零基础大数据就业课程

Hadoop 面试过关必备

主讲人：Gerry

上海育创网络科技有限公司

HADOOP 相关面试题

Hadoop 相关面试内容

1、简述 MapReduce 架构组成

2、有不同的数据库，比如 oracle、mysql 等，它们存储着不同格式的数据，如果想对不同来源的数据进行清洗和分析，请写出一个设计方案

3、什么是分布式缓存

4、有 1T 的文本文件，放在 HDFS 上，怎么实现排序，每个节点的内存只有 1G，考虑性能

5、MapReduce 数据清洗的具体过程

HADOOP 相关面试题

6、以前用的什么版本的 Hadoop

7、简要描述如何安装配置 Hadoop 集群，简单描述即可，无需列出完整步骤，能列出完整步骤更好

8、请列出正常工作的 Hadoop 集群中，Hadoop 分别需要启动哪些进程，它们的作用分别是什么，尽可能写的全面些

9、请列出你所知道的 Hadoop 调度器，并简要说明其工作方法

10、请列举你以前的工作中所使用过的开发 map/reduce 的语言

11、你认为 Java ，Streaming，pipe 方式开发 map/reduce 各有哪些优缺点

HADOOP 相关面试题

12、请简述 Hadoop 如何实现二次排序

13、请简述 MapReduce 中 combiner，partition 的作用

14、HDFS存储数据的replication如何定位

15、Hadoop 框架中怎么来优化

16、我们在开发分布式计算 job 任务的时候，是否可以去掉 reduce()阶段

17、描述一下 Hadoop 中，有哪些地方使用了缓存机制，作用分别是什么

18、使用 MapReduce 来实现下面实例

--现在有 10 个文件夹,每个文件夹都有 1000000 个 url.现在让你找出top1000000url？(可能存在重复数据)

HADOOP 相关面试题

19、谈谈数据倾斜，如何发生的，并给出优化方案

20、MapReduce 基本执行过程

21、MapReduce 出现单点负载多大，怎么负载平衡

22、MapReduce 怎么实现 Top10

23、HadoopHA 集群中，各个服务的启动和关闭的顺序

24、在 Hadoop 开发过程中使用过哪些算法？其应用场景是什么

25、MapReduce 程序如何发布？如果 MapReduce 中涉及到了第三方的 jar 包，该如何处理

HADOOP 相关面试题

26、假设 Hadoop 一个 job 产生 100 个 task，并且其中的一个 task 失败了，Hadoop 会怎样处理

27、MapReduce 中排序发生在哪几个阶段？这些排序是否可以避免？为什么

28、如何使用 MapReduce 实现两个表 join，可以考虑以下几种情况：

（1）一个表大，一个表小

（2）两个表都是大表

29、MapReduce 的调度模式

30、combine 出现在那个过程

31、MapReduce 优化

HADOOP 相关面试题

32、用 MapReduce 实现 sql 语句 select count(x) from a group by b

33、用 MapReduce 如何实现两张表连接，有哪些方法

34、Hadoop 运行的原理

35、HDFS 存储的机制

36、Hadoop 中 Combiner 的作用

37、MapReduce 作业，不使用 reduce 来输出，用什么能代替 reduce 的功能

38、Hadoop 集群运行过程中启动那些线程，各自的作用是什么

39、Hadoop 的调度器有哪些，工作原理

HADOOP 相关面试题

40、MapReduce 的 join 方法有哪些

41、列举几个配置文件优化

42、RPC 原理

43、Hadoop 体系结构（HDFS 与 MapReduce 的体系结构）、 Hadoop 相比传统数据存储方式（比如 mysql）的优势

44、Hadoop 推测执行时如何实现的

45、hdfs 的 client 端，复制到第三个副本时宕机， hdfs 怎么恢复保证下次写第三副本?block块信息是先写 dataNode 还是先写 nameNode?

HADOOP 相关面试题

46、请写出以下执行命令

--如何杀死一个 job

--删除 hdfs 上的/tmp/xxx 目录

--加入一个新的存储节点和删除一个计算节点，需要刷新集群状态命令

47、简述一下 hdfs 的数据压缩算法，工作中用的是那种算法，为什么？

48、Datanode 在什么情况下不会备份

49、三个 datanode，当有一个 datanode 出现错误会怎样

50、hdfs 原理，以及各个模块的职责

51、hdfs 的体系结构

HADOOP 相关面试题

52、 datanode 首次加入 cluster 的时候，如果 log 报告不兼容文件版本，那需要 namenode 执行格式化操作，这样处理的原因是

HBASE 相关面试题

HBase 相关面试内容

1、简述一下 HBase 数据库架构组成部分

2、怎么实现 HBase 的预分区

3、Hbase 设计表的时候 rowkey 和分区考虑哪个？还是都考虑

4、Hbase 过滤器实现原则、介绍一下 hbase 过滤器

5、如何提高 HBase 客户端的读写性能？请举例说明

6、基于 HadoopHA 集群进行 MapReduce 开发时， Configuration 如何设置

7、hbase.zookeeper.quorum 属性的值

HBASE 相关面试题

8、 描述 HBase 搭建过程

9、 HBase 写数据的原理是什么/步骤

10、Hive 跟 HBase 的区别是

11、HBase 接收数据，如果短时间导入数量过多的话就会被锁，该怎么办

12、HBase 宕机如何处理

13、如果让你设计，你觉得一个分布式文件系统应该如何设计，考虑哪方面内容

14、每天百亿数据入 HBase ，如何保证数据的存储正确和在规定的时间里全部录入完毕，不残留数据

15、请问下，HBase 直接将时间戳作为行健，在写入单个 region 时候会发生热点问题，为什么

HBASE 相关面试题

16、Hbase 中的 metastore 用来做什么的

17、HBase 怎么给 web 前台提供接口来访问（HTABLE 可以提供对 HTABLE 的访问，但是怎么查询同一条记录的多个版本数据）

18、htable API 有没有线程安全问题，在程序中是单例还是多例

19、HBase 的并发问题

20、你们的 HBase 在公司业务中大概有几个表，几个表簇，都存什么样的数据

21、怎样将 mysql 的数据导入到 HBase 中

22、简述 HBase 的瓶颈

HBASE 相关面试题

23、HBase 如果只向一个 RegionServer 写入数据，有什么优点

HIVE 相关面试题

Hive 相关面试内容

1、hive 的 join 有几种方式，怎么实现 join 的

2、hive 内部表和外部表的区别

3、hive 表关联要注意什么，是任意两张表都可以关联吗

4、hive 是如何实现分区的

5、hive 支持 not in 吗

6、hive 有哪些方式保存元数据，各有哪些优缺点

7、hive 如何优化，列举说明

HIVE 相关面试题

8、 hive 底层与数据库交互原理

9、 hive 如何权限控制

10、hive 能像关系数据库那样，建多个库吗

11、Hive 的 sort by 和 order by 的区别

12、hive 中的压缩格式 RCFile、 TextFile、 SequenceFile 各有什么区别

13、hive 的两张表关联，使用 mapreduce 是怎么写的

14、hive 相对于 Oracle 来说有那些优点

SQOOP 相关面试题

Sqoop 相关面试内容

1、sqoop 如何实现增量导入数据？

2、sqoop 在导出数据到 mysql 中，如何让数据不重复导出？如果存在重复数据问题 sqoop 如何处理？

ZOOKEEPER 相关面试题

Zookeeper 相关面试内容

1、写出你对 zookeeper 的理解

2、描述 zookeeper 搭建过程

3、你在什么地方用到过zookeeper，起到什么作用

JAVA 相关面试题

java 相关面试内容

1、什么是队列和栈

2、List 与 Set 的区别

3、数据库的三大范式

4、java 实现非递归二分法算法。

5、jvm 优化

6、Java 写一个冒泡程序

7、多线程

JAVA 相关面试题

8、 集合 HashTable 和 HashMap 区别

9、 Java 如何实现高并发

10、HashMap、 TreeMap 区别，以及 TreeMap 原理

LINUX 相关面试题

linux 相关面试内容

1、请随意使用各种类型的脚本语言实现：批量将指定目录下的所有文件中的$HADOOP_HOME$替换成/home/ocetl/app/hadoop

2、find 用法

3、shell 如何判断文件是否存在，如果不存在该如何处理

4、在线安装 ssh 的命令以及文件解压的命令

5、把公钥都追加到授权文件的命令？该命令是否在 root 用户下执行

6、linux 如何合并文件

LINUX 相关面试题

7、 假设有 10 台主机，H1 到 H10，在开启 SSH 互信的情况下，编写一个或多个脚本实现在所有的远程主机上执行脚本的功能，例如：runRemoteCmd.sh "ls -l"，期望结果：

H1:

XXXXXXXX

XXXXXXXX

XXXXXXXX

H2:

XXXXXXXX

……

LINUX 相关面试题

8、 给出一个文本：比如 http://aaa.com http://bbb.com http://bbb.com http://bbb.com http://ccc.com http://ccc.com

让写 shell 统计，最后输出结果：

aaa 1

Ccc 2

Bbb 3

要求结果还要排序

THANK YOU

上海育创网络科技有限公司