

**CHENNAI – PONDICHERRY**

**HEADS-JOIN:EFFICIENT EARTH MOVER'S DISTANCE SIMILARITY**

**JOINS ON HADOOP**

**ABSTRACT**

The Earth Mover's Distance (EMD) similarity join has a number of important applications such as near duplicate image retrieval and distributed based pattern analysis. However, the computational cost of EMD is super cubic and consequently the EMD similarity join operation is prohibitive for datasets of even medium size. We propose to employ the Hadoop platform to speed up the operation. Simply porting the state-of-the-art metric distance similarity join algorithms to Hadoop results in inefficiency because they involve excessive distance computations and are vulnerable to skewed data distributions. We propose a novel framework, named HEADS-JOIN, which transforms data into the space of EMD lower bounds and performs pruning and partitioning at a low cost because computing these EMD lower bounds has constant or linear complexity. We investigate both range and top-k joins, and design efficient algorithms on three popular Hadoop computation paradigms, i.e., MapReduce, Bulk Synchronous Parallel, and Spark. We conduct extensive experiments on both real and synthetic datasets. The results show that HEADS-JOIN outperforms the state-of-the-art metric similarity join technique, i.e., Quickjoin, by up to an order of magnitude and scales out well.