Data clustering is an inherent problem in data analysis. It is the principle work horse of many more targeted data analysis and retrieval methods, has been studied extensively in computing and mathematics. As the ever changing computing landscape moves forward, multi-processor and distributed computing is rapidly overtaking sequential computing prompting a fundamental restructuring in algorithmic design. In this proposal we present a novel method for very large scale distributed approximate data clustering. Random Projection Hash(*RPHash*) was expressly created for minimizing parallel communication overhead while providing algorithmic scalability. In the proposal we discuss the shortcomings of various other parallel methods. Though some have been converted to function efficiently on parallel and distributed systems, they often have potential issues in asymptotic scalability or are limited in application space. We propose our RPHash algorithm as a solution to these two problems.

The principle concept behind *RPHash* clustering is the trade-off of individual core redundancy and approximation to decouple the data processing streams over a many core distributed system. Stream synchronization is provided probabilistically by performing redundant multi-probe random projections into a mathematically generative object. The Leech Lattice object provides a high dimensional space quantizer, with universally applicable region IDs. At the system level, *RPHash*'s sequential bottleneck is no worse than the well know, highly scalable, parallel logarithmic reduction.

Our preliminary objective is to develop a sequential version of the proposed *RPHash* algorithm. Though the sequential version will likely not exceed any current clustering algorithms in any way, it will be an important step in the development of the parallel RPHash algorithm. The sequential algorithm will provide initial validity to our claims through a reasonable comparison in regards to overall accuracy with other clustering methods. The sequential algorithm will also give insight into the differences introduced when RPHash is scaled and deployed on a parallel system.

The principle objective of proposal is to develop the parallel RPHash algorithm. This step will require us to explore a variety of variants for different steps of the algorithm. In general, we will favor empirical evidence to discover the set of variants that provide the best scalability. Though some theoretical work will be provided, the wide application space of our algorithm favors such empirical methods. Furthermore, we will deploy the system on the Hadoop Map Reduce (*MR*v2) core framework. The Hadoop framework will allow us to scale our algorithm using commercially available distributed computing resources.

The redundancy introduced in the multi-probe projection step will adversely effect the sequential asymptotic complexity of the algorithm, making it less efficient than most sequential clustering algorithms. For this reason, the focus of this proposal will not be on theoretical sequential efficiency, or even the theoretical complexity of parallel variants of clustering algorithms, rather it will be on real world empirical tests with both real and synthetic data, on real distributed systems. The purpose of our emphasis on real world testing is to encompass all aspects of distributed systems without requirements on exotic networking and shared memory architecture systems running under ideal conditions. We will test *RPHash* on both synthetic and real world data. The synthetic data will provide parameter sampling rates well above those found in real world data, such as dimensionality and number for clusters. Furthermore, noise and density rates can be controlled to better characterize the strengths and weaknesses of the algorithm.

An additional feature of the RPHash algorithm is its intrinsic data transmission security. Due to the random projection's destructive effect on vector data, RPHash adds a reasonable amount of data anonymization. Furthermore, inter node communication consists only of Leech Lattice region IDs and aggregate centroids. In this proposal we will demonstrate the decoupling of original vector data from the communicated and aggregate centroid data through experimental analysis of simulated and real world data.