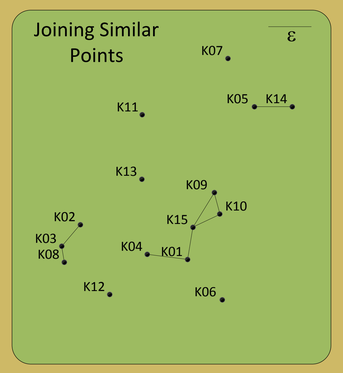
Cloud Similarity Join for Multi-Dimensional Data

Faculty: Yasin N. Silva Student: Jason M. Reed
New College of Interdisciplinary Arts & Sciences,
Arizona State University

Similarity Joins & Cloud-Based Systems

Similarity Joins:

Similarity Joins in metric space join elements that within a specified distance of each other. These elements can be high dimensional and are often represented in vector form.
Similarity Joins are used in:
- Data mining
- Data cleansing

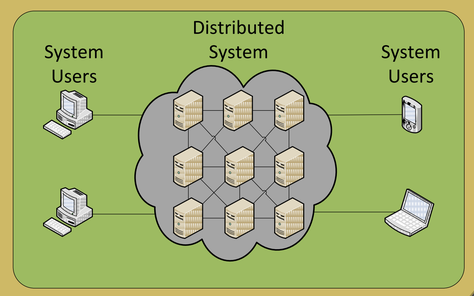


Cloud-Based Systems:

Cloud-based systems are a large number of commodity computer systems joined together through a network to perform shared tasks.

Cloud systems are used to process massive amounts of data and make us of the MapReduce programming interface which divides a large job into many smaller subtasks.

Cloud systems are used in scientific projects and internet companies like Google and Facebook.



Proposal and Challenges

Proposal:

We propose to implement an efficient similarity join process for multidimensional data onto the Hadoop Map Reduce framework

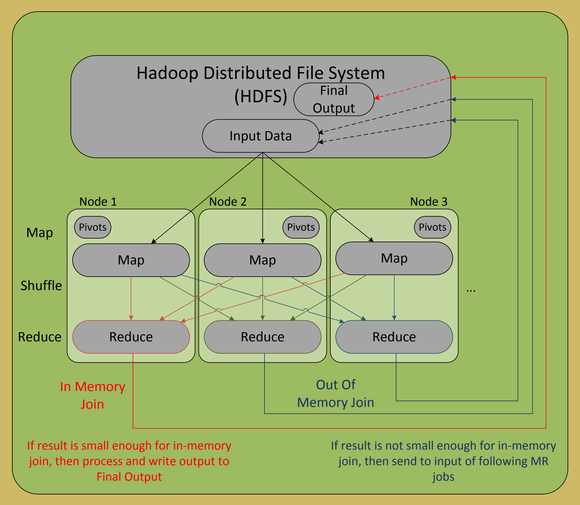
Goal:

The primary goal of this project is to provide useful and meaningful similarity join operations over a cloud-based system. The operators will support multi-dimensional data.

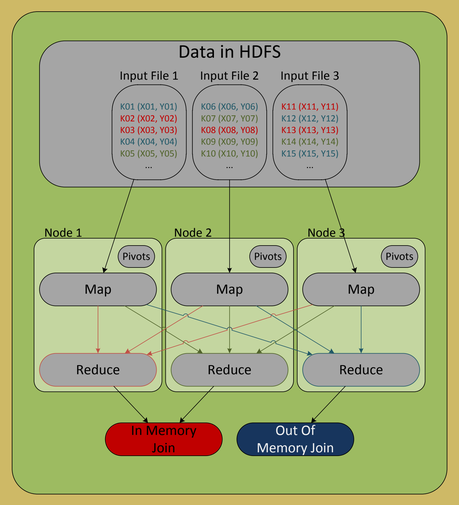
Challenges:

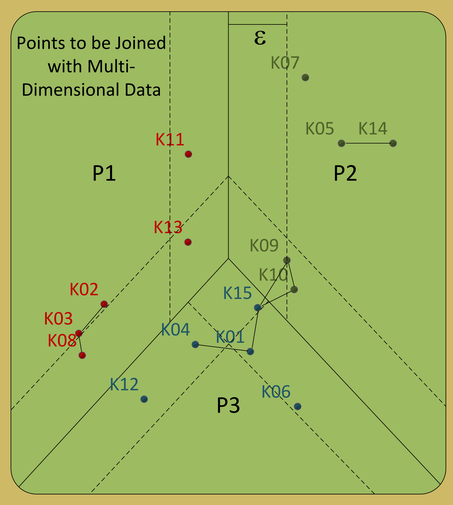
1) MapReduce is not iterative
2) How to partition Data
3) Ensuring proper data grouping

General Solution



Details of Single Iteration





Multiple Iterations