Vivo外包

做完笔试以后，直接从笔试题切入问窗口函数的用法、区别以及展示

然后再问我求的指标中有没有类似求topn的？怎么实现的？

你们公司对日活跃、周活跃、月活跃的理解

大概分别是多少？

一天数据量大概多少？

中信科技银行外包

介绍最熟悉的一个项目

项目人员配置与分工？

为什么要离职？

Hive的数据迁移一般花多少时间，从dw->dm？

如果一个hive比较慢，那要怎么处理？

如果发生数据倾斜，怎么处理？

开发中你遇到过吗？怎么处理的？

max()与max() over()的区别

行转列、列转行

项目的调度用的是什么？

每天的数据量大概多少？

数据仓库的建模了解多少？

知道哪些数据建模的工具？

为什么公司要引进Spark？

Linux会操作吗？

最后笔试

亿达信息公司（腾讯外包8楼）

第一个面试官

15分钟写5个sql的测试题

sort by与order by的区别

count(\*)与count(A)的区别

mr的具体流程

hdfs的读写过程

元数据存储在namenode的哪个地方

hdfs的文件格式

小文件过多怎么处理

数据倾斜是什么

数据倾斜怎么处理

打随机数怎么操作

Mr的Shuffle过程的作用、发生在哪个地方

业务中的月活跃指标怎么实现的

sql怎么进行去重操作

第二个面试官

7日留存率怎么实现的

为什么要用Kafka

Zk的工作机制

Zk文件系统的增删改查

Left join与right join的区别

Ab两表联立以后怎么求在a表但不在b表的值

雁联科技

自我介绍

介绍一下项目中运用到了哪些底层工具（项目架构）

如果在跑hive脚本的时候出现了错误，需要怎么定位？

如果联立的很多张表呢？要怎么定位

Hive的运行原理跟流程？

Spark core的运行原理

你们用spark streaming是用来干嘛的

讲一下你们用spark streaming怎么处理热数据的

Spark的具体提交流程

项目中zk的作用

了解es多少？

你们etl用的是什么

Hive你们分桶分区吗

你们数据量大概多少？

数仓中最大的表大概多大

Spark shuffle的优化有做过吗？

Hive的呢？设置的参数大概多少？

Java用得多不多？用java处理过数据吗？

请开始手写mr（wordcount）