Aayushi International Interdisciplinary Research Journal (AIIRJ)
UGC Approved Sr.No.64259
Vol - V Issue-III MARCH 2018 ISSN 2349-638x Impact Factor 4.574

Big Data Classification and Clustering Using Hadoop Environment

Rajesh S. Walse
Ph.D.Scholer , School of Computational Sciences, S.R.T.M. University, (Research Centre)Nanded, Assistant Professor (Computer Science),College of Dairy Technology Warud (Pusad), Maharashtra Animal and Fishery Sciences University

Dr. G.D. Kurundkar
Assistant Professor, Department of Computer Science (guide), Shri Gurubuddhi Swami Mahavidyalaya, Purna district Parbhani , S.R. T. M. University, Nanded Maharashtra State, India

Abstract:

Today, organizations in every industry are being showered with imposing quantity of new information. As there are many more data channels and categories available along with traditional sources. Therefore the rate of data growth is increasing more and more which results in a very large volume of data. These vastly larger volumes and new assets are known as Big Data. Technologies such as MapReduce & Hadoop are used to extract value from Big Data. Hadoop is well adopted, standard-based, open source software framework build on the foundation of Google's MapReduce. There are also new data storage techniques that have arisen to bolster these new architectures, including very large file system running on commodity hardware. This new data storage technology is HDFS. This file system is meant to support enormous amount of structured as well as unstructured data.

***Keywords**: Big Data, Hadoop, HDFS, Map Reduce Hadoop Clustering, Classification of data, Data Mining.*

Introduction:

This document is a template. Unstructured data mining and learning methods are different than those used for structured data. The structured methods to retrieve values of fields and information do not work for unstructured data. During our research just take an example of classification of human behaviours at big processions from security perspective. There are millions of people visiting for religious purpose. Hence, safety and security are very important. Behavioural analysis based on sampled data may not serve the purpose, since a single anomalous behaviour

Without scanned leads to security hazard. Hence, it becomes important to collect all data of behaviours from various inputs. Those may include videos captured across the place, human interactions, social media exchanges, phone calls, and many other sources. This builds huge heterogeneous data. Then we need methods those can classification and clustering as well as associate this data to understand security hazards.

Big Data:

The definition of Big Data contains three different terms. We can say it Power of 3V's of Big Data which are defined as -

*1) **Volume of Data:***

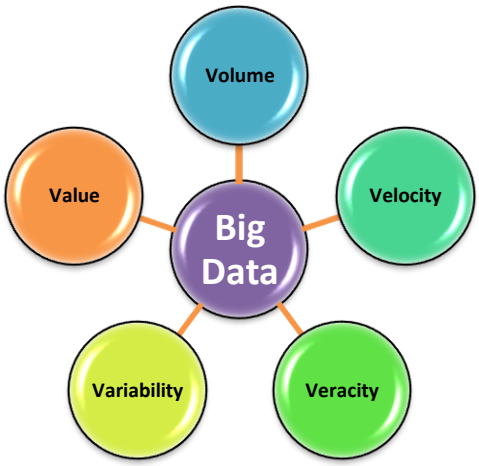
Numerous independent market and research studies have found that data volumes are doubling every year. On top of all this extra new information, a significant percentage of organizations are also storing three or more years of historic data.

*2) **Variety of Data:***

Studies also indicate that 80 percent of data is unstructured (such as images, audio, tweets, text messages, and so on). And until recently, the majority of enterprises have been unable to take full advantage of all this unstructured information.

*3) **Velocity of Data:***

Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.



Human Collaboration:

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

Hadoop: Solution for Big Data Processing

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in following points

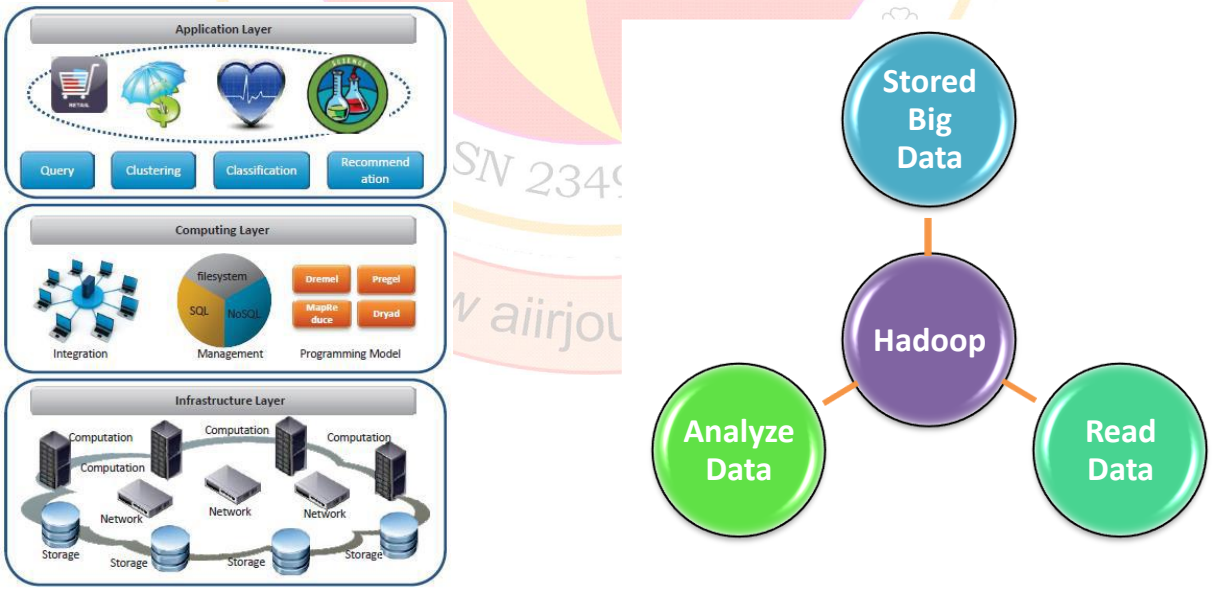


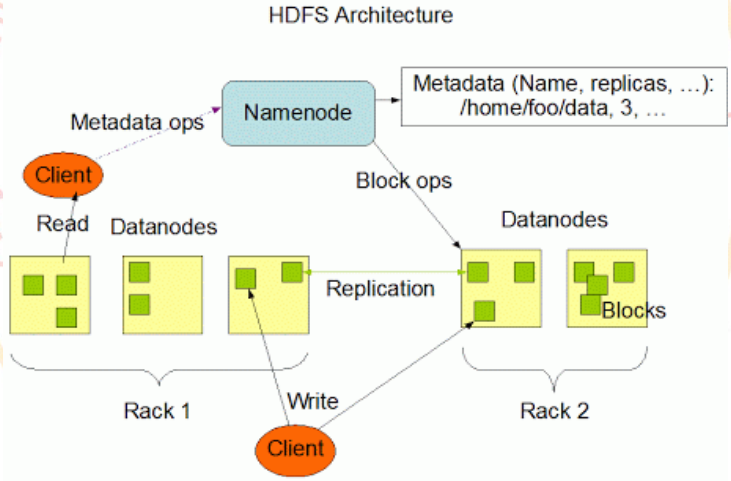
Figure : Layered Architecture of Big Data System

Hadoop—Computational & Storage Solution:

To address the above mentioned issues, the Hadoop framework is designed to provide a reliable, shared storage and analysis infrastructure to the user community. The storage portion of the Hadoop framework is provided by a distributed file system solution such as HDFS, while the analysis functionality is presented by MapReduce. Several other components are part of the overall Hadoop solution suite. The MapReduce functionality is designed as a tool for deep data analysis and the transformation of very large data sets. Hadoop enables the users to explore/analyze complex data sets by utilizing customized analysis scripts/commands. In other words, via the customized MapReduce routines, unstructured data sets can be distributed, analyzed, and explored across thousands of shared-nothing processing systems/clusters/nodes. Hadoop's HDFS replicates the data onto multiple nodes to safeguard the environment from any potential data-loss.

A: HDFS Architecture (Hadoop Distributed File System):

An HDFS cluster encompasses two types of nodes (Name and DataNodes) that operate in a master slave relationship. In the HDFS design, the NameNode reflects the master, system namespace, maintains the file system tree as well as metadata for all the files and directories in the tree. All this information is persistently stored on a local disk via two files that are labelled the namespace image and the edit log, respectively. The NameNode keeps track of all the DataNodes where the blocks for a given file are located. That information is dynamic (and hence is not persistently stored), as it is reconstructed every time the system starts up. Any client can access the file system on behalf of a user task by communicating with the NameNode and the DataNodes, respectively. The DataNodes store and retrieve blocks based on requests made by the by clients or the NameNode, and they do periodically update the NameNode with lists of the actual blocks.



B. Map Reduce:

MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers. The MapReduce framework works in two main phases to process the data which are the Map phase and the Reduce phase. [20]

1) Map Reduce Design:

It takes data set as input which is divided into splits (Split 1, Split 2…. Split N).

- Map: Mapping is done on those splits. After mapping some sorting and shuffling algorithms are applied to splits.
- Reduce: Reduce phase is used to reduce the splits and store them into the centroids files on distributed cache.

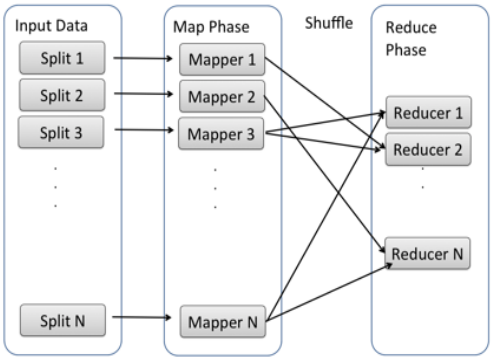
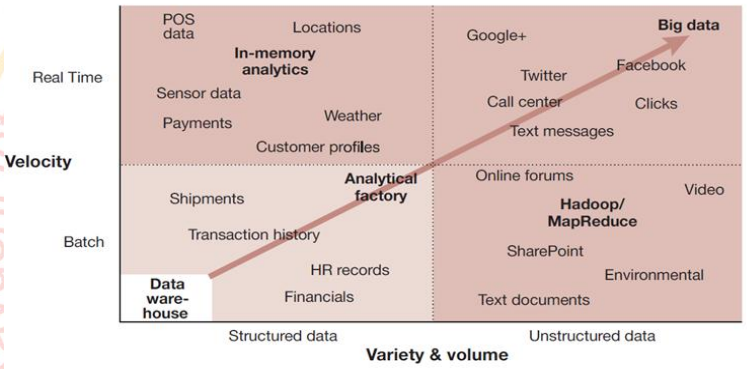


Fig. : Map Reduce Function Block Diagram

Hadoop Cluster:

A hadoop cluster is a special type of a computational cluster designed specially for sorting and analyzing huge amount of unstructured data in a distributed environment. Such cluster run Hadoop's open source distributed processing software on low-cost commodity computers.



[10]

Data Processing Engine for Hadoop

	Current stable release (as of June 1, 2015)	Execution model	Supported languages	Associated ML tools	In-memory processing	Low latency	Fault tolerance	Enterprise support
MapReduce	2.7.0	Batch	Java	Mahout	x	x	✓	x
Spark	1.3.1	Batch, streaming	Java, Python, R, Scala	MLlib, Mahout, H_2O	✓	✓	✓	✓
Flink	0.8.1	Batch, streaming	Java, Scala	Flink-ML, SAMOA	✓	✓	✓	x
Storm	0.9.4	Streaming	Any	SAMOA	✓	✓	✓	x
H_2O	3.0.0.12	Batch	Java, Python, R, Scala	H_2O, Mahout, MLlib	✓	✓	✓	✓

A. **Purpose of Clustering:**

1) Hadoop clusters are known for boosting the speed of data analysis applications.
2) They are used to increase the throughput.

3) Hadoop clusters are highly resistant to failure because each piece of data is copied onto other cluster node which ensures that the data is not lost if one node fails.

The following are the high-level steps involved in configuring Linux cluster on Redhat or CentOS:

- Install and start RICCI cluster service.
- Create cluster on active node.
- Add a node to cluster.
- Add fencing to cluster.
- Configure failover domain.
- Add resources to cluster. Sync cluster configuration across nodes.
- Start the cluster.

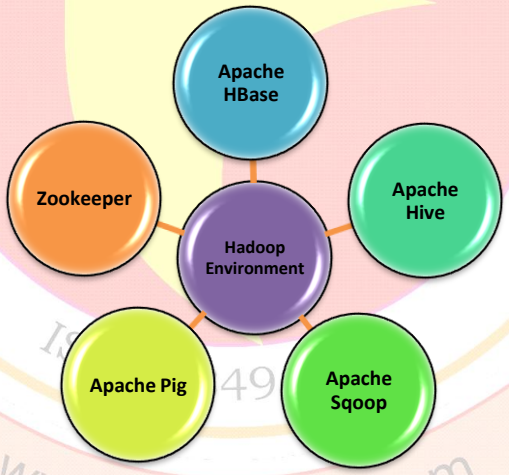
Today everywhere different types of data are getting generated, Data never sleeps, every minutes data is generated, it required to be process or analyzing data for eg reasons for epidemic diseases, cancer cell identification, analyzing super market data, what kind of items sold ? Analyzing weather data – precautionary measures can be taken against floods, Tsunami etc.

Multi Computer Systems:

Data intensive computing – Type of parallel computing tasks are executed by transferring them to the systems whether the data is available and executing these tasks in parallel.

Advantages of Distributed System (DS)

- Scalability
- Reliability
- Availability
- Communication



Literature Review

Harshawardhan S. Bhosale1, Prof. Devendra Gadekar, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, (10-15 October, 2014), a review on Big Data and Hadoop the paper describes the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data.

Shilpa, Manjeet Kaur, LPU, Phagwara, India, a review on Big Data and Methodology (5-10 October, 2013) illustrated that there are various challenges and issues regarding big data. There must support and encourage fundamental research towards these technical issues if we want to achieve the benefits of big data. Big-data analysis fundamentally transforms operational, financial and commercial problems in aviation that were previously unsolvable within economic and human capital constraints using discrete data sets and on-premises hardware. Centralizing data acquisition and consolidation in the cloud, and by using cloud based virtualization infrastructure to mine data sets efficiently, big-data methods offer new insight into existing data sets.

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk big data analytics with Apache Hadoop" Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns.

S. Vikram Phaneendra & E. Madhusudhan Reddy *et.al*. Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as "big data". In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc .

Kiran kumara Reddi & Dnvsl Indira *et.al*. Enhanced us with the knowledge that Big Data is combination of structured , semi-structured ,unstructured homogenous and heterogeneous data .The author suggested to use nice model to handle transfer of huge amount of data over the network .Under this model, these transfers are relegated to low demand periods where there is ample ,idle bandwidth available . This bandwidth can then be repurposed for big data transmission without impacting other users in system. The Nice model uses a store –and-forward approach by utilizing staging servers. The model is able to accommodate differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms.

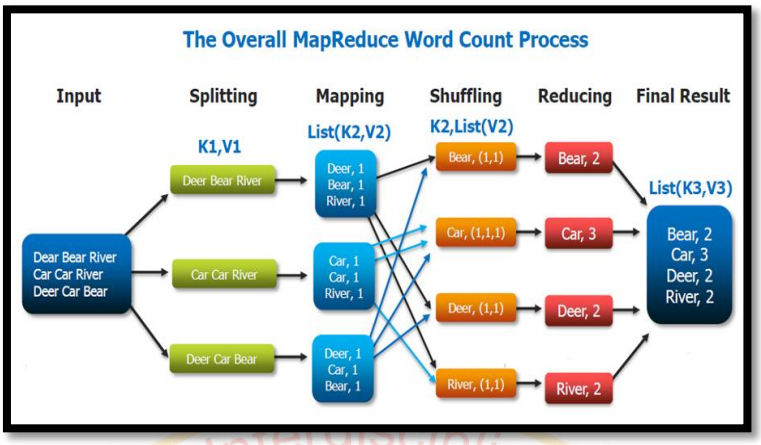
Conclusion:

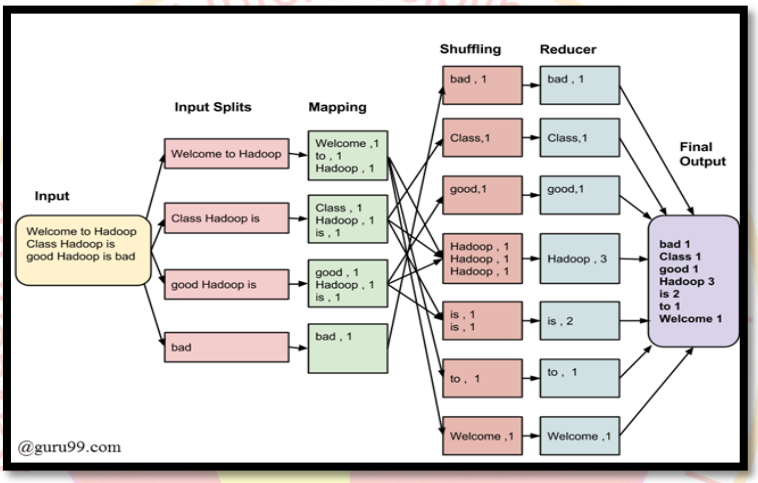
Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. The Hadoop architecture is used to process unstructured and semi-structured using map reduce to locate all relevant data then select only the data directly answering the query. There have been extensive studies on various clustering methods and especially the k-means clustering has been given a great attention. The advent of Big Data has posed opportunities as well challenges to business.

Big Data has become a recent trend in technology, whether it is network, whether data mining or even data management we begin to talk in terms of Big Data. Big data is typically huge size data dealing with different aspects – it may be data generated through social networking sites, or about a big event, worldwide interactions. Since it is huge, coming from all sources and dealing with larger landscape, it is assorted mix where major portion is unstructured and semi –structured. Machine learning for big data is different than

traditional machine learning. Having more data is disposal is a challenge – can it be harmful for delivering results? Some of the researcher say that they are not interested in too much of Big Data.

Word Count MapReduce Paradigm





References:

1. Bakshi, K.,(2012)," Considerations for big data: Architecture and approach"
2. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"
3. Harshawardhan S. Bhosale1, Prof. Devendra Gadekar, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, a review on Big Data Aditya B. Patel, Manashvi Birla, Ushma Nair,(6-8 Dec.
4. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"
5. Shilpa, Manjeet Kaur, LPU, Phagwara, India, a review on Big Data and Methodology
6. Yu Li; Wenming Qiu; Awada, U. ; Keqiu Li,,(Dec 2012)," Big Data Processing in Cloud Computing Environments"
7. Garlasu, D.; Sandulescu, V; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing
8. Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),"Big Data: A Review"
9. Grosso, P. ; de Laat, C. ; Membrey, P.,(20-24 May 2013)," Addressing big data issues in Scientific Data Infrastructure"
10. Source: http://www.strategyand.pwc.com/media/file/Strategyand_Benefiting-from-Big-Data_A-New-Approach-for-the-Telecom-Industry.pdf