面试官，您好，我叫XXX，来自XXX，之前在XXXX（在广州天河区金慧街那边八十八号大院那里）做bi工程师的，之前的是做XXX的一个项目（是在XXXX路，XXXX路，XXXX站点，那边的XXXX上班的），主要是写存储过程，负责将上游的数据同步到数仓层，因为在那边做了挺久的，想换个环境，刚好深圳这边朋友比较多，就过来了，谢谢。

我的任务主要是将上游的数据同步到数仓层，也就是dw层，根据业务需求对事实表和维度表进行建模。

事实表，一般来讲，就是我们的基金或者黄金的交易额，维度的话，就是，我们是从哪个方便去看待的，比如，地区，业务类别都可以看做是一个维度。

我们有客户表，基金表，产品表，保险表，流量表，一般我们会选择表名加字段的方式命名，具体的字段，也不是很清楚，就是一些基金编号，基金收益率，还有风险这些，然后就是风险控制的一些维度维度：资金分类啊，资金来源啊，资金去向啊什么的，或者说对账户，产品限额的一些控制，就比如说，一个人买了什么类别的基金，然后我们相对来说，推送的也是差不多金额的，但是会加一点点。

Xxx时间-xxx时间 xxx做xx项目，有xx多了吧，做了十多个版本。Xxx时间-xxx时间在中国联通做用户积分的项目，做了一年也是十几个版本。

有在华为做前端的，也有外包在腾讯的，感觉这边大公司很多，也期待有一个更好的学习平台吧。

之前写的存储过程，一般情况下就是六百到八百行的样子吧，但是有一些比较复杂的情况下，会写到一千到两千行的样子。

一般我们会先调用日志，查看那个部分占用的时间比较多，然后再通过执行计划，看看哪个执行过程占用的资源比较多，然后先考虑能不能从SQL语句上进行直接优化，比如少用一些方便，但是占用资源的查询方式，比如distinct等，再考虑别的方式，比如建索引。

一般情况下，我会选择查询次数比较多的列或者经常用在连接的列上去创建索引，有时候也会根据需要对经常进行范围查询的列创建索引，因为索引语句排序，这样可以利用索引的排序加快排序查询时间。

数据库是面向事务的，直接用来捕获和存储在线数据，数据仓库一般是面向主题需求设计的，一般存储的是离线数据。

对于那些查询中很少使用的列不应该创建索引，对于那些只有很少数据值或者说，重复数据很多的列，比如说性别之类的，也不应该增加索引。

表分区可以改善查询性能，对分区对象可以搜索增加关系的分区，提高检索速度。

增强可用性，如果表的某个分区出现故障，表在其它分区的数据不会受到影响，

相对应的，维护也相对方便，当出现问题的生活，只需要修复出现故障的分区就可用了，可用把同的分区映射到磁盘以平衡I/O，改善整个系统性能。

星型模型，所有的维度表都能直接跟事实表关联，存在冗余，一般来说，性能会更好。

雪花模型，有维度表不能直接跟事实表关联，需要通过关联其它维度表才能关联到事实表，去除了冗余数据，一般要进行多次关联，效率不一定有星型模型好。

项目上线的时候，主要会先进行一些代码检查，整理上线清单之类的事情，确保提交的代码是最新的，并且没有拉下的，上线清单，就是一些脚本和一些表结构还有映射关系之类的。

因为每个源系统都有自己的功能划分呀，虽然都可以做同样的事情，但是，会对系统的性能有影响，而且，为了避免源系统频繁交互，也需要建立数仓呀。

一些Linux命令：切换目录：cd； 匹配当前目录唯一的文件名：tab；列出目录和文件：ls；创建目录：mkdir；删除目录：rmdir；显示文件结尾，（-n显示最后n行）：tail（-n）；改变文件或目录名：mv；杀死进程：kill；获得计算机网络名称：hostname；显示文件内容：cat。

项目业务系统：以客户为中心，产品模型化，业务流程化，架构组件化，海内外一体化，风控体系化，管理信息化，对业务数据进行统一管理规划。

银行内部数据：客户信息，账户信息，交易流水，信用评估等。

存储过程主要是对一些表的增删改操作的，自定义函数是对一些指标进行技术，必须要返回一个结果。存储过程可以通过begin，end直接调用，自定义函数一般接在查询语句里面。

帆软：一般我们会先将上游的数据进行指标计算，根据需求可能还需要些一些自定义函数，然后使用帆软连接到数据库，在根据需求进行一个柱状图、饼图、线图的输出。

Kettle：我们会先根据需求，创建一些转换和作业，再把数据同步到ods层。

浏览量，客户交易量啊，还有客户选择新产品的预期值。然后统计一些途径，比如是从交易页面点击的，还是从广告页面进入的。