International Journal of Future Generation Communication and Networking
Vol. 13, No.2s, (2020), pp. 1739–1746

Review of Big data: Recent Tools and Technologies

Prasad Ligade, Prof. P.P.Jorvekar

Computer Department, NBN Sinhgad School of Engineering

Abstract

Today Big data becomes very important concept in IT world. There is rapid rise in volume of data which can be in a structured, unstructured and semi-structured form. Now a dayssources of big data are through social media like Facebook, twitter, search engines, etc. Big data is largely growing very fast at exponential rate so it necessarily becomes important to develop new tools and technologies to handle with it. This paper represents various fields where Big data is used largely and also gives the briefinformation of different frameworks available to process a Big data.

Keywords— *Big data, Hadoop, NoSQL, Analytics, Spark.*

I. INTRODUCTION

1. BIG DATA

Big Data is one single word that has become popular in our generation based on the increase in the rate at which unstructured data is been generated regularly. Few years ago, Organizations or Systems were using all Structured Data only which having some specific schema and can be presented in row/column form. It was very easy to use Relational Databases and old Tools to store, manage, process and report this data.

However recently, nature of data is changed as new technologies, devices, communications are growing day by day. So the amount of data in variety of formats produced by mankind is increasing every year. Meaning of thisdata is not structured data and it does not have any proper format. It is very hard to use old tools to store, manage, process and report this data.Then how to solve this problems? Here BigData Solutions come into picture.

Big Data characteristics with the help of six v's:

Fig-1:Big data Characteristics

1. **Volume**: Volume describes as a huge amount of data produced by any organization like healthcare, education institutes and financial in terms of petabytes & zettabytes.
2. **Variety**: It describes data types which may be structured,semi-structured, unstructured.
3. **Velocity**: It describes the speed of accumulation of Big data. Telecommunication produces 35TB of data on per day.
4. **Veracity**: Veracity means the inconsistencies and uncertainty of captured data.
5. **Value**: Having endless amounts of data but it can be move into value.

ISSN: 2233-7857 IJFGCN

6. **Variability**: It refers to how fast and to what extent meaning or shape of data is constantly changing.

Different Big data Sources :
Big data are coming from various organizations which includes one of three primary sources of big data which are:

- Social Networks(Social data).
- Traditional Business System.
- Internet of Things(IoT).

The data from these sources can be structured, semi-structured or unstructured, or any combination of these varieties. Social Networks give human sourced information from Facebook and Twitter, Instagram, Flickr, Pinterest, etc. Traditional Business Systems provide customers services or products like Commercial Transactions, E-commerce, Banking Records, Credit Cards, Medical Records and Internet of Things include Sensors, Traffic, Weather, Mobile phone location, etc. Security, surveillance videos, and images Satellite images, Data from computer systems (logs, weblogs, etc.).

Big data Advantages :

- Big data is helpful for understanding and targeting customers.
- Big data is useful for understanding and optimizing business process.
- Big data is improving science and research .
- Provides financial trading.
- It ishelpful for improvement of healthcare and Public Health.
- Improving security and law Enforcement.
- Increases sport performance.

2. BIG DATA ANALYTICS:

It is a process of analysing huge amount of data. The main objective of examining is to reveal unseen patterns,market trends, customer preferences and useful business information. It enables data scientists to analyse a huge amount of data that may not be harnessed using traditional tools. It requires tools and technologies that can transform huge amount of semi-structured, structured and unstructured data into a understandable metadata as well as data format for analytical process[19]. There are various algorithms used in these analytical tools for discovering patterns, correlations and trends over a variety of time horizons in the data. Efficient decision making takes place after examining the data, these tools visualize the findings in tables, charts and graphs. It processes consume considerable time to provide guidelines and gives feedback to users, whereas only a few tools can handle large data sets within reasonable amount of processing time[19].

Tools for Big Data Analytics:

1.**Hadoop**: It is distributed data processing environment developed by apache software foundation[11]. It is open source framework for big data. Hadoop is reliable and fault tolerant with no rely on hardware for these properties. It has master-slave principle as it takes large data sets as input, processes it and produces the output.

2.**HPCC**: HPCC is an open source big data tool which is part of LexisNexis Risk Solutions. It is flexible to use and delivers on a single platform as well as a single architecture[20 21]. HPCC platform implemented on commodity computing clusters to give higher performance and data parallel processing for Big data applications.

3.**Cassandra**: Success of Facebook because of the tool Cassandra. Apache Cassandra could be a NoSQL database ideal for highspeed, on-line transactional data. The advantages of this tool areopen

ISSN: 2233-7857 IJFGCN

source, decentralized, fault-tolerant, highly available and elastically scalable.It is in use at eBay, GitHub, Netflix, Reddit, Hulu,etc.

4.**Hive**: Hive started at Facebook to manage lots of data but now it is part of Apache Hadoop project. It is Scalable, fast and provides the SQL like interface to analyse data. It works on high throughput and high latency principle. It has ability to plug-in custom Map reduce programs.

5.**MongoDB**: It is one of the open source NoSQL document database. This tool is useful for real-time analytics and high speed logging.Itcreate globally distributed clusters that provides low latency read and write data access to users anywhere in the world. It can be written in C++,Go, JavaScript, Python.

6.**RapidMiner**: It is powerful and Robust Graphical User Interface developed in the year 2001. It is open-source software implementing tools for data mining, knowledge discovery, forecasting, machine learning and predictive analytics. It is fully transparent end-to-end data science Platform.

7.**KAFKA**: Kafka is designedfor distributed high Streaming Platform developed by Apache Software Foundation in the year 2011.It was created for transferring data from one application to another with largest throughput and low latency.

8.**Spark**: It is Cluster computing tool developed by Apache Software Foundation. It is a powerful open source distributed computing engine for data processing and data analytics. It is an ultra-fast cluster computing technology designed for fast calculations which is based on Hadoop Map Reduce. It provides linear scalability in the distributed environment.

9.**Pig**: It is a high level scripting language and useful for analysing large set of data set. It uses HDFS for storing as well as retrieving data and Hadoop map reduce for processing Big data. Pig is developed by Facebook and yahoo. It provide the facilities to define easy programming environment and permit the system their execution automatically.

10.**Sqoop**: For moving bulky data between Hadoop and Relational databases Sqoop is designed. It supports plugin, so new external system can be integrated. Sqoop uses JDBC Connector to connect with Relational databases and it provides a command line interface to end user.

Application of big data:

1. **Big data in Healthcare**:
   Healthcare is one of the sectors where Big data involves asit can monitor patients and send reports to the associated doctors. It helps for maintaining electronic health records EMR/HER which serve the customer. It maintain accurate information about patient which can reduce mistakes. Cost optimization through efficiency of new e-health facilities.It is starting to play an important role in supporting the improvements , efficiency of health system. Since healthcare monitoring system deal with large quantities of data, a low latency is needed in data capturing while simple query methods can be implemented to process a large quantity of data.

ISSN: 2233-7857 IJFGCN

Fig – 2: Big data in Healthcare

2. **Big data in agriculture:**
Big data helps for crop monitoring through sensors and management of crop health. It helps in early issue detection to counteract during growing season. In agriculture Big data helps for accurate crop prediction, agricultural automation, monitoring natural trends and risk assessment. Data analytics must be targeted, customizable and must helps farmers carry out targeted scouting to be a value. It gives discovering knowledge and co-relations from historical records. Latest technologies are used to collect data from the field and data science helps to drive decision making abilities.

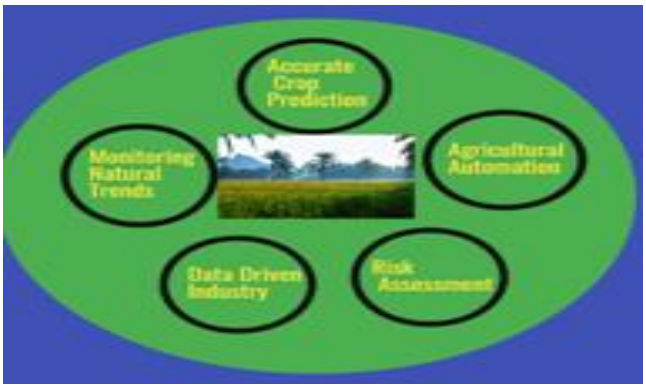
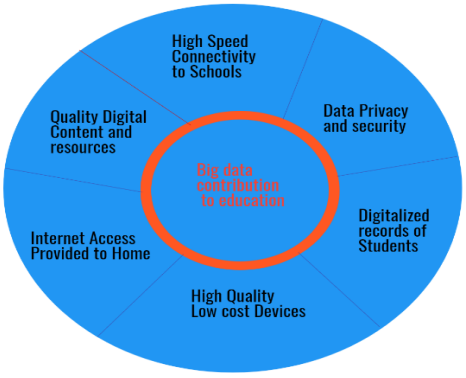


Fig – 3: Big data in agriculture

3. **Big data in education**: Big data can manage, store and analyse the large datasets like student records. Appropriate study and analysis of each and every student's records will help in understanding each student's progress, strengths, weaknesses, interests and more. It would also help in maintaining digitalized records of students and gives data privacy and security.



ISSN: 2233-7857 IJFGCN

Fig – 4: Big data in education

4. **Big data in Government field**: Governments,have to keep track of various records and databases regarding their citizens, their growth, tax evaders, electricity, energy resources, geographical surveys, insurance and many more.Big data helps to monitor the decisions taken by the government and also evaluate the results.

Fig-5: Big data in Government Field

5. **Big datain Media and Entertainment** : Big data increases day by day in media and entertainment industry through access to various electronic devices, social networking sites like Facebook, Twitter, Instagram, etc Big data applications helps media and entertainment industry by predicting as well as updating what the audience wants, increasing acquisition, retention, content and new product development.

Fig-6: Big data in Media and Entertainment

3. Comparison of latest tools:

Big data various tools	Hadoop	Storm	HPCC	Pig	Hive	Spark
Community Service	Distributed File System.	Distributed Stream Processing.	Data Processing	Analysing large data sets	data summarization	Cluster computing

ISSN: 2233-7857 IJFGCN

Developers	Apache Software Foundation	Canonical Ltd.	HPCC systems, LexisNexis Risk solution	Apache Software Foundation	Netflix, Financial Industry Regularity Authority	Apache Spark
Programming languages	JAVA	JAVA, Clojure	C++,ECL	JAVA	JAVA	SCALA
Current Version	3.2.1	0.4.3	7.4.18-1	0.17.0	1.2.1	3.0.0
Operating System	Cross Platform	Cross Platform	LINUX	Microsoft Windows, Linux	Cross Platform	Microsoft Windows, Linux
Organization	Facebook, Twitter, Amazon, Yahoo, Adobe	Yahoo,Twitter	Google	Yahoo, Facebook	Facebook	Apache Software Hardware

Table1:Comparison between Big Data Tools

Discussions:

Big data can be described in the form of unstructured, structured or semi-structured also different organizations and the system at various rates is used to generate the data. Big data refers to large data sets growing rapidly. Several organizations from different sectors generate large data sets which are in the form of heterogeneous formats. Traditional Big data management techniques are less efficient to handle the heterogeneous data. They provide a slow response, lower performance, lack of accuracy and scalability. Current big data platforms are more efficient to extract knowledge and value from large volumes of data. Hadoop provides scalability and Apache Hadoop contains two main components, Hadoop is used forstoring a large amounts of data in its very own distributed file system called HDFS in a fault tolerant manner that is suitable for commodity hardware. Big data plays a important role in different fields like Healthcare, Internet of things, Education, Agriculture, Government Sector, Media and Entertainment etc. Hadoop is designed for processing of large data sets and solves the problems related to analytics on big data scale. This whole paper gives brief description about the Big data tools in tabular as well as in well explained manner. Most of the tools can be used in private or public sectors, helps to save time and increase production of company bigger and better.

Conclusion: This paper concludes that necessity of Big data increases day by day in today's world. There are different fields in which Big data is used largely. These fields are Healthcare, Internet of Things, Education, Government sector, Agriculture, Media and Entertainment and so on. For Handling large data there are different Big data tools are available like Hadoop, HPCC, Storm, Pig, Hive, MongoDB, RapidMiner, Kafka, Cassandra, Sqoop, etc. Big data is the future so currently a lot of research is going on in this field. As data is increasing at faster ratethere is demand of such tools are increasing in market day by day thus there is a must need of such tools and technology which can deal with it.

References:

ISSN: 2233-7857 IJFGCN

[1] Pradeep S and Jagadish S Kallimani.(2019). The Different Tools and Technique to Handle Challenges in Big Data. Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019).

[2] Toshifa, Aniruddh Sanga, Shweta Mongia*(2019). Big Data Hadoop Tools and Technologies: A Review. International Conference on Advancements in Computing & Management (ICACM-2019).

[3] Taiwo Kolajo, Emeka Ogbuju, Sunday Eric Adewumi(2017). TRENDS AND TECHNOLOGIES IN BIG DATA ANALYTICS: A REVIEW. Confluence Journal of Pure and Applied Sciences (CJPAS).

[4] Dharminder Yadav, Umesh Chandra(2017). Modern Technologies of BigData Analytics: Case study on Hadoop Platform. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).

[5] Jaskaran Singh, Varun Singla(2015). Big Data: Tools and Technologies in Big Data. International Journal of Computer Applications (0975 – 8887) Volume 112 – No 15, February 2015.

[6] J. Vijayaraj, R. Saravanan, P. Victer Paul, R. Raju.(2016). A COMPREHENSIVE SURVEY ON BIG DA TA ANALYTICS TOOLS. 2016 Online International Conference on Green Engineering and Technologies (IC-GET).

[7] Varsha B.Bobade.(2016). Survey Paper on Big Data and Hadoop. International Research Journal of Engineering and Technology (IRJET).

[8] J.Archenaa and E.A.Mary Anita.(2015). A Survey Of Big Data Analytics in Healthcare and Government. 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[9] https://cassandra.apache.org/

[10] http://sqoop.apache.org/

[11] https://hadoop.apache.org/

[12] https://pig.apache.org/

[13] https://hive.apache.org/

[14] https://spark.apache.org/

[15] https://storm.apache.org/

[16] Rahul Kumar Chawda, Dr. Ghanshyam Thakur.(2016). Big Data and Advanced Analytics Tools. 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

[17] https://www.wikipedia.org/

[18] Hsi-Yuan Chang1, Jyun-Jie Wang2, Chi-Yuan Lin2,ī, and Chin-Hsing Chen1. (2018). An Agricultural Data Gathering Platform Based on Internet of Things and Big Data. 2018 International Symposium on Computer, Consumer and Control (IS3C).

ISSN: 2233-7857 IJFGCN

[19] MOHSEN MARJANI1, FARIZA NASARUDDIN2, ABDULLAH GANI1, (Senior Member, IEEE), AHMAD KARIM3, IBRAHIM ABAKER TARGIO HASHEM1, AISHA SIDDIQA1, AND IBRAR YAQOOB1.(2017). Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. Digital Object Identifier 10.1109/ACCESS.2017.2689040.

[20]Dhumane, A., & Prasad, R. (2015). Routing challenges in internet of things. CSI Communications.

[21] Dhumane, A. V., Prasad, R. S., & Prasad, J. R. (2017). An optimal routing algorithm for internet of things enabling technologies. *International Journal of Rough Sets and Data Analysis*, *4*(3), 1–16.

ISSN: 2233-7857 IJFGCN