Chapter 31
Ameliorating the Privacy on Large Scale Aviation Dataset by Implementing MapReduce Multidimensional Hybrid k-Anonymization

Stephen Dass A.
Vellore Institute of Technology, Vellore, India

Prabhu J.
Vellore Institute of Technology, Vellore, India

ABSTRACT

In this fast growing data universe, data generation and data storage are moving into the next-generation process by generating petabytes and gigabytes in an hour. This leads to data accumulation where privacy and preservation are certainly misplaced. This data contains some sensitive and high privacy data which is to be hidden or removed using hashing or anonymization algorithms. In this article, the authors propose a hybrid k anonymity algorithm to handle large scale aircraft datasets with combined concepts of Big Data analytics and privacy preservation of storing the dataset with the help of MapReduce. This published anonymized data are moved by MapReduce to the Hive database for data storage. The authors propose a multi-dimensional hybrid k-anonymity technique to solve the privacy issue and compare the proposed system with other two anonymization methods such as BUG and TDS. Three experiments were performed for evaluating classifier error, calculating disruption value and p% hybrid anonymity and estimation of processing time.

DOI: 10.4018/978-1-7998-8954-0.ch031

INTRODUCTION

The global community is experiencing rapid growth in a huge number of data generated by all sensitive personal information (Madden, 2012). When data generation is rapid, the data holder faces a very challenging scenario in holding each and every data which lead into lack of data privacy on sensitive information. A data holder faces a huge compromise in data hide and handling a huge variety of data. Big Data analytics is one of the advanced analytical technologies used on large scale datasets. Big Data plays a vital role in this field leading to a data privacy breach. As owing to the huge technological enhancement and advancement, data streaming has been huge. Google, YouTube, Facebook, and WhatsApp collect personal and sensitive data of the user and they are archived by the social media organization (Kavanaugh et al., 2012). In research, Big Data includes mobile data, healthcare, traffic multimedia data, and aircraft data. The data generated by airline transportation is more challenging for big data analytics. These generated archives are used for analysis of the personal information for their profit. Therefore, the privacy of information is very important for one's private and public data. Hence preserving the privacy of large datasets is ponderous. So many corporate organizations, customers, end-users, hesitate to take Cloud privacy and security due to its insecure and virtual storage and security on large scale datasets.

Anonymization

Anonymization is one of the information bits which are referred to as the extraction of sensitive data intent to privacy protection. Data anonymization helps in sharing from one server source to the destination client across the boundary without any side attack. Data anonymization based on k-anonymity is extremely used for this purpose in data hide or data sharing. With these structures, we combine the data processing categories in order to process large datasets in an efficient manner. Two broad anonymization methods such as bottom-up generalization (BUG) and top-down specialization (TDS) play a vital role in data privacy and data hiding of sensitive attribute in the dataset. The first BUG generalizes the data from bottom to up taxonomy (Wang et al., 2004) whereas the latter method, TDS specializes from the top down taxonomy of data flow processing (Fung et al., 2005). Nevertheless, these two methods fit only traditional data, but do not function on large scale data with a lack of efficiency and scalability. With this as the base of these two techniques, it is categorized as parallel BUG, Hybrid BUG, TDS and Two way TDS, Mondrian TDS, etc.

Data Anonymity

Data anonymity is one or more techniques used to hide the information such as blanking, hashing or marking of sensitive data files making it impossible to identify any sensitive data. Privacy of this sensitive data can be confidential by upholding particular amassed information. Data anonymization is preferred for use on k-anonymity methods. However, anonymization is lacking on big data application datasets. In general, existing anonymization algorithms are four types such as simple anonymization, anonymization based on vertex clustering, anonymization based on edge clustering, and anonymization based on the bigraphy.