International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

Mapreduce Based K-Means Clustering Over Large-Scale Dataset

Sajitha N
Asst. Professor, Dept. of CSE
BNM Institute of Technology,
Bengaluru, India

Bhagyalakshmi K C
Student, Dept.of CSE
BNMIT,
Bengaluru, India

Bhagyashree
Student, Dept.of CSE
BNMIT,
Bengaluru, India

Chaitra C M
Student, Dept.of CSE
BNMIT,
Bengaluru, India

Kasthuri R
Student, Dept.of CSE
BNMIT,
Bengaluru, India

***Abstract* -The enlarging volumes of information emerging by the progress of technology, makes clustering of very large scale of data a challenging task. In order to deal with the problem, a parallel k- Means clustering algorithm based on MapReduce is proposed which is a simple yet powerful parallel programming technique.**

A major trend to handle a clustering over large-scale datasets is outsourcing it to HDFS. This is because it offers reliable services with performance guarantees. However, as datasets used for clustering may contain sensitive information, e.g., patient health information, commercial data, and behavioural data, etc., directly outsourcing them to storage inevitably raise privacy concerns.

The proposed system takes files as input and applies clustering algorithm. Clustering enhances the easy accessibility of particular data. MapReduce method increases the speed of the data storage and it also reduces the amount of space consumed by the file. The data at the end is stored in Hadoop file system which is a distributed file system hence transmission of file is faster. Hash code and DNA encryption methods are used to ensure the security of the data.

Keywords: MapReduce, Clustering, Data mining, Large data sets, DNA Encryption

1. INTRODUCTION

Clustering is one major task of exploratory data mining and statistical data analysis, which has been ubiquitously adopted in many domains, including healthcare, social network, image analysis, pattern recognition, [1] etc. Meanwhile, the rapid growth of big data involved in today's data mining and analysis also introduces challenges for clustering over them in terms of volume, variety, and velocity [3].

To efficiently manage large-scale datasets and support clustering over them, public cloud infrastructure is acting the major role for both performance and economic consideration. Nevertheless, using public cloud services inevitably introduces privacy concerns. This is because not only many data involved in data mining applications are sensitive by nature, such as personal health information, localization data, financial data, etc., but also the public cloud is an open environment operated by external third-parties.

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. MapReduce is a core component of the Apache Hadoop software framework. Hadoop enables resilient, distributed processing of massive unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage. The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples.

The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job [3].

2. PROBLEM STATEMENT

With the explosion of data in today's big data era, searching for a particular file in huge dataset becomes more tedious job. The concept of map reduce is used, it divides the file into blocks and search the block while retrieving in particular cluster, which saves time and enhances the performance.

3. LITERATURE SURVEY

[1] Clustering techniques have been widely adopted in many real world data analysis applications, such as customer behavior analysis, targeted marketing, digital forensics, etc. With the explosion of data in today's big data era, a major trend to handle a clustering over large-scale datasets is outsourcing it to public cloud platforms. This is because cloud computing offers not only reliable services with performance guarantees, but also savings on in-house IT infrastructures. However, as datasets used for clustering may contain sensitive information, e.g., patient health information, commercial data, and behavioral data, etc, directly outsourcing them to public cloud servers inevitably raise privacy concerns. In this paper, a practical privacy-preserving K-means clustering scheme is proposed that can be efficiently outsourced to cloud servers.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

This scheme allows cloud servers to perform clustering directly over encrypted datasets, while achieving com-parable computational complexity and accuracy compared with clustering over unencrypted ones. Secure integration of MapReduce is also investigated into this scheme, which makes this scheme extremely suitable for cloud computing environment.

[2] It is attractive for an organization to outsource its data analytics to a service provider who has powerful platforms and advanced analytics skills. However, the organization (data owner) may have concerns about the privacy of its data. In this paper, a method is represented that allows the data owner to encrypt its data with a homomorphic encryption scheme and the service provider to perform k-means clustering directly over the encrypted data. However, since the ciphertexts resulting from homomorphic encryption do not preserve the order of distances between data objects and cluster centers, an approach is proposed that enables the service provider to compare encrypted distances with the trapdoor information provided by the data owner. The efficiency of this method is validated by extensive experimental evaluation.

[3] Big Data has come up with aureate haste and a clef enabler for the social business; Big Data gifts an opportunity to create extraordinary business advantage and better service delivery. Big Data is bringing a positive change in the decision making process of various business organizations. With the several offerings Big Data has come up with several issues and challenges which are related to the Big Data Management, Big Data processing and Big Data analysis. Big Data is having challenges related to volume, velocity and variety. Big Data has 3Vs Volume means large amount of data, Velocity means data arrives at high speed, Variety means data comes from heterogeneous resources. In Big Data definition, Big means a dataset which makes data concept to grow so much that it becomes difficult to manage it by using existing data management concepts and tools. Map Reduce is playing a very significant role in processing of Big Data. This paper includes a brief about Big Data and its related issues, emphasizes on role of MapReduce in Big Data processing. MapReduce is elastic scalable, efficient and fault tolerant for analyzing a large set of data, highlights the features of MapReduce in comparison of other design model which makes it popular tool for processing large scale data. Analysis of performance factors of MapReduce shows that elimination of their inverse effect by optimization improves the performance of Map Reduce.

[4] Data clustering has been received considerable attention in many applications, such as data mining, document retrieval, image segmentation and pattern classification. The enlarging volumes of information emerging by the progress of technology, makes clustering of very large scale of data a challenging task. In order to deal with the problem, many researchers try to design efficient parallel clustering algorithms. In this paper, a parallel k-means clustering algorithm based on MapReduce is proposed, which is a simple yet powerful parallel programming technique. The experimental results demonstrate that the proposed algorithm can scale well and efficiently process large datasets on commodity hardware.

4. PROPOSED SYSTEM

A system called K-means clustering is proposed that uses Map Reduce technique over Large-scale Dataset [4]. First the trained data sets are initialized for every different cluster which is related to Reuter's collection Information .After, the clustering algorithm divide file into number of chunks and for every chucks hash code is generated for the security purpose [2]. Before storing into HDFS System, classification algorithm classifies that file belong to which cluster category [1]. Advantages of proposed system are:

- Hadoop is a distributed file system and provides fast storage of files.
- Security is more because of Hash code generation
- Speed of Transmission is high because of duplication concept is used while uploading file to the HDFS storage

5. PERFORMANCE EVALUATION

Classification	Actual Storage	MapReduce Storage (KB)	MapReduce storage %	Saved %
Cluster 1	520	340	65.38461538	34.61
Cluster 2	450	250	55.55555556	44.44
Cluster 3	636	323	50.78616352	49.21
Cluster 4	341	212	62.17008798	37.82
Cluster 5	250	142	56.88888888	43.12

Table 4.1: Performance of MapReduce

6. CONCLUSION

With the explosion of data in today's big data era, searching for a particular file in huge dataset becomes more tedious job. Clustering is one major task of exploratory data mining and statistical data analysis, which has been ubiquitously adopted in many domains, including healthcare, social network, image analysis, pattern recognition, etc. Meanwhile, the rapid growth of big data involved in today's data mining and analysis also introduces challenges for clustering over them in terms of volume, variety, and velocity.

In this work, privacy-preserving MapReduce based K-means clustering scheme is proposed, which allows the better classification of the files into relevant clusters. Whenever we are outsourcing the data, security is the main concern because there may be a malicious hacker who may hack our data. Encryption is required to securely protect data that we don't want anyone else to have access to. Privacy-preserving is achieved through the DNA encryption which involves the cooperation of both the sender and the receiver and also provides corresponding decryption method. Considering the support of large-scale dataset securely integrated MapReduce framework is designed, MapReduce is an innovative tool which is used in order to reduce the storage space required. Speed of transmission is very low when we are dumping the files into big data, MapReduce improves the efficiency by reducing the space and the time used for computation and makes it extremely suitable for parallelized processing in HDFS environment. In the end files are stored on Hadoop storage which is cost free and provides good capability for storage. The user can now download the required decrypted file from the Hadoop.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

5. REFERENCES

[1] Jiawei Yuan, Yifan Tian, "Practical Privacy-Preserving MapReduce Based K-Means Clustering over Large-Scale Dataset", IEEE Transaction On Cloud Computing (Vol: PP, Jan 2017)

[2] Dongxi Liu, Elisa Bertino, Xun Yi, "Privacy of Outsourced K-Means Clustering", Cyber Center Publications, June 2014, Paper 590

[3] Shweta Pandey, Dr.Vrinda Tokekar, "Prominence of Mapreduce in Big Data Processing", Fourth International Conference on Communication Systems and Network Technologies, 978-1-4799-3070-8/14, 2014 IEEE

[4] Weizhong Zhao, Huifang Ma, and Qing He, "Parallel K-Means Clustering Based On Mapreduce". CloudCom 2009, LNCS 5931, pp. 674-679, 2009

[5] Guangzhou Cui, Limin Qin, Yanfeng Wang, Xuncai Zhang, "An Encryption Scheme using DNA Technology" 978-1-4244-2724-6/08, BIC-TA 2008 IEEE