**学习日志**

姓名：陈雨舒 学号：1813065 日期：6月23日

**学习内容安排：**

了解MapReduce以及Hadoop分布式文件系统（HDFS）

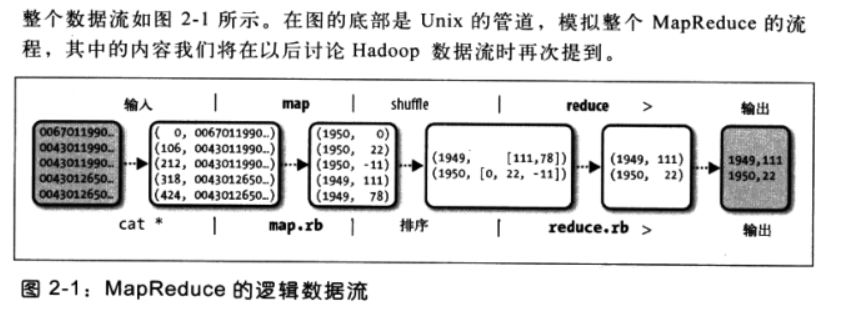
**学习反馈：**

已掌握知识：

MapReduce：

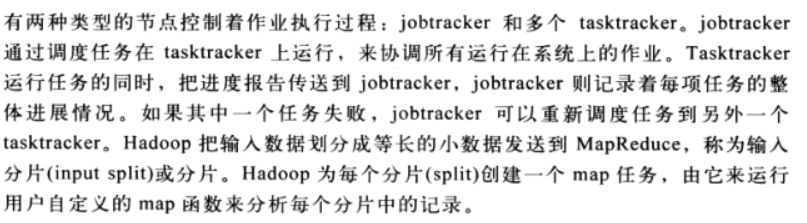
MapReduce的工作过程分为两个阶段：map和reduce

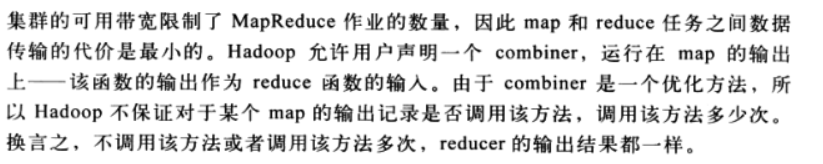
在map阶段输入原始的NCDC数据，选择文本输入格式



数据流：

Hadoop通过把作业分为若干个小任务来工作，包括map任务和reduce任务

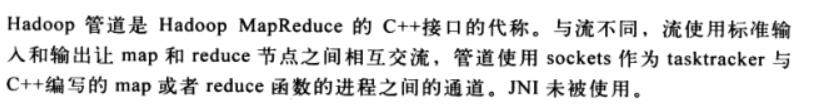




combiner是用reducer接口定义的，应用程序的实现与MaxTemperatureReducer中的函数相同。

Hadoop流使用Unix标准作为Hadoop和程序之间的接口，所以可以用任何语言。

Hadoop管道



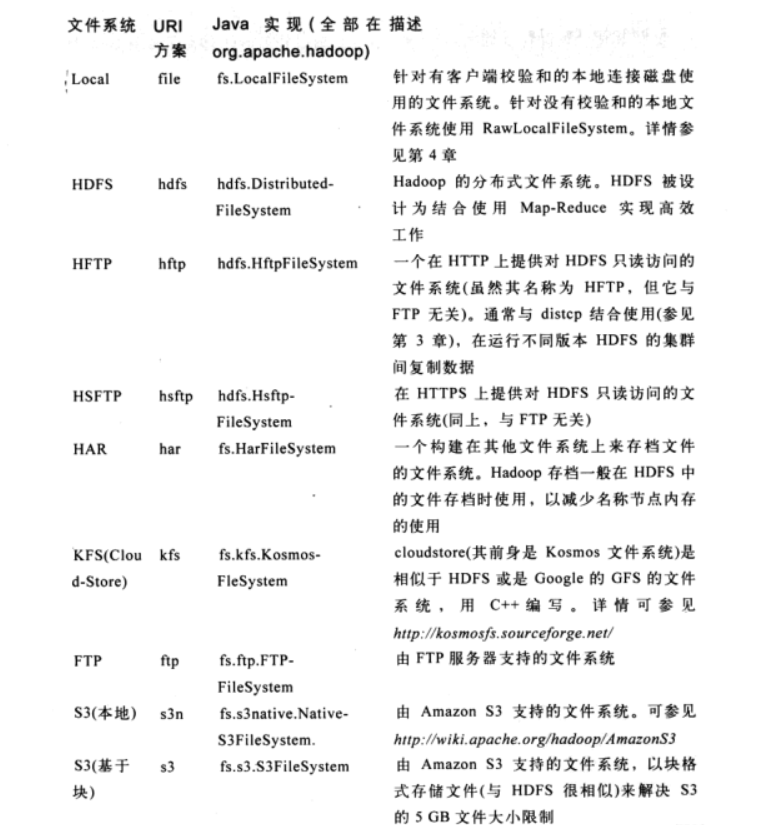
Hadoop分布式文件系统：

Blocks：

物理磁盘中有块的概念，磁盘的物理Block是磁盘操作最小的单元，读写操作均以Block为最小单元，一般为512 Byte。文件系统在物理Block之上抽象了另一层概念，文件系统Block物理磁盘Block的整数倍。通常为几KB。Hadoop提供的df、fsck这类运维工具都是在文件系统的Block级别上进行操作。

HDFS的Block块比一般单机文件系统大得多，默认为128M。HDFS的文件被拆分成block-sized的chunk，chunk作为独立单元存储。比Block小的文件不会占用整个Block，只会占据实际大小。例如， 如果一个文件大小为1M，则在HDFS中只会占用1M的空间，而不是128M。

Hadoop文件系统：



命令行接口：

HDFS提供了各种交互方式，例如通过Java API、HTTP、shell命令行的。命令行的交互主要通过hadoop fs来操作。例如：

hadoop fs -copyFromLocal // 从本地复制文件到HDFS

hadoop fs mkdir // 创建目录

hadoop fs -ls // 列出文件列表

Hadoop中，文件和目录的权限类似于POSIX模型，包括读、写、执行3种权限：

读权限（r）：用于读取文件或者列出目录中的内容

写权限（w）：对于文件，就是文件的写权限。目录的写权限指在该目录下创建或者删除文件（目录）的权限。

执行权限（x）：文件没有所谓的执行权限，被忽略。对于目录，执行权限用于访问器目录下的内容。

每个文件或目录都有owner，group，mode三个属性，owner指文件的所有者，group为权限组。

Java接口：

从Hadoop URL读取数据

java.net.URL类提供了资源定位的统一抽象，任何人都可以自己定义一种URL Schema，并提供相应的处理类来进行实际的操作。hdfs schema便是这样的一种实现。

InputStream in = null;

try {

in = new URL("hdfs://master/user/hadoop").openStream();

}finally{

IOUtils.closeStream(in);

}

写数据：

在HDFS中，文件使用FileSystem类的create方法及其重载形式来创建，create方法返回一个输出流FSDataOutputStream，可以调用返回输出流的getPos方法查看当前文件的位移，但是不能进行seek操作，HDFS仅支持追加操作。

创建时，可以传递一个回调接口Peofressable，获取进度信息。

append(Path f)方法用于追加内容到已有文件，但是并不是所有的实现都提供该方法，例如Amazon的文件实现就没有提供追加功能。

目录操作：

使用mkdirs（）方法,会自动创建没有的上级目录

HDFS中元数据封装在FileStatus类中，包括长度、block size，replicaions，修改时间、所有者、权限等信息。使用FileSystem提供的getFileStatus方法获取FileStatus。exists()方法判断文件或者目录是否存在；

列出文件（list），则使用listStatus方法，可以查看文件或者目录的信息

public abstract FileStatus[] listStatus(Path f) throws FileNotFoundException,

IOException;

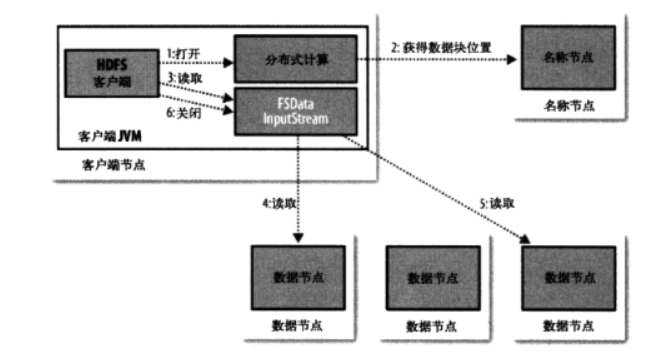
删除数据：

使用FileSystem的delete()方法

public boolean delete(Path f , boolean recursive) throws IOException;

recursive参数在f是个文件的时候被忽略。如果f是文件并且recursice为true，则删除整个目录，否则抛出异常。

数据流：



未掌握知识：

初步了解MapReduce和HDFS，对一些语句的实际应用还不是很熟练。

学习心得：

初步了解了Hadoop的MapReduce和HDFS，对接下来的学习有了很好的帮助。