**Физическая организация БД**

Объекты БД в целом имеют схожую структуру – большинство из них хранят в себе данные и имеют строчную организацию (последовательность – однострочная таблица, материализованные представления – таблица, которая помнит запрос и т.д.).

Состоят из слоев: 1) Основной слой – хранит в себе данные, индексные или табличные строки. Имеется у всех отношений, кроме представлений, т.к. они не хранят данных.

2) Слой инициализации – существует только у нежурналируемых таблиц и индексов, записи которых не попадают в журнал предзаписи (работа с ними быстрее, но потеряют данные в случае сбоя).

3) Карта свободного пространства (free space map) – Слой, в котором отмечено наличие пустого места внутри страниц. Помогает избегать фрагментации данных.

4) Карта видимости (visibility map) – слой в виде битовой карты, в котором отмечены страницы, содержащие только актуальные версии строк (то есть на данный момент не редактируются ни в одной транзакции).

Каждый слой представлен файлом до 1Гб, после заполнения этого пространства добавляется новый файл (сегмент). Каждый файл разделен на страницы по 8Кб (стандарт, может быть 16, 32, но нужно пересобрать базу), страница содержит различные версии строк и массив указателей на них.

**Версии строк и снимки**

Каждая версия строки должна храниться в рамках одной страницы, если этого не происходит, то используется либо стратегия сжатия до допустимого размера, либо использование стратегии TOAST – создается служебная таблица и версия строки «режется» на страницы допустимого размера, к этой стратегии также применимо сжатие.

Каждая версия строки содержит служебную информацию – номер транзакции, которая создала строку – поле xmin, транзакция, удалившая строку – поле xmax, помимо номеров транзакций эти поля хранят информацию о состоянии этих транзакций (завершена или продолжается), которая дублируется из журнала статусов транзакций.

Записанная в xmax транзакция со статусом незавершенности – признак блокировки, то есть построчные блокировки располагаются в самих данных, что не ведет к перерасходу памяти

Снимки данных – состояние данных доступное на момент начала транзакции. Состоит из зафиксированных версий строк, то есть транзакции, изменяющие эти строки должны быть завершены, и снимок должен содержать только одну версию определенной строки. То есть снимок данных дает согласованную картину данных на данный момент времени, которая лежит в основе MVCC (Multi-Version Concurrency Control) – метод управления параллельным доступом в БД за счет многоверсионности строк без блокировок.

*Read committed* – снимок данных создается в начале каждого оператора транзакции, *repeatable read и serializable* – снимок создается в начале первого оператора.

xmax – номер следующей (еще не существующей транзакции) xip\_list – список активных транзакций, xmin – номер самой ранней активной транзакции – для оптимизации и определения горизонта событий (По нему определяются версии, которые гарантировано находятся в прошлом, то есть они точно не будут использованы и точно не участвуют в активных транзакциях, что говорит о том, что их можно подвергнуть vacuum). За счет этих параметров можно определить, какие версии доступны, а какие нет.

**Очистка –** удаление невидимых для активных транзакций записей. Необходимость очистки тесно связана с операциями delete и update и зависит от частоты использования этих операций, т.к. update включает в себя delete строки (отметка в xmax) и insert строки с новым значением. Про этом строки продолжают существовать в БД и их физическое удаление происходит только при очистке.

Внутристраничная очистка – работает в рамках одной табличной страницы, не затрагивает индексы, не может быть запущен вручную - активируется, если подразумевается, что место на странице заканчивается (если при update не хватило места для вставки или страница заполнена больше, чем на fillfactor). При этом удаляются все версии строки, не видимые ни в одном снимке (находятся за горизонтом событий).

Vacuum – Выполняется на уровне таблицы, просматриваются только страницы таблицы не отмеченные в visibility map (очистке подлежат только те строки, которые не используются ни в одной транзакции), неотмеченные страницы пропускаются и помечаются для обработки при следующем запуске. Актуализирует информацию о просмотренных страницах в free space map и в visibility map. не блокирует операции чтения и изменения таблицы. Помимо очистки записей таблицы, также очищает и индексные записи. Освобожденное пространство помечается как свободное, но не возвращается операционной системе.

Помимо очистки, vacuum также занимается «заморозкой» строк – помечает строки, существующие достаточно давно, как устаревшие. Механизм нужен из-за наличия цикличной нумерации транзакций – нумерация транзакций заключена в 32 бита, что является оптимальным для хранения в заголовках строк, что хоть и является достаточно большим числом (более 4 млрд), но конечным, поэтому со временем транзакциям приходится использовать уже использованный ранее номер строки. Из-за этого может возникнуть конфликтная ситуация, когда есть достаточно старая строка с неким номером транзакции, который используется повторно уже другой транзакцией. Заморозка строк и позволяет избегать конфликта номеров.

Vacuum full – заново пересобирает таблицу, видимые записи копируются в новый файл и он становится основным при этом «лишнее» свободное место возвращается операционной системе (применяется, если таблица сильно выросла в размерах из-за накопившихся старых записей). Также пересобираются и индексы. Блокирует таблицу и все ее индексы.

Reindex – полностью пересоздает индекс без перестройки таблиц. Блокирует индекс и запрещает запись в таблицу, reindex concurrently – не блокирующий, но перестраивается дольше (с 12 версии).

Vacuum analyze – вакуум вместе со сбором статистики для оптимизации запроса.

**Буферный кэш и журналирование**

Буферный кэш – Разделяемая область в оперативной памяти, используемая для временного хранения данных в целях ускорения доступа к ним. Снижает кол-во обращений к диску за счет того, что запрашиваемые данные сначала ищутся в буфере и только в случае их отсутствия она будут вычитаны с диска сначала в буфер, а затем переданы в запрос.

Когда буфер заполняется, то подключается механизм вытеснения, организованный на основании счетчика обращений (каждое чтение страницы из буфера прибавляет 1 (5 максимум), а каждый проход вытесняющего потока -1 для каждой страницы), вытесняется та страница, счетчик обращений которой равен 0, и она не будет заблокирована другими процессами. Если вытесняемая страница была грязной (была изменена в кэше, но изменения не были сохранены на диск), то подключается поток предзаписи, который запишет данные на диск, а затем страница вытесняется.

Журнал предзаписи – хранит в себе записи, содержащие действия, произведенные над страницами в оперативной памяти (изменения страниц, фиксация или отмена транзакций). Подразумевает, что данные из журнала предзаписи должны попасть на диск до самих данных.

Процесс журналирования так же контролируется отдельным процессом (checkpointer), который занимается постепенным очищением журнала предзаписи путем отправки изменений на диск (контрольная точка).

При возникновении сбоя, записи из журнала вычитываются, и данные, утраченные в оперативной памяти, восстанавливаются посредством повторения действий из журнала. То есть возникновение журнала перезаписи напрямую связано с использованием буферного кэша.

*Вся теория в серии* [*статей*](https://habr.com/ru/companies/postgrespro/articles/442804/) *и* [*видео*](https://www.youtube.com/watch?v=g3NH55nQgX4&list=PLaFqU3KCWw6K3AyBVcZGdXMtM2xvjHA1N&index=2&ab_channel=PostgresProfessional)

**Обработка и исполнение запроса**

*Разбор (Parsing) -* Синтаксический разбор - СУБД анализирует SQL-запрос, чтобы удостовериться, что он соответствует правилам языка SQL. Семантический разбор - проверка существования таблиц и полей, а также правильность их использования. Для проверки используется системный каталог.

*Трансформация –* замена представлений на таблицы, преобразование SEARCH и CYCLE для рекурсивных запросов.

*Планирование (Query Planning) -* Планировщик запросов создает дерево выполнения запроса и выбирает вариант исполнения с наименьшей стоимостью, используя статистику из системного каталога.

*Выполнение* - СУБД выполняет запрос в соответствии с выбранным планом выполнения.

Данные этапы выполняются для каждого запроса.

Prepare - Подготовленные запросы в SQL позволяют СУБД сохранять оптимизированный план выполнения запроса и повторно использовать его при последующих выполнениях запроса с разными параметрами. Это позволяет уменьшить накладные расходы на планирование и оптимизацию запроса каждый раз при его выполнении. (отсекается часть работы – не выполняется повторно).

**База данных и схема**

База данных представляет набор связанных данных (в виде отношений), хранящиеся в структурированном виде (в виде таблиц). Помимо таблиц обычно содержит другие объекты – индексы, процедуры, представления. То есть это хранилище для всех объектов.

Схема данных логически сгруппированные объекты БД, т.е. схема обеспечивает не организацию данных в каждой отдельной таблице, а объединяет сущности БД в отдельную логическую группу – схему. Так же позволяет обеспечивать контроль доступа к объектам и безопасность.

Т.е. БД — это хранилище данных для всех объектов, а схема — это набор данных внутри БД, сгруппированных логически.

Схема – для разграничения пространства имен и ограничения прав доступа.

Проектирование БД – [1](https://www.youtube.com/watch?v=HnRXzrg3Sd4&list=PLf30vI0hEi1v435cBmZSHkr1QAJdOk9mb&index=5&ab_channel=RclassTech), [2](https://www.youtube.com/watch?v=tNQlFW3BRUc&list=PLLACapFZr4XvyQgxZaIOK-MJwblGVCL41&index=7&ab_channel=ListenIT)