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

ABSTRACT

Clustering techniques have been widely adopted in many real world data analysis applications, such as customer behavior analysis, targeted marketing, digital forensics. , we propose a practical privacy-preserving K means clustering scheme that can be efficiently outsourced to cloud servers. Our scheme allows cloud servers to perform clustering directly over encrypted datasets, while achieving comparable computational complexity and accuracy compared with clustering over unencrypted ones. We also investigate secure integration of MapReduce into our scheme, which makes our scheme extremely suitable for cloud computing environment.

EXISTING SYSTEM

In existing 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. 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

DRAWBACKS

- It doesn't provide security.
- Threat to the personal information.

PROPOSED SYSTEM

In this paper, we propose a practical privacy-preserving Kmeans clustering scheme that can be efficiently outsourced to cloud servers. Our scheme allows cloud servers to perform clustering directly over encrypted datasets, while achieving comparable computational complexity and accuracy compared with clusterings over unencrypted ones. We also investigate secure

Integration of MapReduce into our scheme, which makes our scheme extremely suitable for cloud computing environment. Thorough security analysis and numerical analysis carry out the Performance of our scheme in terms of security and efficiency.

ADVANTAGES

- It provides security to the personal information.
- Performance is good on large scale datasets.
- The computational cost of the dataset owner shall be minimized,

SYSTEM REQUIREMENTS

H/W System Configuration:-

- Processor - Pentium –IV
- RAM - 4 GB (min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

S/W System Configuration:-

- Operating System : Linux
- Application Server : Tomcat5.0/6.X
- Backend coding : Java
- Tool : Virtual Box
- Environment : Ubuntu
- Technology : Hadoop