Part1：

1. 自我介绍
2. 说一下hadoop生态圈，以及它一些原理性的东西
3. 说下MR程序工作调度的过程
4. 两个表做一个join操作，针对这种场景，这个map过程是怎么 ，接着又问这里join的key是怎么样定义的之类的问题

以这份简历为例： 你问完它SQL熟练吗

然后它简历上有说hive处理的，肯定有做数据清洗的，问问平常怎么清洗数据的

可以，清洗完的数据；我们公司一般都是放在数据仓库各层的

你们公司数据分层怎么实现就行了

或者有用hive SQL分析过一些业务指标吗

抽问他其中一个业务指标怎么 分析的就行了

|  |
| --- |
| Part2：  编写SQL  有一个表 ，字段： 视频uid（string），视频类型type（int），视频播放量（vv）  1、求所有uid中vv播放量top10的 uid列表  2、求各种类型下vv播放量top10的 uid列表（输出一个type，uid这样的格式）  （我使用开仓函数解答）  3、有一个表：日期字段day（天），vv（播放量）  假如有连续三天播放量>一亿，那么这些天就是 流量高峰期， 把这些满足高峰期的天数 求出来。  这里可以使用表自联解决 |

|  |
| --- |
| Part3——场景分析   1. 前一天视频的播放量是 1亿，然后第二天的播放量为8千万，这里丢失20%   现在让你来排查原因，你怎么分析  （也就是证明给老板看这是正常现象）  我说了些（A：是否因为前一天是周末，或者节假日，访问量多一些  B:与其他同行竞争产品数据分析，例如看看爱奇艺这些是否也是这样  C:也与上一个月这两天，或者上一年这两天的 视频播放量进行比较  ）   1. 像刚刚这一类问题你一般会怎么分析 ？   3、市场中有很多APP应用，对应APP刷下载量（如：对于APP下载量排名这些），  你如何找出其中刻意刷量的人员  我：  对于APP的下载量在某一天，某一个时间段下载量是否有暴增，异常的情况。  对于用户行为进行后续分析，是否只在当天下载使用过，类似沉默用户数，或者短期的流式用户数这些、  包括对本身下载行为的区域、ip这些等等  反正一直问还有什么其他的，然后说不知道了，我再反问面试官，你觉得还有很原因  （结果和我说的差不多，醉了，男人何苦为难男人呢） |

其他：

1. 你自己对工作规划是怎么样的
2. 算法提升指哪些呢
3. 管理层面是怎么提升呢
4. 然后你在项目中，有参与过什么管理的吗

然后我问问它一些问题

1. 对于腾讯云,类似k8s等等一些问题
2. 您负责的这个大数组项目，主要是负责怎么样的业务或者说这个项目组的发展战略是怎么样的
3. 问了他们的可视化平台

总结：

重点是SQL和 场景业务分析 这两部分