自我介绍

面试官好，我叫孙昌勇，本科就读于北京邮电大学软件工程专业，院内排名前7%，保研本校，目前在北邮计算机学院软件工程专业读研二，实验室研究方向为深度强化学习，项目是做算法的。我在本科大四的时候有过一年的实习经历，在一家数据库公司做数据库底层开发，主要贡献是执行器优化、参与资源管理器的设计和数据库开源的一些工作。

项目经历

1、遇到的最大困难是什么，怎么解决的？

- 实习中遇到的最大困难的话，印象最深的是线上报出的两个bug，一个bug的场景发生在tpch和tpcds的测试scale在180g的场景下，出错的那个sql非常长，定位到出错位置后，发现它干了这么一件事，有一张表假设叫它old_table，表里存储了大量的unicodechar字符串，类型是varchar，然后它用`insert into new_table select * from old_table`的方式又新建了另一张表new_table，然后它用`EXCEPT`operator取两个表的差集，即`select * from old_table except select * from new_table`。理论上将返回的结果集应该是空集，因为两个表里是相同的数据。但是结果集返回了很多条数据，180g规模下表里共有600w行数据。当时测试人员和组件owner给的问题定位是新执行器的插入功能出错了，没有插入正确的数据，但是180g规模下无法调试，只好缩小问题规模，这样才能在自己的mac上调试嘛，然后一步步调试发现插入过程完全没有问题。然后这个思路就被否了，于是转变思路，看except操作符的原理，它就是把左右两边数据的哈希值做比对，如果某一条数据的哈希值对不上，就把这条数据抽出来。然后就发现出错问题的数据后面都有trailing blanks，于是猜测算哈希值时把后面的空格trim掉了，从而导致哈希结果不一致，然后进一步验证发现猜想正确。
- 还有一个线上bug，出错query是一条长查询，这条查询在进入解析器的时候，会先序列化为protobuf，然后后面做查询计划的时候会在进行反序列化，但是反序列化parseFromString()的时候会报错。由于出错query是一个长查询，猜测查询过长有关，于是去github上查找protobuf的源码和文档，发现protobuf序列化的时候有一个recursiveLimit参数，而这个长查询就是超过了默认的recusiveLimit所以才导致报错，将这个参数改大就好了。
- 实验室项目的最大困难是如何处理动态负载场景下的任务分配和负载均衡，如何同时应对I/O密集型任务和计算密集型任务，采用的方法就是利用深度强化学习对环境的自适应能力，从而实现任务调度器对环境的变化进行感知，从而及时调整自己的任务分配策略。

> tpch和tpcds是国际著名的数据库测评组织（TPC）发布的两项数据库性能评测基准，主要用于测试数据库的查询响应时间和多个并发用户提交查询时的查询吞吐量。

项目闪光点：

- 实习经历的话，基于SIMD技术进行的查询引擎优化、基于Gossip协议的分布式资源管理器
- 实验室项目的话，使用了比较前沿的技术比如深度强化学习技术和联邦学习技术

一句话介绍深度强化学习：

通俗来讲，强化学习就是训练这样一个智能体，这个智能体可以根据环境给自己的反馈做出不同的动作，如果做出的动作符合我们的预期，我们就给予其奖励，然后通过奖励值的大小调整动作，通过这样不断的训练，这个智能体最终会按照我们预期的那样做出响应的动作。而传统强化学习的弊端是处理高维输入数据时，会存在维度爆炸问题，比如Q-learning算法在处理高维数据时Q表会爆内存，而深度强化学习就是解决这个问题而提出的，通过将Q表提供Q值，改为使用神经网络预测Q值，由于神经网络在处理高维数据时的优势，可以在该场景下获得更好的效果。

一句话介绍联邦学习：

数据不动模型动，数据可用不可见。通过模型参数平均算法获得全局模型，从而充分利用各个不同机构的数据，提高模型的泛化性能，同时可以避免数据泄露的问题。