Week9 Summary

General idea:

MapReduce is a programming model designed to process and generate large data sets. Users can realize such function by “map” and “reduce”. “Map” is used to partite the original large data sets into smaller clusters. “Reduce” means merging the results of these smaller clusters into the final result. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The programmers can utilize the resources of a large distributed systems with any experience on distributed systems.

Details about the paper:

In the first part, the author gave a introduction to the MapReduce system. The abstraction of MapReduce is inspired by the map and reduce primitives present in Lisp and many other functional languages. The author then gave a simple example of how to use MapReduce model to count words in a large document. More examples like Distributed Grep, Reversing Web-Link Graph is also mentioned.

In the second part, the author talked about the exact process of how MapReduce is realized in a distributed computer system. More detailed information about the master data structures, fault tolerance, master failure, semantics in the presence of failures, locality, task granularity decision and backup task implementation.

In the third part, the author talked about some useful extensions for the MapReduce system to enhance its performance. The techniques include partition function, ordering guarantees, combiner function, input and output types, side-effects handling, skipping bad records, local execution, status information, counter, performance, cluster configuration, grep, sort, effect of backup task, machine failures, experience, large-scale indexing. The author also provided some detailed analysis of the sorting performance based test data.

In the last part, the author summed up the functional parts of MapReduce and talked about the wide-spread applications of MapReduce in different Google products and other places.

My Personal Thoughts:

The MapReduce model is useful to handle large size data. It just parses it into smaller tasks and merge the smaller scale results into the final result. This model designed by Google just realize a something like a black box and we can realize our functions based on this model and detailed task scheduling is handled by MapReduce itself. I am not so familiar with this model before and it is a technology with increasingly popularity. I think I should do some real practice to enhance my understanding of this model.