大数据开发面试题下载

面试——走进企业的唯一途径，无论是知名企业还是一般的小企业，都是要经过面试官层层把控的，尤其是对于大数据技术人才的招聘，更是要好几轮，要面对不同级别的面试官。

想要从众多的应聘者中脱颖而出还是有一定的难度的。那话又说回来，怎么才能赢得面试官的青睐？怎么样才能拿下心仪的工作？不怕，千锋小编来帮忙，千锋独家秘制的大数据开发面试题免费下载，让你笑傲江湖。

1.hadoop 运行原理

包括 HDFS 和 Mapreduce 两部分。

1)HDFS 自动保存多个副本 ,移动计算。缺点是小文件存取占用 namenode 内存，写入只支持追加，不能随机修改。

它存储的逻辑空间称为 block，文件的权限类似 linux。整体架构分三种节点，NN,SNN,DN

NN 负责读写操作保存 metadata(Ownership Permission blockinfo)

SNN 负责辅助 NN 合并 fsimage 和 edits，减少 nn 启动时间

DN 负责存数据，每个数据（文件）分割成若干 block，每个 block 默认 3

个副本。启动后像 NN 发送心跳保持联系

NN 保存的 metadata 在 hdfs 启动后加载到计算机内存，除 block 位置信息的 metadata 保存在 OS 文件系统中的 fsimage 文件中，对 metadata 的操作日志保存在 OS 文件系统中的 edits 文件中。block 位置信息是 hdfs 启动后由 DN 上报 NN 再加载到内存的。

HDFS 的安全模式：直到 NN 完全加载完 metadata 之前的这段时间。期间不能写入文件，DN 检查各个 block 完整性，并修复。

2）MapReduce

离线计算框架，过程分为 split map shuffle reduce 四个过程

架构节点有：Jobtracker TaskTracker

Split 将文件分割，传输到 mapper，mapper 接收 KV 形式的数据，经过处理，再传到 shuffle 过程。

Shuffle 先进行 HashPartition 或者自定义的 partition，会有数据倾斜和 reduce 的负载均衡问题；再进行排序，默认按字典排序；为减少 mapper 输出数据，再根据 key 进行合并，相同 key 的数据 value 会被合并；末后分组形成

（key,value{}）形式的数据，输出到下一阶段

Reduce 输入的数据就变成了，key+迭代器形式的数据，再进行处理

2.MapReduce 原理

逻辑上：

1、split

2、map

3、shuffle

4、reduce

四个过程

物理上：

JobTracker 节点 :JobTracker 创建每一个 Task(即 MapTask 和 ReduceTask ）并将它们分发到各个 TaskTracker 服务中去执行。负责调度 Job 的每一个子任务 task 运行于 TaskTracker 上。

TaskTracker 节点：运行在多个节点上的 slaver 服务。TaskTracker 主动与 JobTracker 通信，接收作业，并负责直接执行每一个任务。TaskTracker 都需要运行在 HDFS 的 DataNode 上

3.hdfs 存储机制

1） client 端发送写文件请求，namenode 检查文件是否存在，如果已存在，直接返回错误信息，否则，发送给 client 一些可用 namenode 节点

2）client 将文件分块 ,并行存储到不同节点上 datanode 上 ,发送完成后，client 同时发送信息给 namenode 和 datanode

3） namenode 收到的 client 信息后，发送确信信息给 datanode

4） datanode 同时收到 namenode 和 datanode 的确认信息后，提交写操作。

4.用 mr 设计一个分组排重计数算法

输入文件格式:二级域名,一级频道,二级频道,访问 ip 地址,访问者 id

需求:按照二级域名,一级频道,二级频道分组,计算 pageview 数,计算独立 ip 数和独立访问者 id 数。

大数据产业已进入发展的“快车道”,急需大量优秀的大数据人才作为后盾。能够在大数据行业崛起的初期进入到这个行业当中来,才有机会成为时代的弄潮儿。技术在手，天下任我走；面试题我有，打遍天下无敌手！千锋大数据开发面试题免费下载，快去寻找自己心仪的工作吧！