ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 5, Issue 4, April 2018

# Applications of Big Data Analytics: A Systematic Review

[1] K. Shailaja, [2] B. Seetharamulu, [3] M.A. Jabbar
[1] M. Tech Scholar, [2][3] Professor
[1][3]Centre for Data Science, [2] Department of CSE
[1][2][3] Vardhaman College of Engineering, Hyderabad, Telangana

**Abstract: Big data is a phrase which is used to report collection of data that is vast in size and still growing exponentially with time. In precise, such a data is extremely massive and complex. It covers structured, unstructured and semi-structured data. It provides an interconnection between people and things which is generally referred as a machine to machine interconnectivity. Big data Defines input rate growth and challenges in three dimensions, which is volume, variety, velocity. Big data has so many applications in different fields such as healthcare, banking, agriculture, marketing, telecoms, fraud detection, finance, media, and entertainment, etc. This paper discusses various applications of big data.**

**Keywords: Big Data, Structured, Unstructured, Semi-structured, Machine to Machine Interconnectivity.**

## I. INTRODUCTION

Big data is relentless. It is moderately developed on a massive scale [1]. Big data is a class of information sets so complex that it becomes tough to process using standard methods of data processing. The problems of such data include collection, storage, search, sharing, transfer, visualization, and analysis [2]. An important advantage of analysis of big data is the additional information that can be obtained from one giant set as opposition separate smaller sets. Big data allows correlations to be found, for example, to identify business trends [3].

Big knowledge is often represented by 5v's they are: Volume- The amount of information that is generated is extremely essential. Variety- It is the class huge knowledge belongs to a terribly essential proven fact that has got to be known for knowledge analytics. Velocity: It specifies the speed of information generation or how briskly the information is made and handled. Variability: It specifies the inconsistency which might be represented by the information occasionally. Veracity- The number of the information being captured will differ to a good extent and hence it does the accuracy [5]. To method this data, huge knowledge tools are used, that evaluate the information and method it according to the necessity. Five v's of big data diagrammatically shown as follows

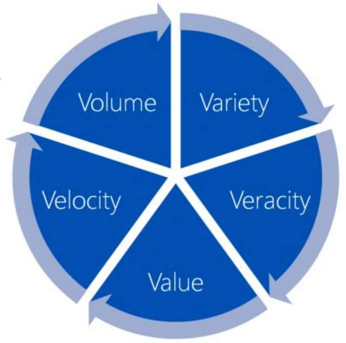

***Figure 1: Big data represented as  Five V's [4]***

Big knowledge analytics is the procedure of examining complex knowledge sets that have a variety of data types that is big data to discover all personal patterns, market trends, client preferences, and alternative effective business data.

The goal of big data analytics is to assist companies to create an additional informative business to investigate massive volumes of transactional information [6]. Big Data Analytics offers a chance to construct unmatched company edges and appropriate service distribution. It additionally wants advanced facilities and a complicated manner of considering the corporate and IT retail market works.

Big data tools which are used for analyzing the huge amount of data [2]. This large amount of information collected and classified into suitable trends and patterns. Thus it must be stored, designed and processed. Widely used big data training tools are discussed in the following section:

***[1] HADOOP:*** HADOOP is not the single software it is a platform which consists of multiple services and it interprets with each other. HADOOP uses a distributed user-level file system, to store resources across the cluster. This file system is called as HADOOP Distributed File System (HDFS), it is written in Java and it is running on the map-reduce model. The following diagram explained in brief about the architecture of a HADOOP ecosystem [7].

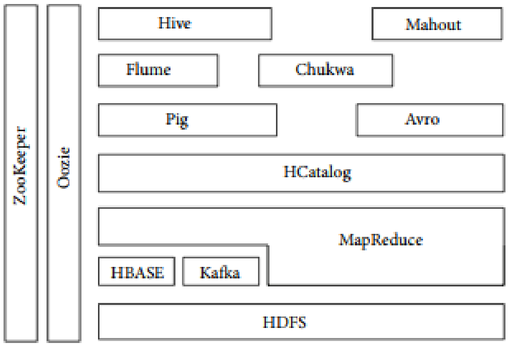

***Figure2: Hadoop 2.0 core components***

Hadoop is the collection of Hbase, pig, hive, Hcatalog, Oozie, zookeeper, and Kafka. It does a solution and give two things that are if we can load any kind of data and process, it allows us to write processing algorithms. Hence it collects data from multiple sources, stores, and process the data into Hadoop, and also create the data, reports the data into Hadoop.Big data Hadoop is an analytics software used for deriving hidden part of the data. HADOOP Distributed File System (HDFS) is a lot of advanced then alternative file systems given the complexities and uncertainties of networks. HADOOP cluster contains 2 nodes. The First node is name node and a second node is information node, name node operated as the master node and information node operated as a slave node.

***HBase:*** Hadoop base defines series of commands and that we will inject the info from the explicit unit into HDFS layer. It is accessible through application programming interfaces like java, Thrift and figural state transfer.

***Zookeeper:*** It manages, configures and names the huge amount of information, provides group synchronization. It is a distributed service that has master and slave nodes and stores contour data.

***Oozie:*** Oozie means workflow coordinational system. It merges actions and organizes Hadoop tasks by using DAG(directed acyclic graph).

***Pig:*** It is a framework and generates Pig Latin which is high level scripting language and manufacture runtime platform that implements users to execute MapReduce on Hadoop.

***Hive:*** Hive could be a sub-platform within the Hadoop scheme and creates source language as its own as Hive query language(HiveQL), HQL is dead by MapReduce. It is supported tables, partitions, and buckets.

***Mahout:*** Its intention is to supply free functions of scattered and ascendable machine learning algorithms and it belongs to the set which may be compiled by MapReduce.

***Hcatalog***: It is managed by Hadoop Distributed File System (HDFS) it stores and produce tables for the huge amount of information. It depends on Hive and other services. It clarifies user connection using HADOOP Distributed File System (HDFS).

***Flume:*** It consists of services under Hadoop as a platform it utilizes two channels they are sources which include system logs, files and sinks include Hadoop Base, HDFS.

***MapReduce:*** It is the framework which processes the large amount of data which is structured and unstructured stored in HDFS. It is useful for batch processing of data stored in Hadoop.

***[2] Apache drill& Dremel:*** Apache drill and Dremel makes it accessible to execute giant scale, ad-hoc queries with low latency's. It scans petabytes of data in terms of few seconds. The business likes apache drill and the Dremel tool [8]. In distinction to workflow-based analysis, business-driven business applications, analytic queries square measure are interactive, ad-hoc, and low latency analysis. It executes ad-hoc queries with low latencies, and it defines that apache drill and Dremel is best than apache HADOOP [9].

***[3] Hive:*** It is a distributed data management for HADOOP. Hive supports Structured Query Language

(SQL) like HIVESQL to access complex data. Hive runs on top of the HADOOP [10].

***[4] Cloud era Impala:*** Cloud era is an open source. It brings protractible and parallel information to apache HADOOP is giving experience to users with low latency for SQL queries hold in HDFS without exploitation of knowledge transformation. By using this tool, real-time analytics that is performed by knowledge scientists and analysts via SQL and knowledge keep in HADOOP [11].

***[5] Mongo DB:*** It is also a useful tool to store and analyze huge knowledge, additionally as facilitate to create applications. Mongo DB contains a distributed information at its core, and it is designed as high availability, geographic distribution hence it's straightforward to use. It was absolutely designed to support large databases with its name Mongo DB and it's written in C++ with document-oriented storage, replication and it has no SQL information [12].

***[6] Rapid Miner:*** Rapid miner offers progressive analytics through template primarily based frameworks. Rather than a native software package, it needs users to write any code and it's offered as a service. Rapid Miner formally referred to as yet another learning environment [13].

Big knowledge application refers to the distributed applications that are usually huge in scale and typically works with massive volumes of knowledge sets. However, it's troublesome for the traditional processing applications to handle such a huge and big datasets. However, if the analytics going to be done in exhausted real time, the numerous quantities of benefits are going to be achieved. Massive information has such a lot applications in numerous fields like banking, marketing, healthcare, education, e-commerce, agriculture, natural disasters, transportation [14] etc.

## [2] LITERATURE SURVEY:

Big Data Analytics Applications are a newly discovered software package applications that leverage large-scale information, that is often too complicated to suit in memory, and big data analytics is used to uncover useful knowledge. The massive information will come from sources like runtime information concentrating on traffic, IOT, twitter, Facebook pages [15].

***[A] Banking:*** The big knowledge, either non-heritable from some supply or internally generated knowledge is to be employed in the way that's incorrect with the structure vision and mission. The banks ought to be ready to use this knowledge therefore on meet the predetermined objectives which may be either to cut cost, minimize the time taken within the process, to launch a new product to call many of these, all factors and variable ought to ultimately cause the higher call making within the organization.

By using data science and technology, the banks have some benefits that are, it finds out the reason for causing issue and failure. Identifies the most valuable customer, and it also prevents the fraudulent behavior [16].

***[B] Healthcare:*** The healthcare business has generated the massive quantity of knowledge obtained from record keeping, compliance, and patient-related information. In today's digital world, it's necessary that this information ought to be digitized. To boost the standard of healthcare considers the following:

- Providing patient central services: To produce quicker relief to the patients by providing proof primary based drugs detecting condition at the sooner stages based on the clinical knowledge obtainable [17]. This helps in decreasing admission rates thereby reducing the worth for the patients.
- Detecting spreading diseases earlier:

(1) Predict the viral diseases.
(2) This may be known by analyzing solid networks [18].
(3) This helps healthcare professionals to take necessary preventive steps.

***[C] Agriculture:*** A biotechnology company uses device information to enhance crop potency. It plants check crops and breaks simulations to live however trees react to numerous modifications in any situation. Its information setting constantly fixes to modify the attributes of varied information it collects, together with temperature, water levels, soil composition, growth, output, and factor sequencing of every plant within the workplace. This simulation permits to find out the optimal environmental conditions for specific factor types.

***[D] In Marketing:*** A recent study revealed that the Harvard Business Review checked out what kinds of advertisements compelled viewers to continue to look and what turned viewers off. A system analyses facial expressions what a viewer is feeling. The analysis was designed to find out what sorts of promotions evoked attaches to share the ads with their social network, serving to marketers produce ads possibly to “go viral” and improve sales. The other applications of big data discussed in the following table [19].

***Table: Applications Of Big Data in various fields.***

| **DOMAIN** | **AREA** |
|---|---|
| 1]consumer goods | ▪ Purchases<br>▪ Weblogs<br>▪ Online retailers |
| 2] In Telecom | ▪ Subscribers rely on over the top service providers focused on revenue reducing opex. |
| 3] In Finance | ▪ Savings.<br>▪ Credit cards.<br>▪ Mortgage. |
| 4]In Credit Cards | ▪ Fraudulent transactions.<br>▪ Suspicious activity. |
| 5] Fraud Detection | ▪ Identify large-scale patterns.<br>▪ Detects anomalous behavior. |

## [3] CONCLUSION:

In this paper, we explored the role of big data and its applications in various fields. Analysis of open source big data tools is also discussed. Big data used in various fields like banking, healthcare, agriculture, smartphones, marketing, telecoms, finance etc. In future, Big Data analytics will play an important role.

## REFERENCES:

[1] https://www.guru99.com/what-is-big-data.html

[2] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, “Interactions with big data analytics,” interactions, vol. 19, no3, pp. 50–59,May 2012.

[3]http://www.crraoaimscs.org

[4]https://www.omnivex.com/company/blog/what-is-big-data

[5]http://en.wikipedia.org/wiki/Big-data Definition

[6]http://searchbusinessanalytics.techtarget.com/big-data-analytics

[7]A.Hadoop,“Hadoop,”,http://hadoop.apache.org/2009.

[8]http://www.mapr.com/support/community-resources/drill

[9]www.dremel.com/

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

[11]http://www.cloudera.com/content/cloudera/en/home.html

[12]https://www.mongodb.com

[13] https:/en.wikipedia.org/wiki/RapidMiner

[14] F.C.P Muhtaroglu, Demir S, Obali M, and Girgin C. “Business model canvas perspective on big data applications”. In IEEE International Conference on Big Data, pages 32–37, Silicon Valley, CA, Oct. 2013.

[15] N. Wingfield, “Virtual product, real profits: Players spend on zynga’s games, but quality turns some off,” Wall Street Journal.

[16] Banking on big Data analytics. Availablefrom http://www.livemint.com/Industry//Banking on Big Data analytics.html.

[17] Yanglin Ren,“Monitoring patients via a secure and mobile healthcare system”, IEEE Symposium on wireless communication,2011

[18] Feldman B, Martin EM, Skotnes T: “Big Data in Healthcare Hype and Hope.” October 2012. Dr. Bonnie 360; 2012. http://www.west-info.eu/files/big-data-inhealthcare.

[19] Ari Banerjee senior analyst, heavy reading, “Big data and advanced analytics in Telecom: A Multi-Billion-Dollar Revenue Opportunity,” December 2013.