A SECURE HASH BASED DEDUPLICATION SYSTEM IN HADOOP ARCHITECTURE

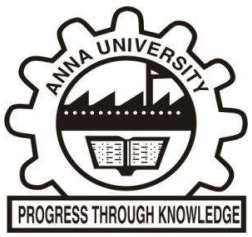
A THESIS

Submitted by

RAMYA P

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY



FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING

ANNA UNIVERSITY

CHENNAI 600 025

DECEMBER 2020

ANNA UNIVERSITY
CHENNAI 600 025

BONAFIDE CERTIFICATE

The research work embodied in the present thesis entitled **"A SECURE HASH BASED DEDUPLICATION SYSTEM IN HADOOP ARCHITECTURE"** has been carried out in the Department of Computer Science and Engineering, Christian College of Engineering and Technology, Oddanchatram. The work reported herein is original and does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion or to any other scholar.

I understand the University's policy on plagiarism and declare that the thesis and publications are my own work, except where specifically acknowledged and has not been copied from other sources or been previously submitted for award or assessment.

RAMYA P
RESEARCH SCHOLAR

Dr. P. BABU
JOINT SUPERVISOR
Professor
Department of Computer Applications
PSNA College of Engineering and Technology,
Dindigul – 624 622.

Dr. C. SUNDAR
SUPERVISOR
Professor
Department of Computer Science and Engineering
Christian College of Engineering and Technology,
Oddanchatram - 624 619.

ABSTRACT

The digital world has greatly shifted towards a distributed environment where the autonomous devices are enabled to do work in a co-operative manner. As more demand for distributed computing blossomed many devices are connected and able to share their resources among themselves in a disciplined manner. In such an environment data has emerged as a valuable asset, since more data sources extraordinarily produce information. The volume of data gets increased along with the velocity of generation and this huge amount of data comes from a variety of sources. The handling, processing, and storing of the huge volume of data has introduced a new concept called Big Data. Big Data is a methodology that employs different techniques to extract more accurate information from the huge volume of data.

The data has changed as a more important asset since they provide valuable information, so they have to be stored and managed properly. The data are stored in an environment like Hadoop which is capable of processing and handling huge amount of data which are scattered among clusters of computers. The exponential growth in the volume of data has threatened the availability of the storage space. To tackle the situation many storage optimization techniques are proposed, one such technique is data deduplication and it eliminates redundant or duplicate data and stores a unique copy of data. The deduplication reduces the storage space requirement considerably. The big data ecosystem is an open environment where the deduplication is carried out in a decentralized manner where the security issues will arise.

The deduplication is a space-saving technique that divides the given file into many fixed or variable-size blocks and a fingerprint for each block is calculated using the hash algorithm. The fingerprints are unique in nature so it can be used to compare with other blocks to eliminate redundant blocks. The fingerprints are stored in an index like structure which facilitates the search

operation. However, each operation has its own limitation like the variable-size data blocks will give a high deduplication ratio at a higher memory and computational cost. The selection of the fingerprint generation algorithm determines the efficiency of redundancy determination and the fingerprint index becomes more complex along with the growth of data. Data deduplication also suffers from data security which can be achieved by encrypting the data. However, the encryption-based deduplication does not improve the deduplication ratio and query performance.

The proposed research work A Secure Hash Based Deduplication System in Hadoop Architecture consists of four methodologies namely Trusted Third Party Vendor based access control using Elliptic Curve Cryptography (TTPV-ECC), Hash Based Deduplication (HBD), Fuzzy C-Means Cluster assist Indexing (FCMI) and MapReduce-Particle Swarm Optimization (MR-PSO).

The proposed method Trusted Third Party Vendor based access control using Elliptic Curve Cryptography (TTPV-ECC) is used to provide authentications to the system and this method is more scalable. The HBD performs deduplication in series of steps by dividing the data obtained from the trusted client followed by hashing, indexing with the help of SHA-3 and FCMI respectively which provides a better deduplication elimination ratio, shorter deduplication detection time and also it consumes less memory when compared to other methods. The final step MapReduce-Particle Swarm Optimization (MR-PSO) is used to find a suitable data node to store deduplicated data in Hadoop Distributed File System (HDFS) which will enhance the data retrieval process, the overall system performs better in terms of throughput and computation time.

ACKNOWLEDGEMENT

I express my gratitude to the Almighty God for his grace to complete this research work. I express my profound and honest gratitude to my guide **Dr. C. Sundar**, Professor, Department of Computer Science and Engineering, Christian College of Engineering and Technology, Oddanchatram and my joint supervisor **Dr. P. Babu**, Professor, Department of Computer Applications, PSNA College of Engineering and Technology, Dindigul, who helped me to complete this dissertation. Without their guidance, support, and encouragement I could not have finished it.

I am also extremely indebted to my Doctoral Committee member **Dr. A.P Janani**, Professor, Department of Information Technology, Dr.Mahalingam College of Engineering and Technology, Pollachi, for her valuable advice, constructive criticism and extensive discussions on my work. I warmly thank my Doctoral Committee member **Dr. R.S Vetrivel,** Professor, Department of Computer Science, NPR Arts and Science College, Natham, for his unending support.

I take this opportunity to thank **Mr. Jacob Thomas**, Chairman, Christian College of Engineering and Technology, Oddanchatram, for all the facilities he provided for carrying out this work and my gratitude's goes to **Dr. B. Justus Rabi**, Principal, Christian College of Engineering and Technology, Oddanchatram, for his constant encouragement and guidance during the course of my research work.

I also thank the faculty and non-teaching staff members of the Christian College of Engineering and Technology, Oddanchatram, for enabling me to concentrate on my research work. I thank my family members and friends for their support at all times.

RAMYA P