海量文件的分布式计算处理方案计算机等级考试 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/646/2021_2022__E6_B5_B7_E9_87_8F_E6_96_87_E4_c97_646078.htm Hadoop 是Google MapReduce的一个Java实现。MapReduce是一种简化的分布式编程模式，让程序自动分布到一个由普通机器组成的超大集群上并发执行。就如同java程序员可以不考虑内存泄露一样，MapReduce的run-time系统会解决输入数据的分布细节，跨越机器集群的程序执行调度，处理机器的失效，并且管理机器之间的通讯请求。这样的模式允许程序员可以不需要有什么并发处理或者分布式系统的经验，就可以处理超大的分布式系统得资源。 一、概论 作为Hadoop程序员，他要做的事情就是： 1、定义Mapper，处理输入的Key-Value对，输出中间结果。 2、定义Reducer，可选，对中间结果进行规约，输出最终结果。 3、定义InputFormat 和OutputFormat，可选，InputFormat将每行输入文件的内容转换为Java类供Mapper函数使用，不定义时默认为String。 4、定义main函数，在里面定义一个Job并运行它。 然后的事情就交给系统了。 1.基本概念：Hadoop的HDFS实现了google的GFS文件系统，NameNode作为文件系统的负责调度运行在 master，DataNode运行在每个机器上。同时Hadoop实现了Google的MapReduce，JobTracker作为 MapReduce的总调度运行在master，TaskTracker则运行在每个机器上执行Task。 2.main()函数，创建JobConf，定义Mapper，Reducer，Input/OutputFormat 和输入输出文件目录，最后把Job提交给JobTracker，等待Job结束。 3.JobTracker，创建一

个InputFormat的实例，调用它的getSplits()方法，把输入目录的文件拆分成FileSplist作为Mapper task 的输入，生成Mapper task加入Queue。 4.TaskTracker 向 JobTracker索求下一个Map/Reduce。 Mapper Task先从InputFormat创建RecordReader，循环读入FileSplits的内容生成Key与Value，传给Mapper函数，处理完后中间结果写成SequenceFile. Reducer Task 从运行Mapper的TaskTracker的Jetty上使用http协议获取所需的中间内容（33%），Sort/Merge后（66%），执行Reducer函数，最后按照OutputFormat写入结果目录。 TaskTracker 每10秒向JobTracker报告一次运行情况，每完成一个Task10秒后，就会向JobTracker索求下一个Task。 Nutch项目的全部数据处理都构建在Hadoop之上，详见Scalable Computing with Hadoop。 二、程序员编写的代码 我们做一个简单的分布式的Grep，简单对输入文件进行逐行的正则匹配，如果符合就将该行打印到输出文件。因为是简单的全部输出，所以我们只要写Mapper函数，不用写Reducer函数，也不用定义Input/Output Format。 package demo.hadoop public