**Hadoop面试题汇总**

[一、知识点 1](#_Toc396210800)

[二、经典题 1](#_Toc396210801)

[三、选择题 2](#_Toc396210802)

一、知识点

了解MapReduce大致流程，map, shuffle, reduce  
了解combiner, partition作用，设置compression  
搭建hadoop集群，master/slave 都运行那些服务  
Hadoop参数调优，cluster level: JVM, map/reduce slots, job level: reducer #,  
memory, use combiner? use compression?  
pig latin,Hive简单语法  
HBase,zookeeper 搭建  
next generation MR2框架  
高可靠性,namenode: avoid single point of failure.  
数据流系统：streaming storm(twitter).

wordcount字典同位词

翻译sql语句 select count(x) from a group by b;

二、经典题

一、现有1亿个整数均匀分布，如果要得到前1K个最大的数，求最优的算法。­（先不考虑内存的限制，也不考虑读写外存，时间复杂度最少的算法即为最优算法）­­

二、编写一只爬虫，要求：1、可配置要爬取的网页URL格式 2、可定制要爬取的深度3、对爬取下来的页面可由后期调用的程序进行存储（即事件）  
三、现有大批量url需要爬取，其中url的解析以及n层抓取已有服务端实现（多级深度），现在给定若干台服务器以及不断增加的客户机，各服务端的url任务已有机制保证平衡，爬虫url任务由客户机向服务器请求并完成。请设计一个分布式框架，以完成单层的ur1抓取，并且每个服务器都能尽可能平均的获取客户机资源。注意：服务器可能当机。

四、设计一套系统，使之能够从不断增加的不同的数据源中，提取指定格式的数据。  
要求：1、运行结果要能大致得知提取效果，并可据此持续改进提取方法；  
    2、由于数据来源的差异性，请给出可弹性配置的程序框架；  
    3、数据来源可能有Mysql,sqlserver等；  
    4、该系统具备持续挖掘的能力，即，可重复提取更多信息；  
五、编写一个工具，该工具能够根据不同的文档模板，生成提取格式化数据的正则表达式。

三、选择题

1.下面哪个程序负责 HDFS 数据存储。

a)NameNode b)Jobtracker c)Datanode d)secondaryNameNode e)tasktracker

答案 C datanode

2. HDfS 中的 block 默认保存几份？

a)3 份 b)2 份 c)1 份 d)不确定

答案 A 默认 3 份

3.下列哪个程序通常与 NameNode 在一个节点启动？

a)SecondaryNameNode b)DataNode c)TaskTracker d)Jobtracker

答案 D

分析：hadoop 的集群是基于 master/slave 模式，namenode 和 jobtracker 属于 master，datanode 和tasktracker 属于 slave，master 只有一个，而 slave 有多个SecondaryNameNode 内存需求和 NameNode 在一个数量级上，所以通常 secondaryNameNode（运行在单独的物理机器上）和 NameNode 运行在不同的机器上。

**JobTracker 和 TaskTracker**

JobTracker 对应于 NameNode，TaskTracker 对应于 DataNode，DataNode 和 NameNode 是针对数据存放来而言的，JobTracker 和 TaskTracker 是对于 MapReduce 执行而言的。mapreduce 中几个主要概念，mapreduce 整体上可以分为这么几条执行线索：jobclient，JobTracker 与 TaskTracker。

1、JobClient 会在用户端通过 JobClient 类将应用已经配置参数打包成 jar 文件存储到 hdfs，并把路径提交到 Jobtracker,然后由 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。

2、JobTracker 是一个 master 服务，软件启动之后 JobTracker 接收 Job，负责调度 Job 的每一个子任务 task 运行于 TaskTracker 上，并监控它们，如果发现有失败的 task 就重新运行它。一般情况应该把 JobTracker 部署在单独的机器上。

3、TaskTracker 是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。

4. Hadoop 作者

a)Martin Fowler b)Kent Beck c)Doug cutting

答案 C Doug cutting

5. HDFS 默认 Block Size

a)32MB b)64MB c)128MB

答案：B

6. 下列哪项通常是集群的最主要瓶颈

a)CPU b)网络 c)磁盘 IO d)内存

答案：C 磁盘

首先集群的目的是为了节省成本，用廉价的 pc 机，取代小型机及大型机。小型机和大型机有什么特点？

1.cpu 处理能力强

2.内存够大，所以集群的瓶颈不可能是 a 和 d

3.如果是互联网有瓶颈，可以让集群搭建内网。每次写入数据都要通过网络（集群是内网），然后还要写入 3 份数据，所以 IO 就会打折扣。

7.关于 SecondaryNameNode 哪项是正确的？

a)它是 NameNode 的热备 b)它对内存没有要求

c)它的目的是帮助 NameNode 合并编辑日志，减少 NameNode 启动时间

d)SecondaryNameNode 应与 NameNode 部署到一个节点

答案 C。

**多选题：**

8. 下列哪项可以作为集群的管理？

a)Puppet b)Pdsh c)Cloudera Manager d)Zookeeper

答案 ABD

具体可查看什么是 Zookeeper，Zookeeper 的作用是什么，在 Hadoop 及 hbase 中具体作用是什么。

9. 配置机架感知的下面哪项正确

a)如果一个机架出问题，不会影响数据读写

b)写入数据的时候会写到不同机架的 DataNode 中

c)MapReduce 会根据机架获取离自己比较近的网络数据

答案 ABC

具体可以参考hadoop 机架感知--加强集群稳固性，该如何配置 hadoop 机架感知

10. Client 端上传文件的时候下列哪项正确

a)数据经过 NameNode 传递给 DataNode

b)Client 端将文件切分为 Block，依次上传

c)Client 只上传数据到一台 DataNode，然后由 NameNode 负责 Block 复制工作

答案 B

分析：Client 向 NameNode 发起文件写入的请求。NameNode 根据文件大小和文件块配置情况，返回给 Client 它所管理部分 DataNode 的信息。Client 将文件划分为多个 Block，根据 DataNode 的地址信息，按顺序写入到每一个DataNode 块中。具体查看HDFS 体系结构简介及优缺点。

11. 下列哪个是 Hadoop 运行的模式

a)单机版 b)伪分布式 c)分布式

答案 ABC 单机版,伪分布式只是学习用的。

12. Cloudera 提供哪几种安装 CDH 的方法

a)Cloudera manager b)Tarball c)Yum d)Rpm

答案：ABCD 具体可以参考Hadoop CDH 四种安装方式总结及实例指导。