北京大学 PostgreSQL 内核开发从入门到进阶 V6.2

世界最强大的开源关系型数据库

PostgreSQL

CHINA POSTGRESQL ASSOCIATION

目录

PostgreSQL 数据加密与存取开源开发实践

项目概要

主要人员：

PostgreSQL 分会社区技术专家 1 人、瀚高研发工程师 2 人、北大学生小组预估 6 人以上；

项目周期：

16 周（时间自主，建议每周固定时间，每周至少保证 2 小时）

开展形式：

企业资深研发人员与学生以线上协作为主，推进学习计划及作业打分，鼓励学员参与开源 IvorySQL 数据库项目和参与 PostgreSQL 国际社区贡献。

注：IvorySQL 项目是一个具有广泛生态基础和中国特色的 PG 开源衍生项目，是瀚高股份设计研发的一款具备强大 Oracle 兼容能力的开源数据库。具备高兼容性和高可用性，并致力于遵守'the open-source way'。

官方网址：https://www.ivorysql.org/zh-cn/

社区仓库：https://github.com/IvorySQL/IvorySQL

背景意义

本次课程在前两次课程基础上进行了优化，面向**开源 PostgreSQL 内核开发从入门到进阶**组织主题内容，并将优化后的内容作为 PostgreSQL 内核开发培训认证体系发布运营，助力数据库内核开发人才培养。课程同时保留 **PostgreSQ 数据存取、PostgreSQ 数据加密**主题内容，同学们能够在线跟随 PG 社区专家讲师在线授课的同时，自主选择观看前两期积累的课程视频。

企业导师

吕海波

联系方式：lvhb@mchz.com.cn

	杭州美创科技技术研究院研究员，"IT 老兵，国内最早一批 PG ACED 之一，27 年 IT 软件领域从业经历，近二十年数据库经验，惯看 IT 江湖风起云涌。曾在中国、美国多家巨头型互联网公司（阿里巴巴、京东、ebay）从事数据库管理与研究工作。 多次在 PG Conf.ASIA、中国数据库大会（DTCC）等技术大会中，以个人、独立身份发表演讲。 在 2019 DTCC DM8 数据库发布会中，作为国内企业界杰出数据库专家代表，与中国工程院院士倪光南、方滨兴共话中国数据库技术的自主可控话题。 出版技术书籍《Oracle 内核技术揭密》，被誉为国内最深度解密 Oracle 算法原理的技术书籍。 现从事计算机体系结构与数据库内核开发的融合领域。

David Zhang

联系方式：david.zhang@highgo.ca

	David 是瀚高软件北美研究院（加拿大）的资深软件架构师，在加入瀚高之前，他在智能电网，计量领域，网络安全，资安方面开发创新软件解决方案已有 20 年以上的行业经验。David 于加拿大滑铁卢大学（UW）获得电气与计算机工程硕士学位，并在以下技术方面拥有丰富的实践经验：网络安全，数据安全，身份验证，软件设计与开发，PostgreSQL 数据库功能及内核开发，嵌入式系统，功能和体系结构设计等

Cary Huang

联系方式：cary.huang@highgo.ca

	Cary 是瀚高软件北美研究院（加拿大）的高级软件开发人员，在加入瀚高之前，他在智能电网和计量领域以 C / C ++开发创新软件解决方案已有 8 年以上的行业经验。 他于 2012 年在加拿大温哥华的英属哥伦比亚大学（又译"不列颠哥伦比亚大学"，UBC）获得电气工程学士学位，并在以下技术方面拥有丰富的实践经验：高级网络，网络和数据安全性，智能计量创新，Docker 部署管理，软件工程生命周期，拓展，身份验证，加密，PostgreSQL 和非关系数据库，Web 服务，防火墙，嵌入式系统，RTOS，ARM，PKI，Cisco 设备，功能和体系结构设计等。

实践安排

方向一：PostgreSQ 数据存取（2 课时/每周*16 周）

（本部分是 2023 年 PostgreSQL 内核开发实践课主要内容，由中国 PostgreSQL 分会社区专家吕海波老师直播授课，其内容已作为 PostgreSQL 内核开发培训认证课程体系发布，涵盖 PG 内核开发初中级内容，具体可参见课程计划章节。相关文章：*https://mp.weixin.qq.com/s/ch1mzYnYTUfQziR5-Pebvg*）

课题概要：

数据库 I/O 读写最基本单元是 page，page 大小通常是 8K、16K 等数值。但 OS 中 page 大小是 4K。数据库的 page 大小是 OS page 的两倍或更多倍，这带来一个经典的 partial write 问题（有些资料也称为页断裂）。各种数据库为了解决 partial write 各出奇招。比如，MySQL 使用了 double write 特性（双写）。PostgreSQL 使用了 Full Pages Write（简称 FPW，整页写）。

PostgreSQL FPW 特性能很好的解决 partial write 问题，但是 FPW 同时带来了巨大的性能负担，使 PostgreSQL 性能下降至少 30%。在写较多且 I/O 是瓶颈的环境中，FPW 甚至可以使性能下降 200%。MySQL 的 double write 也有同样的问题，目前主流保证数据一致性的、完整性的数据库中，只有 Oracle 较好的解决了这一问题，既避免 partial write 又对性能没有影响。

PostgreSQL Non-FPW 就是针对这一情况，使用类似 Oracle 的方式，在百分百保证数据一致性、完整性的条件下，既去除了 PostgreSQL 的 FPW 特性，又解决了 partial write 问题，使数据库性能大幅提升。

内容概要：

1. 介绍 PostgreSQL 源码编译、代码调试技术、内核开发流程、工具；启动过程、进程跟踪。
2. 介绍 SQL 执行流程跟踪、事务流程跟踪。
3. 学习 XLOG、VACUUM 流程跟踪，阅读理解源代码：PG 内存池与内存上下文原理。
4. 学习如何跟踪 MemoryContext 创建流程、分析其数据库结构、相关代码原理。
5. 阅读理解源代码：ReadBuffer（）与 Shared Buffer、ReadBuffer()与逻辑读原理；分析 HASH 表数据结构。
6. 讲解分析 Shared Buffer 相关数据结构；源代码阅读理解方法论总结；PG 插件扩展方式及 pg_stat_statements 插件代码分析。
7. 了解数据库开发前沿方向、基础技术；学习 SQL 性能分段统计、功能模块的分阶段性能分析。
8. 学习 Patch 开发：MVCC 核心增强的背景知识；熟悉数据库可见性原理分析。

9. 学习快照的作用，各种数据库快照的实现方式。
10. 学习数据库的可见性原理，快照如何可控制可见性。
11. 讲解 27 种可见性模拟与原理分析：11-27 种可见性模拟；代码阅读理解：快照创建 GetSnapshotData() 函数源码阅读理解与逻辑分析。
12. 熟悉使用分段式量化分析框架，寻找 PG MVCC 过程中的性能缺陷点
13. 学习总结事务 commit 详细过程、流程图与相关数据结构
14. 学习修改 ProcArrayEndTransaction()等函数，增加 commit_id 与事务数组逻辑；修改 GetSnapshotData() 逻辑，将“遍历 Session”循环，移至 ProcArrayLock 锁之外。
15. 讲解计算机体系结构：CPU Cache MESI 协议与 fence 类指令在 PG 中的深度应用，修改 PG 源码，调试编译错误。
16. 学习修复 BUG，调试无法启动数据库错误，验证效果、设计测试方案、分析未来的设计方案。

方向二：PostgreSQ 数据加密（2 课时/每周*12 周）

（本部分为拓展可选内容，课程计划中未做体现，可自主在线学习，通过邮件与导师沟通答疑，课程内容是 2021 年录播课程，已上传 bilibli 平台，链接地址：*https://space.bilibili.com/521087625/channel/seriesdetail?sid=2069368*）

课题概要：

PostgreSQL 国际社区一直在进行讨论是否以及如何在 Postgres 中**实现透明数据加密（TDE）**。 许多其他关系型数据库都支持 TDE，并且某些安全标准也要求它。

TDE 最主要的功能就是用来保护存储在磁盘上的数据来避免恶意人员直接读取或窃取数据库文件，或是偷走整个磁盘导致用户信息遭窃。

目前 PostgreSQL 支持许多安全的级别来保证数据安全，比如支持 TLS 网络链接加密来保证客户端和 PostgreSQL 服务器间的网络安全，和强大的用户验证功能来保证数据库用户的真实性，但是 PostgreSQL 缺乏服务器和磁盘间存储的安全加密，这也是为什么 TDE 目前在 PostgreSQL 国际社区里被重视的原因。

内容概要：

1. 讲解 PostgreSQL 数据库系统架构、开发环境及开发流程、编译及调试、源代码管理。
2. 讲解加密和算法的基础知识和基本的安全概念。

3. 介绍目前 PostgreSQL 支持的安全功能与实践。
4. 学习当前 PostgreSQL 国际社区对 TDE 开发的内容及流程。
5. 与 PostgreSQL 国际社区核心成员进行在线交流，讨论国际社区 TDE 方面的进展
6. 了解当前 PostgreSQL 对数据的处理，存储及调用的工作原理代码。
7. 在现有的 PostgreSQL 的基础上，设计，研究并开发 TDE 功能以及代码。
8. 学习如何把自己的工作成果分享给 PostgreSQL 国际社区。

课程计划

以下 2023 年 PostgreSQL 内核开发实践课内容，主要面向方向一：PostgreSQL 数据库存取组织课程内容，由中国 PostgreSQL 分会社区专家吕海波老师直播授课。

第一周：

- PG 源码编译与选项、代码调试技术、PG 内核开发流程、PG 内核开发工具使用方法
- PG 启动过程跟踪、后台进程启动跟踪、backend 进程启动跟踪

作业：

- 配置自己研究环境
- 总结流程图：PG 启动时各进程启动顺序与启动条件

第二周：

- SQL 执行流程跟踪、
- 事务流程跟踪

作业：

- 总结 SQL 执行过程的流程图，标记关键步骤和函数
- 总结事务流程图，标记关键步骤和函数

第三周：

- XLOG 流程跟踪
- VACCUM 流程跟踪
- 代码阅读理解一：PG 的内存池与内存上下文原理

作业：

- 总结 XLOG 和 VACCUM 流程图，标记关键步骤和函数
- 设计内存测试程序，触发 TLB Miss，对比大页与普通页对 PG 的性能影响

第四周：

- 跟踪 MemoryContext 创建流程
- 分析 MemoryContext 的数据结构
- 总结 MemoryContext 相关代码原理

作业：

- 总结 MemoryContext 相关数据结构关系图
- 总结 NUMA 对 PG 性能影响

第五周：

- 代码阅读理解二：ReadBuffer（）与 Shared Buffer、ReadBuffer()与逻辑读原理
- 分析 HASH 表数据结构

作业：

- 设计测试程序，对比 HASH Bucket 数量对整体性能影响

第六周：

- 分析 Shared Buffer 相关数据结构
- 代码阅读理解方法论总结
- PG 的插件式扩展、pg_stat_statements 插件代码分析

作业：

- 修改 PG 源码中 HASH Bucket 数量，重新编译 PG 源码，并调试错误

第七周：

- 数据库开发前沿方向
- 基础技术
- SQL 性能的分段统计、功能模块的分段性能分析

作业：

- 使用动态分析语言（eBPF、Systemtap 等皆可），开发性能分析程序，创建分段式量化分析框架

第八周：

- Patch 开发：MVCC 核心增强的背景知识
- 数据可见性原理分析

作业：

- 使用 SQL，复现 PG 中常见可见性测试（提供测试 SQL）。

第九周：

- 什么是快照
- 快照的作用
- 各种数据库快照的实现

作业：

- 分析 PG/Oracle/MySQL 快照实现的优劣

第十周：

- 数据库的可见性
- 快照如何控制可见性
- 27 种可见性模拟与原理分析：1-10 种可见性模拟

作业：

- 按 PPT 步骤，复现 1-10 种可见性测试，记录并提交测试日志，做为作业结果

第十一周：

- 27 种可见性模拟与原理分析：11-27 种可见性模拟
- 代码阅读理解：快照创建 GetSnapshotData()函数源码阅读理解与逻辑分析

作业：

- 按 PPT 步骤，复现 11-27 种可见性测试，记录并提交测试日志，做为作业结果

第十二周：

- MVCC 模块性能建模
- 热点代码与竞争点分析
- 确认 MVCC 核心模块改进可行性

作业：

- 使用分段式量化分析框架，寻找 PG MVCC 过程中的性能缺陷点

第十三周：

- 事务 commit 流程分析
- 提交重点函数 ProcArrayEndTransaction()代码阅读理解、逻辑分析

作业：

- 总结事务 commit 详细过程、流程图与相关数据结构

第十四周：

- 确认代码修改方案
- 增加事务数组的逻辑
- 增加 commit_id 的逻辑

作业：

- 修改 ProcArrayEndTransaction()等函数，增加 commit_id 与事务数组逻辑
- 按课程思路，修改 GetSnapshotData()逻辑，将“遍历 Session”循环，移至 ProcArrayLock 锁之外。

第十五周：

- 计算机体系结构：CPU Cache MESI 协议与 fence 类指令在 PG 中的深度应用
- 修改 PG 源码
- 调试编译错误

作业：

- 设计程序，测量 CPU 中某 Core 修改变量后，另一 Core 多少周期后可以读取到前一 Core 修改过的变量

第十六周：

- 修复 BUG
- 调试无法启动数据库错误
- 验证效果、设计测试方案、分析未来的设计方案

大作业：

- 完成为 PG 的提交增加 commit_id、并修改 GetSnapshotData()逻辑、增加事务数组，去除 Session 数量对并发提交的影响，并重编译 PG 源码、调试通过。

评分标准

- 出勤：10%
- 课后作业：50%
- **大作业：**Non-FPW 代码实现　40%

预习内容

1. Linux 基础

   Ubuntu Linux 18.04：https://releases.ubuntu.com/18.04/

   Vmware player：*--可选择 vmware，资源站点章节有分享安装包*

   https://www.vmware.com/ca/products/workstation-player/workstation-player-evaluation.html

   Virtualbox：https://www.virtualbox.org/

   Ubuntu 官方 Linux 指令指南：*--可选择 centos7 或 8，资源站点章节有分享安装包*

   https://ubuntu.com/tutorials/command-line-for-beginners#6-a-bit-of-plumbing

2. PostgreSQL 预习

   PostgreSQL github 官方代码下载：https://github.com/postgres/postgres

   PostgreSQL 配置：https://www.postgresql.org/docs/current/runtime-config.html

   https://www.postgresql.org/docs/current/auth-PostgreSQL-hba-conf.html）

   基本 PostgreSQL 指南：https://www.postgresqltutorial.com/

   PostgreSQL 内部架构：http://www.interdb.jp/PostgreSQL/

   PostgreSQL 官方 TDE 和 KMS wiki page：

   https://wiki.postgresql.org/wiki/Transparent_Data_Encryption

3. SQL 语言常用操作
4. GDB 使用
5. 主流数据库 I/O 与 partital write

   https://www.percona.com/blog/2006/08/04/innodb-double-write/

   《Oracle 内核技术揭密》第三章：Buffer Cache 内部原理与 I/O

   《Oracle 内核技术揭密》第五章：Redo 调优与备份恢复原理

资源站点

1. PostgreSQL 国际社区网站： https://www.postgresql.org/
2. PostgreSQL 分会问答网站：https://www.pgfans.cn
3. PostgreSQL 分会资源网站：https://www.postgreshub.cn
4. 瀚高软件官方网站 ：https://www.highgo.com/
5. 瀚高软件北美研究院网站：https://www.highgo.ca/
6. **百度网盘安装软件分享：**

   链接：https://pan.baidu.com/s/1o6waCReP9278g_djqRxiCg

   提取码：fbbj
7. **2021-2022 年课程在线视频：**

   https://space.bilibili.com/521087625/channel/seriesdetail?sid=2069368