International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | Apr-2016 www.irjet.net p-ISSN: 2395-0072

A Survey on Tools, Technologies and challenges in Big Data

Nivetha.D[1], Shelgin.S[2]

[1] *PG Scholar, Dept. of Computer science and Engineering, Valliammai Engineering college, Chennai ,India*
[2] *PG Scholar, Dept. of Computer science and Engineering, Valliammai Engineering college, Chennai, India*

Abstract - *For every industry, organizations, business function data has become very essential. Data that is so large in volume, so distinct in variety or moving with such velocity is called Big data. The big data introduce the statistical challenges that include timeliness, data heterogeneity, scale and complexity. This challenge require computational and statistical prototype. In this paper we review the characteristic, tools, technologies and challenges of big data.*

Key Words: **Volume, Velocity, Variety, Hadoop, Map Reduce, Big data**

1. INTRODUCTION

In the last decade of computing in computer science, the idea of big data was introduced. High volume of data is the key factor in big data and it needs advanced methods to get processed. A new storage concept is introduced when the data gets increased rapidly and it helps in easy data retrieve. Big data is used widely in the world because of enormous increase in world data. Real time analysis is much needed in big data which has masses of unstructured data. To discover new values, many opportunities were created by big data. It helps to understand hidden values in depth. Many major plans were implemented by government agencies and Industries for research and applications in big data because of its high potential [1]. Public media which covers big data often are, The economist [2, 3], new York times [4] and National public radio [5, 6]. Nature and science are the specific premier journals, allocates special columns to debate about improvisation in big data [7, 8].

Many challenging problems demanding appropriate solutions were created because of steady increasing in large datasets. Generating data becomes easier becomes easier because of latest advancement in Information Technology. For example, in you tube for every minute 72 hours of video were uploaded [9]. Collecting and integrating massive data from widely distributed data sources is the main challenge to be faced.

- Sharp growth of data is promoted due to the sudden growth of cloud computing and Internet of things (IoT). Safeguarding, access sites and channels for data asset are the services provided by cloud computing. Likewise in IoT paradigm, Sensors collected from all over the world are transmitted to form data which is to be stored and processed in the cloud. The capacities of the IT architectures and infrastructure of existing enterprises is surpassed by the data which has both quantity and mutual relations and also its available computing capacity is over stressed by its real time requirement.

- Storing and management of heterogeneous datasets with moderate requirements on software and hardware infrastructure is the major problem caused by increase in growth of data.

- To "mine" the datasets at different levels; scalability, complexity, Heterogeneity, real time and privacy of big data is considered. The different levels are during analysis, modelling, visualization and forecasting of datasets. To improve decision making, its intrinsic property is revealed.

2. CHARACTERISTICS OF BIG DATA

Large Hidden values are diverse, complex and of a massive scale. To uncover large hidden values big data requires new forms of integration. Big data is a set of techniques and technology. The big data characteristics are shown in Fig.1

2.1 Volume

It refers to the continuous expansion of enormous amount of different data types generated from various sources. Through data analysis, Hidden information and patterns are created. It is the main advantage in gathering large amount of data. A unique collection of longitudinal data from smart phones is provided by Laurila et al. [10].This collection is also for research community for the future purposes. Nokia makes a aforesaid initiative called mobile data challenge and made a motivation to other companies [10]. More efforts and investments are necessary for collecting longitudinal data. An interesting result is produced by the mobile data challenge which is as same as the predictability examination of Human behavioural patterns.

The result also means in Complex data visualization techniques and sharing Human mobility data.

2.2 Variety

Sensors, smartphones and social networks are the various sources from which different types of data are collected. Video, text, image, audio, run structured format, etc. are some of the data types shared. Unstructured format data types are formed mostly in mobile applications. Text messages, blogs, online games, etc. are some of the unstructured data types which is formed in mobile applications. Extremely diverse set of unstructured and structured data are created by internet users [11].

2.3 Velocity

It refers to the data transferring speed from one device to another. Absorption of complementary data collections, streamed data arriving from multiple sources and introduction of previous archived data or legacy collections are the changes made constantly by the contents of data [12].

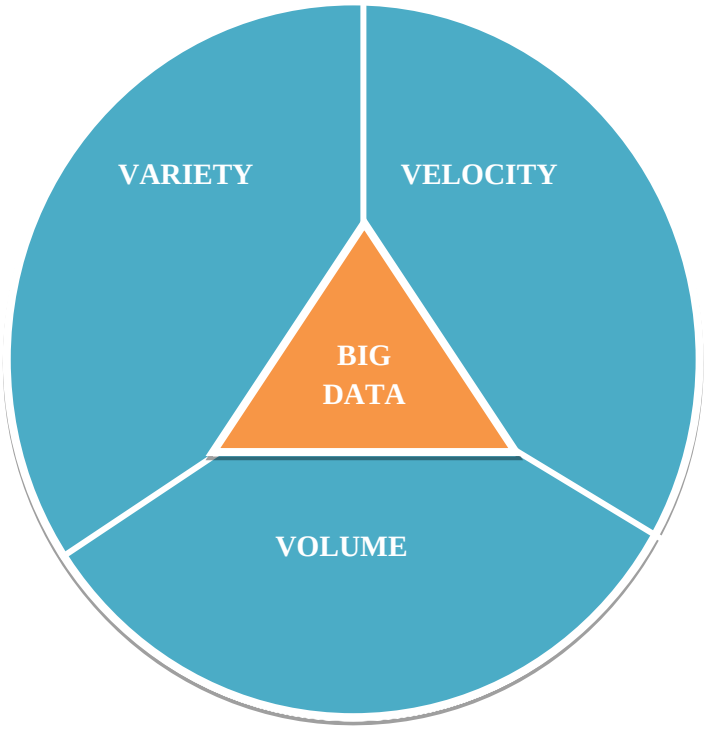

Fig 1 Characteristics of big data

3. TOOLS FOR ANALYSING BIG DATA:

To analyze big data and generating insight, five key approaches are used.

3.1 Discovery tools:

For rapid, intuitive exploration and to analyze information in the information lifecycle, Discovery tools are used. The information is taken from any combination of unstructured and structured sources. Traditional BI source systems are permitted by these tools along with analysis of source systems. It is independent that the users can draw new insights, make informed decisions quickly and come to meaningful conclusions. It is due to no need for up-front modeling.

3.2 BI Tools:

Analysis and performance management, reporting, primarily with transactional data from production information systems and data warehouses are the important factors of BI tools. For Business intelligence and performance management, dashboards, including enterprise reporting, and what-if scenario analysis on an integrated, enterprise scale platform, ad-hoc analysis and scorecards; comprehensive capabilities were provided by BI tools.

3.3 In-Database Analytics:

For finding patterns and relationships in your data, variety of techniques was implemented by In-database analytics. Total cost of ownership is reduced by eliminating data movement from other analytical servers, which boosts information cycle times. It is due to direct application of these tools within the database.

3.4 Hadoop:

To identity macro trends or find nuggets of information, hadoop tools are used. Hadoop is used to pre-process the data required. It enables businesses using inexpensive commodity servers, for unlocking the potential value from new data.

3.5 Decision Management:

Predictive modeling, self-learning to take informed action based on the current context and business rules are the factors of decision management. To maximize the value of every customer interaction, individual recommendations across multiple channels is enabled by this type of analysis.

4. TECHNOLOGIES IN BIG DATA:

To promote the development of storage mechanism for big data, considerable researches are made on Big data. Three bottom-up levels were existed in big data storage architectures:

4.1 File systems

Hadoop Distributed File System is to handle enormous and high volumes of data in any structure, Hadoop is designed which is a data processing engine and distributed file system. It is also independent, programming framework based on java. Large datasets are computed in distributed computing environment using Hadoop. Hadoop is taken as a part of Apache project by sponsors offered by the Apache Software Foundation [13]. The components of Hadoop are classified as shown in Fig 2,

- Hadoop distributed file system
- Map reduce

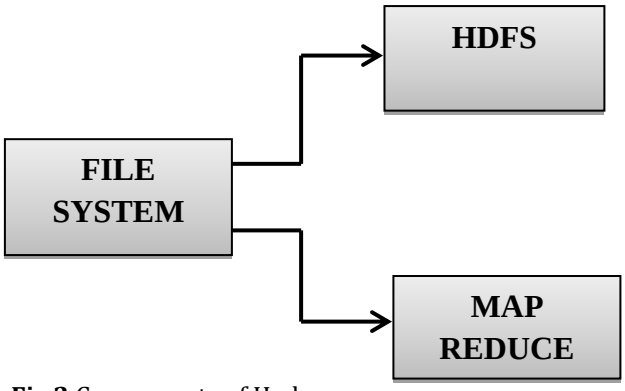

Fig 2 Components of Hadoop

Any form of data like unstructured form, structured relational form or any form in between is supported by HDFS. In Hadoop, an application is splintered down into various small portions called fragments due to Google's Map. In any computer these fragments can run in the cluster. For managing applications on multiple distributed servers, The Map-Reduce programming paradigm is used. Supporting redundancy, parallel processing and distributed architectures is focused. A document is explicitly divided into 64MB "wedges" when a file is copied on HDFS. For authentication, this process is repeated for three times. Given 64MB block exists on three independent nodes when in Hadoop collection all blocks are computed in different systems while distributing. HDFS controls all operation burden of fragmentation, merging and dispensing your data.

4.2 Map Reduce

To store large datasets on commodity hardware, Google introduced Map-Reduce process. In clusters, large scale data records are processed by Map-Reduce. Map() and reduce() are the two basic functions in Map reduce programming model. The input is taken by master node, divided into small sub modules and distributed into slave nodes. This is the task performed by Map function. Hierarchical tree structure is again formed due to division of slave node into sub modules. Result is passed back to master node by processing base problem by slave node. All intermediate pairs are arranged by the map reduce system and produce final output by referring reduce() function. All the results are collected from sub problems and an final output is produced by reduce function which acts as a master node [14].

Map(in_key,in_value)---

>list(out_key,intermediate_value)

Reduce(out_key,list(intermediate_value))---

>list(out_value)

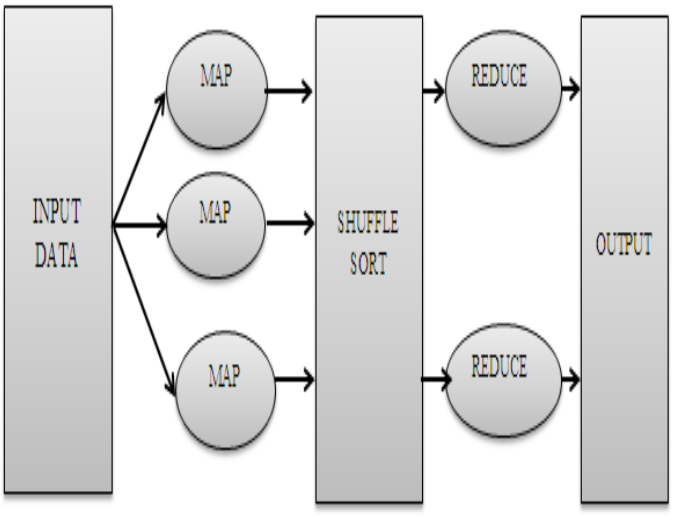

Fig 3 Architecture of Map Reduce

4.3 Hive

A decentralized system for developing applications by networking local system resources is called Hive. It is a distributed agent platform [15]. An element of cloud-based Hadoop ecosystem called HiveQL, offers a query language. It

is also called as Apache hive data warehousing component. Automatically SQL-like queries are converted into map reduce jobs by HiveQL. Map-Reduce-oriented execution and an execution part that receives a query from user or applications for execution are the divisions in this architecture.

4.4 No-SQL

Data management and data design for very large set of distributed data is performed by No-SQL database. In the real time events like process of deploying inbound channels, No-SQL database takes part. It also takes part in relative search applications. The elastic nature of the No-SQL database is the main reason for its participant in these applications. Instead of using advanced developer, Data in scope and domain is used to evolve the queries. To access enormous amount of unstructured data, No-SQL database is used. No-SQL databases have more than one hundred approaches which is specialized in solving very specific challenges of different multi-model data types. Apache Cassandra is the most popular No-SQL database. Open source, Horizontal scalability, Easy to use, store complex data types and very fast for adding new data are the major advantages of No-SQL database. Immaturity, No indexing support, No ACID, Complex consistency models, Absence of standardization are the drawbacks of No-SQL database.

4.5 HPCC

HPCC is an open source and used for computing and handling services of massive big data workflow. According to user end requirements HPCC data model is designed. To manage most complex and data-intensive analytical related problems, HPCC is used. It is basically a single based model (i.e) It is a system of single architecture and single programming language in a single platform used for data simulation. Analyzing gigantic amount of data for solving complex problem of big data is the main purpose of designing HPCC system. Enterprise control language is the basic of HPCC system which is declarative. The main components of HPCC on-procedural nature programming language are:

- HPCC Data Refinery: Parallel ETL engine is used mostly.
- HPCC Data Delivery: The usage of structured query engine is the basic in it.
- Between the nodes in appropriate even load the distribution of workload is made by the Enterprise control language.

5. CHALLENGES

5.1 Failure Handling

It is not an easy process to devise 100% reliable systems. Permitted threshold value should be greater than the probability of failure for devising the systems. Numerous network nodes is involved in start-up of the process and the process becomes heavy while computing. In case of failure in process, it should be restarted. Check points should be retained and the threshold level should be fixed, are major concerns.

5.2 Data heterogeneity

There are unstructured, semi-structured and structured data types in which big data deals with all data types. We should do more research to convert one data form to another (i.e) Computing Unstructured data with structured data.

5.3 Data Quality

A Big asset for both Businessman and IT leaders is the problem due to large amount of data concerns. Amount of relevant data helps in decision making of predictive analysis. The source of derivation is the basic for those relevant data. Source domain helps in accuracy of relevant data and big data depends on this factor because of its huge storage of relevant data. But still there are queries in trusting the data from sources. The solution can be given only if appropriate trust agent filters are fixed.

5.4 Scale and complexity

The main challenge is to manage the data which is huge and rapidly increasing by time. It cannot be controlled by traditional software tools. Due to scalability and complexity there are many other challenges to be analyzed such as data analysis, organization, retrieval and modeling.

5.5 Timeliness

The time taken for analyzing the data will be more due to the size of the data sets which is to be processed. Requirement of analysis of results will be immediate in some cases. For example, Credit card should be flagged before the transaction takes place when it is identified as fraudulent card. But in real time analyzing whole user's purchase

history every time is impossible. In advance, development of partial results is needed to get the determined data by computing small incremental with new data as quick as possible. The classification of big data challenges are shown in Table.1

Table -1: Classification of big data challenges

CLASSIFICATION OF BIG DATA CHALLENGES	
Security	Secure computations in distributed programming frameworks
	Security best practices for non-relational data stores
Data Privacy	Privacy preserving data mining and analytics
	Cryptographically enforced data centric security
	Granular access control
Data Management	Secure data storage and transaction logs
	Granular Audits
	Data provenance
Integrity and reactive security	End point validation and filtering
	Real time security monitoring

6. CONCLUSION

Due to explosion of social network sites, media sharing etc., the amount of data is growing aggressively. In this paper, the tools technologies and challenges of big data are surveyed. HDFS and Map Reduce is the big data analytic tool which helps the organization for better understanding with customers. The main goal of our paper is to make survey on technologies and tool of big data which handle the large amount of data to improve the performance.

REFERENCES

[1] Fact sheet: Big data across the federal government (2012).http://www.whitehouse.gov/sites/default/files/microsites/ostp/big data fact sheet 3 29 2012.pdf

[2] Cukier K Data, data everywhere: a special report on managing information. Economist Newspaper,2010

[3] Drowning in numbers - digital data will flood the planet-and help us understand it better 2011. http://www.economist.com/blogs/dailychart/2011/11/bigdata-0

[4] Lohr S ,"The age of big data,"New York Times, 2012 pp 11

[5] Yuki N, "Following digital breadcrumbs to big data gold."2011,http://www.npr.org/2011/11/29/142521910/th,edigitalbreadcrumbsthat- lead-to-big-data,

[6] Yuki N ,"The search for analysts to make sense of big data,"2011.http://www.npr.org/2011/11/30/142893065/the-searchforanalyststo- make-sense-of-big-data

[7] Big data , 2008. http:// www.nature.com/news/specials/bigdata/index.html

[8] Special online collection: dealing with big data ,2011. http://www.sciencemag.org/site/special/data/

[9] Mayer-Sch¨onberger V, Cukier K,"Big data: a revolution that will transform how we live, work, and think. "Eamon Dolan/Houghton Mifflin Harcourt.2013

[10] J.K.Laurila, D.Gatica-Perez, I.Aad, J.Blom, O.Bornet ,T.-M.-T.Do, O.Dousse, J.Eberle,M.Miettinen, "The mobile data challenge: Big data for mobile computing research, Workshop on the Nokia Mobile Data Challenge, "in: Proceedings of the Conjunction with the 10th International Conference on Pervasive Computing,2012, pp. 1–8.

[11] D.E. O'Leary, "Artificial intelligence and bigdata,"IEEE Intell.Syst.Vol.28,2013,pp-96–99.

[12] J.J. Berman, "Introduction in:Principles of Big Data," Morgan Kaufmann, Boston,2013,pp.xix–xxvi.

[13] http://www.saama.com/what -is-a-hadoop-explaining-big-data-to-the-csuite.

[14] Sagiroglu, S.; Sinanc, D.,"Big Data: A Review,"2013,pp.20-24

[15] Dong, X.L.; Srivastava, D. Data Engineering (ICDE)," Big data integration," IEEE International Conference on , Vol.29,2013,pp.1245–1248