A SURVEY ON BIG DATA PRIVACY USING HADOOP ARCHITECTURE

[1]A. KANIMOZHI, [2]Dr. N. VIMALA

[1]Ph.D(Part Time) Scholar, [2]Assistant Professor
Department of Computer Science, LRG College for Women.

***Abstract*: Big data is the term for any gathering of datasets so vast and complex that it gets to be distinctly troublesome to process using traditional data processing applications. The challenges include analysis, catch, sharing, stockpiling, exchange, perception, and security infringement. It is a set of techniques and technologies that require new forms of integration to uncover huge concealed qualities from substantial datasets that are assorted, complex, and of a huge scale. This environment is used to acquire, organize and analyze the various types of data. For such data-intensive applications, the Apache Hadoop Framework has recently attracted a lot of attention. This framework Adopted MapReduce, it is a programming model and a related execution for preparing and producing large data sets. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. This paper refer privacy and security aspects healthcare in big data and also randomization, theoretical limits associated with privateness-preservation over immoderate dimensional records sets. This hadoop architecture handles data loses and intermediate data capturing by hadoop online prototype. Finally this survey deals with parallel processing with massive data sets and capturing, managing within a time perod.[1]**

Keywords: Big Data, Hadoop, HDFS, MapReduce, Hadoop Components, Hive, NoSQL, Hpcc

1. Introduction

Big data is a biggest popular expressions in space of IT, new advances of individual correspondence driving the big data new trend and internet population grew day by day but it never reach by 100%. The need of Big Data created from the extensive organizations like facebook, hurray, Google, YouTube etc for the purpose of analysis of enormous amount of data which is in unstructured frame or even in organized shape. Google contains the vast measure of data. So there is the need of Big Data Analytics that is the processing of the complex and massive datasets. This information is not quite the same as organized information as far as five parameters – variety, volume, value, veracity and velocity (5V's). The five V's (volume, variety, velocity, value, veracity) are the challenges of big data management [2]

1.1 Characteristics of Big Data

Volume: Information is steadily developing step by step of different types ever MB, PB, YB, ZB, KB, TB of data. The data results into large files. Excessive volume of data is main issue of storage. This main issue is resolved by reducing storage cost. Data volumes are expected to grow 50 times by 2020.

Variety: Information sources are amazingly heterogeneous. The records comes in different configurations and of any sort, it may be structured or unstructured such as text, audio, videos, log files. The assortments are interminable, and the information enters the system without having been measured.

Velocity: The information comes at fast. Now moment is past the point of no return so big data is time delicate.. Some organizations data velocity is main challenge. The social media messages and credit card transactions done in millisecond and data generated by this putting in to databases.

Value: It is a most important v in big data. Value is main buzz for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

5.Veracity: The expansion in the scope of qualities run of the mill of an extensive information set. When we managing high volume, velocity and variety of data, the all of data are not going 100% correct, there will be messy information. Big data and examination innovations work with these sorts of information. Immense volume of data (both structured and unstructured) is management by organization and administration. Unstructured information is an information that is not present in a database. Unstructured data may be text, verbal data or in another form. Textual unstructured data is like power point presentation, email messages, word reports, and moment kneads. Information in another arrangement can be .jpg images, .png images and audio files. The parameters five v's of big data describes in fig 1.

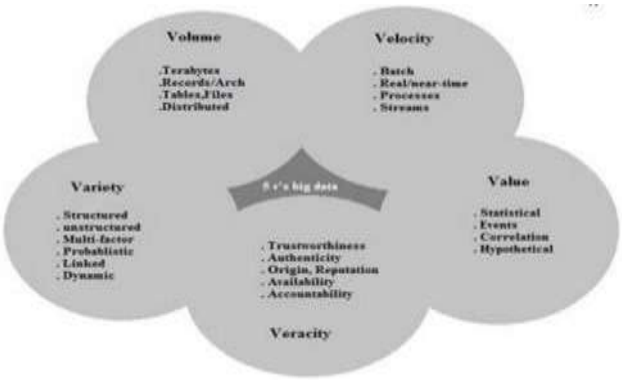

Fig1: Parameters of Big Data

For example, lethal and non-deadly falls, Parkinson's disease, cardio-vascular disorders, stress, etc. We have discussed different human services methods accessible to address those illnesses and numerous other perpetual impediment, like blindness, motor disabilities, paralysis, etc. Moreover, a plethora of commercially available unavoidable social insurance items. [Sagiroglu, 2013][3].

2. Technologies and Methods

All paragraphs must be indented. Big data is a new concept for handling massive data therefore the architectural description of this technology is very new. There are the different technologies which utilize practically same approach i.e. to convey the information among different nearby specialists and diminish the load of the main server so that traffic can be avoided. There are endless articles, books also, periodicals that portray Big Data from an innovation point of view so we will rather center our efforts here on setting out some basic principles and the minimum technology foundation to help relate Big Data to the broader IM domain [4].

A. Hadoop

Hadoop is a structure that can run applications on frameworks with a large number of hubs and terabytes. Hadoop architecture shown in Fig 2. It distributes the file among the nodes and allows to system continue work in case of a node failure. This approach reduces the risk of catastrophic system failure.

In which application is broken into littler parts (sections or blocks).Apache Hadoop comprises of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System consists of three Components: the Name Node, Secondary Name Node and Data Node. The multilevel secure (MLS) issues of Hadoop by utilizing security improved Linux (SE Linux) convention. In which various sources of Hadoop applications run at different levels.

This protocol is an extension of Hadoop distributed file system. Hadoop is commonly used for distributed batch index building; it is desirable to optimize the index capability in ongoing. Hadoop gives segments to capacity and investigation for vast scale handling. Now a day's Hadoop used by hundreds of companies.

The upside of Hadoop is Distributed stockpiling and Computational abilities, to a great degree versatile, Optimized for high throughput, large block sizes, tolerant of software and hardware failure.[5]

Hadoop is a much more vulnerable target – too open to be able to fully protect. Further exacerbating the risk is that the aggregation of data in Hadoop makes it an even more Existing IT security including network firewalls, logging and monitoring, and configuration management Enterprise-scale security for Apache Hadoop Apache Knox used for perimeter security Kerberos used for strong authentication

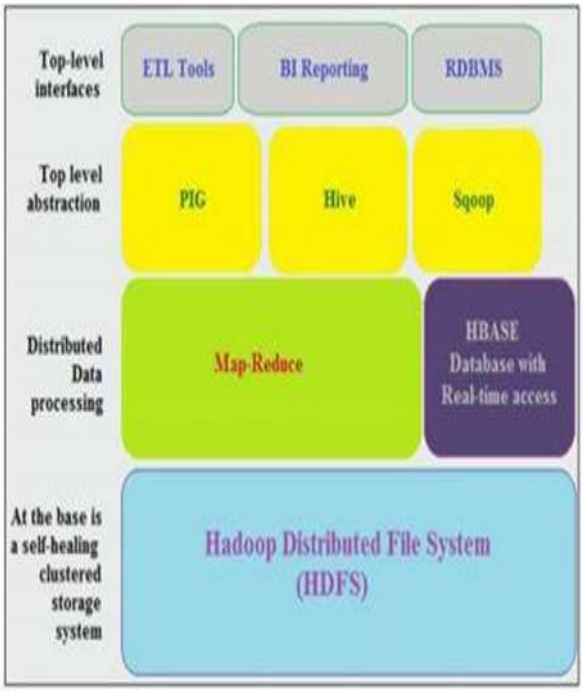

Table 1: The Ecosystem of Hadoop

Components of Hadoop [11]:

HBase: It is open source, circulated and Non-social database framework executed in Java. It runs above the layer of HDFS. It can serve the input and output for the Map Reduce in well-mannered structure.

Oozie: Oozie is a web-application that runs in a java servlet. Oozie use the database to gather the information of Workflow which is a collection of actions. It manages the Hadoop jobs in a mannered way.

Sqoop: Sqoop is an order line interface application that gives stage which is accustomed to changing over data from relational databases and Hadoop or vice versabreak down the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.

Pig: Pig is high-level platform where the MapReduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high as relational model.

Hive: It is application developed for data ware house that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing conclusion, and analysis for respective queries.

Reliable: The software is fault tolerant, it expects and handles hardware and software failures

Scalable : Designed for massive scale of processors, memory, and local attached storage Distributed:[6]

Handles replication. Offers massively parallel programming model, Map Reduce. Hadoop system describes the unstructured data needs to be turned into structured data. Queries can't be reasonably expressed using SQL Heavily recursive algorithms. Complex but parallelizable algorithms needed, such as geo-spatial analysis or genome sequencing in fig 3.

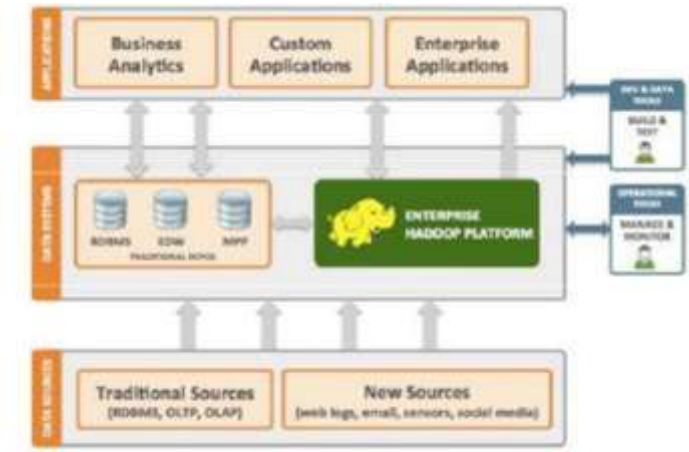

Fig 3: Hadoop System

HDFS Architecture

Hadoop incorporates a fault-tolerant stockpiling framework called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the disappointment of noteworthy parts of the capacity framework without losing information. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive PCs. In the event that one fizzles, Hadoop keeps on working the group without losing information or hindering work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking approaching records into pieces, called "squares," and putting away each of the squares redundantly across the pool of servers. In the common case, Fig 4. HDFS stores three complete copies of each file by copying each piece to three different servers [7].

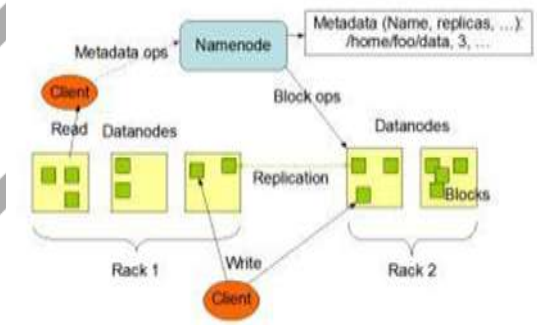

Fig4: HDFS Architecture

B. MapReduce

The preparing column in the Hadoop biological system is the MapReduce structure. The system permits the specification of an operation to be applied to a huge data set, divide the problem and information, and run it in parallel. From an expert's perspective, this can happen on various measurements. For example, a very large dataset can be reduced into a smaller subset where analytics can be connected. In a conventional information warehousing situation, this may involve applying an ETL operation on the data to produce something usable by the analyst. In Hadoop, these kinds of operations are composed as MapReduce employments in Java. There are various larger amount dialects like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or put in a customary information distribution center. There are two capacities in MapReduce as follows [8]:

map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs.

reduce – the function which merges all the intermediate values associated with the same intermediate key Outline plays out the errand as the ace hub takes the information, isolate into littler sub modules and distribute into slave nodes. A slave node further divides the subs again that prompt to the various leveled tree structure. The slave hub forms the base issue and passes the result back to the master Node. The Map Reduce system arrange together all sets in light of the middle of the road keys and allude them to diminish() work for creating the final output. Reduce function works as the master node collects the results from all the sub problems and combines them together to form the output.

Map (in_key, in_value) ---
>list (out_key, intermediate_value) **Reduce** (out_key, list (intermediate_value))---
>list (out_value)
The parameters of map () and reduce () function is as follows:
map (k1, v1)! list (k2,v2) and reduce (k2,list(v2)) ! list (v2)

A Map Reduce framework is based on a master-slave architecture where one master node handles a number of slave nodes . Map Reduce works by first dividing the input data set into even-sized data blocks for equal load distribution. Each data block is then assigned to one slave node and is processed by a map task and result is generated. The slave node interrupts the master hub when it is sit without moving. The scheduler then doles out new assignments to the slave hub. The scheduler takes data locality and resources into consideration when it disseminates data blocks [9,10].

Reduction to MapReduce

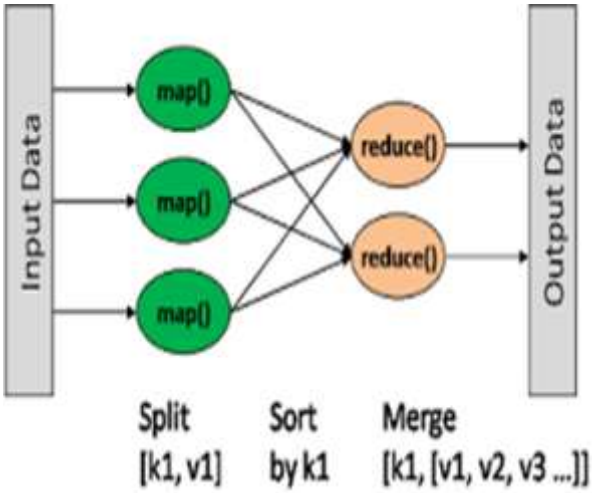

Fig 5 Map reduce framework

Mapreduce is mainly a data processing component of Hadoop. It is a programming model for processing large number of data sets. It contains the task of data processing and distributes the particular tasks across the nodes. It consists of two phases –

Map

Reduce

Map converts a typical dataset into another set of data where individual elements are divided into key/value pairs.

Reduce task takes the output files from a map considering as an input and then integrate the data tuples into a smaller set of tuples. Always it is been executed after the map job is done.

Features of Mapreduce system

Features of Mapreduce are as follows:

1. Framework is provided for Mapreduce execution Abstracts developer from the complexity of distributed programming languages.
2. Partial failure of the processing cluster is expected and tolerable to fulfill the requirements.
3. In-built Redundancy and fault tolerance is available.
4. Mapreduce programming model system is language independent.
5. Automatic parallelization and distribution are in charge.
6. Fault tolerance
7. Enable data local processing
8. Shared nothing than architectural model

Manages all the inter process communication [11]

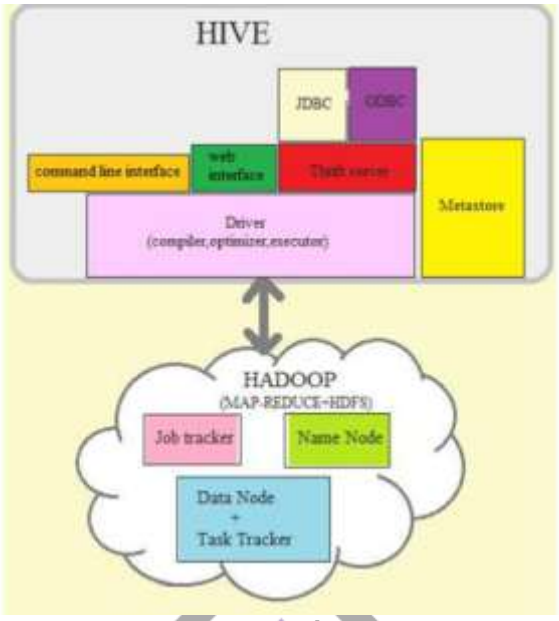

Fig 7: Hive Architecture

D. No-SQL:

No-SQL database is a way to deal with information administration and information configuration that is valuable for extensive sets of distributed data. These databases are in general part of the real-time events that are distinguished in process sent to inbound channels yet can likewise be viewed as an empowering innovation following analytical capabilities such as relative search applications. These are only made attainable in view of the flexible way of the No-SQL show where the dimensionality of an inquiry is evolved from the data in scope and domain rather than being fixed by the developer in advance It is helpful when endeavor need to get to gigantic measure of unstructured information. There are more than one hundred No SQL approaches that specialize in management of different multimodal data sorts (from organized to non-organized) and with the plan to comprehend particular difficulties. Data Scientist, Researchers and Business Analysts in specific pay more attention to agile approach that prompts to earlier bits of knowledge into the information sets that might be covered or compelled with a more formal development process. The most popular No-SQL database is Apache Cassandra. The favorable position of No-SQL is open source, Horizontal adaptability, Easy to utilize, store complex data types, Very fast for adding new data and for simple operations/queries. The disadvantage of No-SQL is Immaturity, No ordering support, No ACID, Complex consistency models, Absence of standardization [12,13].

E.HPCC:

HPCC is an open source stage utilized for registering and that gives the administration to taking care of massive bigdata workflow. HPCC data model is defined by the user end according to the requirements. HPCC framework is proposed and afterward additionally intended to deal with the most mind boggling and information escalated analytical related problems. HPCC system is a single platform having a single architecture also, a solitary programming dialect utilized for the information simulation. HPCC framework was intended to analyze the gigantic amount of data for the purpose of solving complex problem of big data. HPCC framework depends on big business control dialect which has the decisive and on-procedural nature programming language.

The main components of HPCC are:

HPCC Data Refinery: Use parallel ETL engine mostly.
HPCC Data Delivery: It is massively based on structured query engine used.

Enterprise Control: Language distributes the workload between the nodes in appropriate even load.[14]

Literature Review

Charu C. Aggarwal et.al in [15] provided a top level view of the modern techniques for privacy. They talked techniques for randomization, good enough-anonymization, and dispensed privateness-keeping facts mining. They also discussed instances in which the output of records mining packages wants to be sanitized for privateness-maintenance functions. They talked the computational and theoretical limits associated with privateness-preservation over immoderate dimensional records sets.

Vidyasagar S. D[`16] did a survey on Big Data and Hadoopsystem and found that organizations need to process and handle petabytes of Data sets in efficient and inexpensive manner. According to him if there is any node failure then we can lose some information. Hadoop is an Efficient, reliable, Open Source Apache License. Hadoop is used to deal with large data sets. Apache Cassandra and Voldemort and discusses the application's requirements for consistency, availability, partition tolerance, data model and scalability. They explore the enhancements made to Hadoop to make it a more effective real-time system, the tradeoffs they made while configuring

the system, and how this solution has significant advantages over the shared MySQL database scheme used in other applications at Facebook and many other web-scale companies per day.

Tyson Condie[17] et.al. Propose a modified MapReduce architecture in which intermediate data is pipelined between operators, while preserving the programming interfaces and fault tolerance models of other MapReduce frameworks. To validate this design, author developed the Hadoop Online Prototype (HOP), a pipelining version of Hadoop.

Mrigank Mridul,Akash deep khajuria, [18] discussed the analysis of big data and they stated that Data is generated through many sources like business processes, transactions, social networking sites, web servers, etc. and remains in structured as well as unstructured form . Processing or analysing the huge amount of data or extracting meaningful information is a challenging task. The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.
Kyong-Ha Lee Hyunsik choi [19] emphasizes on a prominent data processing tool Map Reduce which will help in understanding various technical aspects of the Map Reduce framework. In this survey, the author expresses different views on Map Reduce framework and introduces its optimization strategies. Author also hands a challenge on parallel data analysis with Map Reduce framework.
Aditya B.Patel, Manashri Birla, [20]have done a lot of experiment on the big data problem. At last he found that the hadoop cluster, Hadoop Distributed File System (HDFS) for storage and map reduce method for parallel processing on a large volume of data.
Mukherjee A Dattu Jorapur, Singhvi Haloi S.Akram[21] defines big data Problem using Hadoop and Map Reduce" reports the experimental research on the Big data problems in various domains. It describe the optimal and efficient solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage data and Map Reduce framework for parallel processing to process massive data sets and records.
D.Rajasekar, C.Dhanamani, SK Sandhya[22] an overview of big data's concept, tools, techniques, applications, advantages and challenges. They used Hadoop technology for the implementation purpose. The authors have briefly discussed about HDFS and Map Reduce technology to process massive data sets and records.
vshma Nair[23] Author explained its need, uses and application. Now days, Hadoop is playing an important role in Big Data. The author concluded that "Hadoop is designed to run on cheap commodity hardware. it automatically handles data replication and node failure, it does the hard work – you can focus on processing data, Cost Saving and efficient and reliable data processing"
Snehasish Dutta, kumar N [24]Pipelining provides several important advantages to a MapReduce framework, but also raises new design challenges. In this demonstration, we describe a modified MapReduce architecture that allows data to be pipelined between operators. This extends the MapReduce programming model beyond batch processing, and can reduce completion times and improve system utilization for batch jobs as well.

4. Conclusion

This paper surveyed various technologies to handle the big data and also, different points of interest and a drawback of these advancements. This paper examined a draftsman using Hadoop HDFS distributed data storage, real-time NoSQL databases, and MapReduce distributed data processing over a cluster of commodity servers. It also covers the survey of various big data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems. 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. This paper concludes data replication, node failure, cost saving and efficient data processing and also time reduce in batch processing within a pipeling architecture. There are different fields of application in the area for Big Data. But what makes data mining more in the spotlight, is the necessities of using data mining techniques in Big data due to its specific properties that makes it more suitable when dealing with current data. In data mining proposed a frame work which is aiming that it will improve the performance of Hadoop MapReduce workloads and at the same time will maintain the decent results in big data. This extends the MapReduce programming model beyond batch processing, and can reduce completion times and improve system utilization for batch jobs and challenge on parallel data analysis with map reduce framework.

References

[1] Sagiroglu, S.Sinanc, D., "**Big Data: A Review"**, International Conference on Collaboration Technoogies and Systems", volume 5, March 13, 2017.
[2] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, **—Survey Paper On Big Data** International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.
[3] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N **" Analysis of Bidgata using Apache Hadoop and Map Reduce"** Volume 4,Issue 5, May 2014"
[4]D. Che, M. Safran, and Z. Peng, "**From Big Data to Big Data Mining: challenges, issues, and opportunities," in *Database Systems for Advanced Applications*,** pp. 1–15, Springer, Berlin, Germany, 2013.
[5]Bijesh Dhyani, Anurag Barthwal, "**Big Data Analytics using Hadoop",** International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014.
[6] Vasiliki Kalavri & Vladimir Vlassov(2013) "**MapReduce: Limitations, Optimizations and Open Issues**", 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications.

[7]Ms. Gurpreet Kaur, Ms. Manpreet Kaur, "**REVIEW PAPER ON BIG DATA USING HADOOP"**, International Journal of Computer Engineering & Technology (IJCET), Volume 6, Issue 12, Dec 2015, pp. 65-71, Article ID: IJCET_06_12_008 ISSN Print: 0976-6367 and ISSN Online: 0976–6375
[8]Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chandler, "**The Hadoop Distributed File System"**, October 2010, 978-1-4244-7153, IEEE
[9]Dr. Sanjay Srivastava, Swami Singh, Viplav Mandal, "**The Big Data analytics with Hadoop",** International Journal of Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue III, March 2016, ISSN: 2321-9653
[10] Ahmed Eldawy, Mohamed F. Mokbel "**A Demonstration of SpatialHadoop:An Efficient MapReduce Framework for Spatial Data**" *Proceedings of the VLDB Endowment, Vol. 6, No. 12 Copyright 2013 VLDB Endowment 21508097/13/10.*
[11]Poonam S. Patil, Rajesh. N. Phursule, "**Survey Paper on Big Data Processing and Hadoop Components"**, International Journal of Science and Research (IJSR), Volume 3 Issue 10, October 2014, ISSN (Online): 2319-7064
[12]Rahul Beakta, "**Big Data And Hadoop: A Review Paper**",RIEECE -2015, Volume 2, Spl. Issue 2 (2015), e-ISSN: 1694-2329, p-ISSN: 1694-2345
[13]Ivanilton Polatoa, Reginaldo Reb, Alfredo Goldman, Fabio Kona, "**A Comprehensive View of Hadoop Research - A Systematic Literature Review"**, Journal of Network and Computer Applications, Volume 46, PP 1-25, November 2014
[14]Dr. Madhu Goel, Suman Arora, "**Survey Paper on Scheduling in Hadoop"**, International Journal of Advanced Research in Computer Science and Software Engineering [IJARCSSE], Volume 4, Issue 5, May 2014, ISSN: 2277 128X
[15] Aggarwal, C.C. and Philip, S.Y., 2008. **A general survey of privacy-preserving data mining models and algorithms.** In Privacy-preserving data mining (pp. 11-52). Springer US. Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertain Fuzz. 002;10(5):557–70.
[16]Vidyasagar S. D, A Study on "**Role of Hadoop in Information Technology era",** GRA - GLOBAL RESEARCH ANALYSIS, Volume : 2 | Issue : 2 | Feb 2013 • ISSN No 2277 –8160.
[17]Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein "**Online Aggregation and Continuous Query support in MapReduce**" *SIGMOD'10,* June 6–11, 2010, Indianapolis, Indiana, USA. Copyright 2010 ACM 978-1-4503-0032-2/10/06.
[18]Mrigank Mridul, Akash deep khajuria is of **"bigdata using apache hadoop and map reduce** "volume 4, Issue 5, May 2004.
[19]Kyong Ha lee Hyunsik choi "**Parellel data processing with map reduce;** A survey" sigmod record dec 2011.
[20]Aditya B.Patel , Manashri Birla Manashvi Birla, Ushma Nair, "**Addressing bigdata problem using hadoop and map reduce**"Dec 2011.
[21] Mukherjee A dattu Jorapur, singhri, haloi, S.Akram W, "**Shared disk big data analytics with apache hadoop"** 2008.
[22]D.Rajasekar, C.Dhanamani, S.K.Sandhya " **A survey on bigdata concepts and tools"**, Voulume 5 , I ssue 2, Feb 2015.
[23]Ushma Nair, Aditya Betal " **Big data with Hadoop Techniques –Survey"** Dec 2015.
[24]Snebasish Dutta N, **"Anaysis of Map reduce Algorithms and Modeling Approach",** Volume 5, Issue 4, May2016.