International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Implementation of Privacy Preserving Methods Using Hadoop Framework

Sarangkumar S. Dubey
Student of Master of Engineering in (CSE)
Sipna's college of Engineering and Technology, Amravati, India

Prof. Seema B. Rathod
Assistant Professor in Department of (CSE)
Sipna's College of Engineering and Technology, Amravati, India

----------***----------

Abstract - *The proliferation of information technology and the internet in the last few years leading to an explosion in the collection of data which requires all the time by organizations and individuals. Organizations push their vast amount of data into big data clusters, but most of these implementations have security concerns. Protection and confidentiality of these huge amount of data is also very important. So to extract the useful information from this vast amount of growing data, in recent year's privacy-preserving big data mining has emerged as a popular research area for the security of sensitive information. This paper presents a privacy preserving method for big data mining, which is novel, effective and efficient method for mining proper information from large and distributed databases using Hadoop framework. This paper shows the Implementation of Hadoop Framework on the Windows Operating System and connect this Hadoop framework with the software application for performing all the activities of storage, retrieve, delete, etc. data files and important documents. It has file system called HDFS that can process data on large scale and also presents the distributed approach which help to preserve privacy at the same time. Our goal is to mining the data from large volumes and at the same time preserving data privacy and confidentiality that perfectly satisfies with Hadoop framework.*

Keywords: Big Data, Data Mining, Hadoop, Hadoop Distributed File System (HDFS), Windows Application, Privacy Preservation, Authentication, Authorization and Accounting.

I. INTRODUCTION

With the increasing world of digitalization and the internet, from last few years it is becoming easy to find anything or any individual's information [1]. This is possible only because this vast amount of data related to all activities of individuals are stored somewhere on the internet, this leads to the explosion of data. The term big data usually describe the exponential growth of data for both structured and unstructured type. The rapid advancement in the technology, the use of this big data and computation performed on it is important to business and society for carrying out large amount of information. Production of data is increase so rapidly such that world's volume of data doubles every eighteen months. This vast amount of data is called as big data having large volumes of information. To extract knowledge from such huge amount of data the term emerges as big data mining. As the data are going on increasing the techniques for retrieving useful information from such big data are also simultaneously being develop.

A new era of research started where existing data mining techniques are considered for privacy requirement as the development and use of internet increases the threat against privacy of are also becoming increasing in the same amount. Some data owners utilized data mining techniques to extract knowledge and compromise the privacy of their client but doing this for increasing amount of big data along with privacy protection is also a research question [2]. The organizations or the companies have their big data clusters on the cloud about the ongoing activities of their clients. It is very challenging to mine knowledge while protecting individuals' privacy from such a huge amount of data sets containing sensitive information also [3]. We propose a method for preserving privacy in mining big databases using Hadoop framework which fits proper for this task.

Hadoop [5] is a scalable and distributed data processing, open-source project having their own standard for big data processing and contributed to the proliferation of massive data analysis [6]. This large size of data processing utility is enhanced through its combination with data mining algorithm employed in rule mining and pattern recognition. And thus, numerous studies have been conducted to apply existing mining techniques to the Map Reduce programming model [7]. As this vast amount of data contains some sensitive

information leading to the research about privacy preserving data mining for storing and using distributed massive datasets. Using the Technology of Hadoop and HDFS we try to perform Distributed data storing and mining proper information whenever required.

We are using one of the framework of Hadoop that is HDFS which is mainly useful for processing our big data. Hadoop provides a distributed file system and a framework for the analysis and transformation of very large data sets using the MapReduce paradigm. An important characteristic of Hadoop is the partitioning of data and computation across thousands of hosts, and the execution of application computations in parallel close to their data. A Hadoop cluster scales computation capacity, storage capacity and I/O bandwidth by simply adding commodity servers. Hadoop clusters at Yahoo! span 40,000 servers, and store 40 petabytes of application data, with the largest cluster being 4000 servers. One hundred other organizations worldwide report using Hadoop. This suggest for processing and storing big data Hadoop is most useful framework [4].

As the usage of data is growing on increasing so its proper storage is very much important so that it can be made useful later as efficiently as possible. We propose the privacy preserving approach for Data mining by implementing the approach based on Hadoop Map Reduce framework [8]. We store the database on HDFS framework to achieve proper mining from big data in distributed manner [9]. This framework uses the external services for data providers that want to limit privacy violation. This HDFS framework provide distributed service which is useful for providing security mechanism for processing big data and mining useful patterns.

As the Big data technique that adds noise to data results in a problem for privacy and utility of data tradeoff. Thus, high privacy protection degree reduces data utility, and the high data utility requirement increases the possibility of privacy violations. In this paper, we propose a privacy-preserving data mining that prevents privacy violation without deteriorating data utility [10]. The proposed technique adopts proper data storage using Hadoop framework. This Hadoop framework can efficiently handle proper data mining along with privacy concern. We have also use the method of AAA i.e. Authentication, Authorization and Accounting technique that also adds security to PPDM technique for big data [11]. The remaining paper is organized as, Section II contains the important definitions of the terms that we have used in this paper. Section III gives the Hadoop installation process on windows platform. Section IV defines the system architecture. In the next section V gives the stepwise working process of our proposed system. Finally section VI conclude the paper.

II. IMPORTANT DEFINITIONS

A. Hadoop: Apache Hadoop is an open-source framework for distributed storage and processing of large sets of data on commodity hardware that enables businesses to quickly gain insight from massive amounts of structured and unstructured data.

B. HDFS: The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS is a Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers.

C. MapReduce:
MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

D. Data Mining:
Data mining is the extraction of hidden predictive information from large databases, invented as a powerful new technology with great potential to help companies and organizations to focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans.

E. Association rule mining:
Association rule mining is a method for mining interesting relations between different parameters in large databases. It is intended to identify some interestingness measure using strong rules discovered from the large databases.

F. Frequent Itemset Mining (FIM):
Frequent Itemset Mining (FIM) has been an essential part of data analysis and data mining. FIM tries to extract information from databases based on frequently occurring events, i.e., an event, or a set of events, is interesting if it occurs frequently in the data, according to a user given minimum frequency threshold

III. HADOOP INSTALLATION PROCESS

Basically there are following steps to be followed for installing Hadoop framework to any computer that has Windows operating system:

Step -1 Windows path configuration

- set HADOOP_HOME path in environment variable for windows
- Set hadoop bin directory path

Step 2 – Hadoop configuration

- Edit - hadoop-2.7.1/etc/hadoop/core-site.xml
- Edit - hadoop-2.7.1/etc/hadoop/mapred-site.xml
- Edit - hadoop-2.7.1/etc/hadoop/hdfs-site.xml
- Edit - hadoop-2.7.1/etc/hadoop/yarn-site.xml
- Edit - hadoop-2.7.1/etc/hadoop/hadoop-env.cmd

Step 3 – Start everything

Before starting everything you need to add some [dot].dll and [dot].exe files of windows platform please download bin folder from my github repository - sardetushar_gitrepo_download bin folder
– this contains .dll and .exe file (winutils.exe for hadoop 2.7.1)

Now delete you existing bin folder and replace with new one (downloaded from my repo)
Open cmd and type 'hdfs namenode -format'

- Open cmd and point to sbin directory and type 'start-all.cmd'

It will start following four process at the same time which is the indication of our installation is successful

- Name node
- Data node
- YARN resource manager
- YARN node manager

This four process after successfully starting are seen as shown in the fig 1 below:

Figure 1: After successful installation starting 4 processes of Hadoop

Step 4 – namenode GUI, resource manager GUI

After successfully starting Hadoop on the windows platform user can also see the GUI interface for Name node and Recourse manager on the following address.

- Resource manager GUI address - http://localhost:8088
- Namenode GUI address – http://localhost:50070

IV. SYSTEM ARCHITECTURE

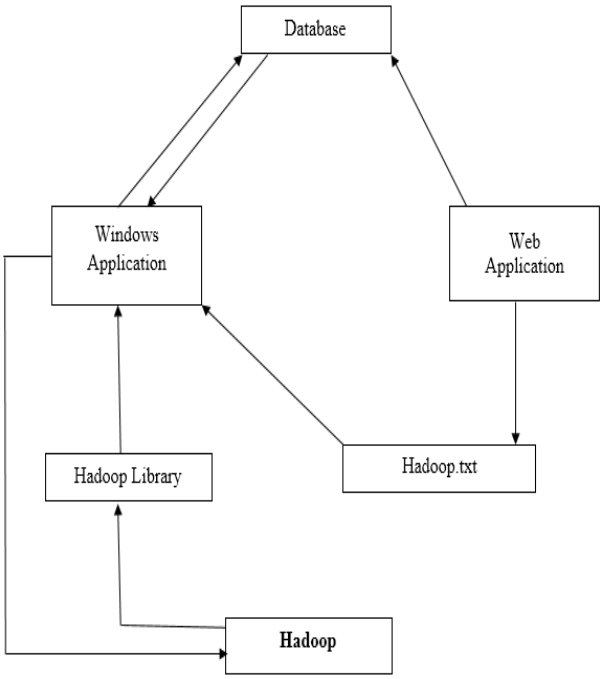

Figure 2: Architecture diagram of Working System

In this paper, we are performing storage and operations on big data and mainly on the Windows operating system. The above figure 2 shows the architecture diagram of our proposed system. With this system both windows application and web application are able performing computation on Hadoop framework. The first step starts with the Hadoop installation and its configuration to our windows based system properly. After the successful installation we have to build the Hadoop library for inserting, accessing or performing any transaction with Hadoop from our windows platform. As we have to handle large amount of data in terms of files and other storage repositories it is necessary to develop a web application that communicate with Hadoop. But web application is unable to directly communicate with Hadoop

framework. For that, we have develop the architecture as, first of all building the library for communicating the Hadoop with windows application. Then develop a windows application which is able to perform transactions with files storage and insert the library into it. Then perform transaction with windows application on Hadoop and its result is stored in the database. And now web application is able to fetch the data that is stored in the database and perform different transactions also with large amount of data storage. In this way the complete architecture performs its functionality.

This architecture suites mostly for our designed functionality. For handling large amount of data with Hadoop on the windows platform required all the things shown in the architecture to be work co-operatively. The arrow head shown in the architecture shows the stepwise implementation process. As we have to handle large amount of database our applications interface should be User-friendly. As we are dealing with big data it may always contains some sensitive information so it's necessary to be access with proper privacy and security concern.

V. STEPWISE WORKING OF SYSTEM

Now after performing Successful installation of Hadoop, it's the time to build the system architecture as shown above. The following are the steps that is useful for properly building our proposed system.

- Configure and Install the Hadoop for manipulating Big Data
- Used HDFS for storing the Big Data sets in the Distributed manner
- Develop the Library for manipulating commands with Hadoop.
- Develop the windows application that is necessary for performing transaction with Hadoop successfully installed on our system.
- Now, develop a web application which is useful for large amount of data manipulations and at the same time requires the Providers and Users Privacy.
- This leads to huge amount of data generation and Storage in proper manner and make arrangement for further use. We store the data in HDFS.
- Now, at the time of retrieval we apply the Data Mining techniques. Here, we use the Association rule mining and frequent item sets mining techniques that fits suitable for our application
- For security and privacy concern, we apply Authentication, Authorization and Accounting (AAA) Framework

VI. CONCLUSION

Here, we are going to apply the Privacy preserving data mining techniques on big data. And we analyzes that Hadoop is the most suitable platform for performing this task. Now in business and actual working environment, having access to the right information means making the right decision critical to surviving is important. Business need to protect their information as it accumulates much faster and secure because of important big data. As the era of big data begin, companies databases also increases so it is possible to take one by one data and analyzed it based on some rules made. Therefore we used data mining technique on big data to extract interesting knowledge. Data mining process work on data and data contain information about individuals. Our approach of using Hadoop framework is very much useful for such huge amount of data storage. The functionality of Hadoop i.e. HDFS used for distributed file Storage system is used for distributed processing of application along with support to privacy and security concern for such a big data also. Here, for proper preserving privacy of such huge amount of data we use Authentication, Authorization and Accounting (AAA) framework. The AAA is the system in which only authorized person is able to enter his authentication information to get access to use our designed system.

REFERENCES

[1] Nasrin Irshad Hussain, Bharadwaj Choudhury, Sandip Rakshit, "A Novel Method for Preserving Privacy in Big-Data Mining", *International Journal of Computer Applications (0975 –8887),* Volume 103–No16, October 2014.

[2] Gartner, Post event brief, Gartner IT Infrastructure, Operations and Management Summit 2009, Orlando, FL. [Online]: available at www.gartner.com. June 23–25 2009.

[3] Arie Friedman, "Privacy preserving data mining", pp.4, January 2011.

[4] "Big security for big data", available at www8.hp.com/ww/en/secure/pdf/4aa4-4051enw.pdf

[5] Apache Hadoop. http://hadoop.apache.org/.

[6] http://hortonworks.com/hadoop/hdfs/

[7] *Ullman, J. D.* (2012). *"Designing good MapReduce algorithms"*. XRDS: Crossroads, The ACM Magazine for Students (*Association for Computing Machinery*) 19: 30. *doi*:*10.1145/2331042.2331053*.

[8] Data mining Articles [Online] available: http://www.dataminingarticles.com/info/data-mining-introduction/

[9] Malik, M.B. Ghazi, M.A. ; Ali, R., "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", *Computer and Communication Technology (ICCCT)*, Third International Conference on 23-25 Nov. 2012

[10] Yehuda Lindel, Benny Pinkas, "Privacy Preserving Data Mining" [online] available: http://www.pinkas.net/PAPERS/id3-final.pdf

[11] Blue Coat, "Technology Primer: Authentication, Authorization, and Accounting" online pdf.