Optimizing Hadoop’s Mapreduce Framework For High Scalable EHR Sharing

T.Krithika¹, Dr.J.Vellingiri²
¹PG Student, ²Professor,
Department of Computer Science and Engineering,
Kongunadu College of Engineering and Technology, Anna University, Tamilnadu.

ABSTRACT
Big data is an enticing technology that provides large number of storage capacity and by using a large number of computers together it enable the users to deploy applications in an shared environment. Personal Health Records (PHR) are user friendly online solution that helps patients to manage their health details. One of the major problems in moving to cloud computing is its security and privacy issues. The encryption technique that is used is not suitable for data that is processed and shared frequently and anonymizing big data and managing anonymized data sets are still challenges for securing details of patients. Thus, a scalable two-phase bottom up generalization (BUG) approach has been proposed to anonymize large-scale data sets using the dynamic MapReduce framework on cloud. In both the phases of approach, group of innovative dynamic MapReduce algorithms is designed to concretely accomplish the specialization computation in a highly scalable way and ensure efficiency.

Keywords: Electronic Health Records, BUG approach, HDFS, Dynamic MapReduce

I. INTRODUCTION
The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modellers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications.

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. The big data is the term of datasets so large and complex that it becomes difficult to process using hand database management tools or traditional data processing applications. Big Data processing is performed through a programming paradigm known as Map reduces. Typically, implementation of the Map Reduce paradigm requires networked attached storage and parallel processing. The fact is that with so much data being generated by so many organizations and users, storage and security simply have to become critical business issues.

Ninety per cent of the total data in the world today has been created in the past two years, and 2014 and beyond will see us generating exponentially larger levels of data. So with more data comes greater threat of attack and greater need for security. Privacy is one of the most concerned issues in cloud computing, and the concern aggravates in the context of cloud computing although some privacy issues are not new to adopt such frameworks to address the scalability problem of anonymizing large-scale data for privacy preservation. Today's databases contain a lot of sensitive personal data. So it's crucial to design information systems which might limit the revealing of personal data. As an example, think about a hospital that maintains patient records. The hospital desires to disclose data to a company in such some way that the company cannot infer that patients have that diseases. One technique to formally specify privacy policies is to specific sensitive data as queries and enforces excellent privacy, an awfully sturdy notion of privacy.

II. RELATED WORK AND PROBLEM ANALYSIS
A wide variety of privacy models and anonymization approaches have been put forth to preserve the privacy sensitive information in data sets. Data privacy is one of
the most concerned issues because processing large scale privacy-sensitive data sets often requires computation power provided by public cloud services for big data applications. We studied the scalability issues of existing BUG approaches when handling big data-data sets on cloud. Most exiting algorithms exploit indexing data structure to assist the process of anonymization. To address the scalability problem of the Top-DownSpecialization (TDS) approach for large scale data set used a widely adopted parallel data processing framework like Map Reduce. In first phase, the original datasets are partitioned into group of smaller. In second phase, these intermediate results are integrated into one and further anonymized to achieve consistent k-anonymous dataset.

Map Reduce is used to split up the large input data into chunks of more or equal size, spinning up a number of processing instances for the map phase apportioning data to each of the mappers. It is easy to scale up MapReduce to handle bigger jobs or to produce results in a shorter time by simply running the job on a larger cluster. When Map reduce is not used the process fails in distribution system. By creatively applying Map Reduce on cloud to Bottom Up Generalization (BUG) for data anonymization and deliberately design a group of innovative Map Reduce jobs to concretely accomplish the generalizations in a highly scalable way. Secondly, introduce a scalable Advanced BUG approach, which performs generalization on different partitioned data set and the resulting intermediate anonymizations are merged to find final anonymization which is used to anonymize the original data set. Electronic health record sharing is provided using hybrid architecture.

III. BUG ARCHITECTURE

BUG architecture provides services including anonymization and specialization approaches. Electronic health records are obtained from the website and patient identifying information is stored at the Hadoop framework.

IV. SYSTEM IMPLEMENTATION

4.1 Datasets Acquisition

Privacy is one of the most concerned issues in cloud computing. Personal data like financial transaction records and electronic health records are extremely sensitive although that can be analyzed and mined by research organization. Data privacy issues need to be addressed before data sets are shared on cloud for analysis purpose. Data anonymization refers to as hiding sensitive data for owners of data records. So in this module importing health records. An electronic health record (EHR), or electronic medical record (EMR), is a systematic collection of electronic health information about an individual patient or population.

4.2 Data Chunk Creation

Effective management and analysis of large-scale data poses an interesting but critical challenge. Recently, big data has attracted a lot of attention from academia, industry as well as government. So in this module all data records are split into data chunks, e.g. the type of chunk, comments, size etc. In the middle there is a variable area containing data which are decoded by the program from the parameters in the header. Chunks may also be fragments of information which are downloaded or managed by distributed programs. In distributed computing, a chunk is a set of data which are sent to a processor or one of the parts of a computer for processing.

4.3 Mrbug Driver

In this module implement Bottom-Up Generalization (BUG) approach to improve the scalability and performance of BUG in advance fashion. The function of this approach is based on the two levels of parallelization provisioned by MapReduce on cloud. Basically, MapReduce on cloud has two levels of parallelization, i.e., job level and task level.

Job level parallelization means that multiple MapReduce jobs can be executed simultaneously to make full use of cloud infrastructure resthises. Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data partitions.

4.4 Anonymized Datasets

Data anonymization is widely used method for Privacy Preserving of data in non-interactive data publishing scenario Data anonymization refers to the hiding the identity or sensitive data for owners data record. The privacy of individual can be effectively preserved while some aggregate information is shared for data analysis and mining. Several models of security can improve Data Anonymization include k-anonymity and l-diversity. Data anonymization enables the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics post-anonymization. In the context of medical data, anonymized data refers to data from which the patient cannot be identified by the recipient of the information. The name, address, and full post code must be removed together with any other information which,
in conjunction with other data held by or disclosed to the recipient, could identify the patient.

**System Architecture**

![System Architecture Diagram]

**Figure 1: System Architecture**

V. CONCLUSION

The scalability problem of data anonymization for big data applications on cloud has been studied using Bottom Up Generalization (BUG) and proposed scalable Advanced Bottom Up Generalization. The proposed BUG performed as Data partitioned, executing the driver producing the intermediate results. Then, the intermediate results are merged and generalization is applied to produce anonymized data without violating k-anonymity. The MapReduce Framework is effectively applied on cloud for data anonymization and shows that scalability and efficiency of centralized BUG are improved significantly over existing approaches. In future optimized balanced scheduling strategies are expected to be developed towards overall scalable privacy preservation aware dataset scheduling. And also this method is designed for achieving k-anonymity; it can be modified to adopt the privacy model in order to accommodate for the high-dimensional data.

**REFERENCES**

Journal paper:


