International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585
Volume 4, Issue 5, May-2018

# A Review Paper on Big Data Analytics Tools

Ms. Komal[1]
[1] *Department of Computer Science, Amity University Haryana, Komal.sang@gmail.com,*

***Abstract*— Big Data analytics has become the need of the hour for academia, research and IT industry. The exponentially growing digital information is moving at a lightning fast speed over the internet infrastructure and is mainly in the unstructured form including Facebook posts/likes, tweets, blogs, news, articles, YouTube videos, website clicks etc. Everyday billions of people fetch, upload and share information on social media and other platforms through mobile phones, laptops, PDAs. The information comprises of pictures, blobs, goggle map locations, videos, text, voice messages that are collection of structured, unstructured and complex data objects. Traditional data processing techniques are insufficient to handle this enormous, heterogeneous and fast-paced data. E-commerce and digital marketing has gained so much popularity over these years that business industry has become more dependent on online transactions and services. Big data analytics has proven to be a boon for such an industry as it helps to extract useful patterns and unknown correlations of potential consumer market, client preferences, buying attributes and lot of other information from intricate data sources. This paper aims to provide a detailed review and comparative assessment of latest tools and frameworks used for big data analytics.***

***Keywords*— Big Data, Data Analytics, Hadoop, MapReduce, Cassandra, MangoDB.***

## I. INTRODUCTION

The term 'Big Data' is characterized by three things- a) it is highly voluminous 2) it is created, shared and removed online in fraction of seconds 3) it is in varied forms i.e. collection of structured, unstructured and complex datasets. Big data analytics has quickly drawn the attention of IT industry due to its application in majority of areas like healthcare, business firms, social media, education, banking [1] etc. Traditional means of processing and analyzing data mainly rely on limited data set organized in a structured form. Such tools and techniques fail to add any value in big data aspects. The six parameters of big data [2]- volume, variety, velocity, veracity, variability and complexity make the data processing cumbersome for old data management tools and techniques.

Volume of data can still be managed as the digital storage capacities have increased over the period of time leading to cheap hard disks, large and extensible storage in mobile phones and above all cloud services supported by many service providers. Management of this huge repository of data is another challenging aspect. Though cloud storage has eased out data storage issues, it has additional risk of information security associated with it. The biggest challenge still remains to be the analysis and mining of the unstructured data on the go as it is generated over internet. Thus, big data analytics play a crucial role in today's scenario.

This promising field of big data analytics comes together with many challenges for the professionals. Data inconsistency, integrity, privacy, timeliness, storage & representation, unstructured heterogeneous data sources pose lot of challenges. Efficient organization and representation of this huge repository of data is quite challenging. Various data pre-processing techniques such as filtering, noise elimination, classification and transformation have their own challenges [4].These aspects make the field of big data analytics even more interesting. Lots of tools and techniques have been developed so far to ease out the process of data analysis. The paper provides a summarized review of these tools.

The paper is further organized as follows: Section II explains the lifecycle of Big Data analytics. Section III contains comparative analysis of tools used at various stages of Big Data analytics. Section IV concludes the work with the findings.

 

## II. METHODOLOGY OF BIG DATA ANALYTICS

This section explains various stages of lifecycle of big data analytics [3]-

A. *Data identification and collection*- In this phase, wide variety of data sources are identified depending upon the severity of problem. More data resources mean more chances of finding hidden correlations and patterns. Tools are needed to capture keywords, data and information from these heterogeneous data sources.

B. *Data storage*- The captured structured and unstructured data need to be stored in databases/ data warehouse. NoSQL databases are needed to accommodate Big Data. Various frameworks and databases have been developed by organizations like Apache, Oracle etc. that allow analytics tools to fetch and process data from these repositories.

C. *Data filtering and noise elimination*- This phase is dedicated to removal of replicated, corrupt, null and irrelevant data objects from the gathered information. However, filtered and removed data might be of some importance in another context or analysis. Hence, it is advisable to keep a copy of original data sets in compressed form to save storage space [3].

D. *Data classification and extraction*- This phase is responsible for extracting incongruent data and converting it into a common data format that the underlying analytics tool can use for its purpose. This may also involve extracting relevant fields or texts to reduce the volume of data to be submitted to analytics engine.

E. *Data cleansing, validation and aggregation*- This stage applies validation rules based on the business case to confirm the necessity and relevance of data extracted for analysis. Although it may be difficult sometimes to apply validation constraints to the extracted data due to complexity. Aggregation helps to combine multiple data sets into fewer numbers based on common fields. This simplifies further data processing.

F. *Data analysis and processing*- This stage carries out actual data mining and analysis to establish unique and hidden patterns for making business decisions. Data analytics technique may vary depending upon the scenario i.e. exploratory, confirmatory, predictive, prescriptive, diagnostic or descriptive [3].

G. *Data visualization*- This phase involves representation of analysis results into visual or graphical form that makes it easier to understand for the audience.

## III. COMPARATIVE ASSESSMENT OF BIG DATA ANALYTICS TOOLS

Since the advent of big data, a number of tools have been developed by programmers and agencies to assist in the process of data analysis. These tools have been categorized into different stages of big data lifecycle based on their usage and implementation. This section classifies and compares some of the most popular and widely used tools.

### *A. Data Collection tools*

Though data collection is dependent on business case scenario and type of data sources identified. Unstructured data is captured mostly from social networking. There are some popular tools to collect data from embedded websites with the help of semantic and text analysis. Below table compares such data collection tools.

TABLE I
COMPARISON OF POPULAR DATA COLLECTION TOOLS [4]

| **Tool** | **Characteristics of tool** | | | |
|---|---|---|---|---|
| | **Type of analysis** | **Analysis Engine** | **Deployability** | **Open/License/ Enterprise solution** |
| Semantria | Text and sentiment analysis | NLP based | Web, cloud API, Excel | Proprietary License |
| Opinion Crawl | Sentiment analysis | SenseBot | Web | Open website |
| OpenText | Content management and analysis | Red Dot, Captiva | Window based server application | Enterprise |
| Trackur | Influence and sentiment analysis | Trackur | Web (social media) | Proprietary License |

*B. Data Storage tools and frameworks*

Most of the data processing and analysis tools work on top of a database framework. Hence, some of the popular companies have come into the league of providing database solutions and frameworks. Following table provides a summarized assessment of these popular NoSQL databases.

TABLE III
COMPARISON OF POPULAR DATA STORAGE TOOLS [5]

| **NoSQL Databases** | **Characteristics of tool** | | | |
|---|---|---|---|---|
| | **Data Model** | **Zero Downtime (on node failure)** | **Concurrency** | **Secondary Indexes** |
| Apache HBase (Hadoop database) | Column-oriented | Yes | Yes (optimistic concurrency) | No |
| CouchDB | Document-oriented | No | Yes (optimistic concurrency) | Yes |
| MangoDB | Document-oriented | No | Yes | Yes |
| Apache Cassandra | Column-oriented | Yes | Yes | No |
| Apache Ignite | Multi-model | Yes | Yes | Yes |
| Oracle NoSQL Database | Key-value based | No | Yes | Yes |

*C. Data filtering and extraction tools*

Data filtering and extraction tools are used to create structured output from unstructured data gathered in previous stages. Some of these tools are compared below.

TABLE IIIII
COMPARISON OF POPULAR DATA FILTERING AND EXTRACTION TOOLS [7]

| **Tool** | **Characteristics of tool** | | | |
|---|---|---|---|---|
| | **Free/ Paid version** | **Extensible** | **Feature** | **Output** |
| Pentaho | Both free and enterprise paid version | Yes | ETL and data mining capabilities | Structured data |
| OctoParse | Both free and paid version | No | Web scrapping | Structured spreadsheets |
| ParseHub | Both free and paid version | No | Cloud-based desktop app | Excel,CSV, Google sheet |
| Mozenda | Paid Enterprise and Professional version | Yes | Web scraper | Structured data (JSON,XML and CSV) |
| Content Grabber | Paid version | Yes | Web scrapping with debugging and error handling | Structured data (XML, CSV and databases) |

*D. Data cleaning and validation tools*

Data cleaning tools are extremely helpful in reducing the processing time and computational speed of data analytics tools and engines. Though, they are not used as often as other tools. A significant comparison of latest data cleaning tools is provided in the table below.

TABLE IVV
COMPARISON OF POPULAR DATA CLEANING TOOLS [6]

| **Tool** | **Characteristics of tool** | | |
|---|---|---|---|
| | ***Processing model*** | ***Additional features*** | ***Data Source*** |
| DataCleaner | Record and field processig | Data transformation, validation and reporting | Integration with Hadoop database |
| MapReduce | Parallel data processing | Searching, sorting, clustering and translation | Hadoop database |
| Rapidminer | GUI and batch processing | Filtering, aggregation and merging | Internal database integration |
| OpenRefine | Batch processing | Transforming data from one form to another | Web services and external data |
| Talend | Streaming, batch processing | Data integration | Numerous databases |

*E. Data analysis tools*

Most of the tools in this category are not only analysis tools but perform other functions too. However, they deploy data mining, artificial intelligence and other techniques for data analysis. A summarized review of these tools is provided in the table below.

TABLE V
COMPARISON OF POPULAR DATA ANALYSIS TOOLS [8][9]

| **Tool** | **Characteristics of tool** | | |
|---|---|---|---|
| | ***Processing model*** | ***Language support*** | ***latency*** |
| Hive | Streaming | SQL-like | high |
| Apache Spark | Mini/ micro batches, streaming | Scala,Java, Python | seconds |
| Apache Storm | A record at a time | Any | milli-seconds |
| MapReduce | Parallel Processing | Java, Ruby, Python, C++ | More (seconds) |
| Qubole | Stram processing,ad-hoc queries | Python,Scala, R, Go | seconds |
| Flink | Batch and stram processing | Scala,Java, Python | seconds |

*F. Data Visualization tools*

There are plenty of data visualization tools available in the market. Most of them are integrated version of data extraction, analysis and visualization. Following table compares most popular and widely used data visualization tools.

TABLE VV
COMPARISON OF POPULAR DATA VISUALIZATION TOOLS [11][12]

| **Tool** | **Characteristics of tool** | | | |
|---|---|---|---|---|
| | ***Licensed/ Open-source*** | ***Data Source compatibility*** | ***Coding/Programming Language need*** | ***Output features*** |
| DataWrapper | Open-source | CSV,PDF,Excel, CMS | Ready-to-use codes | Bar chart,line chart,map, graphs |
| Tableau | Open-source | Database , API | No coding | Maps, Bar charts, Scatter plots |
| Orange | Open-source | files, SQL tables, and data tables or can paint random data | No programming needed | scatter plots, bar charts, trees, dendrograms, networks and heat maps |
| Qlik | Licensed | Database, spreadsheet, website | Programming language and SQL knowledge needed | Dashboard, Apps |
| Google Fusion tables | Google's web service | Comma-separated value file formats | No programming needed | pie charts, bar charts, lineplots, scatterplots, timelines |
| CartoDB | Open-source | Location data, plenty of data types | CartoCSS language | Maps |
| Chartio | Open-source | Multiple Data sources | Own visual query language | Line/bar/ pie charts, dashboard sharing as pdf reports |
| Gephi | Open-source | CSV, GraphML, GML,GDF Spread-sheet | No programming | Graphs and networks |

## IV. CONCLUSIONS

The rate of development of information processing tools is comparatively much slower than the rate of development of information. Currently available tools in the market do not address all the issues of Big Data analytics. Even the most high-tech tools and techniques like Hadoop, Cassandra and Ignite can't justify real-time analysis in true sense. Though they have fairly increased the ease of handling diverse data sets and reduced the time of data processing. There are still some unaddressed issues related to effective storage, searching, analysis, sharing and security. This paves a way for future improvements and developments of Big Data analytics tools.

.

## REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, "Big data: a survey", *Mobile Networks and Applications*, vol. 19, No. 2, pp. 171–209, 2014.

[2] S. Mujawar, S. Kulkami, "Big Data: Tools and Applications", *International Journal of Computer Applications*, vol. 115, No. 23, pp. 7-11, 2015.

[3] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals: Concepts, Drivers & Techniques*, Prentice Hall, India, pp. 65-88, 2015.

[4] N. Khan et. al, "Big Data: Survey, Technologies, Opportunities, and Challenges", *The Scientific World Journ*al, vol.2014, Issue.4, pp.1-18, 2014.

[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 Applicatio*ns, 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.