学习日志

姓名：辛昊洋　　学号：1813090　　日期：6.24

学习内容安排：

上午：
Hadoop 的 I/O
下午：
MapReduce 应用开发

已掌握知识：
I/O：

1. 数据完整性：任何语言对 IO 的操作都要保持其数据的完整性。hadoop 当然希望数据在存储和处理中不会丢失或损坏。检查数据完整性的常用方法是校验和。
HDFS 的数据完整性：客户端在写或者读取 HDFS 的文件时，都会对其进行校验和验证，当然我们可以通过在 Open（）方法读取之前，将 false 传给 FileSystem 中的 setVerifyCheckSum() 来禁用校验和。本地文件系统，hadoop 的本地文件系统执行客户端校验，这意味着，在写一个 filename 文件时，文件系统的客户端以透明方式创建了一个隐藏的文件.filename.crc,块的大小做为元数据存于此，所以读取文件时会进行校验和验证。ChecksumFileSystem：可以通过它对其数据验证。
2. 压缩：压缩后能够节省空间和减少网络中的传输。所以在 hadoop 中压缩是非常重要的。
hadoop 的压缩格式

压缩格式	算法	文件扩展名	多文件	可分割性
DEFLATEa	DEFLATE	.deflate	no	no
gzip（zip）	DEFLATE	.gz(.zip)	no(yes)	no(yes)
bzip2	bzip2	.bz2	no	yes
LZO	LZO	.lzo	no	no

3. 序列化：将字节流和机构化对象的转化。hadoop 是进程间通信（RPC 调用），PRC 序列号结构特点：紧凑，快速，可扩展，互操作，hadoop 使用自己的序列化格式 Writerable。

Mapreduce 应用开发：
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。
MapReduce 的工作机制：可通过一个简单的方法调用来运行 MapReduce 作业：Job 对象上的 submit()。也可调用 waitForCompletion()，它用于提交以前没有提交过的作业，并等待它的完成。submit()方法调用封装了大量的处理细节。Hadoop 2.0 引入了一种新的执行机制，这种机制建立在一个名为 YARN 的系统上。通过设置 mapreduce.framework.name 属性值选择执行的框架。local 表示本地的作业运行器，classic 表示经典的 MapReduce 框架（使用一个

jobtracker 和多个 tasktracker），yarn 表示新的框架。不同的执行框架表示运行 MapReduce 程序的不同途径。

学习心得：最近几天停留在理论上面，搭建出了很多问题。还需要更多操作才可以熟练掌握。