1. HDFS的底层逻辑是什么？
2. 常见的Linux命令
3. 存储过程的调用（非begin end）
4. 笛卡尔积优化
5. 常见窗口函数（开窗函数）
6. 你们ETL用的啥工具
7. 大数据ETL过程

---------------------------------------------------------------------------------

1. 协助导师搭建数仓，搭建数仓模型，你能介绍下吗？就是你们仓库的架构啊，搭建模型的方法和规范什么的。

答：传统数仓的三层架构。建模方式以星型模型为主，雪花模型用的少。

面：第二层建那些模型？有没有分主题？你举个例子啊？有哪些事实表和临时表？你建了哪些事实表，是宽表吗？

答：没答好，主要是没有听清有没有按主题分。我说的是雪花模型怎么建设的。我说是按事实表和维度表来的，后来他跳过了这段，问道大数据方面。

1. Hadoop这块有了解.....

答：回答不佳，说的少了。我说，Hadoop了解一些，接触时间不长。我主要在这个大数据这块做的数据抽取工作，用HIVE，没有怎么用过，只是知道基本的流程。

1. 调优这块你能介绍介绍吗？

时间戳：03:23.75 没听懂他说的啥

答：数据宝典调优策略，他指的调优不是这个。我提到这个，他就顺口提了句，语句怎么优化

1. 那SQL语句怎么优化呀？

答：数据宝典SQL调优

1. 汇丰银行改造项目，花了多长时间

答：半年左右，项目在进行，我离开了。

-----------------------------------------------------------------------------------------

京北方

1. kettle 数据同步的逻辑吗？

答：不好意思，我没听清楚

面：比如它跨库同步，具体是怎么做这个同步的，通过什么方式去访问，通过什么方式把它抽取出来，在怎么加载到目标库的？

老师解答：

kettle里面有个lib文件夹，里面有Mysql、oracle的包，就可以进行同步了。

1. 你的实际工作中有用到Linux吗？

答：有用到，只是用到的比较少.....噼里啪啦。（面试官打断）

1. 那你是用来干嘛呢?

答：常见Linux命令。

1. 有亿级数据处理经验嘛，大表怎么优化？包括它怎么去存储，怎么去优化，怎么去设计？

答：最常见的分区，如索引优化，分区表优化，执行计划优化。

1. 索引的问题，你说让索引失效，应该是XX光同步是才能这样做吧，如果是增量同步的话，用调度工具去同步的话，应该怎么去处理啊？

答：增量同步啊。我们的增量同步数据量可能没那么大哈，我们增量同步数据差不多30万到50万。都没咋设计索引这一块，直接取个时间戳弄过来就行了。---我不会。多半讲错了

老师解答：增量按会计期取增量。

1. 索引的问题，比如问我对A字段建立索引，在看执行计划时发现索引没有运行，请问是什么原因呢？

答：（索引失效没回答好）--索引进行运算，索引使用函数，对索引使用隐式转换，用！=和null进行判断，like ’%\_’ ，百分号在前。

1. oracle数据加载，有几种方式，手工的就行。

答：我们都是使用工具弄得，如kettle,DS,sqoop等

面：那你知道oracle有自带批量导入工具，你知道吗？

答：这个可能没涉及到....

答：sqlldr

1. 大表，对大表进行过滤，3个where条件，现在要判断哪个where条件要放前面，哪个放后面，排序的规则有什么需要考虑的地方吗?

答：瞎扯。我能学习下吗？

面：百度上挺多的，你可以自己去查一下。

1. 报表用到哪些聚合函数

答：聚合函数

1. 分组排序有用过吗？

答：没想起来，你那边可以给我个具体的例子吗？

面：row\_number用过吗？

答：用过，还有rank，dense\_rank.

1. 行列转换有用过吗？

答：case when 或者decode。位移函数如lag,lead.最笨的办法如切片。Oracle11G之后可以用它自带的行列转换函数。Pivot()和unpivot().

1. 学生成绩表，三个字段，一个是学生编号，一个是科目，一个是成绩。要两个字段，比如一个字段是学生，另一个字段是科目和成绩拼起来的字符串。

答：没回答好，说了个最笨的办法，先切片，把科目从行放到列上去，union all 在进行一个字符串的拼接。

老师解答：WM\_CONCAT函数

1. 调度工具有配置过吗？

答：调度工具不就是在kettle上配置吗？

老师解答：调度工具配置有专人进行

14：脚本语言有写过吗？

答：shell涉及的会少一点点哈，有看过，但自己也写过一点点，但是还是感觉，自己写的和别人写的有差距啊

1. shell写过什么功能？

答：其实前面有个项目是这样的，传统数仓改造成大数据，我们就把传统数仓的SQL放脚本里跑一下，其实shell脚本不就相当于Oracle中的存储过程嘛。

面：就是容器是吧？

答：嗯

-----------------------------------------------------------------------------------------------

1. 贷款业务主要是个人贷款还是对公贷款？

答：个人贷款。

1. 也做ETL开发吧？

答：也有做，但是涉及少，我们主要是根据数据字典和mapping文档，用工具进行抽数。

1. Oracle的存储过程写过吗？

答：在同事离职期间，顶替了一个月。

1. 分区表一般在什么情况下使用?

答：数据量太大了，可能它两个小时，三个小时没跑出来，我们就会考虑建设一个分区表。

老师解答：避免全表扫描

1. 我现在有一个存储过程，它运行时间非常长，请问从哪些方面去优化？

答：我首先看看这个存储过程的SQL语句嘛，如果涉及delete，update，我会考虑用查找的方式实现。看看它是增量跟新还是全量跟新，看看有没有能用动态SQL解决的问题，能用就尽量用。然后我会看索引，是不是有索引，索引失效没，还是索引过多。然后数据量太大了我看看有没有分区，没有分区，就和领导商量分区。之后看执行计划，看执行顺序，全表扫描，还是部分扫描，耗费。在可能的情况下，看看能不能设置并行。设计表关联，把大表放前头，小表放后面。会快些。

1. 你们常用的并行有几种方式啊？

答：/\*+ PARALLEL(表名,并行数) \*/

面：有没有用过强开并行的那种？

答：这个没涉及到，你们在用吗？

面：我之前项目上有人用过，我见他用过，实际上我也没去研究过，他是整个存储过程，在开头加了个强制并行，整个存储过程都会并行去跑。

答：感叹了下。回头在找看

1. Linux常见命令

答：常见基础Linux命令

面：其实咋们会用基础就行了

答：我们之前项目组也是这么讲的，所以了解不是很深

面：笑

---------------------------------------------------------------------------------------------

1. 有用kettle语句进行复杂加工吗？

答：您指的复杂加工时什么呢？

面：比如说类型存储过程那种相对复杂的聚合运算啊，您会在kettle中实现吗？

答：这个我们是没涉及到的。因为我们的kettle主要是保证源数据过来嘛，我们最多就是建立个零时表，把源数据放进去，然后对临时表的数据进行一个数据清洗工作。我们主要是这么干的，因为我不清楚你们那边是怎么做的

1. 你这边对SQL语句熟悉吗？

答：基本上的函数都有涉及到，存储过程写的可能没别人干练，大不了多套几层的那种感觉

面：（笑）哦哦...

1. 数据库优化有了解吗?

答：面试宝典

1. 大表小表关联怎么优化

答：大表放前面，关联条件看需要，如果需要建索引就建，建索引更快一些

1. 索引失效

答：面试宝典

1. 你接触过存储过程，你对PLSQL这个工具熟吗？

答：没听懂，问了下

面：就是你再什么环境下编写存储过程？

答：专门有测试环境啊

面：呃呃，就是用什么工具？

答：楞了三秒，说，我们是怎么写的SQL，问他们怎么做的

面：我们一般是借助工具啊，一般是PLSQL那个工具。

1. 您对Linux操作系统熟吗？

答：涉及到的比较浅，说常见的Linux常见命令

1. 脏数据的处理

答：面试宝典

1. 有使用到调度工具吗？

答：调度，我们工作的时候一般是在kettle中进行调度的，建一个工作流嘛，调度很方便的

1. 除了这些，你还有其他的技能吗?

答：同事离职期间顶替了一个月时间。会基本的帆软和Tableau 的操作。Tableau主要呈现给我们自己用的，帆软主要是客户看的。会用，但不精通。

-----------------------------------------------------------------

1. 前面霹雳吧啦一大堆，没问出个所以然
2. 那你对什么索引啊，主键啊的概念有了解吗？

答：可能是在听他们讲优化方案时有了解，具体了解不是很多

1. 长时间的沉默。我看你有些Hadoop，你在这方面怎么做的？

答：我在Hadoop中也是负责抽数工作，.........（怎么做的）

1. 你觉得Hadoop和ORACAL有什么区别

答：面试宝典

11.16 答疑小结：

调度不用kettle，用（CTM）。

存储过程调用，调度工具kettle。

银行的数据是看不到的。拿到的是脱敏数据。

三套环境，本地环境（开发环境），测试环境（查询权限），生产环境（正式环境）

数据从正式环境来（业务系统来）

数据量多大？听说......

测试环境的数据量造出来差不多是正式环境的十分之一。

开发规范：

1.数据分层规范，比如ODS层，DW层，DM层各执行什么功能，放什么东西。

2.表方面：三范式，命名规范，代码规范。

命名规范如：如事实表用DW\_表名，维度表用DIM\_表名。

代码规范如：如关键字右对齐，代码注释详尽等等。