暑期实训学习报告

1813012　柳鑫

一．学习内容

① Hadoop 的 I/O

② MapReduce 应用开发

二．学习结果

① Hadoop 的 I/O

（1）数据完整性

HDFS 会对写入的所有数据计算校验和，并在读取数据时验证校验和。datanode 负责在收到数据后存储该数据及其校验和之前对数据进行验证。datanode 的管线中最后一个 datanode 执行校验。HDFS 存储着每一个数据块的复本，因此可以通过数据复本来修复锁坏的数据块。可以用 hadoop 的命令 fs -checksum 来检查一个文件的校验和，可用于检查 HDFS 中两个文件是否具有相同的内容。

（2）文件压缩

通用的压缩工具是 gzip，是否可切分列表示对应的压缩算法是否支持切分，也就是说，是否可以搜索数据流的任意位置并进一步往下读取数据。可切分压缩格式尤其适合 MapReduce。

（3）序列化

将结构化对象转化为字节流以便在网络上传输或写到磁盘进行永久存储的过程，反序列化是指将字节流转回结构化对象的逆过程。序列化用于分布式数据处理的两大领域：进程间通信和永久存储。

Hadoop 使用的是自己的序列化格式 Writable，它紧凑、速度快，但不太容易被 Java 以外的语言进行扩展和使用，因为 Writable 是 Hadoop 的核心(大多数 MapReduce 程序都会为键和值类型使用它)。

② MapReduce 应用开发

（1）MapReduce 编程流程：

首先写 map 函数和 reduce 函数，使用单元测试确保函数的运行符合预期，然后写一个驱动程序来运行作业（可在本地 IDE 中用一个小数据集进行测试），最后将通过测试的程序放到集群上运行。

（2）资源文件：

configuration.xml、core-default.xml、core-site.xml。

（3）MapReduce 的工作流：

JobControl 的实例表示一个作业的运行图，可以加入作业的配置，告知 JobControl 实例作业之间的依赖关系。在一个线程中 JobControl 将按照依赖顺序执行这些作业。如果一个作业失败，JobControl 将不执行与之有依赖关系的后续作业（JobControl 在客户端运行并提交作业）。

（4）Apache Oozie：

Oozie 作为服务器运行，客户端提交一个立即或稍后执行的工作流定义到服务器。在 Oozie 中，工作流是一个由动作（action）节点和控制流节点组成的有向无环图。动作节点执行工作流任务，控制流节点通过构建条件逻辑或并行执行来管理活动之间的工作流执行情况。当工作流结束时，Oozie 通过发送一个 HTTP 的回调向客户端通知工作流的状态（Oozie 工作流）。