个人信息

- **基本信息** | 高国彬 | 男 | 1996 | **目标岗位**| 分布式数据库| 基础架构 | 后端研发
- **硕士** | 华东师范大学软件工程 | 2020~2023 | **本科** | 西南交通大学软件工程 | 2016~2020
- **联系方式** |13648111678 | guobin.gao@foxmail.com | Github.com/TaurusGGBOY

工作经历

DolphinDB 数据库内核研发工程师（2023年3月~2023年10月）

- **数据再平衡**:
  - **需求**：设计并实现模块rebalanceChunksAmongDataNodes。需求提出的场景为用户积累的数据量日益增多，磁盘占用率长期处于高点，客户希望实现功能，在插入新盘后，调用接口，使数据能均匀分布，各磁盘占用打平，达到负载均衡的目的
  - **设计**：该模块分为三个阶段，预处理阶段，信息收集阶段，决策阶段。预处理阶段主要的工作为校验等。信息收集阶段需要收集的信息包括机器信息，磁盘信息，数据块信息等。决策阶段分为预分配阶段，真分配阶段，搬运阶段。调度目标为全系统磁盘占用率打平，调度依据为磁盘占用率
  - **优化**：采样估算文件块大小，提高信息收集阶段的效率；优先向同机器搬运，降低网络负载；任务队列洗牌，增加搬运并发度
  - **测试**：实现自动化测试框架，在搬运同时执行DDL和DML，注入故障如数据节点宕机，控制节点宕机，安全关机，磁盘写满，高可用频繁切主，实现OLAP和TSDB引擎随机测试，校验阶段校验元数据完整性，恢复任务状态，数据块版本信息等，减少手动操作，提高稳定性测试效率
  - **结果**：该模块能够高效完成多场景下的数据平衡工作，如添加空磁盘，单磁盘占用过高，存储负载均衡，真分布式场景，在2000文件块，80G倾斜数据情况下，能够在100ms内计算出搬运任务，80s内完成搬运任务
- **TSBS时序数据库性能对比测试**
  - **TSBS框架介绍**：该框架是Timescale在2021年开源的时序数据库性能对比框架，主要分为测试数据生成，测试数据加载，查询数据生成，查询数据四个模块
  - **测试流程**：加载测试需要根据框架修改代码，生成本数据库能加载的数据格式，数据量在1800w到18000w条，每条数据有10个标签值和10个测量值，以及一个时间戳，测试输出加载时间和磁盘占用等指标。查询测试根据生成的点查，简单聚合，复杂查询等语句进行执行，输出查询时间，磁盘占用，内存占用等指标
  - **涉及优化**：设计cache大小，任务队列大小，排序列选择，分区策略，查询改写，batch大小设计，插入冲突解决
  - **测试结果**：测试出加载数据时cpu占用较高，时间较长，加载之后存储文件较大。查询部分，部分聚合排序场景下，内存占用较大，并发时出现OOM问题
- **其他工作**
  - **多表Join**：跟踪SQL执行流程，修复语句改写后列名错误等缺陷
  - **异步复制**：针对分段传输部分以及传输管理器部分，设计不同传输场景，增加单元测试，提高稳定性
  - **算子添加**：线性回归函数，三次样条采样函数，Oracle功能替代函数，Python Pandas Index替代函数

实习工作

字节跳动 火山引擎关系型数据库基础架构实习生（2022年4月~2022年7月）

- **白名单**：对接生命周期框架，完成火山引擎**Mysql Shard**产品白名单V2版本更新开发，基于**ZK**以及**Ingress**，提供白名单的创建，删除，更改，查询，与实例进行关联与取消关联等功能
- **缺陷追踪**：优化内部查询实例资源接口速度至原消耗时间一半，并保留一段时间缓存加速后续查询；定位**从库宕机**，运维面查询实例信息等待的问题

微软上海 C+AI R&D部门SWE Intern（2022年7月~2022年10月）

- **Airflow**：了解开源组件Airflow，了解DAG，Operator，Task等组件含义，实现在本机部署并运行，对标自有产品，总结两者相同点及差异，对接微软Teams API进行二次开发
- **Azure kubernetes Service**：部署Airflow组件到AKS上，探索方案对其进行改造利用，用以同功能替换基于内部服务的**Workflow**，以期实现私有化部署

华宇万益能源 数据分析实习生 （ 2018年8月 ~ 2018年10月 ）

- **可视化**：负责**数据监控可视化**，使用第三方库可视化**Bokeh**，完成基于**Django**框架的可视化界面，从后台获取数据，显示图表等功能 ，最终部署在内网服务器上
- **自动化**：负责自动化报表生成系统，从客户，采集，开发，研发部门调研需求，使用Python生成**Markdown**报表，转为PDF，自动发送邮件给客户，可显示生成进度及报错

比赛项目

PingCAP Tinykv分布式数据库lab（2022年2月~2022年4月）

- **一致性**：实现**Raft**分布式一致性算法，实现领导选举，日志复制，Propose等功能，在日志Commit之后，执行日志中的内容，能够响应垃圾回收请求，生成快照；实现在某些机器需要下线时进行**领导转移**，并提供**成员添加和删除**功能等
- **事务模型**：实现**MVCC**，并在此基础上实现快照隔离级别，通过Default，Write，Lock三个列族实现2PC分布式事务模型Percolator，实现快照读，事务状态检查，批量回滚等

阿里巴巴 Oceanbase数据库大赛（2021年10月~2021年12月）

- **初赛**：开发**MiniOB**项目，负责赛题分析，任务分配，进度管理，规范流程等工作， 优化缓冲池实现LRU算法，支持Date字段，支持Null类型，支持Unique索引，支持多列索引等
- **复赛**：负责优化**Oceanbase开源版本**Nested Loop Join部分的内存分配问题

学习项目

- **MIT6.824**：阅读GFS，BigTable等论文，实现论文MapReduce以及一致性算法Raft
- **CMU15445**：实现前缀树，页面缓冲池，B+树索引，查询，聚合算子，并发控制功能
- **Mysql45讲**：学习了解redo，undo log，事务，索引，锁，隔离等级，存储，排序等知识
- **GoLevelDB源码阅读**：学习了解memtable和sstable的内容，存取的过程和压缩等知识

校园相关

- **英语6级572 4级594 计算机四级网络工程师**
- **国家励志奖学金** “电算七九”奖学金 综合奖学金
- 2019ACM校赛三等奖 新秀杯校赛三等奖
- 校级优秀共青团员 优秀学生干部