A Survey on Big Data privacy using Hadoop Architecture

Priyank Jain[1], Manasi Gyanchandani[2], Nilay Khare[3], Dhirendra Pratap Singh[4], Lokini Rajesh[5]

CSE Department [1,2,3,4,5], MANIT, Bhopal M.P India[1, 2, 3, 4, 5]

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, curation, look, sharing, stockpiling, exchange, perception, and security infringement. Big data 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. Big data environment is used to acquire, organize and analyze the various types of data. Data that is so substantial in volume, so differing in assortment or moving with such speed is called big data. Analyzing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. 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. To begin with, we introduce the meaning of enormous information and discuss big data challenges. Hadoop is the core platform for structuring Big Data, and tackles the issue of making it helpful for examination purposes. Hadoop is an open source programming project that enables the distributed processing of large data sets across clusters of commodity servers. It is intended to scale up from a solitary server to a great many machines, with an extremely high degree of fault tolerance. This paper refer privacy and security aspects healthcare in big data. Next, we present Existing techniques of anonymization using MapReduce framework of big data privacy is also done as well.

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 are [1]:

1.1 Characteristics of Big Data

1. 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.

2. 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 and then some. The assortments are interminable, and the information enters the system without having been measured.

3. Velocity: The information comes at fast. Now and then 1 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.

4. 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, administration 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

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

files [Sagiroglu, 2013]. The parameters five v's of big data describes in fig 1.

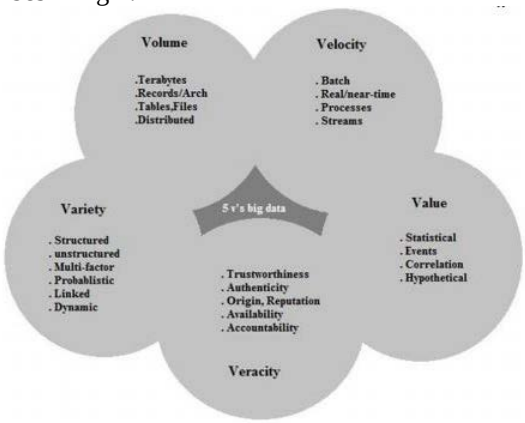

Fig1: Parameters of Big Data

1.2 Privacy and security aspects in big data

The new influx of digitizing therapeutic records has seen an outlook change in the human services industry. As a result, healthcare industry is witnessing an increase in sheer volume of data in terms of complexity, diversity and timeliness [3–5]. The term "big data" refers to the agglomeration of large and complex data sets, which exceeds existing computational, storage and communication capacities of routine strategies or frameworks. In human services, a few variables give the necessary impetus to harness the power of big data [6]. The harnessing the power of big data investigation and genomic look into with continuous access to patient records could permit specialists to make informed decisions on treatments [7]. Big data will compel insurers to reassess their prescient models. The continuous remote observing of indispensable signs through installed sensors (attached to patients) permits human services suppliers to be alarmed if there should arise an occurrence of an irregularity. Social insurance digitization with integrated analytics is one of the next big waves in healthcare Information Technology (IT) with Electronic Health Records (EHRs) being an essential building hinder for this vision. With the introduction of HER incentive programs [8], healthcare organizations recognized EHR's esteem suggestion to encourage better access to finish, precise and sharable medicinal services data that eventually lead to improved patient care. With the ever-changing risk environment and what's more, presentation of new rising dangers and vulnerabilities, security infringement are normal to grow in the coming years [9].

Big Data introduced a far reaching overview of various instruments and methods utilized as a part of Pervasive healthcare in a disease-specific manner. It covered the major diseases and disorders that can be immediately recognized and treated with the utilization of innovation, 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. It gives comprehension of the different parts of unavoidable healthcare with respect to different diseases [10].

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 [11].

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 [11].

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.

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

alluring target for hackers and data thieves. Hadoop presents brand new challenges to data risk management: the potential concentration of vast amounts of sensitive corporate and personal data in a low-trust environment. New methods of data protection at zettabyte scale are essential to prevent these potentially huge big data exposures.

Current security measures may be insufficient to protect sensitive data in Hadoop from new, advanced threats. These measures include the following:

- 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
- Apache Argus monitoring and management

Several traditional data de-identification approaches can be deployed to improve security in the Hadoop environment, such as storage level encryption, traditional field-level encryption, and data masking.

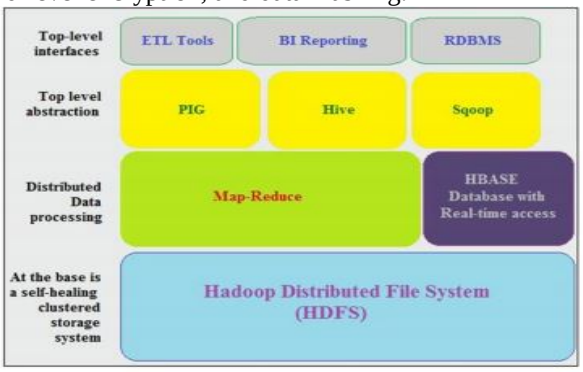

Fig 2: Hadoop Architecture

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 ajava 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 versa.

Avro: It is a framework that gives usefulness of information serialization and administration of information trade. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.

Chukwa: Chukwa is a structure that is utilized for information gathering and investigation to handle and break 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 level language.

Zookeeper: It is a centralization based administration that gives conveyed synchronization and gives group services along with maintenance of the configuration information and records.

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.

Table 1: The Ecosystem of Hadoop

Elements\Ecosystems	Hadoop
Distributed File Systems	HDFS, FTP File system, Amazon-S3, Windows Azure Storage Blobs
Distributed Resource Management	YARN framework
SQL Query	**HIVE**: A data warehouse component
Machine Learning	**Mahou**t: A Machine learning component
Stream Processing	**Storm**: real-time computational engine
Graph Processing	**Giraph:** A framework for large-scale graph processing
Management Interface	**Zookeeper**: A management tool for Hadoop cluster
Stream tool	**Flume**: a service for efficiently transferring streaming data into the Hadoop Distributed File System (HDFS).
Pluggable to RMDB	**Sqoop**: transfer data between Relational Database Management System (RDBMS) and Hadoop
Data Flow Processing	**Pig**: a high level scripting data flow language which expresses data flows by applying a series of transformations to loaded data [12].
NoSQL database	**HBase**: based on Big Table, and column-oriented

Hadoop was made by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed bolster dispersion for the Nutch web index extend. Hadoop is open-source programming that enables reliable, scalable, distributed computing on clusters of inexpensive servers [11].

Hadoop is:

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:

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.

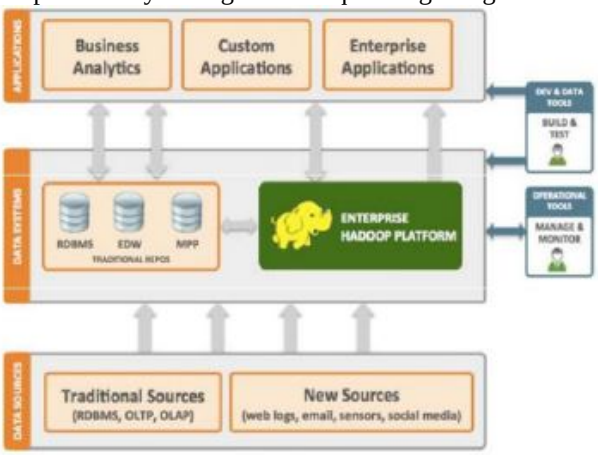

Fig 3: Hadoop System

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 [13].

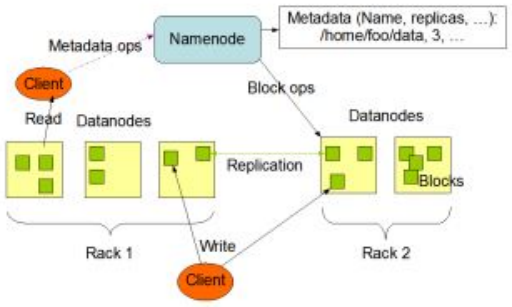

Fig4: HDFS Architecture

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 [14]:
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 sub modules 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 [14,13].

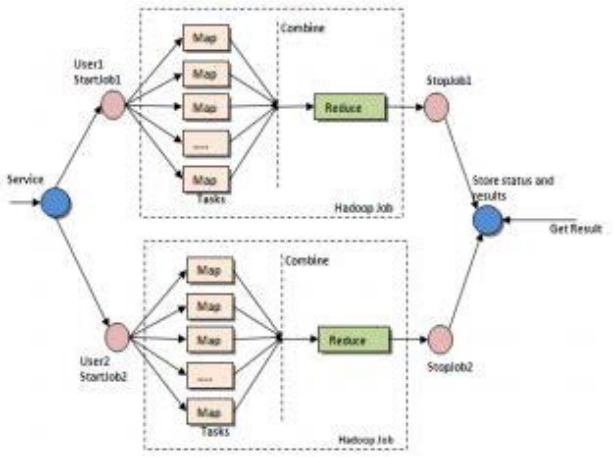

Fig 5: MapReduce Architecture

Fig5: shows the Map Reduce Architecture and Working. It generally figures out how to distribute a neighborhood information block to a slave node. If the effort fails, the scheduler will assign a rack-local or random information piece to the slave hub rather than nearby information square. Whenever outline () finish its task, the runtime system gather all intermediate pairs and launches a set of condense tasks to create the last yield. Huge scale information preparing is a troublesome undertaking, overseeing hundreds or thousands of processors and managing parallelization and distributed environments makes is more troublesome. Delineate gives answer for the said issues, as is backings conveyed and parallel I/O scheduling, it is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data. It is way of approaching and solving a given problem. Using Map Reduce framework the efficiency what's more, an ideal opportunity to recover the information is very sensible. To address the volume viewpoint, new procedures have been proposed to enable parallel processing using Map Reduce framework. Data aware caching (Dache) system that rolled out slight improvement to the first guide decrease programming model and framework to enhance processing for big data applications using the map reduce model [15,16].

The benefit of guide decrease is an expansive assortment of issues are effortlessly expressible as Map reduce computations and cluster of machines handle thousands of nodes and fault-tolerance. The burden of guide diminish is Real-time preparing, not generally simple to execute, shuffling of data, batch processing.

Map Reduce Components:

1. Name Node: manages HDFS metadata, doesn't deal with files directly.

2. Data Node: stores blocks of HDFS—default replication level for each block: 3.

3. Job Tracker: schedules, allocates and monitors job execution on slaves—Task Trackers.

4. Task Tracker: runs Map Reduce operations.

MapReduce Framework

Programs written in this useful style are naturally parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the information, booking the program's execution over an arrangement of machines, taking care of machine failures, and managing the required inter-machine communication. This allows programmers without any involvement with parallel and appropriated frameworks to effortlessly use the assets of a vast distributed system. Our implementation of MapReduce [17] runs on a large cluster of commodity machines and is exceptionally adaptable: a regular MapReduce calculation forms numerous terabytes of data on thousands of machines. Programmers find the system easy to use: hundreds of MapReduce programs have been actualized and upwards of one thousand MapReduce employments are executed on Google's clusters every day.

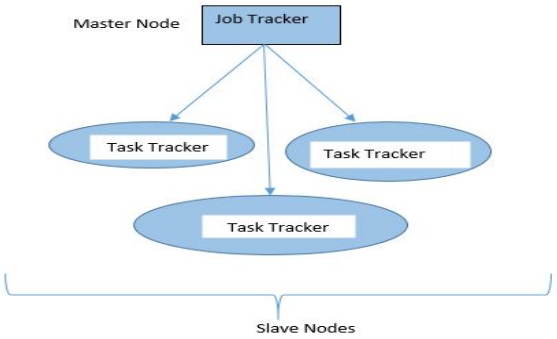

Fig 6: Hadoop Master Slave Architecture

C. Hive:

Hive is an appropriated operator stage, a decentralized framework for building applications by networking local system resources. Apache Hive data warehousing component, an element of cloud-based Hadoop biological system which offers an inquiry dialect called Hive SQL that interprets SQL-like inquiries into Map Reduce jobs automatically. Applications of apache hive are SQL, oracle, IBM DB2. Architecture is separated into Map-Reduce-situated execution, Meta information data for information stockpiling, and an execution part that receives a query from user or applications for execution. The advantage of hive is more secure and usage are great and very much tuned. The inconvenience of hive is only for ad hoc queries and performance is less as compared to pig its shown in Fig 7.

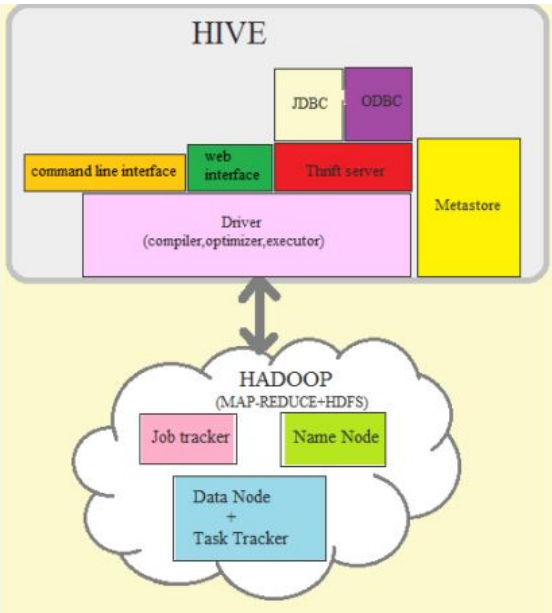

Fig 7: Hive Architecture

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 [20,21].

E.HPCC:

HPCC is an open source stage utilized for registering and that gives the administration to taking care of massive big data 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 [21]:

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.

3. Existing anonymization Techniques of Big Data Privacy using MapReduce Architecture

MapReduce-based anonymization: For productive information preparing MapReduce structure is proposed. Bigger information sets are taken care of with large and distributed MapReduce like frameworks. The data is split into equal sized chunks which are then encouraged to separate mapper. The mapper's procedure its pieces and give combines as outputs. The pairs having the same key are transferred by the framework to one reducer. The reducer output sets are then used to produce the final result [22, 23].

K-anonymity with MapReduce: Since the information is consequently part by the MapReduce system, the k-anonymization algorithm must be insensitive to data distribution across mappers. Our MapReduce based algorithm is reminiscent of the Mondrian calculation. For better sweeping statement and all the more imperatively, lessening the required iterations, each equivalence class is split into (at most) q equivalence classes in each iteration, rather than only two [24].

MapReduce-based l-diversity: The expansion of the security show from k-obscurity to l-differences requires the combination of sensitive values into either the output keys or values of the mapper. Thus, pairs which are produced by mappers and combiners should be fittingly altered. Not at all like the mapper in k-anonymity, the mapper in l-diversity, receives both quasi-identifiers and the sensitive attribute as input [24].

Research Gap

1. Anonymization techniques such as generalization, bucketization, and multi-set based generalized, one-attribute-per column slicing and slicing are well designed for improving accuracy in privacy preservation. Slicing

with suppression is an innovative data Anonymization technique which can improve the privacy preservation in current scenario and make it more difficult for intruder to retrieve information. In future direction, addressing the scalability problem in other types of privacy models by exploiting the Hadoop MapReduce framework. [25 26 27 29].

2. In the necessities of using data mining techniques in Big Data due to its specific properties that makes it more suitable when dealing with current health data. There will be some challenges and issues when applying data mining techniques in healthcare, such as algorithm performance, information reliability, data quality, and variety of methods [28, 29].

3. In Big data 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 output results in Big data.

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. The paper portrays Hadoop which is an open source programming utilized for preparing of Big Data. Big data holders have shown more attention to data mining techniques in the past decades, as these techniques can help them to extract very useful information from their gathered data. These information can be used to improve the health services and deliveries, find the unknown relationship between diseases, and make the organization cost efficient. Traditional data de-identification approaches can be deployed to improve security in the Hadoop environment, like storage level encryption, traditional field-level encryption, and data masking. 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.

References

[1] Sagiroglu, S.Sinanc, D.,‖Big Data: A Review‖,2013, 20-24.

[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] Haferlach T, Kohlmann A, Wieczorek L, Basso G, Kronnie GT, Bene M-C, De Vos J, Hernandez JM, Hofmann W-K, MillsKI, Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Liu WM, Williams PM, Fo R. Clinical utility of microarray-based gene expression profiling in the diagnosis and sub classification of leukemia: report from the international microarray innovations in leukemia study group. J Clin Oncol. 2010;28(15):2529–37.

[4] Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Tollenaar R. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol. 2011;29(1):17–24.

[5] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–7.

[6] Groves P, Kayyali B, Knott D, Kuiken SV. The 'big data' revolution in healthcare. New York: McKinsey & Company; 2013.

[7] Public Law 111–148—Patient Protection and Affordable Care Act. U.S. Government Printing Office (GPO); 2013.

[8] EHR incentive programs. 2014. [Online]. https://www.cms.gov/Regulations-and Guidance/Legislation/EHRIncentive-Programs/index.html.First things first—highmark makes healthcare-fraud prevention top priority with SAS. SAS; 2006. 63. Acampora G, et al. Data analytics for pervasive health. In: Healthcare data analytics. ISSN:533-576. 2015.

[9] Amogh Pramod Kulkarni, Mahesh Khandewal, ―Survey on Hadoop and Introduction to YARN‖, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[10] E. Yu and S. Deng, "Understanding software ecosystems: A strategic modeling approach," in Proceedings of the Workshop on Software Ecosystems 2011, 746(IWSECO2011), 2011, p. 6-6.

[11] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).

[12] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.

[13] Aditya B. Patel, Manashvi Birla and Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce," in Proc. 2012 Nirma University International Conference On Engineering.

[14] Jimmy Lin ―Map Reduce Is Good Enough?‖ The control project, IEEE Computer 32 (2013).

[15] Apache HDFS. Available at http://hadoop.apache.org/hdfs

[16] Apache Hive. Available at http://hive.apache.org

[17] dhruba,jssarma,jgray,kannan,nicolas,hairong,krangana than,dms,aravind.menon,rash,rodrigo,amitanand.s
[18] "Apache Hadoop Goes Realtime at Facebook" SIGMOD .'11, June 12.–16, 2011, Athens, Greece. Copyright 2011 ACM 978-1-4503-0661-4/11/06
[19] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N ―Analysis of Big Data using Apache Hadoop and Map Reduce‖ Volume 4, Issue 5, May 2014‖.
[20] Suman Arora, Dr.Madhu Goel, ―Survey Paper on Scheduling in Hadoop‖ International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.
[21] Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.
[22] Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertain Fuzz. 2002;10(5):557–70.
[23] HessamZakerdah CC, Aggarwal KB. Privacy-preserving big data publishing. La Jolla: ACM; 2015.
[24] Morey, Timothy, Theodore Theo Forbath, and Allison Schoop. "Cus-tomer Data: Designing for Transparency and Trust," Harvard Busine ss Review 93.5 (2015): 96-+.
[25] Friedman.A, R. Wolff, and A. Schuster, "Providing k-Anonymity in Data Mining," Intl J. Very Large Data Bases, vol. 17, no. 4, pp. 789-804, 2008.
[26] Fung, Benjamin, et al. "Privacy-preserving data publishing: A sur-vey of recent developments," ACM Computing Surveys (CSUR) 42.4 (2010):14.
[27] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, 'Health big data analytics: current perspectives, challenges and potential solutions', Int. J. Big Data Intell., vol. 1, no. 1/2, pp. 114– 126, 2014.
[28] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacypreserving data publishing: A survey of recent developments," ACM Computing Surveys, vol. 42, no. 4, pp. 1–53, June 2010.

[29] Priyank Jain, Manasi Gyanchandani, Nilay Khare, "Big data privacy: a technological perspective and review", *Journal of Big Data*, vol. 3, pp. , 2016, ISSN 2196-1115.

Mr. Priyank Jain is working as a PhD Research Scholar. He is having 8 years' Experience as an Assistant professor & in research field. Mr. Priyank Jain has experience From Indian Institute of Management Ahmedabad India (IIM A) in research field. His Educational Qualification is M.Tech & BE in Information Technology. Mr. Priyank Jain's areas of specialization are Big data, Big Data Privacy & Security, data mining, Privacy Preserving, & Information Retrieval. Mr. Priyank Jain has publications in various International Conference, International Journal & National Conference. He is a member of HIMSS.

Mansi Gyanchandani working as Assistant Professor in MANIT Bhopal. She is having 20 years' experience, Her Educational Qualification is PhD in Computer Science & Engineering. Dr. Manasi Gyanchandani area of Specialization in Big data, Big Data Privacy & Security, data mining, Privacy Preserving, Artificial Intelligence, Expert System, Neural Networks, Intrusion Detection & Information Retrieval. Dr. Manasi Gyanchandani, publications in 04 International Conference, 04 International Journal & 04 National Conference. She is Life member of ISTE.

Nilay Khare is having more than 21 years' experience, His Educational Qualification is PhD in Computer Science & Engineering. Dr. Nilay Khare area of Specialization in Big data, Big data privacy & security, Wireless Networks, Theoretical computer science. Dr. Nilay Khare, publications in 55 National and International conference He is Life member of ISTE.

Dhirendra Pratap singh received his PhD degree in computer science and engineering from the Maulana Azad National Institute Technology, Bhopal, India in 2015. After his Post graduation, he has worked as Software Developer in NIIT Technologies Ltd., New Delhi, India. Currently he is working as assistant professor in the department of computer science and engineering, Maulana Azad National Institute Technology, Bhopal, India.