Assessment of the Various Techniques and the Latest Tools for the Big Data Analytics

Manu Raj Moudgil, Anil Kumar Lamba

Abstract: Big data analytics plays a vital role in today’s environment and a need for the researchers, academicians and industry people. The digital information over the internet is spreading with a very high speed by facebook likes/posts, blogs, tweets, articles, news, website clicks, youtube videos etc in an unstructured form. On the daily basis people around the globe which are in billions fetching, uploading and sharing the information through laptops and mobiles on social media platforms. The data includes structured and unstructured information in the form of blogs, google locations, pictures, videos, voice messages and text etc. The question arises how to process this huge information, because the basic techniques of data processing are not enough to handle the heterogeneous, enormous and Exponential data. Online marketing and E-Commerce has become very famous in recent times because all types of businesses are mostly depend on the online transactions and services provided by the seller. Big data analytics has demonstrated to be an aid for such an industry as it removes valuable examples and obscure connections of potential buyer showcase, customer inclinations, purchasing traits and part of other data from mind boggling information sources. The different problems specified above can be resolved by using latest tools available. This paper focuses to provide detailed information about the latest tools and frameworks which are used for big data analytics with comparative assessment.

Keywords— Big Data, Data Analytics, Cassandra, Hadoop, MongoDB.

I. INTRODUCTION

The data that is highly voluminous which is created, removed as well as shared in a jiffy. It also varied in different forms such as collection of structured, unstructured and complex data sets. All above mentioned features are termed as ‘big data’ such data or we can say big data draws the attention of IT industry because of its application in majority of areas such as health care, banking, social media etc. Processing and analyzing of data in traditional means rely on structured form of data which is organized by limited data set. Such techniques and tools are failed to put any value to big data aspects. Variety, volume, veracity, velocity, complexity and variability are the six parameters of big data which make the data processing heavy for old data management tools and techniques.[1][2].

Over the period of time, the management of data volume storage as digital storage capacity has been increased by chip, hard disk, extensible storage in mobile phone and specially cloud service supported by many service providers. It is also challenging aspect to manage this huge repository of data. Although the cloud storage has made it easy to deal with data storage issues, yet it has risk also for the security associated with information.

It is still remain to analysis the biggest challenge and mining of data which is unstructured on the go as it generate over internet. Therefore, in today scenario big data analysis plays a crucial role.

For the professional big data analytics field is very promising. Data storage, consistency, privacy, timeliness, heterogeneous data sources, representation and integrity creates a lot of challenges. Also, the representation and the organization is quite challenging as it contains the huge repository of data. Different pre-processing data techniques like noise elimination, filtering, transformation and classification has its own challenges [3]. Because of these characteristics makes the big data analytics field interesting. Various types of techniques and tools are created to make the data analysis process easier. This paper delivers a brief analysis of those tools.

This paper is ordered in such a manner as the Section-II describes the Lifecycle of Big Data analytics, Section-III shows the various stages of Big Data analytics and the comparative analysis of the tools, Section-IV completes the work with the conclusions.

II. BIG DATA ANALYTICS METHODOLOGY

This section describes the different phases of lifecycle of big data analytics [4]

A) Data Collection and Identification: On the basis of severity of problem the wide variety of data storage is identified in this phase. More the data resources will give more chances of finding concealed patterns and correlation. There is need of such tools to pick up the keywords, information and data from the data sources.

B) Data Storage: The collected structure as well as unstructured data used to be stored in database. To accommodate the Big Data there is no SQL database needed. The organization like oracle, apache has developed various database and framework that permit the analytic tools to collect and process the data from the repositories.

C) Noise Elimination and Data Filtering: In this phase replicated, null, irrelevant and corrupt data objects are removed from the gathered information. Though, in the analysis or context there is importance of removed and filtered data. So it is necessary to make a copy of the actual data sets to storage device in the compressed form.[4]

D) Data extraction and classification: In this phase extraction of incongruent data and convert the format of data so that different analytic tools can be used. It also includes text and other relevant fields for reducing the data volume which is submitted to the engine.

E) Data, validation, cleansing & aggregation: In this phase the data which is extracted. from analysis are validated by the rules on the business to confirm its need. It is also very difficult
to apply constraints due to complexity on the data. With the help of aggregation multiple data combines into fewer numbers based on the common fields. Further processing of data is simplified.

F) Data analysis and processing: In this stage analysis and data mining carries out to make hidden and unique patterns for business decisions. The technique of data analytics depends upon exploratory, prescriptive, predictive, diagnostic, descriptive or confirmatory.[4]

G) Data visualization: This phase represents analysis results into the graphical and visual forms which makes it very easier for the audience to understand.

III. TOOLS FOR COMPARATIVE ASSESSMENT OF BIG DATA ANALYTICS

By the arrival of Big Data, set of tools designed by the programmers and different organisations to support the development of data analysis. Based on the usage and implementation, set of tools have been categorized into diverse phases of Big Data Life-cycle. This part categorizes and compares few of the utmost standard and extensively used tools.

(a) Tools used for Data Collection:
However type of data sources recognized and business case scenario mainly dependent on the data collection. Unstructured data is taken frequently from social-networking. With the help of semantic and text analysis that is already embedded in different websites, we can collect data from the different tools available. Below mentioned table compares such tools for data collection.

Table I
Comparison of Popular Data Collection Tools [3]

| Tools Used | Open/License / Enterprise solution | Type of analysis | Analysis Engine | Deploy ability |
|------------|-----------------------------------|------------------|-----------------|---------------|
| Open-Text  | Enterprise                        | Content Management & Analysis | Red Dot, Captiva | Window Based Server Application |
| Trackur    | Proprietary License               | Influence & Sentiment Analysis | Trackur          | Web (Social Media) |
| Opinion- Crawl | Open Website                     | Sentiment Analysis  | Sensebot        | Web |
| Semantria  | Proprietary License               | Text & Sentiment Analysis | NLP Based       | Web, Cloud API, Excel |

(b) Tools Used For Data Storage Tools & Frameworks:
Database framework contains maximum of the tools used for data processing and analysis. Database solutions and frameworks are providing by few of the standard companies. Subsequent table offers abridged assessment of these widespread NoSQL databases.

Table II Comparison of Popular Data Storage Tools [5]

| NoSQL Databases | Characteristics of Tools Used |
|-----------------|------------------------------|
|                 | Zero Downtime (on node failure) | Secondary Indexes | Data Model | Concurrency |
| Apache-Hbase (Hadoop-Database) | Yes | No | Column-Oriented | Yes (Optimistic Concurrency) |
| Couch-DB | No | Yes | Document-Oriented | Yes (Optimistic Concurrency) |
| Mango-DB | No | Yes | Document-Oriented | Yes |
| Apache-Cassandra | Yes | No | Column-Oriented | Yes |
| Apache-Ignite | Multi-Model | Yes | Yes | Yes |
| Oracle-NoSQL Database | Key-Value Based | No | Yes | Yes |

(c) Tools used for Data Filtering & Extraction:
When we want to create structured output from unstructured data collected from previous, few Data filtering and extraction tools are used. Few of these tools are equated under neath.

Table III Comparison of Popular Data Filtering & Extraction Tools [7]

| Tools Used | Exten sible | Free/ Paid version | Feature | Output |
|------------|-------------|--------------------|---------|--------|
| Content Grabber | Yes | Paid Version | Web Scrapping With Debugging & Error Handling | Structured Data (XML, CSV & Databases) |
| Octo-Parse | No | Both Free & Paid Version | Web Scrapping | Structured Spreadsheets |
| Parse-Hub | No | Both Free & Paid Version | Cloud-Based Desktop App | Excel, CSV, Google Sheet |
| Mozenda | Yes | Paid Enterprise &Professional Version | Web Scraper | Structured Data (JSON, XML & CSV) |
| Pentaho | Yes | Both Free & Paid Version | ETL &Data Mining Capabilities | Structured Data |

DOI: 10.35940/IJITEE.B6679.129219
(d) Tools used for Data Cleaning and Validation: When we are dealing with data analytics tools and engines, Data cleaning tools are extremely helpful in reducing the processing time and computational speed for the same. However, these are used occasionally not as frequently as other tools. An important contrast of modern data cleaning tools is given in the table underneath.

### Table IV
**Comparison of Popular Data Cleaning Tools** [6]

| Tools Used          | Characteristics of Tools Used                  | Additional features               |
|---------------------|------------------------------------------------|----------------------------------|
| Talend              | Numerous Databases                             | Streaming, Batch Processing      |
|                     |                                                 | Data Integration                 |
| Open-Refine         | Web Services And External Data                 | Batch Processing                 |
|                     |                                                 | Transforming Data From One Form To Another |
| Rapidminer          | Internal Database Integration                  | GUI & Batch Processing           |
|                     |                                                 | Filtering, Aggregation & Merging |
| Data-Cleaner        | Integration With HadoopDatabase                | Record &Field Processing         |
|                     |                                                 | Data Transformation, Validation & Reporting |
| Map-Reduce          | HadoopDatabase                                 | Parallel Data Processing         |
|                     |                                                 | Searching, Sorting, Clustering & Translation |

(e) Tools used for Data Analysis: Maximum of tools in this class are not simply analysis tools but carry out other purpose too. Though, they organize artificial intelligence, data mining & other methods for data analysis. A bridle of these tools is given in the table underneath.

### Table V
**Comparison of Popular Data Analysis Tools** [8][9]

| Tools Used          | Characteristics of Tools Used                  |
|---------------------|------------------------------------------------|
| Qubole              | Languages Supported Python,Scala,R, Go         |
|                     | Delay: Seconds                                 |
|                     | Processing Representation: Stream Processing,Ad-Hoc Queries |
| Hive                | Languages Supported SQL-Like                   |
|                     | Delay: High                                   |
|                     | Processing Representation: Streaming           |
| Map-Reduce          | Languages Supported Java, Ruby, Python, C++    |
|                     | Delay: More (Seconds)                          |
|                     | Processing Representation: Parallel Processing |
| Flink               | Languages Supported Scala,Java,Python          |
|                     | Delay: Seconds                                 |
|                     | Processing Representation: Batch,&Stram Processing |
| Apache-Storm        | Languages Supported Any                        |
|                     | Delay: Milli-Seconds                           |
|                     | Processing Representation: A Record At A Time  |
| Apache-Spark        | Languages Supported Scala,Java,Python          |
|                     | Delay: Seconds                                 |
|                     | Processing Representation: Mini/ Micro Batches, Streaming |

(f) Tools Used For Data Visualization:

In the market, a lot of data visualization tools are available and maximum of them are incorporated of data analysis, visualization & extraction. The below mentioned table compares nearly all the accepted and broadly used tools for data visualization.

### Table VI
**Comparison Of Popular Data Visualization Tools** [11][12]

| Tools Used          | Characteristics of Tools Used                  |
|---------------------|------------------------------------------------|
| Gephi               | Licensed/ Open-source                           |
|                     | Coding/Programming Language need                |
|                     | Output features: Graphs & Networks              |
| Carto-DB            | Open-source                                    |
|                     | Own Visual Query Language                       |
|                     | Output features: Line/Bar/ Pie Charts, Dashboard Sharing As Pdf Reports |
| Tableau             | Open-source                                    |
|                     | Coding is not required.                         |
|                     | Output features: Scatter Plots, Bar Charts, Scatter Plots, Trees, Dendrograms, Networks And Heat Maps |
| Orange              | Open-source                                    |
|                     | No need for Programming                         |
|                     | Output features: Scatter Plots, Bar Charts, Trees, Dendrograms, Networks And Heat Maps |
| Qlik                | Licensed                                       |
|                     | Need for Programming Language & SQL Knowledge   |
|                     | Output features: Dashboard, Apps                |
| Google’s Web Service| No need for Programming                         |
|                     | Output features: Pie Charts, Bar Charts, Line Plots, Scatter Plots, Timelines |
| Data-Wrapper        | Open-source                                    |
|                     | Ready-To-Use Codes                             |
|                     | Output features: Bar Chart,Line Chart, Plans Graphs |

IV. CONCLUSION

In today’s environment the issues of Big Data Analytics are not addressed fully by the presently available tools. The information development rate is much higher than the processing tools of information development. Ignite Hadoop & Cassandra are the high-tech tools and methods that can’t justify the real-time investigation in truly means. However they have reasonably improved the simplicity of conducting different data sets and also reduces the data processing time. Still there are few unexplained topics associated to efficient storage, security, sharing, searching and analysis. This paper covers the latest developments, future enhancements and improvements of tools for Big Data Analytics. We have discussed different latest tools for data analysis, visualization, cleaning, cleaning and validation and storage tools in this way the different problems related to structured and unstructured data can be easily resolved.
REFERENCES

1. S. Mujawar, S. Kulkarni, “Big Data: Tools and Applications”, International Journal of Computer Applications, vol. 115, No. 23, pp. 7-11, 2015.
2. M. Chen, S. Mao, and Y. Liu, “Big data: a survey”, Mobile Networks and Applications, vol. 19, No. 2, pp. 171–209, 2014.
3. N. Khan et. al, “Big Data: Survey, Technologies, Opportunities, and Challenges”, The Scientific World Journal, vol.2014, Issue.4, pp.1-18, 2014.
4. T. Eri, W. Khattak, and P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Prentice Hall, India, pp. 65-88, 2015.
5. Online source, [Available] https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/, 2018.
6. Online source, [Available] https://www.guru99.com/big-data-tools.html, 2018.
7. Online source, [Available] https://www.octoparse.com/blog/yes-there-is-such-thing-as-a-free-web-scraper/, 2018.
8. https://data-flair.training/blogs/apache-storm-vs-spark-streaming/.
9. A. Narang, “A review-Cloud and cloud security”, International journal of Computer Science and mobile Computing, vol. 6, issue 1, pp. 178-181, 2017.
10. K. Komal, “Cognitive Science: Bridging the Gap between Machine and Human Intelligence”, International Journal of Computer Applications, vol. 114, issue 5, pp. 16-19,2015.
11. S Kaushal, J.K. Bajwa, “Analytical Review of User Perceived Testing Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, issue 10, 2012.
12. S. M. Ali et.al, “Big Data Visualization: Tools and Challenges”, 2nd International Conference on Contemporary Computing and Informatics,2016.

AUTHORS PROFILE

Dr. Manu Raj Moudgil is working as a Professor in the department of Computer Science & Engineering at Chandigarh Group of Colleges, Technical Campus, Jhanjeri, Mohali, (Punjab). He has rich experience of more than 15 years in teaching of graduate and post graduate classes of Engineering students, also currently guiding and guided many M.Tech thesis and Phd Students for the research work. He has more than 50 quality research publications in International Journals/Conferences and his area of research is Natural Language Processing, Machine translation systems, High level Languages and Computer Networks. Beyond this he has got many awards like Teacher of the Year for the session 2008-09 and 2010-11 for the Excellency in Teaching. He is also awarded excellent research paper many times in the international conferences, recent one in Melbourne, Australia in 2018.

Dr. Anil Kumar Lamba is working as a Professor & Head in the department of Computer Science & Engineering at Chandigarh Group of Colleges, Technical Campus, Jhanjeri, Mohali (Punjab). He has rich experience of more than 22 years in teaching of graduate and post graduate classes of Engineering students, also currently guiding and guided many M.Tech thesis and PHD Students for the research work. He has more than 20 quality research publications in International Journals/Conferences and his area of research is Security in mobile adhoc networks, High level Languages. He had written international book on load distribution in peer to peer networks and some more in progress.