

**CHENNAI – PONDICHERRY**

**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, 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, 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. Experimental evaluation over a 5 million objects dataset further validates the practical performance of our scheme.