Hadoop面试题系列(6/11)

一. 问答题

1. 简单说说map端和reduce端溢写的细节

2. hive的物理模型跟传统数据库有什么不同

3. 描述一下hadoop机架感知

4. 对于mahout，如何进行推荐、分类、聚类的代码二次开发分别实现那些接口

5. 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，为什么呢？

二. 计算题

1. 比方:如今有10个文件夹, 每个文件夹都有1000000个url. 如今让你找出top1000000url。
方法一：
运用2个job，第一个job直接用filesystem读取10个文件夹作为map输入，url做key，reduce计算url的sum，
下一个job map用url作key，运用sum作二次排序，reduce中取top10000000

方法二：
建hive表A，挂分区channel，每个文件夹是一个分区.
select x.url,x.c from(select url,count(1) as c from A where channel ='' group by url) x order by x.c desc limit 1000000;

2.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？