西安工业大学
XI'AN TECHNOLOGICAL UNIVERSITY

本科毕业设计(论文)附件

题目：大规模分布式分析型数据库中的Pipeline 执行引擎实践研究

院　（系）：计算机科学与工程学院

专　　业：计算机科学与技术

班　　级：18060314

学　　生：赵长乐

学　　号：18030513121

指导教师：杨国梁

2023 年 6 月

西安工业大学
XI'AN TECHNOLOGICAL UNIVERSITY

本科毕业设计(论文)附件

题目：大规模分布式分析型数据库中的Pipeline执行引擎实践研究

院　（系）：　计算机科学与工程学院

专　　业：　计算机科学与技术

班　　级：　18060314

学　　生：　赵长乐

学　　号：　18030513121

指导教师：　杨国梁

2023年 6月

目录

西安工业大学毕业设计（论文）任务书

院(系) 计算机学院 专业 信息对抗技术 班 18060314 姓名 赵长乐 学号 18030513121

1.毕业设计（论文）题目： 大规模分布式分析型数据库中的 Pipeline 执行引擎实践研究

2.题目背景和意义：

3.设计(论文)的主要内容（理工科含技术指标）：

主要内容：

（1）研究 OLAP 数据库主要技术特点和查询执行引擎性能提升要点

（2）实现一款 OLAP 数据库上可用的 Pipeline 执行引擎

（3）测试该引擎性能并确保其稳定性。

4.设计的基本要求及进度安排（含起始时间、设计地点）：

基本要求：

（1）Pipeline 引擎支持传统数据库查询引擎全部功能，在大型分布式数据库上完美适配

（2）Pipeline 引擎要达到实际性能提升的效果

（3）引擎通过 fuzzy 测试和长稳测试

进度安排：

（1）使用 1 个月时间对现有开源产品进行调研

（2）使用 2-3 个月时间进行项目开发

（3）使用 1 个月时间进行测试及 bug 修复

起始时间：2022.12 地点：工程实验室

5.毕业设计（论文）的工作量要求

① 实验（时数）*或实习（天数）：300 学时

② 图纸（幅面和张数）*：

③ 其他要求：

指导教师签名： 年 月 日

学生签名： 年 月 日

系（教研室）主任审批： 年 月 日

说明：1 本表一式二份，一份由学生装订入论文，一份教师自留。

西安工业大学
XI'AN TECHNOLOGICAL UNIVERSITY

毕业设计(论文)开题报告

题目：大规模分布式分析型数据库中的 Pipeline 执行引擎实践研究

院　（系）：计算机科学与工程学院

专　　业：计算机科学与技术

班　　级：18060314

学　　生：赵长乐

学　　号：18030513121

指导教师：杨国梁

2023 年 1 月 5 日

开题报告填写要求

1．开题报告作为毕业设计（论文）答辩委员会对学生答辩资格审查的依据材料之一。此报告应在指导教师指导下，由学生在毕业设计（论文）工作前期内完成。

2. 开题报告内容必须按教务处统一设计的电子文档标准格式(可从教务处网页上下载)填写并打印（禁止打印在其它纸上后剪贴），完成后应及时交给指导教师审阅。

3．开题报告字数应在1500字以上，参考文献应不少于15篇（不包括辞典、手册，其中外文文献至少3篇），文中引用参考文献处应标出文献序号，“参考文献”应按附件中《参考文献“注释格式”》的要求书写。

4．年、月、日的日期一律用阿拉伯数字书写，例：“2005年11月26日”。

撰写内容要求（可加页）：

1. 毕业设计（论文）综述（题目背景、研究意义及国内外相关研究情况）

1.1 题目背景

随着大数据时代的到来，互联网以及依托互联网技术而发展的企业。正在面对越来越高强度的信息处理需求。部分企业每日所需处理的新增数据量，从以往的 GB 级别一跃提升为 PB 级别。面对如此繁重的信息处理压力，对高效的数据分析处理工具的需求已经成为制约行业发展与科技进步的关键环节。

在这种发展趋势下，以往的单机、少量数据的即时查询工具已经不能满足业界的绝大部分需求。能够针对海量数据高效完成数据查询及分析的数据库引擎是市场及产业发展急需的工具。这也就催生了 OLAP（On-line Analytical Processing，联机分析处理）型数据库的市场。相比于传统的如 Mysql、Oracle 等数据库，OLAP 型数据库大多采用了强大的分布式并行架构和存储、聚合优化等技术，极大地提升了分析型查询语句的执行效率，使得企业进行大规模数据分析成为了实际可能。

数据库市场需求正在逐渐转向复杂化、规模化。因此，OLAP 数据库，尤其是基于大规模分布式架构的大型 OLAP 数据库对于整个互联网行业的发展具有着重大作用。而对于这样的一个高性能的 OLAP 数据库，查询引擎是其中最为关键的核心部件，是整个数据库的核心与灵魂。它的性能与可靠性直接决定了一款 OLAP 数据库的性能与可靠性。经过数据库领域的多年发展，现已得到证明的是：基于数据的流式处理的 Pipeline 引擎在可靠性、性能等方面明显优于传统的 Volcano 执行引擎。因此，对于一款高性能的 OLAP 数据库，Pipeline 引擎的实现至关重要。可以说，它奠定了一款 OLAP 数据库的竞争力基础。

1.2 研究意义

在面对海量数据增长的互联网新业态，OLAP（联机分析处理）数据库对于互联网企业的重要性愈发凸显。它通过多维数据模型、数据立方体计算和聚合查询优化等技术，为用户提供高效的数据分析和决策支持能力。在实践中，相对于 OLTP（联机事务处理）场景中重要的高并发点查，OLAP 类数据库的使用场景往往更为复杂，需要处理大表、宽表下的复杂聚合查询。查询场景通常涉及多个表，需要对特定列进行全表扫描。因此，OLAP 类数据库通常使用列式存储代替传统 OLTP 场景中使用的行式存储，通过提供多种聚合优化、实现物化视图、表达式复用、Runtime Filter 等方式优化性能。

查询引擎是数据库系统运转的核心。随着 OLAP 场景需要处理的数据量快速攀升，传统的 Volcano 引擎由于无法灵活调度查询、难以实现资源隔离，加之本

身因为无法规避的虚函数调用而产生的额外性能开销等原因，已经成为了业务性能瓶颈。这种前提下，新的 Pipeline 查询执行引擎遵照“小数据块的数据驱动”原则，将查询计划树拆分为多个非阻塞的数据流，高度契合现代 CPU 架构的缓存与指令流水线加速需求，实现了更好的查询性能和更加灵活的调度控制，尤其适用于大规模数据集的分析查询。

1.3 国内外相关研究情况

1.3.1 国外研究现状

2.3.2 国内研究现状

2. 本课题研究的主要内容和拟采用的研究方案、研究方法或措施

3. 本课题研究的重点及难点，前期已开展工作

3.1 课题重难点

开发一款 OLAP 数据库中的 Pipeline 执行引擎需要关注以下重点和难点：

1. 数据流处理：Pipeline 执行引擎需要设计和实现高效的数据流处理机制。难点在于确保数据流的高吞吐量和低延迟，以支持实时数据处理和流水线操作

2. 并行执行：Pipeline 执行引擎应支持并行执行多个阶段的操作，以提高查询性能和响应时间。难点在于有效地划分和调度阶段，并处理并发操作和资源管理，以保证数据的一致性和正确性。

3. 高效的数据传输和通信：Pipeline 执行引擎需要实现高效的数据传输和通信机制，以确保阶段之间的数据流畅和快速。难点在于设计和实现低延迟的数据传输协议和通信框架，减少数据传输和序列化开销。

4. 算子优化：Pipeline 执行引擎需要对各个算子进行优化，以提高执行效率和资源利用率。难点在于选择合适的算法和数据结构，优化算子的计算和存储方式，减少不必要的计算和数据移动。

5. 内存管理：Pipeline 执行引擎需要合理管理内存资源，以提高查询的执行效率和吞吐量。难点在于设计和实现高效的内存分配和回收机制，避免内存碎片和内存泄漏问题。

6. 错误处理和容错机制：Pipeline 执行引擎应具备良好的错误处理和容错机制，以应对各种异常情况和故障。难点在于设计和实现可靠的错误检测、恢复和

重试策略，保证系统的稳定性和可靠性。

7. 性能调优和测试：Pipeline 执行引擎的性能调优和测试是一个持续的过程。难点在于深入理解系统瓶颈和性能瓶颈，通过优化算法、调整参数和进行负载测试，不断提升执行引擎的性能和效率。

3.2 已开展工作

前期作者已经对国内外相关项目进行了充分调研，了解了 OLAP 数据库的核心架构与设计模式。对于使用 C++语言开发项目，掌握了充分的性能调优经验和技巧。

对于选择恰当的 OLAP 数据库，作者做了充分调研，对 Apache Doris 这款数据库有了充分的了解，掌握了进行执行引擎改造的必要基础。

4. 完成本课题的工作方案及进度计划（按周次填写）

2022 年 15-18 周：

1. 熟悉 Apache Doris 数据库
2. 研读论文了解 Pipeline 引擎及其优化

2023 年 1 月：

1.学习友商竞品实现

2023 年 1 月-2023 年 4 月：

1. 进行算法开发与调试
2. 完成中期答辩

2023 年 4 月-2023 年 6 月：

1. 完成算法调试与改进工作
2. 进行详细性能与稳定性测试

2.撰写毕业论文

3.毕业论文答辩

注：1、正文：宋体小四号字，行距 22 磅。

2、开题报告装订入毕业设计（论文）附件册。

参考文献

[1] 前瞻产业研究院. 2019 中国大数据行业研究报告. http://pdf.dfcfw.com/pdf/H3_AP201911251371103072_1.pdf, 2019

[2] 中国信息通信研究院. 大数据白皮书. http://www.caict.ac.cn/english/research/whitepapers/202303/P020230316608528378472.pdf, 2023-01

[3] 一个会写诗的程序员. 主流的 OLAP 引擎介绍. https://cloud.tencent.com/developer/article/1924583, 2021-12-24

[4] Apache Doris. Doris 简史-为分析而生的 11 年. https://xie.infoq.cn/article/4bdf3da72bc868ad78cf6bf4b, 2021-03-24

[5] 魏祚. 最佳实践: Apache Doris 在小米数据场景的应用实践与优化. https://doris.apache.org/zh-CN/blog/xiaomi_vector/#新旧架构性能对比, 2022-12-08

[6] 梁程加, 陈俊云, 许英博. 数据库：企业数字化支撑，大数据时代基石. 中信证券前瞻研究系列报告, 2021(86)

[7] 中金公司. 数智中国之二：数据库商业市场五问五答. https://research.cicc.com/frontend/recommend/detail?id=3100, 2022-06-08

[8] 朱良. Apache Doris 在美团外卖数仓中的应用实践. https://tech.meituan.com/2020/04/09/doris-in-meituan-waimai.html, 2022-04-09

[9] Apache Doris. Doris 介绍. https://doris.apache.org/zh-CN/docs/dev/summary/basic-summary/, 2023

[10] QIN. 物理执行引擎之火山引擎. https://www.qin.news/wu-li-zhi-xing-yin-qing-zhi-huo-shan-yin-qing/, 2022-02-20

[11] OceanBase. 数据库查询引擎的进化之路. https://zhuanlan.zhihu.com/p/41562506, 2018-08-15

[12] caroly. 分布式数据库（九）. https://caroly.fun/archives/分布式数据库九, 2021-05-15

[13] Viktor Leis, Peter Boncz, Alfons Kemper, and Thomas Neumann. Morsel-driven parallelism: a NUMA-aware query evaluation framework for the many-core age. 2014 ACM SIGMOD International Conference, 2014, 743–754.

[14] G. Graefe. Volcano— An Extensible and Parallel Query Evaluation System. IEEE Trans, 1994, 120–135.

[15] G. Graefe,W. J. McKenna. The Volcano optimizer generator: extensibility and efficient search. IEEE 9th International Conference on Data Engineering, 1993, 209-218

[16] Thomas Neumann. . Efficiently compiling efficient query plans for modern hardware. VLDB Endow, 2011, 539-55

西安工业大学毕业设计（论文）开题检查表

计算机科学与工程学院

<table>
<tr><td>姓名</td><td>赵长乐</td><td>班级</td><td>1806
0314</td><td>学号</td><td>18030513
121</td><td>专业名称</td><td>信息对抗技术</td></tr>
<tr><td>设计（论文）题目</td><td colspan="3">大规模分布式分析型数据库中的 Pipeline 执行引擎实践研究</td><td>检查方式</td><td colspan="3">☐ 开题答辩
☐ 审阅开题报告</td></tr>
<tr><td rowspan="6">检查内容</td><td>选　题</td><td colspan="6">☐ 合理　☐ 建议修改</td></tr>
<tr><td>工作量</td><td colspan="6">☐ 过大　☐ 饱满　☐ 偏少</td></tr>
<tr><td>设计任务理解</td><td colspan="6">☐ 全面理解　☐ 基本理解　☐ 未完全理解</td></tr>
<tr><td>设计方案</td><td colspan="6">☐ 合理　☐ 可行　☐ 不可行</td></tr>
<tr><td>设计内容及表达</td><td colspan="6">☐ 准确　☐ 基本准确　☐ 部分正确</td></tr>
<tr><td>工作态度</td><td colspan="6">☐ 认真　☐ 一般　☐ 不认真</td></tr>
<tr><td>检查结论</td><td colspan="7">☐ 同意开题
☐ 不同意开题
检查人（小组成员）签字:
年　月　日</td></tr>
<tr><td>建议与要求</td><td colspan="7"></td></tr>
</table>

注：1 “检查方式”“检查结论”栏内可在相应方框内划“√”。2　本表装订入附件册。

西安工业大学
XI'AN TECHNOLOGICAL UNIVERSITY

毕业设计(论文)中期报告

题目：大规模分布式分析型数据库中的Pipeline执行引擎实践研究

院　（系）：计算机科学与工程学院

专　　业：计算机科学与技术

班　　级：18060314

学　　生：赵长乐

学　　号：18030513121

指导教师：杨国梁

2023年3月25日

撰写内容要求（可加页）：

1. 设计（论文）进展状况

当前，作者已经按预期进度完成了论文研究工作，主要进展如下：

1. 对 Apache Doris 数据库进行了熟悉，对该项目进行了功能贡献与 bug 修复。功能贡献主要集中在执行引擎向量化性能提升方面。成为了 Apache Doris 社区 contributer。

2. 作者调研了 Pipeline 引擎的实现，主要参考了 Starrocks 和 Clickhouse 的相关代码实现，对他们的 Pipeline 引擎实现方式进行了详细的调研，摸排了它们的性能表现，对其代码进行了细致研读，产出了大量调查报告。在这个基础上，作者已经完成了对一款合适的 Pipeline 引擎的设计。

3. 具体开发上，Pipeline 引擎的雏形已经设计完成，对其进行了性能测试。测试结果表明，该引擎在 Apache Doris 数据库上取得了预期表现，明显超过了现有的 Volcano 引擎性能。

2. 存在问题及解决措施

目前，开发工作上主要遇到的问题是，由于引入了 Pipeline 引擎的调度机制与额外的抽象层，代码的调试难度显著上升，出现了大量多线程同步问题。这一类问题难于调试是项目的主要痛点。

目前规划的解决方案如下：

1. 多与开发社区进行交流，共同协作解决问题
2. 与 Clickhouse、Impala 等社区保持合作，积极寻求国外开发者的帮助
3. 为 Pipeline 引擎增加更多的 Tracing 机制，使 debug 过程更加顺畅
4. 为 Pipeline 引擎增加 Profiling 机制，方便使用者进行调试

3. 后期工作安排

1. 继续按进度进行 Pipeline 引擎的开发。
2. 继续关注友商进展，吸取它们的技术闪光点。
3. 补足相关的测试 case，完成性能保障工作。
4. 保证 Pipeline 引擎能够通过 P0 测试，尽快上线 P1 测试及 fuzzy 测试，在相关测试中解决对应问题。在结题前保证 Pipeline 引擎能够通过长稳测试及压力测试。

注：1、正文：宋体小四号字，行距 22 磅。

2、中期报告装订入毕业设计（论文）附件册。

西安工业大学毕业设计（论文）中期检查表

计算机科学与工程学院

<table>
<tr><td>姓名</td><td>赵长乐</td><td>班级</td><td>18060314</td><td>学号</td><td>18030513121</td><td>专业名称</td><td>信息对抗技术</td></tr>
<tr><td>设计（论文）题目</td><td colspan="7">大规模分布式分析型数据库中的 Pipeline 执行引擎实践研究</td></tr>
<tr><td rowspan="3">资料情况</td><td colspan="2">选题是否变化</td><td colspan="5">☐ 有 ☐ 无</td></tr>
<tr><td colspan="2">中期报告</td><td colspan="5">☐ 有 ☐ 无</td></tr>
<tr><td colspan="2">外文翻译</td><td colspan="5">☐ 优 ☐ 良 ☐ 中 ☐ 差</td></tr>
<tr><td>工作进度</td><td colspan="7">☐ >60% ☐ 40~60% ☐ 30~40% ☐ <30%</td></tr>
<tr><td>工作态度</td><td colspan="7">☐ 认真 ☐ 一般 ☐ 不认真</td></tr>
<tr><td>工作质量</td><td colspan="7">☐ 优 ☐ 良 ☐ 中 ☐ 差</td></tr>
<tr><td>检查结论</td><td colspan="7">☐ 优秀 ☐ 通过 ☐ 警告 ☐ 终止毕业设计（论文）</td></tr>
<tr><td colspan="8">存在的问题与建议：

指导教师（签名）：

年 月 日</td></tr>
</table>

注：1 指导教师在相应项目方框内划“√”。

2 中期检查结果应与是否有资格参加答辩相挂钩。

3 本表装订入毕业设计（论文）附件册。

西安工业大学毕业设计（论文）指导教师评分表

计算机科学与工程学院

姓名	赵长乐	班级	18060314	学号	18030513121	专业名称	信息对抗技术
毕业设计（论文）题目			大规模分布式分析型数据库中的Pipeline执行引擎实践研究				

序号	评价项目（分值）	评价内容	评分
1	文献应用（10）	查阅中英文文献，了解研究动态及发展趋势；结合文献进行研究、分析，得到有效结论的能力。	
2	问题分析（10）	综合运用所学知识对复杂工程问题进行分析，并获得有效结论。	
3	设计能力（30）	综合运用所学知识，结合工程实际，设计完成任务书规定的软件、硬件及计算机相关系统；设计中体现创新意识。	
4	工具使用（10）	能够选择与使用恰当的技术、现代工程工具和信息技术工具进行设计、分析、研究复杂工程问题。	
5	工程与社会（5）	具有社会责任感，理解并遵守工程职业道德及规范；正确认识、评价设计中涉及到的社会与文化、安全与健康、环境与法律等因素的能力。	
6	沟通能力（20）	具备运用专业知识进行书面或口头沟通的能力：包括规范撰写毕业论文、英文摘要，翻译英文文献；按照专业规范编制相关技术文档。	
7	项目管理（5）	具备项目管理能力和团队协作精神，并能够在多学科环境中应用。	
8	学习能力（10）	具有自主学习和终身学习的能力，能够有效解决工程实际问题。	
总　分			

评语：

结论：☐ 同意按期答辩　　☐ 不同意答辩

指导教师 ：

年　月　日

西安工业大学毕业设计（论文）评阅教师评分表

计算机科学与工程学院

姓名	赵长乐	班级	18060314	学号	18030513121	专业名称	信息对抗技术
毕业设计（论文）题目			大规模分布式分析型数据库中的 Pipeline 执行引擎实践研究				

序号	评价项目（分值）	评价内容	评分
1	文献应用（10）	查阅中英文文献，了解研究动态及发展趋势；结合文献进行研究、分析，得到有效结论的能力。	
2	问题分析（10）	综合运用所学知识对复杂工程问题进行分析，并获得有效结论。	
3	设计能力（35）	综合运用所学知识，结合工程实际，设计完成任务书规定的软件、硬件及计算机相关系统；设计中体现创新意识。	
4	工具使用（10）	能够选择与使用恰当的技术、现代工程工具和信息技术工具进行设计、分析、研究复杂工程问题。	
5	工程与社会（5）	具有社会责任感，理解并遵守工程职业道德及规范；正确认识、评价设计中涉及到的社会与文化、安全与健康、环境与法律等因素的能力。	
6	沟通能力（20）	具备运用专业知识进行书面或口头沟通的能力：包括规范撰写毕业论文、英文摘要，翻译英文文献；按照专业规范编制相关技术文档。	
7	项目管理（5）	具备项目管理能力和团队协作精神，并能够在多学科环境中应用。	
8	学习能力（5）	具有自主学习和终身学习的能力，能够有效解决工程实际问题。	
总　分			

评语：

结论： ☐ 同意按期答辩　　☐ 不同意答辩

评阅人 ：

年　　月　　日

西安工业大学毕业设计（论文）答辩暨综合评分表

计算机科学与工程学院

姓名	赵长乐	班级	18060314	学号	18030513121	专业名称	信息对抗技术
毕业设计（论文）题目			大规模分布式分析型数据库中的 Pipeline 执行引擎实践研究				

序号	评价项目 (分值)	评价内容	评分
1	文献应用（10）	查阅中英文文献，了解研究动态及发展趋势；结合文献进行研究、分析，得到有效结论的能力。	
2	问题分析（10）	综合运用所学知识对复杂工程问题进行分析，并获得有效结论。	
3	设计能力（30）	综合运用所学知识，结合工程实际，设计完成任务书规定的软件、硬件及计算机相关系统；设计中体现创新意识。	
4	工具使用（5）	能够选择与使用恰当的技术、现代工程工具和信息技术工具进行设计、分析、研究复杂工程问题。	
5	工程与社会（5）	具有社会责任感，理解并遵守工程职业道德及规范；正确认识、评价设计中涉及到的社会与文化、安全与健康、环境与法律等因素的能力。	
6	沟通能力（30）	具备运用专业知识进行书面或口头沟通的能力：包括规范撰写毕业论文、英文摘要，翻译英文文献；按照专业规范编制相关技术文档。	
7	项目管理（5）	具备项目管理能力和团队协作精神，并能够在多学科环境中应用。	
8	学习能力（5）	具有自主学习和终身学习的能力，能够有效解决工程实际问题。	
总　分			

毕业设计（论文）综合成绩：_____分，成绩评定结论：_____

答辩委员会（小组）负责人（签字）：

答辩委员会（小组）成　员（签字）：

年　　月　　日