

PostgreSQL并发处理方式——MVCC



ZYJ2016 (/u/d1ef74921117) [+ 关注](#)

(/u/d1ef74921117)

PostgreSQL的特色之一是它的并发控制机制，在维护一致性和完整性的同时，尽量避免读写的堵塞。

对于传统数据库，为了维护一致性和完整性，避免一个事务看到其它并发事务更新而到会不一致的数据，通常采用的是LOCK机制。这样付出的代价是，当锁请求无法被响应时，待处理的请求必须进入等候队列，甚至等待超时不被处理。

MVCC通过避开传统数据库的LOCK机制，最大限度的减少锁竞争以允许合理的多用户环境中的性能。

恰当地使用MVCC总会提供比LOCK更好的性能。对于那些无法轻松接收MVCC行为的应用，PostgreSQL也提供了表和行级别的LOCK机制。

(<https://doi.org/10.1002/for.2596>)

PostgreSQL存储结构

PostgreSQL中，一个表对应一个逻辑文件，一个表被分割成若干个物理段文件（relation segment），除最后一段外默认大小40M。

文件页（磁盘块）是物理段文件的基本储存单位，也是内存和磁盘交换的单位。文件页大小限制了表元组的大小并影响磁盘操作效率，缺省大小8192字节，最大可设置为 2^{15} 字节（这是由磁盘块索引是15位决定的）。

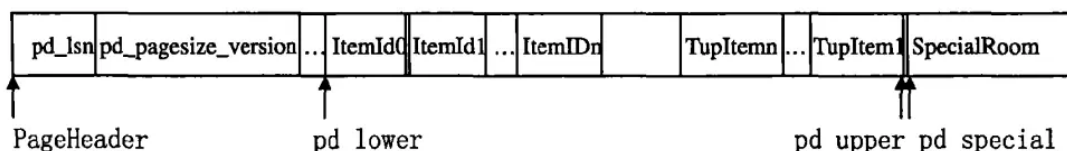
一个文件页空间被逻辑分割为三个部分：

- PageHeader：页描述区
- 记载页的使用情况, 如页分布格式版本, 元组数据空间和特殊空间的起始位置以及文件页相关的事务日志记载点等信息
- Tuple Item space：元组数据空间
- 实际记录元组数据的地方
- Special space：特殊空间

每个记录的元组 (Tuple) 称为一项，每项由描述ID和元组数据构成。

项描述ID描述了元组存储位置，大小以及一些状态标识。

项描述ID和项数据分别在元组数据空间的两头往中间存放，最早的项存在最两侧，越晚的数据越靠中间。



元组的写过程：先写到文件页的内存缓冲区（ Buffer ），再更新到磁盘中

- 从缓冲页的元组数据存储区分配空间
- 构造元组描述ID，写入低端处
- 把实际数据写到高端处，并设置缓冲区的脏标记
- 更新到磁盘
- 写元组不会立即更新到磁盘，而是推迟到所在的缓冲区被替换（ Replace ）时进行
- Replace时，判断缓冲区是否脏：
- 如果脏，启动实际磁盘IO进行写；
- 如果不脏，直接回收再用该Buffer。

文件页的写过程：

- 先更新该文件页的事务日志，事务日志由页头部的页描述符指出
- 把文件缓冲页写到指定磁盘块

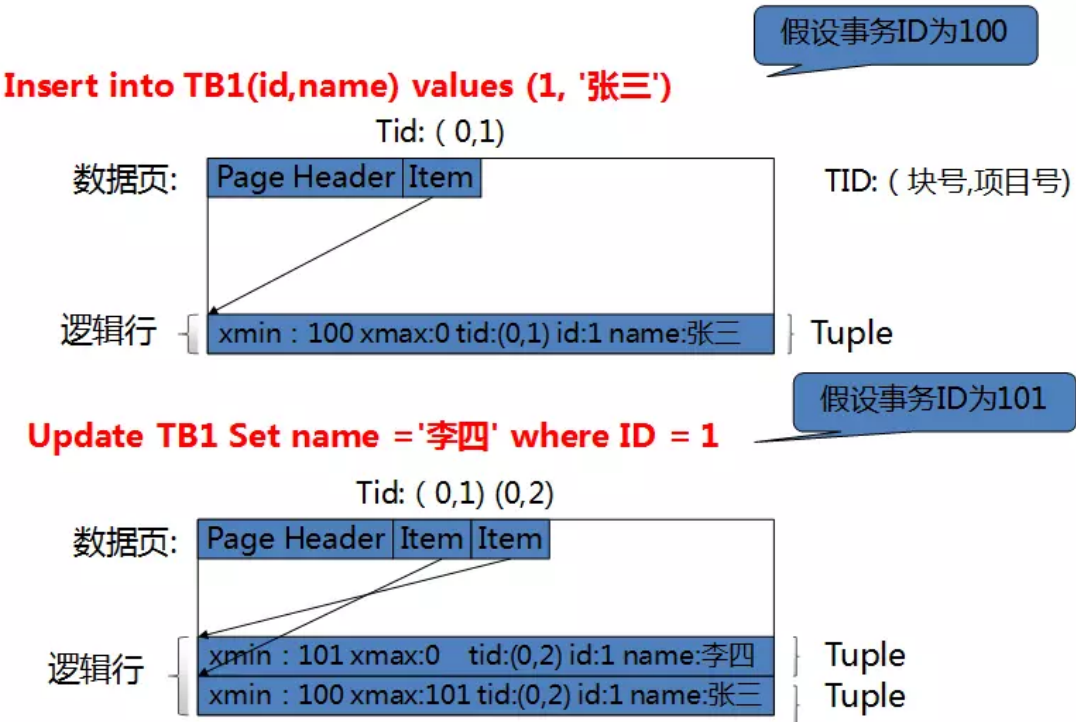
MVCC

MVCC(Multiversion Concurrency Control)，多版本并发控制。

举一个简单的例子来理解它的机制

(https://dsp
click.youde
slot=30edc
1950-4f20-
4db06549
59683067`

```
inset into T1(id,name) values (1,'zhangsan');  
updata T1 set name = 'lisi' where id =1;
```



- 每个事务都会得到一个XID（称为事务ID），当一个新事务开始，递增XID，然后把它赋予这个新事务。
- 把一个元组（ Tuple ）称作同一逻辑行的一个**行版本**，数据文件中存放同一逻辑行的多个行版本
- 每个行版本的头部，记录该行版本的创建和删除的事务ID（分别称为xmin和xmax）
- 每个事务的状态（ running, abort 或 commit ）记录在pg_clog文件中
- 运用**一定的规则**，使每个事务只会看到一个特定的行版本（快照）

(/apps/redi
utm_sourc
banner-clic

举个例子，当 insert 一行记录时，只有那些已提交的、并且xmin比当前事务XID小（ xmin<XID ）的行记录 对当前事务才是可见的。

这意味着你可以创建一个新事务然后插入记录，直到commit之前，这些记录对其他事务永远都是不可见的；commit之后，其他后创建的新事务就可看到这行新记录了（ xmin<XID ）。

对于 delete 和 update ，机制也是类似的，不同的是要用xmax值来判断数据的可见性。

隔离级别

SQL标准定义了四个级别的事务隔离。最严格的是**可串行化**，是通过标准定义，即保证并发执行和顺序执行的结果相同；其他三个级别是通过现象定义的。

(https://dsp
click.youda
slot=30edc
1950-4f20-
4db065493
59683067`

隔离级别	脏读	不可重复读	幻读
读未提交（ read uncommitted ）			
读已提交（ read committed ）	避免		
可重复读（ repeatable read ）	避免	避免	
可串行化（ serializable ）	避免	避免	避免

- 幻读：重新执行一个查询，由于最近另一个事务的提交，返回的结果（一批数据）和刚才不同；
- 不可重复读：针对同一个数据，一个事务内多次查询，由于期间另一个事务的提交，导致结果（同一个数据）不同；
- 脏读：一个事务读取了另一个事务还未提交的改动。

PostgreSQL中：

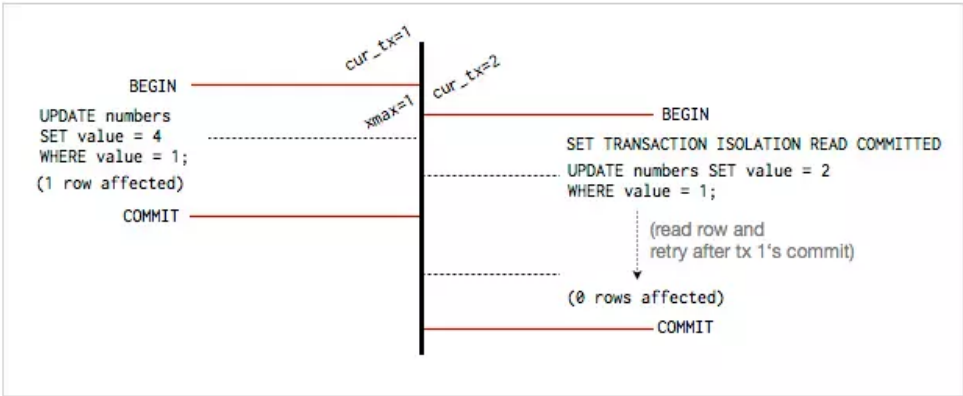
- 默认隔离级别是**读已提交**（ read committed ）；
- 可以请求四中级别的任意一种，但对于内部其实只有读已提交、可重复读、可串行化三种级别；
- 选择读未提交时，实际上用的是读已提交；
- 选择重复读时，不会发生幻读；

SQL标准只定义了那种现象不能发生，但是没定义哪种现象一定发生。



- 读已提交
- BEGIN TRANSACTION ISOLATION LEVEL READ COMMITTED;
- PostgreSQL里的缺省隔离级别

- 看到的是**当前查询开始时的快照**
- 当有两个事务同时修改同一行数据时，后发生的事务在初始事务提交前就可以进行查找，然后不执行进入等待，待初始事务提交后retry，检验查找条件是否仍然满足，如果满足，后续操作才会被执行；
- 例如下图中的例子：
在事务1提交之前，通常LOCK机制会让事务2进入等待，到事务1提交后才可以扫描查找；
而MVCC允许事务2在事务1提交之前就可以进行扫描查找工作，当事务1提交之后，事务2retry，检验where条件是否仍然满足；
若不满足，则不会执行任何操作（保证了一致性）；
若满足，则相比LOCK机制节省了扫描查找所消耗的时间（在LOCK机制等待commit时MVCC就开始扫描查找了）。



读已提交

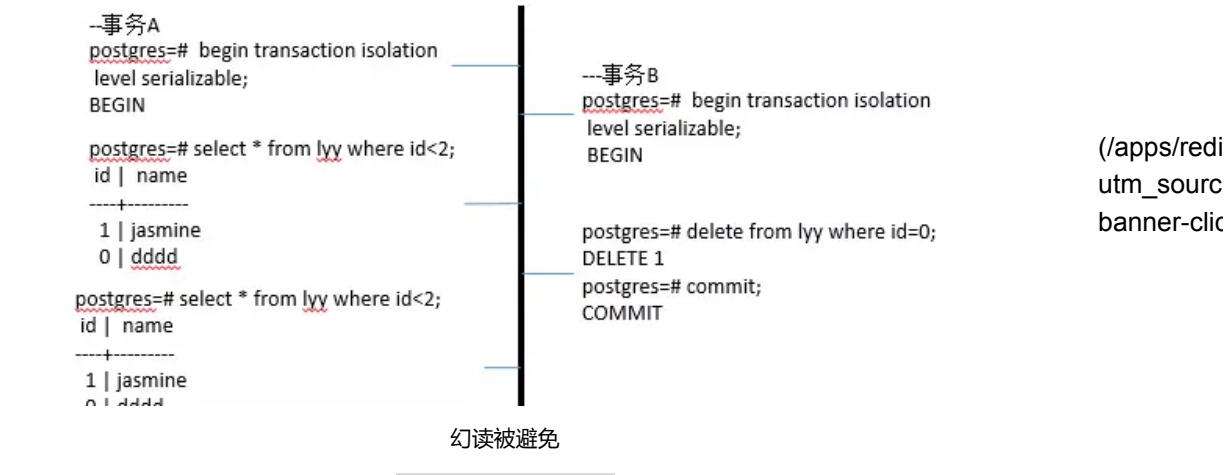
(https://dsp
click.youda
slot=30edc
1950-4f20-
4db065493
59683067'

- 可重复读
- BEGIN TRANSACTION ISOLATION LEVEL REPEATABLE READ;
- 看到的是**当前事务开始时的快照**
- 使用这个级别需要准备好重试事务，因为串行化可能失败
- 在第二张图的例子中，按照一般的LOCK机制，在可重复读的级别下，在事务B提交后，查询结果应该不同（即幻读），但是在MVCC机制中，查询结果是相同的（幻读也被避免了）



可重复读





- 可串行化
- BEGIN TRANSACTION ISOLATION LEVEL SERIALIZABLE;
- 是严格意义上的可串行化
- 可重复读级别已经避免了幻读，达到了SQL标准约定的“可串行化标准”，但能避免幻读并不等于严格意义上的可串行化
- 在这种策略下，不同事务同时修改同一数据的行为会直接失败，并返回错误信息
- 需要准备好重启事务



MVCC实现方法

MVCC的实现方法有两种：

- 写新数据是，把旧数据移到一个专门的地方（如回滚段），其他人读数据时，从回滚段中把旧数据读出来
- 写数据时，旧数据不删除，把新数据插入

PostgreSQL使用的是第二种方法，Oracle数据库和MySQL innodb引擎使用一种

比较：

- 优点：
 - 回滚可以立刻完成，无论进行了多少操作
 - 数据可以进行逆很多更新，不必担心需要保证回滚段不被用完
- 缺点：
 - 旧版本数据需要清理
 - 旧版本数据过多导致查询变慢



存在的问题及解决方法

MVCC实现了一种期待：读永远不堵塞写。但是也带来了一些问题：

- 1. 因为不同的事务会看到不同版本的记录，所以PostgreSQL连那些可能过期的数据也要保留着；
当 UPDATE 时，真正地创建了一行新记录，而 DELETE 时，并不会真正地删除一行旧记录；
最终数据库中会存在一些对有事务永远不可见的记录，称作dead rows。
- 2. 事务ID只能增加，它是个32bit，支持大约40亿个事务，达到最大值会从0重新开始；
这样带来一个逻辑问题：突然所有记录都变成了发生在将来的事务所产生的，而所有新事物也都没有办法访问这些旧记录了。

• 解决方法：VACUUM
PostgreSQL自带了auto_vacuum守护进程会在一个可配置的周期内自动执行清理，解决了这两个问题；
使用者需要留意这个auto_vacuum，以免发生不想要的结果；
vacuum命令也可以手动执行。

小礼物走一走，来简书关注我

赞赏支持

(/apps/redi
utm_sourc
banner-clip

(https://dsp
click.youda
slot=30edc
1950-4f20-
4db065493
59683067

postgres (/nb/8832913) 举报文章 © 著作权归作者所有



ZYJ2016 (/u/d1ef74921117)

写了 39525 字，被 53 人关注，获得了 51 个喜欢

(/u/d1ef74921117)


+ 关注

中科院研究生毕业，关注大数据架构，分布式系统。百度入职新人。

喜欢 | 5




更多分享



下载简书 App ▶

随时随地发现和创作内容



(/apps/redirect?utm_source=note-bottom-click)

↑





登录 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-comr


(/apps/redi
utm_sourc
banner-lic

评论

智慧如你，不想发表一点想法 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-nocomments-text)咩~

《MySQL技术内幕：InnoDB存储引擎(第2版)》书摘 (/p/3eca0b18cf51?utm...

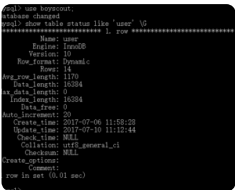
MySQL技术内幕：InnoDB存储引擎(第2版) 姜承尧 第1章 MySQL体系结构和存储引擎 >> 在上述例子中使用
了mysqld_safe命令来启动数据库，当然启动MySQL实例的方法还有很多，在各种平台下的方式可能会又...

 沉默剑士 (/u/6ca93a173ea2?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio slot=30edd


(https://dsp
click.youde
slot=30edd
1950-4f20-
4db06549?
59683067'

(/p/bb13f741a7a0?



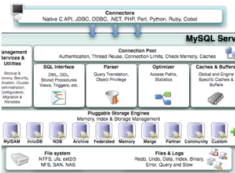
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
第一章 MySql架构与历史 (/p/bb13f741a7a0?utm_campaign=maleskine&...

为了充分发挥MySQL的性能并顺利地使用，就必须理解其设计。MySQL的灵活性体现在很多方面。例如，你
可以通过配置使它在不同的硬件上都运行得很好，也可以支持多种不同的数据类型。但是，MySQL最重要...

 李文文、 (/u/45fdd15a1b3e?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/bd8675e5c7b2?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
《高性能MySQL》 & 《MySQL技术内幕 InnoDB存储引擎》笔记 (/p/bd867...

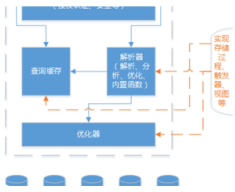
《高性能MySQL》 & 《MySQL技术内幕 InnoDB存储引擎》笔记 第一章 MySQL架构与历史 MySQL的架构
从上图可以看出，MySQL数据库区别于其他数据库的最重要的一个特点就是其插件式的表存储引擎。需要...

 xiaogmail (/u/59cac9fff87c?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio




(/p/bb6bdebf9ce?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
本来以为mysql即将消亡，没想到它却越来越强大 (/p/bb6bde9ce?utm...

(/apps/redi
utm_sourc
banner-clic

前言 前几天跟一个朋友聊天，忽然就说到IT技术上来。作为IT从业者的我，说起自己的本行自然滔滔不
绝、天花乱坠。我的这个朋友精通数据库技术和数据处理，而这两个方面正好是我的短板。说到这两个技...

 白昔月 (/u/81ade25855c1?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/9523669cc82e?

last_name	first_name
stark	tony
tom	hiddleston
morgan	freeman
jeff	dean
donald	trump

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
事务及锁的实现 (/p/9523669cc82e?utm_campaign=maleskine&utm_con...

什么是事务 事务是一条或多条数据库操作语句的组合，具备ACID，4个特点。原子性：要不全部成功，要不
全部撤销 隔离性：事务之间相互独立，互不干扰 一致性：数据库正确地改变状态后，数据库的一致性约束...

 jiangmo (/u/de31051e96e1?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio


(https://dsp
click.youde
slot=30edc
1950-4f20-
4db065493
59683067'

(/p/3fd4ecf65750?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
追随自己的本心，做自己喜欢的事 (/p/3fd4ecf65750?utm_campaign=mal...

这几日一直在研究怎么在一些网站上发表文章，发文的目的很简单，就是冲着新手期后的收益而去。这种带
有强烈目的性的写作，让自己很不安。单纯的去写文章，可以毫无顾虑，只是用笔抒发内心想说的话即可...

 二月雨丝 (/u/0ddb6e4045a?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/e03f5e5d40e5?

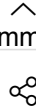


utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
花开旅行：世界顶级妖孽建筑，见过2个算你厉害 (/p/e03f5e5d40e5?utm_c...

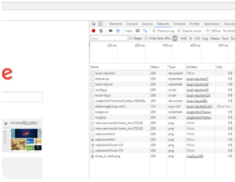
平日里见惯了高楼大厦 也没觉得它们有什么特别的地方 但在建筑界，总有一些不一样的存在 各种妖孽级别
建筑 让你大开眼界 请点击此处输入图片描述 你没有看错 这是在波兰真实存在的屋子 不是哈哈镜，也没有...

 花开旅行 (/u/1d13c045dff2?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio




(/p/d9d09d17ad0d?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
chrome调试js (zt) (/p/d9d09d17ad0d?utm_campaign=maleskine&ut...

原帖点这里平常在开发过程中，经常会接触到前端页面。那么对于js的调试那可是家常便饭，不必多说。最近一直在用火狐的Firebug,但是不知道怎么的不好使了。网上找找说法，都说重新安装狐火浏览器就可以了...

 夏的背影 (/u/ad770aadceab?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/b24fdf6f2260?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
东风悦达.起亚安全气囊不安全 突然爆炸致人受伤 (/p/b24fdf6f2260?utm_c...


本网讯7月7日，接群众举报:市民张女士在许昌市长葛市长社路正常行驶，在没有任何撞击的情况下，安全气囊突然爆炸弹出，当场造成左车窗炸飞了出去，前挡风玻璃炸裂鼓出三十厘米大包，车内物品连同车玻璃...

 人物河南 (/u/a7ef184fc285?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

多余的一节 (/p/65ce09dd4c60?utm_campaign=maleskine&utm_content...

父亲节，对有些人来讲是个极好的日子，他们多数是子女的身份。我想有句广告词或者鸡汤金句是说“父亲节是所有爸爸的生日”。但对于不需要特意过父亲节的人，这一天就像是陌生人的生日一样不重要。此外，对...

 叶武青城 (/u/1df511cb9388?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/apps/redi
utm_sourc
recommendatio
banner-clic

(https://dsp
click.youda
slot=30edc
1950-4f20-
4db065493
59683067`