# Hadoop面试题系列(1/11)

## 一. 问答题

1.请说说hadoop1的HA如何实现？

2.列举出hadoop中定义的最常用的InputFormats。那个是默认的？

3.TextInputFormat和KeyValueInputFormat类之间的不同之处在于哪里？

4.hadoop中的InputSplit是什么？

5.hadoop框架中文件拆分是如何被触发的？

6.hadoop中的RecordReader的目的是什么？

7.如果hadoop中没有定义定制分区，那么如何在输出到reducer前执行数据分区？

8.什么是jobtracker？jobtracker有哪些特别的函数？

9.hadoop中job和task之间是什么关系？

10.假设hadoop一个job产生了100个task，其中一个task失败了，hadoop会如何处理？

11.hadoop推测执行是如何实现的？

12.关系型数据库有什么弱点？
很难进行分布式部署，I/O瓶颈显著，依赖于强大的服务器，需要花更大的代价才能突破性能极限
很难处理非结构化数据

13.什么情况下使用hbase？
适合海量的，但同时也是简单的操作（例如：key-value）
成熟的数据分析主题，查询模式已经确定并且不会轻易改变。
传统的关系型数据库已经无法承受负荷，高速插入，大量读取

## 二. 分析题

1.有一千万条短信，有重复，以文本文件的形式保存，一行一条，有重复。请用5分钟时间，找出重复出现最多的前10条。
分析：
常规方法是先排序，在遍历一次，找出重复最多的前10条。但是排序的算法复杂度最低为nlgn。
可以设计一个 hash_table, hash_map ，依次读取一千万条短信，加载到hash_table表中，并且统计重复的次数，与此同时维护一张最多10条的短信表。这样遍历一次就能找出最多的前10条，算

法复杂度为 O(n)。