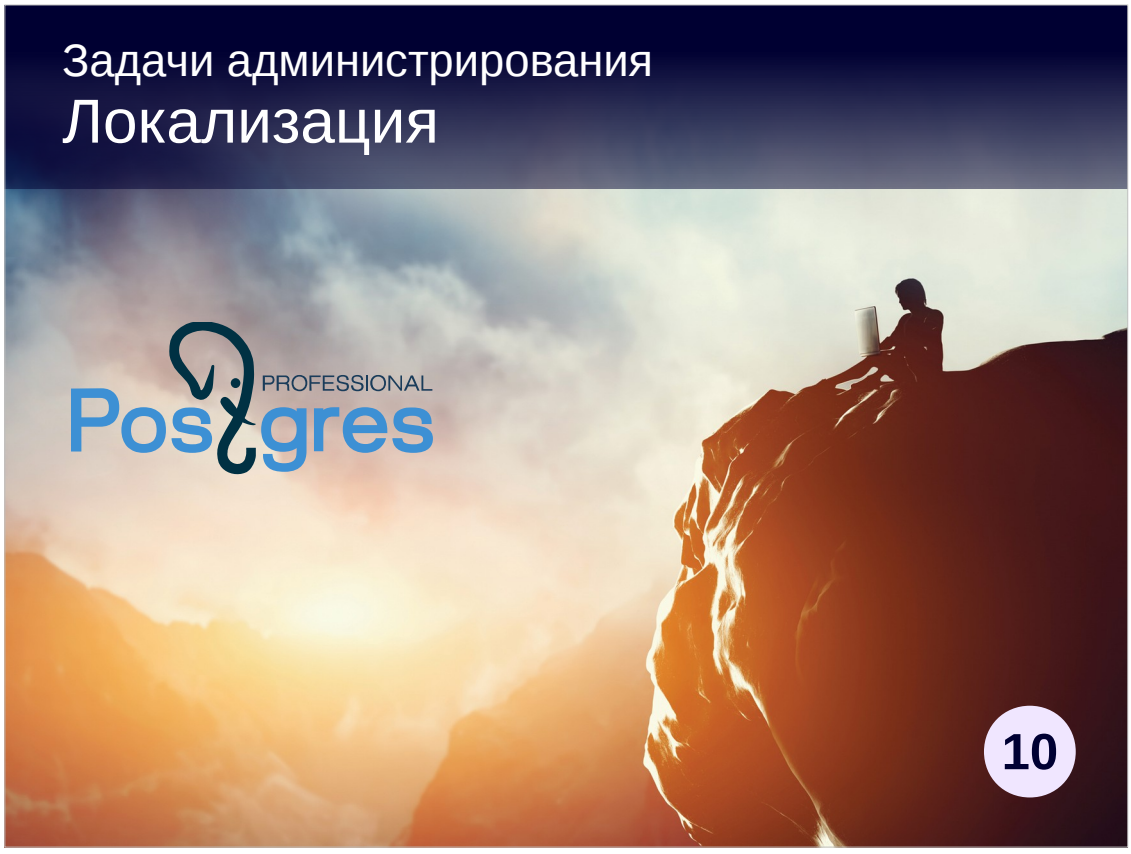


Авторы: Егор Рогов, Павел Лузанов

Обратная связь
Отзывы, замечания и предложения направляйте по адресу:
edu@postgrespro.ru

Темы

Назначение локализации

Локали и категории

Правила сортировки

Настройка сервера и клиента

Настройка сообщений сервера и клиентских утилит

Работа с датами, числами, денежными единицами

Назначение

Поддержка кодировок

- выбор кодировки символов для различных языков
- перекодировка символов между клиентом и сервером

Функционал, зависящий от локализации

- сортировка и сравнение символов
- функции upper, lower, initcap
- поиск по шаблону
- функции to_char с датами, числами, денежными единицами
- язык сообщений сервера и утилит

Возможности локализации в PostgreSQL позволяют хранить текст в различных кодировках. Для русского языка поддерживаются все основные кодировки символов, включая UTF8, WIN1251, KOI8R, ISO_8859_5.

Клиентское приложение может работать в кодировке, отличной от кодировки сервера. В таком случае настраивается преобразование символов между клиентом и сервером.

Кроме поддержки различных кодировок локализация влияет на работу следующего функционала сервера:

- сортировка символов, например в предложениях ORDER BY или в операциях сравнения (>, <);
- преобразование букв в верхний и нижний регистр в функциях upper, lower, initcap;
- поиск по шаблону в регулярных выражениях, операторах LIKE, SIMILAR TO, включая поиск без учета регистра символов;
- форматирование дат, чисел и денежных единиц в функциях to_char;
- выбор языка сообщений сервера и утилит.

https://postgrespro.ru/docs/postgresql/10/locale

Локали и категории

Локали

определяют язык, территорию и кодировку символов, например `ru_RU.UTF8`

установлены в операционной системе

Категории локали

lc_ctype	классификация символов
lc_collate	правила сортировки символов
lc_messages	язык сообщений
lc_monetary	формат денежных единиц
lc_numeric	формат чисел
lc_time	формат даты и времени

PostgreSQL использует возможности локализации, предоставляемые операционной системой. Поэтому в ОС следует предварительно настроить локали, которые потребуются для работы СУБД.

Обычно локали в ОС задаются в формате «`язык_территория.кодировка`». Например, `ru_RU.UTF8` определяет локаль с русским языком (ru), на котором говорят в России (RU), и кодировкой UTF8. В Windows используются развернутые имена: `Russian_Russia.1251`.

Язык и территория определяют такие национальные особенности, как порядок символов, формат даты, разделитель десятичных разрядов и т. п.

Иногда бывает необходимо комбинировать поведение некоторых функций из разных локалей. Например, вместе с правилами сортировки русского языка использовать английские сообщения сервера. PostgreSQL поддерживает раздельную установку категорий локали через одноименные параметры конфигурации.

Правила сортировки

Объекты базы данных

- разные правила сортировки в одной БД
- начальное наполнение при создании БД
- указываются для столбцов таблиц и в выражениях

Провайдеры

- libc
- ICU

Специальные правила сортировки

- «C» и «POSIX»

Чтобы в одной базе данных можно было по-разному сортировать и сравнивать текстовые данные, в PostgreSQL существуют специальные объекты — правила сортировки (collation). Они позволяют использовать порядок сортировки, отличный от установленного по умолчанию для базы данных.

При создании правила сортировки указываются внешняя библиотека, реализующая сортировку (провайдер), и классификация символов. Начиная с PostgreSQL 10 можно выбирать между библиотеками ICU и libc; до этого всегда использовалась libc.

Начальный список объектов — правил сортировки — формируется при инициализации кластера. Для всех имеющихся в ОС локалей загружаются правила сортировки в базу данных. В дальнейшем при добавлении локалей в ОС можно дозагрузить их в базу данных.

Правила сортировки можно использовать при создании текстовых столбцов таблиц, при определении доменов и просто в выражениях, где сортируются или сравниваются текстовые строки.

Специальные правила сортировки «C» и «POSIX» работают одинаково и создаются для всех кодировок сервера. Для этих правил буквами будут считаться только латинские символы от A до Z, все остальные знаки будут сортироваться в соответствии со своими кодами в данной кодировке.

Посмотреть имеющиеся правила сортировки можно в таблице системного каталога pg_collation.

https://postgrespro.ru/docs/postgresql/10/collation

Провайдер libc

Порядок символов

- зависит от реализации в операционной системе

Изменения в библиотеке

- не отслеживаются

Использование

- по умолчанию для базы данных или кластера
- явное указание при определении столбцов и в выражениях

Правила сортировки провайдера libc в базе данных работают точно так же, как и в операционной системе. Однако библиотека libc может быть реализована по-разному в разных операционных системах. Как следствие, сортировка для одних и тех же локалей может отличаться в зависимости от ОС сервера базы данных. Если приложение должно поддерживать работу (включая логическую репликацию) СУБД на разных ОС, то следует убедиться, что сортировка везде работает корректно.

Более серьезной потенциальной проблемой является установка новой версии библиотеки libc в ОС. Такое может произойти, например, при переходе на новый сервер с новой версией ОС. Если в новой версии libc изменились используемые в базе данных правила сортировки, то это может привести к некорректной работе индексов и пр. Плохо то, что PostgreSQL ничего не знает о том, какая версия libc используется. Администратор БД должен следить за версиями libc, в том числе на серверах с физическими репликами.

При инициализации кластера баз данных или при создании новой БД можно указать только локали библиотеки libc.

Провайдер ICU

Порядок символов

- зависит от версии библиотеки, но не от операционной системы

Изменения в библиотеке

- предупреждение при изменении версии

Использование

- явное указание при определении столбцов и в выражениях
- нельзя использовать по умолчанию для базы данных или кластера

Дополнительные возможности

- управление порядком сортировки групп символов

Для использования провайдера ICU в определении правил сортировки необходимо, чтобы PostgreSQL был собран с поддержкой этой библиотеки. Библиотека ICU и реализованные в ней правила сортировки работают одинаково на всех операционных системах.

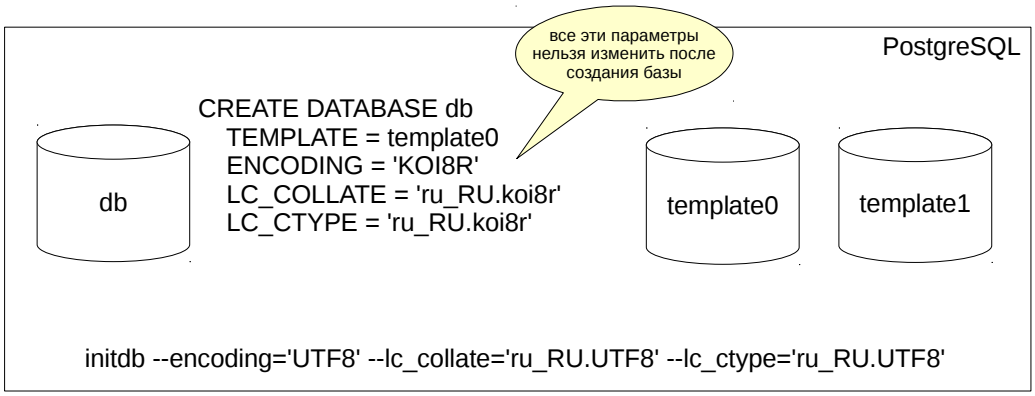
Но изменение самой библиотеки может по-прежнему вызвать проблемы, если поменялись используемые в базе данных правила сортировки. В случае с правилами сортировки ICU PostgreSQL о таких изменениях будет предупреждать. При создании правила, в системном каталоге сохраняется номер его версии. Каждый раз при использовании правила сверяется сохраненный номер версии с установленным в библиотеке ICU. В случае расхождения запросы будут выдавать предупреждение о несоответствии версий.

Если такая ситуация случилась, следует пересоздать индексы, использующие измененные правила сортировки, проверить влияние новых правил сортировки на ограничения CHECK, табличные триггеры. То же самое следует делать и при изменении библиотеки libc. Отличие в том, что о потенциальной проблеме с ICU сразу же сообщит СУБД.

В настоящий момент (для версии 10) локали библиотеки ICU нельзя выбрать при инициализации кластера и создании новой базы данных.

Библиотека ICU допускает видоизменение правил сортировки без смены языка и территории. Можно указать разный порядок сортировки для отдельных групп символов. Например, символы какой группы должны идти раньше: кириллица, латиница или цифры; буквы в верхнем регистре или в нижнем.

Настройка сервера



LC_COLLATE = en_US.UTF8
LC_CTYPE = en_US.UTF8

Установка локали выполняется при инициализации кластера баз данных. Утилита initdb использует локаль из окружения ОС, либо локаль можно указать явно через соответствующие параметры.

При создании новой базы данных можно указать параметры локализации, отличные от тех, что использовались при инициализации кластера баз данных.

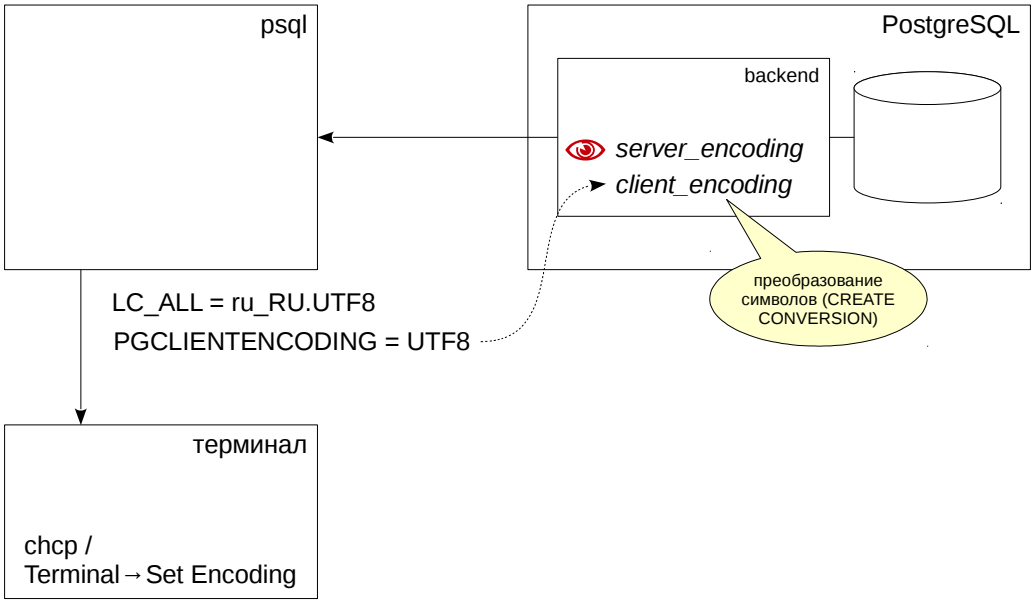
К основным параметрам локализации базы данных относятся: кодировка символов (encoding), а также категории локали lc_ctype и lc_collate.

На эти категории локали накладываются ограничения:

- кодировки символов у lc_ctype и lc_collate должны совпадать с кодировкой базы данных.
- после создания базы данных lc_ctype и lc_collate нельзя изменять.

Второе ограничение объясняется тем, что изменение правил сортировки и классификации символов может нарушить работу существующих индексов. Кроме того, может измениться поведение операций сравнения текстовых строк в ограничениях CHECK и табличных триггерах, что сделает уже сохраненные данные некорректными.

Настройка клиента



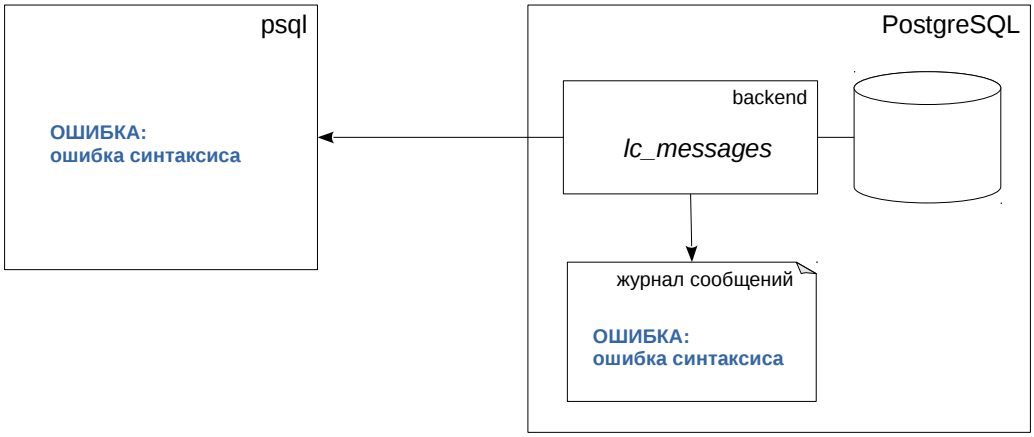
9

Для настройки локализации клиентского приложения нужно сделать следующее.

- Проверить, что настройки сервера корректны. Как минимум, что используется правильная кодировка БД (неизменяемый параметр *server_encoding* покажет кодировку, заданную при создании БД).
- Проверить, что в клиентской ОС установлены нужные локали, и настроить категории локали (переменные среды LC_*) в сеансе ОС.
- Настроить кодировку, с которой работает приложения, и проверить настройки устройства вывода.
  Например, psql в Windows обычно использует кодировку Win-1251, поэтому в терминале (cmd.exe) необходимо выполнить команду `chcp 1251` и установить шрифт true type (например, Lucida Console).
- После подключения к БД проверить параметр *client_encoding*. Этот параметр отвечает за перекодировку символов между клиентом и сервером. При необходимости, установить в значение кодировки приложения. Значение *client_encoding* можно задать и в переменной среды PGCLIENTENCODING.

Обслуживающий процесс автоматически перекодирует символы из кодировки БД (server_encoding) в кодировку клиента (client_encoding). Для большинства кодировок в PostgreSQL преднастроены необходимые процедуры преобразования символов: они находятся в таблице системного каталога pg_conversion. Возможно создание дополнительных пользовательских процедур (CREATE CONVERSION).

Сообщения сервера



configure --enable-nls ...
unset LANGUAGE

Сообщения сервера (и утилит) PostgreSQL переведены на несколько языков. В том числе и на русский.

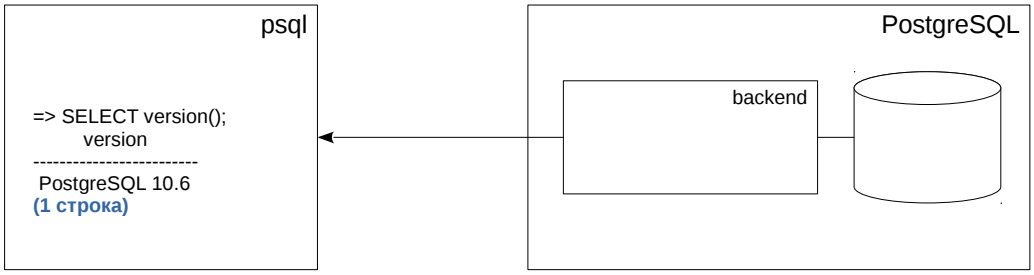
Для того, чтобы сообщения сервера выводились на русском языке, нужно, чтобы сервер PostgreSQL был собран с поддержкой NLS.

Параметр конфигурации *lc_messages* управляет языком сообщений сервера. Однако, если при запуске PostgreSQL была установлена переменная окружения LANGUAGE, то именно она определяет язык и управлять им помощью *lc_messages* будет невозможно.

Сообщения сервера отправляются не только клиенту, но и записываются в журнал сервера. При выборе языка, отличного от английского, следует убедиться, что инструменты работы с журналом сервера понимают этот язык. Например pgBadger понимает только английские сообщения.

Подробнее о переводе сообщений сервера для переводчиков и разработчиков: https://postgrespro.ru/docs/postgresql/10/nls

Сообщения утилит



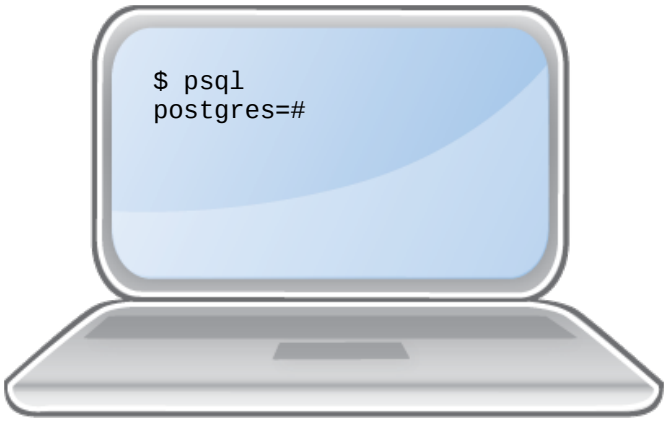
configure --enable-nls ...
export LC_MESSAGES=ru

Утилиты PostgreSQL (psql, pg_dump и пр.) также поддерживают NLS.

Для того, чтобы сообщения утилит выводились на языке, отличном от английского, нужно, чтобы клиент PostgreSQL был собран с поддержкой NLS. Язык вывода устанавливается на клиенте переменной среды LC_MESSAGES (параметр сервера *lc_messages* влияет только на сообщения самого сервера, но не клиента).

Большинство ОС (включая Windows) используют следующий порядок просмотра переменных среды для языка сообщений: LANGUAGE, LC_ALL, LC_MESSAGES, LANG.

Демонстрация



Итоги

Перед инициализацией кластера нужные для СУБД локали должны быть установлены в ОС

Клиент и сервер могут работать в различных кодировках с автоматическим преобразованием символов

Правила сортировки используют внешние библиотеки. Изменения в этих библиотеках могут привести к разрушению индексов и некорректным данным

Явное использование правил сортировки позволяет по-разному сортировать текстовые данные

Сообщения сервера и утилит переведены на несколько языков, включая русский

Практика

1. Перенос данных между базами в разных кодировках.

   Создайте базу данных с кодировкой KOI8R.
   Создайте таблицу и добавьте в нее строки, содержащие символы кириллицы.
   Сделайте копию базы данных утилитой pg_dump.
   Восстановите таблицу из копии в базу с кодировкой UTF8.

2. Получите номер сегодняшнего дня недели.

   Меняется ли номер дня недели в зависимости от настроек локализации?

1. Для создания БД в кодировке KOI8R

   Убедитесь, что в ОС установлена нужная локаль

   В команде CREATE DATABASE используйте шаблон template0 и опции ENCODING, LC_CTYPE, LC_COLLATE

2. Для получения номера дня недели используйте функцию to_char.

   Допустимые форматные маски даты:

   https://postgrespro.ru/docs/postgresql/10/functions-formatting#FUNCTIONS-FORMATTING-DATETIME-TABLE