**学习日志**

姓名：陈雨舒 学号：1813065 日期：6月25日

**学习内容安排：**

上午：MapReduce工作原理

下午：MapReduce的类型与格式

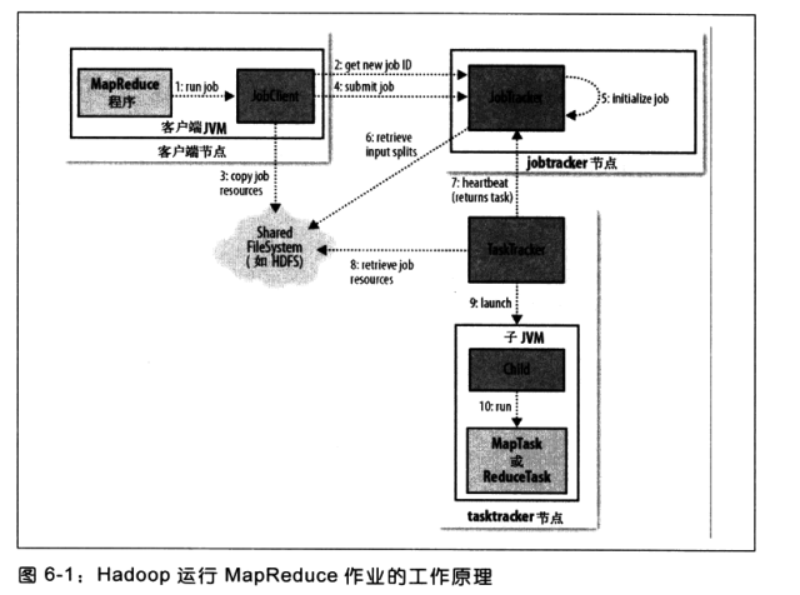
**学习反馈：**

MapReduce工作原理：

将复杂的，运行大规模集群上的并行计算过程高度地抽象两个函数: Map和Reduce。

采用"分而治之”策略,将一个分布式文件系统中的大规模数据集，分成许多独立的分片。这些分片可以被多个Map任务并行处理。

计算向数据靠拢”，而不是数据向计算靠拢。



Hadoop为map和reduce任务提供运行环境的信息。例如，map任务可以知道它处

理的文件的名称，map 和reduce 任务可以得知任务尝试次数。



作业运行流程：

输入分片：根据输入文件计算输入分片（输入分片存储的是一个分片长度和一个记录数据的

位置的数组）

map阶段：在数据存储节点上进行

combiner阶段：在map计算出中间文件前做一个简单的合并重复key值的操作

shuffle阶段：将map的输出作为reduce的输入的过程

reduce阶段：产生输出存储在hdfs上

MapReduce的类型与格式：

输入格式：

分片记录：JobClient生成数据分片InputSplit；InputFormat接口负责生成分片

文件输入：FileInputFormat

CombineFileInputFormat

文本输入：TextInputFormat key存储偏移量, value存储数据内容

KeyValueInputFormat key为行号

NLineInputFormat 设置每个mapper处理的行数

二进制输入：SequenceFileInputFormat

多文件输入：Multiplelnputs

数据库格式输入：DBInputFormat

输出格式：

文本输出：TextOutputFormat key是LongWritable类型的,value是Text类型的

二进制输出：SequenceFileOutputFormat

多文件输出：MultipleOutputFormat

MultipleOutputs

数据库输出：DBOutputFormat

学习心得：

初步了解MapReduce工作原理和MapReduce的类型与格式，了解了输入和输出的格式，应加强实践练习，对今后的MapReduce应用开发有所帮助。