Term Project Proposal Kewen Wang

Topic: MapReduce Architecture and Programming

MapReudce is a distributed computing programming framework, which is designed to solve data analysis problems. As the open source implementation of MapReduce, Hadoop is widely used in practice for its high efficiency in data processing and low requirement in cluster envrionment. To implement user application in MapReduce, user need write Map and Reduce programs for specified task. But it is complicated to write programs in MapReduce for its distributed environment feature. Recently, some tools like Pig and Hive have provided methods to automatically translate SQL languages into MapReduce programs. However, these auto-generated MapReduce programs are inefficient for some tasks compared to optimized MapReduce programs by experienced coders. Moreover, it is not so easy to implement complex algorithm like graph algorithm in MapReduce. Thus designing efficient algorithms and programs in MapReduce will be a great challenge.

In this survey, I will analyze the underlying architecture of MapReduce framework, and how this programming model could improve data processing capability. Especially, I may focus on the parallel computing architecture of MapReduce framework. Besides, I will find the papers about how to translate SQL language to MapReduce program automatically or manually. And I will search for some disciplines or experience of designing efficient programs in MapReduce for certain kind of algorithms. Moreover, based on these methods, I will summarize some guidelines for designing efficient programs in MapReduce.