13-mapreduce的原理-及【mapreduce例子单词统计测试】

----------成都尚学堂-mr-zeng----------

mapreduce概念

分布式计算框架，hadoop的分布式计算的模型，计算需要用到数据和计算程序，这里尽量移动计算而不移动数据。

mapreduce不适合的计算

a）要计算数据量太小 ：GB以下。

b）要计算的复杂程序：不能分割为多个同时执行的程序

1）修房子的步骤-》不可同时执行

x）打地基 x）搭框架 x）码砖头

2）码砖头的步骤-》可以同时执行

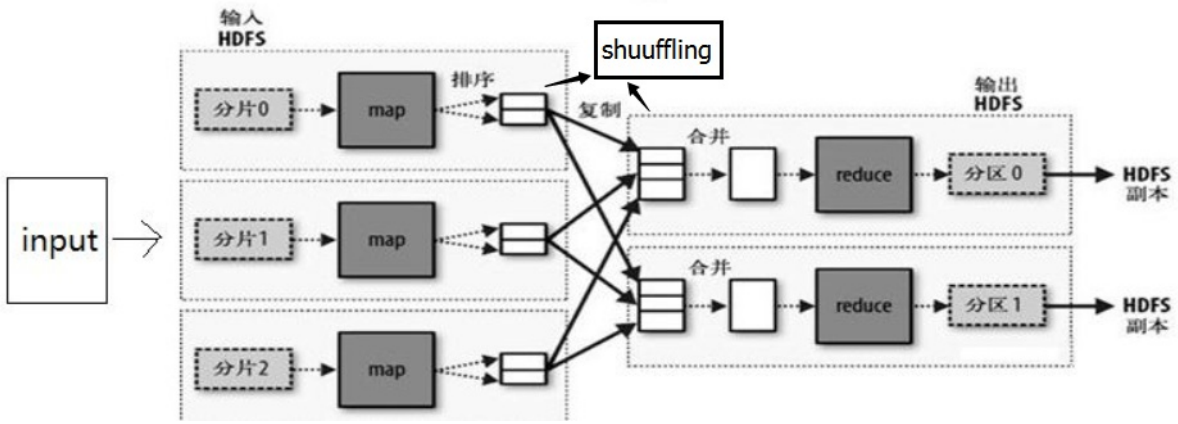
x）刘德华-码砖头 x）志玲姐姐-码砖头 x）柳岩-码砖头

mapreduce的原理

map：分解数据，类似于剁肉

reduce：合并数据，类似于包饺子

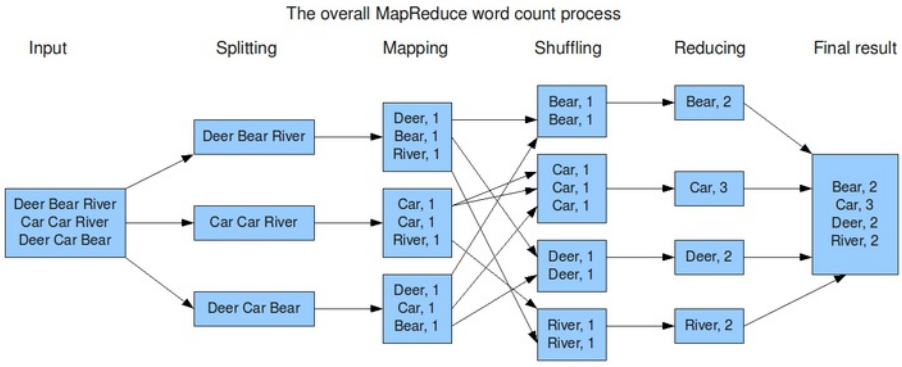
运算流程图（每一个map和reduce都在一个单独节点运行）



运算流程-通俗理解

-》相当于有一个很大的图书馆，有很多书架（分片），书架上很多书（语文，数学，英语---地理，政治，历史））要统计（语文，数学，英语-）数量，每一个书架（分片）交给一个管理员A（map-进程）统计。书架分了（很多列），管理员一次统计一个列里的数据（map方法调用一次），统计之后记录在本子上，另一个管理员B（shuffling）把map（管理员）的本子的数据，进行排序分组，再交给管理员C（reduce-进程），管理员C分别把每一组进行统计（reduce方法调用一次）。

单词统计-模拟流程图（这是一个map一个reduce）



各阶段工作（map和reduce必须编码，其他阶段可以自动进行-如果排序方式，分组方式不满足要求-可以自定义！！！）

input阶段：从hdfs读取需要处理的文件

splitting阶段：把文件按行分割为多个数据-以（key：当前行开头字节数位置，value：行内容）传入给map。

mappping阶段：把value行内容分割为多个单词，输出（key：单词，value：次数）给shuffling

shuffling阶段：把map的数据按key进行升序排列，把相同key合并分到同一个组，输出（key：单词 value：（次数集合：次数1，次数2...））给reduce

reduce阶段：对相同key单词的数据进行（里所有的value次数进行累加）输出到hdfs系统

mapreduce例子单词统计测试

a) 编写待统计的单词文件

cd /usr/hadoop/hadoop-2.7.2/share/hadoop/mapreduce/

vi words.txt
内容如下
hadoop hive pig 周杰伦
hadoop pig 林志玲
hadoop hive pig 周杰伦1
hadoop hive pig 周杰伦 周杰伦 柳岩 柳岩
hadoop pig

b) 通过命令-运行单词统计程序
ps-》如果需要计算多次每次的输出路径不能一致

hadoop jar hadoop-mapreduce-examples-2.7.2.jar wordcount 输入文件路径 输出文件夹路径
输入文件路径：可以是本地 (file:/xxx) 或hdfs((hdfs://namenode域名:9000/xxx))
输出文件夹路径：可以是本地或hdfs

hadoop jar hadoop-mapreduce-examples-2.7.2.jar wordcount
file:/usr/hadoop/hadoop-2.7.2/share/hadoop/mapreduce/words.txt
file:/usr/hadoop/hadoop-2.7.2/share/hadoop/mapreduce/wordsOut

c) 查看输出文件内容
-》cd 切换到输出路径 cat 查看

----------hadoop命令上传下载hdfs文件----------
hadoop fs -put ./words.txt /words.txt
hadoop fs -get /wordsout /wordsout