Homework

OpenMP编程

Name:Li Shibo Student ID:119033910046 Email: ShiboLi@sjtu.edu.cn

MapReduce

Hadoop简介 Hadoop是一个实现了Google云计算系统的开源系统，包括并行计算模型Map/Reduce、分布式文件系统HDFS，以及分布式数据库Hbase，同时Hadoop的相关项目也很丰富，包括ZooKeeper，Pig，Chukwa，Hive，Hbase，Mahout，flume等。本次实验使用Hadoop实现温度统计任务。

MapReduce简介 MapReduce是hadoop的核心组件之一，hadoop要实现分布式需要包括两部分，一部分是分布式文件系统hdfs，一部分是分布式计算框架mapreduce，缺一不可，也就是说，可以通过mapreduce很容易在hadoop平台上进行分布式的计算编程。

MapReduce是一种实现多个节点并行处理事务的编程方式，节点分为Master和Worker，Master负责任务的调度，Worker负责完成任务。Woeker分为两种，一种Worker叫做Mapper，另一种Worker叫Reduc对于一个巨大的数据集，里面有海量的元素，每个元素都需要进行同一个函数处理。于是Master将这些元素分成许多小份，然后每一份分给Mapper来做，Mapper执行完函数，讲结果传给Reducer，Reducer统计汇总各个Mapper传过来的结果，得到最后任务的答案。具体过程如下图所示。

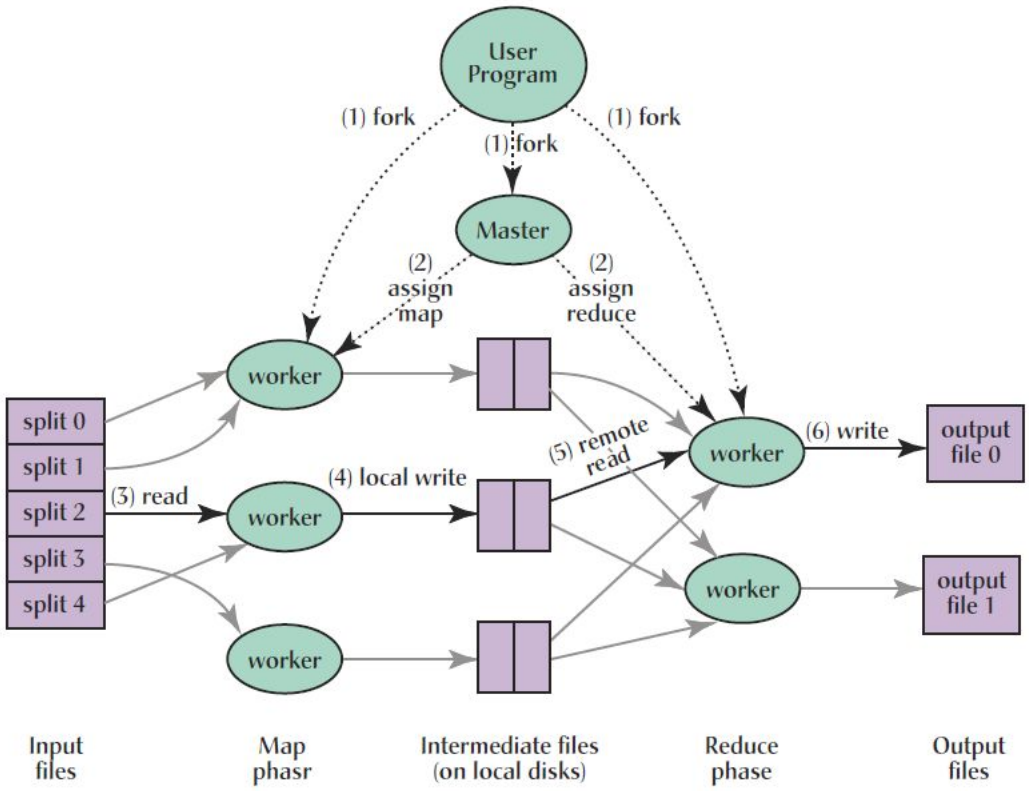

图 1: mapreduce的原理。

实验过程 MapReduce 任务过程分为两个处理阶段：map 阶段和reduce阶段。每个阶段都以键值对作为输入和输出，其类型由程序员自己来选择。程序员还需要写两个函数：map 函数和reduce 函数。在本次实验中map阶段的输入是统计的温度数据，以文本格式作为输入，将数据集的每一行作为文本输入，只需提取出日期和温度信息。通过reduce函数计算每一天的最高和最低气温。部分实验结果如下图所示。

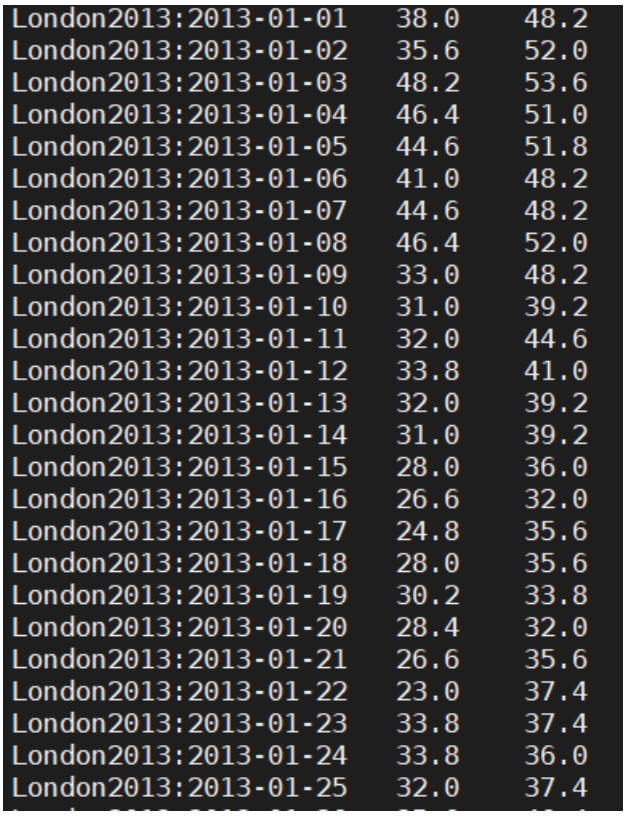

图 2: mapreduce的部分实验结果。

实验总结 MapReduce将很简单的运算逻辑很方便的扩展到海量数据的场景下分布式运算，将很多相同的实现部分封装起来，使得实现海量数据的并行处理变得非常简单。在实现时只需要写业务逻辑来处理数据，使得开发变得快捷方便。MapReduce是比较好的工具和编程手段。