# A MINI REVIEW ON BIG DATA NALYTICS TOOLS, TRENDS AND CHALLENGES

[1]S. Mamatha Upadhya , [2]C. S. K. Raju, [3]G. Madhavi

[1,3] Department of Mathematics, Garden City University, Karnataka, India.

[2]Department of Mathematics, GITAM University, Bangalore, Karnataka, India.

**Abstract:** In the wake of on-going and upcomingdigital revolution, there is enormous advancement in the scale of data routinely generated and gathered through variety of sources including mobile devices, social media , Internet of things (IOT)additionally companies are collecting data from their customers. By 2020, about 1.7 megabytes of data will be generated for every second by every human being. This article provides an overview of Big DataAnalytics, tools available till 2020 implemented by various organizations and industries, its important features and challenges.

**Index terms:** Big Data Analytics, Internet of Things, Hadoop, Apache Spark, Apache Strom, Apache Cassandra.

## 1. INTRODUCTION

There is a proliferation in the data generated which is going to persist for many coming years. These data are generated by online transactions, videos, emails, audios, logs, images, search queries, posts, health records, sensors and mobile phones, social networking interactions and their applications. They are stored in the database which grows massively and thus become difficult to store, capture, form, share, manage, visualize and analyze through typical database software tools. Till 2003, 5 exabytes ($10^{18}$ bytes) of data were generated. In 2012, this data was expanded to 2.72 zettabytes ($10^{21}$ bytes). In 2015, 8 zettabytes of data were created. Today, each day over 2.5 quintillion bytes ($10^{18}$ bytes) of data are generated. It is anticipated by the researchers that, by 2020 1.7 MB of data will be generated for every person for every second according to the researchers [1-5]. It is estimated that personal computer can hold around 500 gigabytes ($10^{9}$ )hence, it would necessitate about 20 billion Personal computers in order to store all the world's data. In the world each day more than 6 billion mobile subscriptions take place and more than 10 billion text messages are sent. Only Google has more than 1 million servers around the world and it is anticipated that by 2020, 50 billion devices would be connected to the network and the internet. In view of this researchers [6-12] have analyzed big data and its challenges.

In the current days, amassive surge in the quantity of data is being created that needs to be analyzed and stored appropriately. Walmart handles more than one million transactions per hourWhile Facebook has more than 2.32 billion monthly active users and 1.52 billion people on an average daily log onto Facebook. Youtube has more than 1.5 billion monthly active users. Instagram has more than 889 million monthly active users while Twitter has more than 328 million monthly active users. However, in the coming years, number of information would enhance by 50 times and the specialists in the field of information technology would also increase by 1.5 times. (See [12-20]).

## II. BIG DATA ANALYTICS TOOLS AND ITS IMPORTANT FEATURES

"Big Data" refers to massive homogeneous volume of both the structured and unstructured data which augments from various sources and which is too large and difficult to process effectively with the traditional database and software techniques.Basically, there are three class of Big Data Analysis. Which are - Descriptive Big Data Analytics, Predictive Big Data Analytics and Prescriptive Big Data Analytics.Descriptive data analytics employ data mining, deep learning, machine learning and data aggregation to analyze the past events. Past events include event occurred at any point of time even those events that have occurred one minute back or one month before. Thus, Descriptive data analytics are helpful since they allow organizations to learn from the past behavior and assist them in analyzing how they may persuade future outcomes. Predictive Big data Analytics employ statistical models to analyze and forecasts the future behavior of the data. Prescriptive Analytics employ simulation and optimization algorithms to predict possible outcomes of the data set. Prescriptive analysis is about providing advice.

Big data analytics scrutinize the huge quantity of data to reveal hidden patterns and correlate the data. Big data analytics assist organizations in understanding the information contained in the data set along with handling those data which are more essential to the business and future business decisions.

Businesses are persistently undergoing a digital transformation owing to an explosion of the data. Voluminous of data is generated with the social media post, every credit card swipe, customer- support call, through myriad devices connected with the Internet of Things (IoT). Raw data alone is not sufficient to drive business growth. Rather, it is the analytics derived from data that construct true value. Emerging Big Data technique facilitates innovation in the product, improves performance, provide service and decision making support. Today, 87% of organizations consider that Big Data techniques will help them to achieve the competitive edge in the market within two –three years and 89% think that if they do not implement Big Data techniques, then they may fall behind their competitors and fail.

For big data processing and analysis organizations are looking forward for open source Big Data tools considering the cost, better time management, free of any licensing overhead, easy to download and use, and other benefits.Below are ten important best open source Big Data tools along with their key features.

**Hadoop**

Apache Hadoop is the most prominent tool in big data industry which has massive capability of processing large-scale data. It is written in Java language which can handle chunk of data sets. Hadoop is a framework which allows to store data in a distributed environment so that one can process it parallel.Originally Hadoop was designed to handle searching and crawling billions of web pages and then collecting this information into a database.It can run on a cloud infrastructure and is 100% open source framework which can run on commodity hardware in the existing data centre. Apache Hadoop consists four parts. (i) Hadoop Distributed File System (HDFS) **-** This is a distributed file system compatible with the large scale bandwidth. (ii)Map Reduce-This is a programming model for processing data.(iii) YARN-This is a platform which is used for scheduling and managing Hadoop's resources in the Hadoop infrastructure.(iv)

Libraries-Help other modules to work with the Hadoop. Some of the key features of Hadoop include quicker data processing, support for POSIX style file system extended attribute,flexible in processing any data,authentication improves when using HTTP proxy server, It offer robust ecosystem which is well suited to meet the analytical needs of developer.

**Apache Spark**

Apache Spark can handle real time data and batch data in a distributed computing environment.Spark executes in-memory computations in order to increase speed of data processing. It processes large scale data much faster as it exploit in-memory computations along with other optimizations. Thus, it entails high processing power. Resilient Distributed Dataset is a fundamental data structure of Spark. Resilient Distributed Dataset contain any type of Java, Python,Scala objects, Spark SQL, including user-defined classes. Thus, Spark is suitable for machine learning, credit card processing system, Internet of Things and security analytics. Spark is flexible to work with data stores such as Apache Cassandra, open Stack Swift and HDFS.

**Apache Strom**

Strom is exceptionally fast it can process per second per node over a million records on a cluster of modest size. It offers distributed real time computation capabilities for processing huge volumes of high velocity data. It supports any programming language. It can be used in handling real time customer service management, continuous computation, online machine learning, operational dashboards, data monetization, threat detection and cyber security analytics. It is scalable, easy to set up and operate, fault-tolerant and guarantees process of every data.

**Apache Cassandra**

Cassandra is an exceptionally popular database which underpins heavy load applications such as Facebook, Google's Big Table, Amazon's Dynamo, Apple, Instagram, Uber etc. It is distributed type databases which manage large amount of structured data across the servers. Cassandra is highly scalable. It provides highly available service with no single failure. It accommodates all structured, unstructured and semi-structured data formats. It provide flexibility in distributing data by replicating the data across multiple data centers.

**Rapid Miner**

Rapid Miner is data science software which offer an integrated environment for deep learning, data preparation, text mining, machine learning, Application development, prototyping and predictive analytics. It is used for research, business and commercial applications, education, machine learning process, data preparation, model validation, results visualization, Evaluation, predictive analytics and Statistical modeling. Rapid Miner follow client/server model the server might e located in a cloud infrastructure or on-premise. Rapid Miner is written in Java it could endow with 99% of advanced analytical solution.

**MongoDB**

MongoDB is classified as NoSQL database program which is compatible with the several built in features. It runs on MEAN software stack, Java platform and NET applications. It is ideal for those businesses which needs real time data for quick decisions. It is flexible and also easily partitions the data across the servers in a cloud –based infrastructure. It can store any form of data such as string, integer, object, array, and date, Boolean etc.

**R Programming Tool**

This is one among the extensively used tool for statistical analysis of data. Though designed for the statistical investigation of the data most positive part of this tool is, as a user one need not be a statistical expert. R tool has got its own public library called CRAN (Comprehensive R Archive Network) consisting more than 9000 modules as well as algorithms for statistical computing and graphics. It can run on Windows, SQL and Linux server .It supports Spark and Hadoop. It is a portable languge thus it can be implemented easily in other servers.

**Neo4J**

Neo4J is graph database management system which is interconnected with node-relationship of data. It is implemented in Java. It preserves key value pattern in data storing. It is flexible since it does not require a schema or data type in order to store the data. It is scalable can integrate with other databases.It supports Cypher and ACID transaction. Companies such as Walmart, eBay, UBS, Cisco, Telenor, Lufthansa and Hewlett-Packard employ the qualities of Neo4J in order to improve their services.

**Apache SAMOA**

Apache Samoa(Scalable Advanced Massive Online Analysis) offer collection of distributed streaming algorithms for machine learning and data mining tasks such as clustering, classification, regression and programming abstractions to develop new algorithms. It is written in Java. It features in pluggable architecture hence as a user one can run Samoa on numerous DSPEs (distributed stream processing engines) such as Apache S4, Apache Storm, Apache Flink and Apache Samza. As a developer one can construct the new algorithms only once and test them in all DSPEs. Its existing infrastructure is reusable thus, one can avoid deploying cycles. It does not need complex backup or update process.

**HPCC**

In big data market HPCC(High-Performance Computing Cluster) is the competitor of Hadoop. Compared to Hadoop the HPCC platform offers less code and less nodes for larger efficiency. It offer a single programming language and single architecture for the efficient processing. It helps in parallel data processing. It runs on commodity hardware. It supports end to end big data workflow management. It comes in binary packages for Linux distributions.HPCC environment would include either Thor and Roxie clusters or only Thor clusters.

## III. DISCUSSION AND OPEN CHALLENGES

The world is moving towards data driven society and currently data are most valuable asset. The proliferation in Big Data and big Computing hasboosted data science and machine learning across the many application domains. For instance, voice driven personal assistance (Amazon Alexa) are available, image recognition systems has surpassed human quality, autonomous vehicles are becoming reality. In order to maintain sustained economic development of societies and countries efficient and effective exploitation and analysis of Data is very essential. Still there are many challengesto fully harness the available data and they include:
Data sharing and Data availability: Big data Analytics survive and diminish with the data. The larger and more varied the data set, better analysis can be made. But in reality, data is sliced, segmented and in the control of individuals, organizations or departments. Thus, it is necessary to motivate all parties to unite and share useful insight/data without the restrictions from patents, copyright or other kind of mechanisms of control.
Proficient distributed implementation mechanisms: In practice, tools used for scientific computing tasks and statistical computation (e.g. Python and R) is memory bounded. Algorithms used for data analysis depend on in-memory data processing mechanisms. Though this sort of approach has several advantages in provisions of speeding up the whole process there could be scalability risk concerned if the process data can not fit in the accessible main memory.
Interoperability: Interoperability the major issue due to the increasing number of services and platform. Standard models and formats are necessary to facilitate interoperability and smooth cooperation among variousservices and platform. Majority of the available big data processing platforms such as Spark and Hadoop are designed on the single cluster setup assuming homogeneous connectivity and centralized management hence, sometimes it becomes infeasible to implement data analytics job on highly distributed data sets.
Explorative character of data analytics process: Data analytics is a highly explorative process. There is no single algorithm or model to handle all varieties of data set. Researchers work hard to get the best algorithm or model which meets their large data set. Currently, Auto-WEKA is attempting to automate this process.
Programming abstractions: Though there are numerous programming approaches availableto implement data analysis applications. Still new programming abstractions areessential to minimize the code developing time, reduce the gap between data analysis algorithms and the scalable computing platforms through which the data are executed.
Model management: Constant increase in models and usage of machine learning, the issues such as model management, model versioning, model sharing and lifecycle management have become important tasks. It is necessary to track the models developed and the differences among them through recording their metadata. In this regard currently ModelHub, ModelDB are available.

## IV. CONCLUSION

In the current article an overview of Big Data concept, technologies implemented by various companies, various tools available in the market with its features and challenges have been reviewed. Additionally,the research efforts are necessary to exploit the data analytics process to make it more usable for solving complex problems.

## REFERENCES

**[1]**.Tankard, C. (2012). "Big Data Security", Network Security Newsletter, *Elsevier*, *ISSN* pp.1353-4858.

**[2]**. Küçükkeçeci, C. and Yazıcı, A. (2017). Big Data Model Simulation on a Graph Database for Surveillance inWireless Multimedia Sensor Networks, *Big Data Res*. https://doi.org/10.1016/j.bdr.2017.09.003.

[**3**].Han, J., Haihong, E. , Le, G., Du, J.,(2011). Survey on nosqldatabase, in: Pervasive computing and applications (ICPCA), 2011 6th *international conference on, IEEE*, pp. 363–366.

[4]. Ashton, K.(2009). That 'internet of things' thing, *RFiD Journal* ,22 (7), pp.97–114.

[5].Mohapatra, S. K., Sahoo, P. K., Wu, S. L.,(2016). Big data analytic architecture for intruder detection in heterogeneous wireless sensor networks, *Journal of Network and Computer Applications*, *66*, pp.236–249.

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

**[7].**Mishra, N., Lin C., Chang, H.(2015). A cognitive adopted framework for iot big data management and knowledge discovery prospective, *International Journal of Distributed Sensor Networks*, pp 1-13

[**8**].Lu, Y. and Xu, X. (2019). Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services. *Robotics and Computer-Integrated Manufacturing*, 57 pp 92-102.

[**9**]. Ismail, A., Shehab, A., El-Henawy, I.M. (2019). Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations. In Security in Smart Cities: Models, Applications, and Challenges, *Springer*, *Cham*. pp. 27-45.

[**10**]. Jaseena, K.U. and David, J.M.(2014) Issues, challenges, and solutions: big data mining. CS & IT-CSCP, 4(13),pp131-140.

[**11**]. Q.P.He, and Wang, J.,(2018). Statistical process monitoring as a big data analytics tool for smart manufacturing, *Journal of Process Control*, 67,pp 35-43.

[**12**]. Sangaiah, A.K., Thangavelu, A., Sundaram, V.M.,(2018) Cognitive Computing for Big Data Systems Over IoT, *Gewerbestrasse*, 11pp.6330.

[**13**]. Del Vecchio, P., Di Minin, A., Petruzzelli, A.M., Panniello, U., Pirri, S., Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges, *Creativity and Innovation Management*, 27(1) (2018)6-22.

[**14**]. Kahng, and Andrew, B.(2018). Reducing time and effort in IC implementation: a roadmap of challenges and solutions, In 2018 55*th* ACM/ESDA/IEEE Design Automation Conference (DAC), *IEEE* pp.1-6.

[**15**]. Evans, M.R, Oliver, D., Yang, K., Zhou, X. , Ali, R.Y., Shekhar, S. (2019). Enabling spatial big data via CyberGIS: Challenges and opportunities. In CyberGIS for Geospatial Discovery and Innovation, *Springer*, *Dordrecht*, pp.143-170.

[**16**]. Azzone, G. (2018) Big data and public policies: Opportunities and challenges. *Statistics & Probability Letters*, 136, pp. 116-120.

[**17**]. Chen, B., Wan, J., Shu, L. P., Li, Mukherjee, M. B.Yin, (2018). Smart factory of industry 4.0: key technologies, application case, and challenges. *IEEE Access*, *6* pp. 6505-6519.

[**18**]. Meoni, M. Perego, R., Tonellotto, N.(2018). Dataset popularity prediction for caching of CMS big data, *Journal of Grid Computing*, 16(2) pp. 211-228.

**[19].** Anawar, M.R. Wang, S., Azam Zia, M., Jadoon, A.K., Akram, U., Raza, S. (2018). Fog computing: an overview of big Iot data analytics. *Wireless Communications and Mobile Computing.*

[**20**]. Arunachalam, N. Kumar, D., Kawalek, J.P. (2018). Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. Transportation Research Part E: *Logistics and Transportation Review*, 114, pp. 416-436.