# CompSci-230: Project Proposal

Member: Jitao Zhang, Fuyao Li, Zhuang Liu

Topics: MapReduce
(Study the distributed solution for a particular application, algorithm, or system software.)

Description:
In the past years, many Google's engineers try to implement hundreds of calculation methods to handle massive amounts of raw data. Most of these data processing operations are conceptually easy to understand. However, due to the huge amount of data input, in order to complete the calculation in an acceptable time, the only way is to put these calculations over hundreds of distributed computers. How to handle parallel computing, how to distribute data, and how to handle errors? All of these problems are combined and require a lot of code processing, which makes the original simple operation more difficult to handle.

MapReduce is a programming model and an implementation of an algorithmic model for processing and generating very large data sets. The user first creates a Map function to process a data set based on the key/value pair, and outputs a data set based on the key/value pair in the middle; then creates a Reduce function to merge all intermediate values with the same intermediate key value. There are many examples in the real world that satisfy the above processing model.

Programs in the MapReduce architecture can be parallelized on a large number of commonly configured computers. In the runtime, this system only cares: how to split the input data, schedule on a cluster of large numbers of computers, error handling of computers in the cluster, and manage the necessary communication between computers in the cluster.

Reference:
Dean J , Ghemawat S . MapReduce: Simplified Data Processing on Large Clusters[C]// Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6. USENIX Association, 2004.