1. shell常用命令至少说五个。

Cd 进一层

Find 查找文件

Vi

Cp 备份

rm-rf 删除

Cat 查看文件名

1. 对hive有什么认识，会哪些常用命令
2. Hive保存的文件格式

大部分是文本格式，二进制文件Hadoop API的SequenceFile

1. 怎么用hive执行sql文件

方法1： hive -f sql文件，方法2： 进入hive shell, 执行source命令

1. Hive建表和分区的命令
2. Json文件解析用什么函数

方法一使用函数get\_json\_object ， 方法二使用函数 json\_tuple

1. 爬虫怎么模仿人点击操作

就是url后面加上parmas， headers就行

1. 爬虫如何获取想要的字段

进行网页解析，如果是json格式用json包解析，掐头去尾留中间，然后通过for循环，按键值对，取你需要的值就行

1. Sql性能优化

我上个项目有用表分区，看优化器执行时间更改sql语句提高查询速度，分区表里面建索引

1. 大表小表关联如何提高性能

大表在前，小表在后，关联字段建索引

1. 你平时如何使用索引

一般就是主键字段，建索引，上个项目就有男女字段建索引，经常用的函数字段会建索引

1. 你们表分区是怎么用的

* 表大小：当表的大小超过2GB时考虑对表进行分区；
* 查询：将经常用到的where条件的字段进行分区；
* 分区上个项目我们有按月进行分区，也有按省份分区。

1. 你在这个项目里面主要作什么

<https://blog.csdn.net/sinat_37935727/article/details/105137725> 逾期用户画像分析链接

我主要做的项目是华夏银行信贷分析里面的一个主题逾期用户画像分析，维度就是初始认证等级、年龄、性别、学历，字段就主要是借款日期，金额，期限，利息，借款类型，已还期数，已还金额，待还期数，待还金额，还款状态这些东西，多的也记不得了  
要问取数的话就是从信贷服务系统和crm系统取数

1. 写sql有遇到什么问题

我写sql经常报错吧，主要是有些外观连导致数据发散，上个项目也有存储过程第一天还好好的第二天就失败了，后来看日志发现是主键冲突

1. 创建索引要多少时间

我有次创建4000万行数据的索引用了12小时，有个明细表创建索引用了3天

1. 优化器的快捷命令是什么

优化器我主要用的F5,mysql里面的长时间不用忘了，他说哦这个Oracle我不了解，MySQL是有快捷命令的

1. sql语句优化，你是怎么做的

少用子查询，能先where就先过滤，尽量等值判断，不用不等于，少用模糊匹配，有用到分析函数的看看能不能替换掉，除非是算排名，算累计，同比，环比

1. 最新的前沿技术你有了解？

Apache Doris是一个现代化的MPP分析型数据库产品。仅需亚秒级响应时间即可获得查询结果，有效地支持实时数据分析分布式架构非常简洁，易于运维，百度的东西

1. 对spark有了解？

Spark是大数据框架，主要用于实时数仓，这个有学过，但是没有做这方面项目时间长忘了。

 只记得原理：**Spark** 是将数据一直缓存在内存中,直到计算得到最后的结果,再将结果写入到磁盘,所以多次运算的情况下, **Spark** 是比较快的

1. **描述性分析，对比分析**是怎么做的

描述性统计分析即是对数据源最初的认知，包括数据的集中趋势、分散程度以及频数分布等就是最大值，最小值，平均值，中位数，众数，四分位数（箱线图）。

对比分析，时间对比，区域对比，同比，环比

1. Grouping怎么用

这个相当于nvl，主要是赋值的

1. 关联机制有哪些？怎么用？

这个就是执行计划，具体怎么关联之前同事讲过点，记不清了，执行计划，基本上不用调整

1. 水位线？

用truncate降低，delete不能，一亿数据删除只剩100万，但是查询还是会查1个亿

99乘法表

定义两个变量下x，y，套两层for循环，第一个变量x是1到9，第二个变量y是1到x变量

打印输出y||\*||x||=||（x\*y）||’ 空格’

第一个循环结束打印个换行

求和1到100

定义两个变量a=1，b=0 for循环a是1到100，b=b+a

循环结束打印b

OLTP和OLAP是什么意思，有什么区别吗？

OLTP：联机事务处理，是实时明细数据

OLAP：联机分析处理，是汇总数据，包含2到3年的历史数据

求连续天数

我们可以先排序，按人名字段和时间排序，再加个rownum，再用人名字段和时间字段减去rownum 分组，加having 条件 count（\*）大于等于2的，

最后查出人名，和count（时间字段减去rownum）

求连续天数起止时间

先进行自关联分A,B表，内关联条件就是人名一样，日期等于日期-1，

分组条件是a表人名，a表时间-rownum

最后查询a表人名，a表最小时间民（a.时间），b表最大时间

如何创建自增序列

用create sequence seq\_ax（序列名）;

再通过存储过程把序列名赋值给创建的表里的列，这个变量作为表里的字段就行