面试总结

问题：存储过程的异常怎么插入日志？

回答：用exception，在exception中把过程数据记录到错误日志表，然后对主过程的事物进行回滚，并且针对异常我们会记录异常的日志，方便后续的问题分析，增量更新。

问：存储过程的结构。

答：创建存储过程名以SP开头，然后声明变量游标，然后删除数据插入数据这些，然后begin end是调用存储过程，我把它给回答到了存储过程结构里面去了，然后对方也没有纠正我。

问：SQL执行计划的标准

答：主要看执行顺序，表扫描方式，关联机制还有耗费这些CPU耗费这些。

问：存储过程怎么优化？

答：他问了存储过程怎么优化，但是我回答到SQL语句优化就是用commit提交数据，然后什么distinct用group by替代这些，然后对方又问了一遍，他说你这个说的是SQL语句的优化，我问的是存储结构怎么优化？然后这个我一下子没反应过来，嗯了半天没回答出来，然后对方就跳过了下一个问题

问：拉链表是怎么操作的

答：我的回答是拉列表分为三种方式，一个是直接覆盖原值，然后一种快照表，对每天的数据进行一次保留的快照表，然后一个是保留全部数据的。这里对方打断我，说你回答的不是我问的问题，你回答的是拉列表的类型，我问的是你平时是怎么操作拉列表的，然后我愣了好久才说在实际建模中我是根据实际情况来决定的，就是看需求，看客户的需求是要要最新的增量更新数据，还是要看所有的数据，客户那边他需要保留多少数据，如果客户那边是只需要最近的数据，那就在目标表中创建最新的数据或者说最近的一批数据 （反正就是乱七八糟扯了一堆，到后面我自己都不知道自己在说啥）

问：怎么创建索引？

答：一下子没找着答案，就说了索引的几种区分，还有索引的优缺点，对方也没打断，直接问了下一个问题

问：遇到过哪些报错

答：存储过程一般报错比较多，像是主键冲突啊，除数为零这些

问：平时用星型模型较多，还是雪花模型较多，

答：都有，用最多的还是星型模式，因为它存在冗余数据，性能比较好一些，效率比雪花模型要高

问：你平时怎么清洗数据

答：我们通常是在从ods到dw层的时候做的数据清洗，因为我们之前做的项目在ods中基本没有做过多的处理，相当于和来源系统一样，所以我们就先选取需要的字段放进DW层，然后过滤掉空值，还有关联不上的数据，编码格式不对的，用case when处理一下，然后保证入仓的时候数据的完整性还有准确性。

问：写过最复杂的存储过程是什么样的，大概描述一下

答：这里要我举一个具体的例子，但是一下子编不出来，就只说了可能大概七八百行吧，那个电商的项目，给客户算订单销售情况的，就这么糊弄了几句，他也没多问

问：kettle调用作业怎么定时

答：kettle有定时功能可以直接设置 （这里应该是要我详细描述哪里设置怎么设置，但是我忘了kettle长什么样了，就回答得很模糊）

问：用过几种报表工具，哪个用得最熟练？

答：用过帆软和tableau，平时用的最方便的还是tableau，它能直接看到报表做出来的效果，帆软要所有东西设置好了之后去网页浏览才能看到效果图，不过在tableau只能做一些比较简单的报表，如果是比较很复杂的多种图表联合的话还是要帆软功能比较全。

然后对方问了报表开发熟练吗（这里问了几个帆软里面我没有听过的功能，我都没回答上来），我回答平时工作做得最多的还是在oracle敲代码，报表这方面做的都是一些表层的最基础的报表制作，再深入就没做过了，因为我们有专门的做报表开发的同事，我就只会做一些最基础表层的报表。

问：做过需求分析吗

答：自己专门做过就没有，我们接项目的时候，项目组会先开会，然后我们会和组长还有专门做需求的需求分析师，一起讨论确认需求的。

问：你们以前一个项目大概是做多久，平时是一个项目结束后才接下一个项目还是几个项目同时接

答：我一下子蒙了，也不知道怎么回答，就说一般就是大项目中间会穿插几个小项目，等项目结束后一般会有几天的休息时间

问：你们项目组里平时干的活怎么分，比如说有一次你同事的进度慢了，组长让你帮忙分担了一部分，今天你同事出问题被领导找了，他和领导说那部分是你做的，领导要你担责，你要怎么办

答：我们写代码时一般都会写备注，备注这一部分是什么时候谁写的，谁什么时候修改过，如果那部分是我做的会有我的备注，有问题就找出来修改，如果没有备注那就不是我做的不应该要我担责，而且是组长安排我帮忙的那组长知道这部分是不是我做的

问：Hadoop的存储原理。

这个我没回答上来

后面问的关于有关大数据的还问了像hive优化，Linux常用命令，shell脚本的结构，怎么编写shell脚本，如何在shell脚本里面调用另一个脚本，我之前列了一个有可能问到的问题的文档，但是一听到大数据的就先慌了，连答案都找不到，答不上来也不像oracle那样还敢扯两句，反正就是沉默，等了好久等不到答案他就问下一个问题，然后问了好几个我都没回答上来，他就问平时工作没接触过大数据吗看我简历项目经验里写着有用过，我说基本不用大数据，我都是在oracle写代码的，他就没问了，就这么结束面试了。

中间出了两道题

第一题：A表里有商品名称，商品ID，商品单价，B表里有用户ID，商品ID，购买数量，订单ID，地址，B表里某些用户有重复购买记录，求不同商品在不同省份的销售情况，哪种商品的回购率最高？

答：首先用商品ID关联两张表，对商品和省份进行分组，销售额用购买数量\*单价，显示商品，省份，销售额（人家还问我这么简短吗），

回购率就是对用户ID进行分组，求ID的个数，取个数最大的就是回购率最高的（面试官指出这里我求的是每个人的购买记录里面回购了几次，其实我漏了一个这个商品不是只有一个人回购，还有其他人也会回购，题目要求的是这些人加起来的回购次数最高的商品）

第二题：给商品销售额排序时，描述一下ROWID，ROWNUM，RANK，DENSE\_RANK，ROW\_NUMBER这几个哪个能用哪个不能用，用这几个排序分布有什么不同

答：我只记得ROWID是一种物理地址，怕说多错多就只说了句ROWID不能用来排序，重点都在描述rank，dense-rank和row-number的区别，rownum我跳过没说