Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Feb 5th. 2019

Please note that assignments in this class may be submitted to

www.turnitin.com, a web-based anti-plagiarism system, for an evaluation of their originality.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Reading review**

**MapReduce: Simplified Data processing on Large Clusters**

MapReduce is a programming model that processes and generates large data sets. This type of computing is highly scalable and can handle data on thousands of computers. And for programmers, this system is easy to use.

In order to complete the processing of a large amount of raw data in a reasonable time, the calculation needs to be distributed on thousands of computers. The main contribution of MapReduce is simple and powerful interface that enables automatic parallelization and distribution of large-scale computing to achieve high performance on PC clusters.

The MapReduce function includes 2 functions: Map and Reduce, which are written by the user. The Map generates a set of intermediate values based on the input, and the MapReduce library combines all intermediate values associated with this intermediate value and passes them to the Reduce function. This feature allows us to handle lists of values that are too large for memory.

MapReduce interfaces can have many different implementations. It should be chosen according to the environment. And because MapReduce helps to process large amount of data by using thousands of computers, it has the ability to tolerate errors.

The MapReduce library can handle different kind of problems and can be widely used. It is used in almost all areas of Google. One of the most important uses of MapReduce is to completely rewrite the production indexing system, which generates data structures for the Google Web Search service. There are several advantages of MapReduce: the index code is simple, small, easy to understand, fault-tolerant, good performance, and easy to operate.

MapReduce has been successfully used for many purposes in Google is because: the model is easy to use, hiding the details of parallelization, fault tolerance, local optimization and load balancing. It is easy to express a variety of problems. It can be extended to large machine clusters, effectively utilizing machine resources.

Through this paper, I learned that the restricted programming model can easily parallelize and distribute calculations and make these calculations fault-tolerant. Network bandwidth is a scarce resource. Redundant execution can be used to reduce the impact of slow machines and handle machine failures and data loss.