data and Distributed Data Mining: An Example of Future Networks, IJARI, Volume1, Issue 2, pp. 36-39, August 2013

[3] Vidyullatha Pellakuri, D Rajeswara Rao, Hadoop Mapreduce Framework in Big Data Analytics, IJCTT, Volume 8, February 2014

[4] Gartner, http://www.gartner.com/it-glossary/big-data.

[5] Zhigao Zhing, Ping Wang and Linux Liu ,Shingli Sun,Real time big data processing framework: Challenges and Solutions,No.6,3169-3190, Nov 2015.

[6] Min Chin, Shiwin Mao, Yunhao Liu Big Data : A Survey, Springer,171-209, 2014

[7] Daniel E O'Leary, Artificial Inteeligence and Big Data, 1541-.1672, IEEE Intelligent Systems, 2013.

[8] All the best big data tools, https://www.import.io/post/best-big-data-tools-use/

[9] Holden Karau, Andy Konwinski, Patrick Windell And Matei Zaharia, Spark- Lightning- Fast Data Analysis.