Contact no: 9008001602

Practical Privacy-Preserving MapReduce Based K-means clustering over Large-scale Dataset

Technofist,

YES Complex, 19/3&4, 2nd Floor, Dinnur Main Road, R.T.Nagar,Bangalore-560032

Ph:080-40969981, Website:www.technofist.com. E-mail:technofist.projects@gmail.com

ABSTRACT:

Clustering techniques have been widely adopted in many real world data analysis applications, such as customer behavior analysis, medical data Analysis, 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 HDFS 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 any Distributed servers inevitably raise privacy concerns. In this paper, we propose a practical privacy-preserving K-means clustering scheme that can be efficiently outsourced to HDFS servers.

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, 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. 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. For example, a promising trend for predicting an individual's disease risk is clustering over existing patients health records , which contain sensitive patient information. Therefore, appropriate privacy protection mechanisms must be placed when outsourcing sensitive datasets to the public cloud for clustering.