Обзор Apache Superset

Apache Superset – открытое программное обеспечение для исследования и визуализации данных, ориентированное на большие данные.

Разработан сотрудником Airbnb Максимом Бошменом, который также является одним из создателей Apache Airflow. В дальнейшем, помимо Airbnb, в проект внесли значительный вклад компании Lyft и Dropbox. В 2021 году стал проектом высшего уровня в Apache Software Foundation.

Реализован, в основном, с помощью Python, TypeScript и JavaScript.

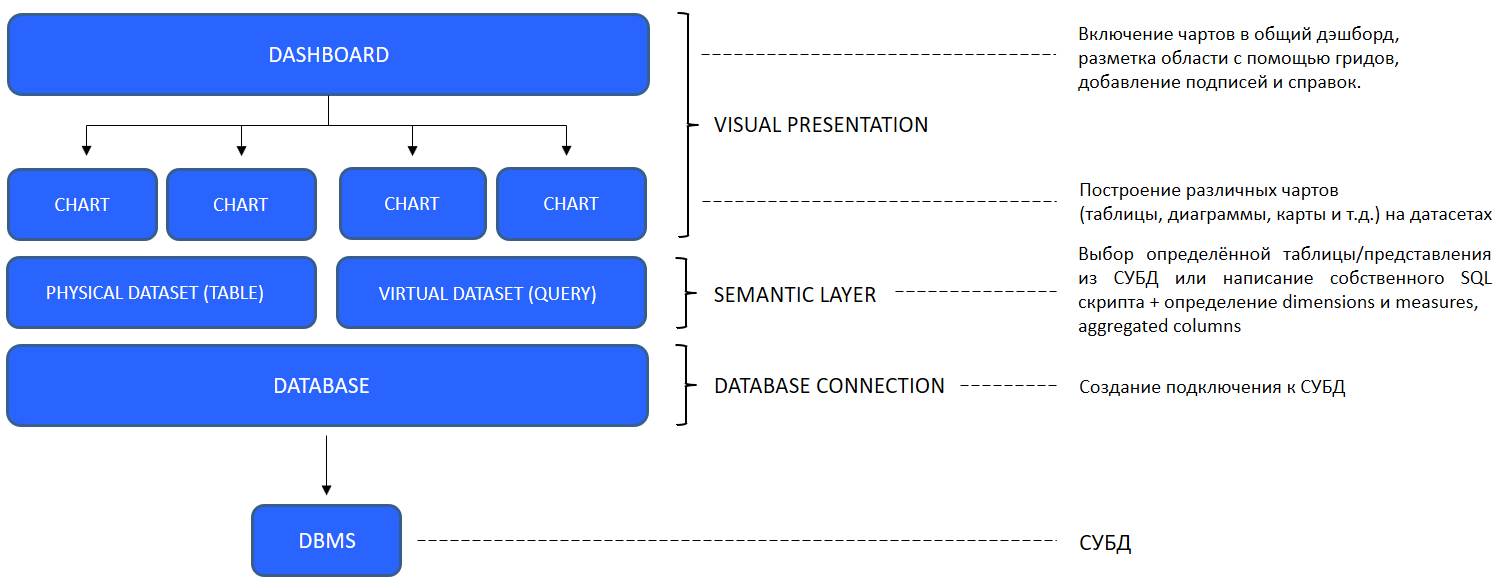
Преимущества и возможности:

* Большой набор доступных визуализаций из коробки и возможность подключить собственные;
* Расширенная модель безопасности (роли, RLS – полномочия);
* Удобный импорт/экспорт объектов (инфопанели, наборы данных, визуальные элементы);
* Можно работать с SQL запросами прямо в интерфейсе приложения;
* Удобный интерфейс для создания метрик и вычисляемых столбцов;
* Поддержка для большинства SQL баз данных;
* Встроенное асинхронное выполнение запросов и кэширование;
* API для программной кастомизации;

Некоторые примеры доступных источников данных:

* PostgreSQL (Greenplum)
* Clickhouse
* Teradata
* Oracle
* SQLite
* MySQL
* SAP HANA
* Google Sheets

Архитектура Apache Superset



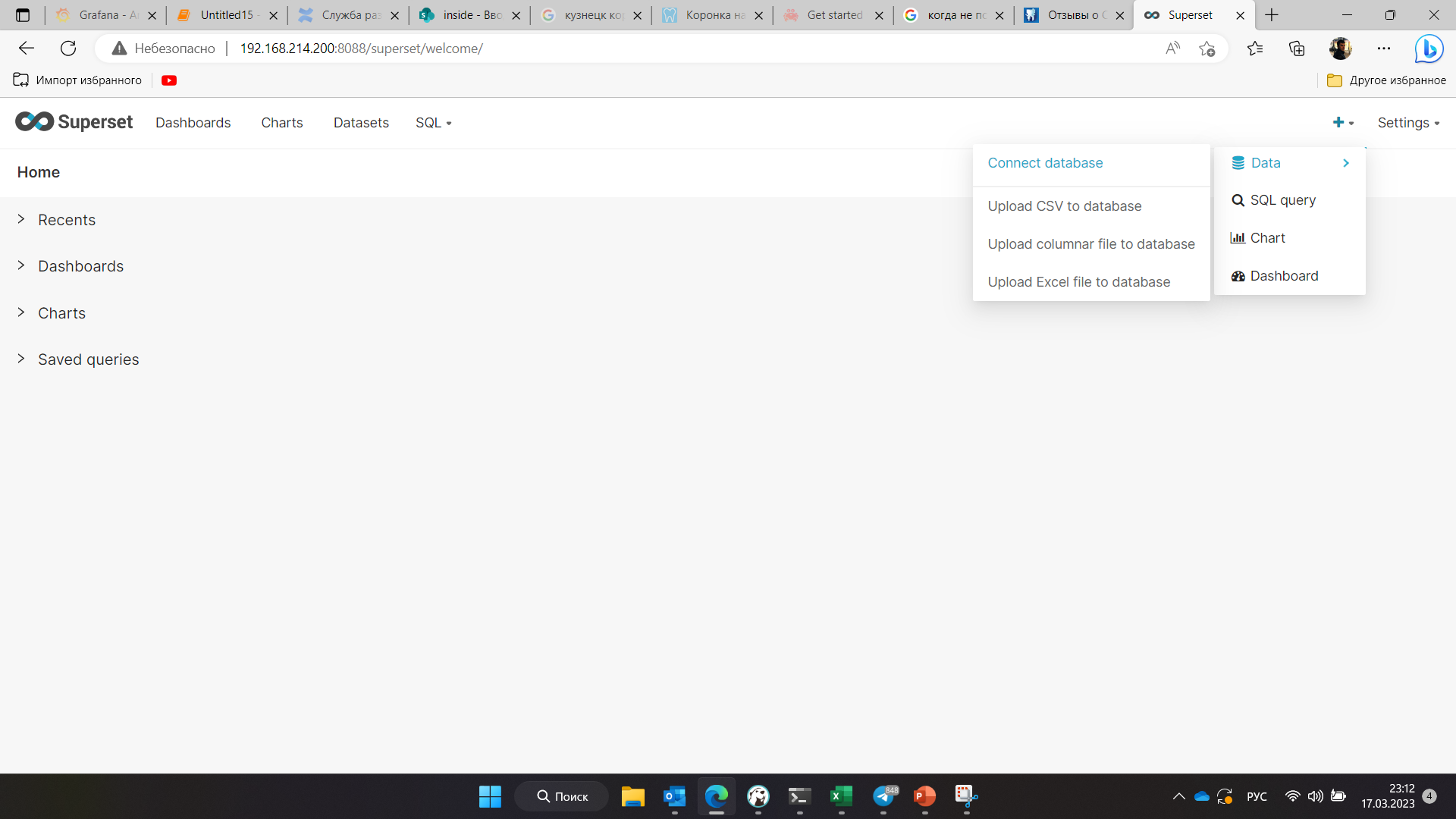
Сначала создаётся соединение к нужно БД на специальной вкладке. Затем, в семантическом слое, можно создать два вида дата сетов для БД: физический на какой-либо таблице или представлении из БД и виртуальный – собственно, SQL-запрос к объектам БД и уже созданным дата сетам в Superset. На этом этапе уже есть возможность разделить поля на dimensions и metrics.

Далее на этих дата сетах строятся чарты на спец. вкладке. Есть различные виды визуализации.

Затем на созданных ранее чартах создаётся дэшборд, где вся его область размечается с пом. гридов определённым образом для включённых в него чартов. Можно организовать работу для пользователей в различных воркспейсах – разграниченных областях для разработки и просмотра всех перечисленных объектов.

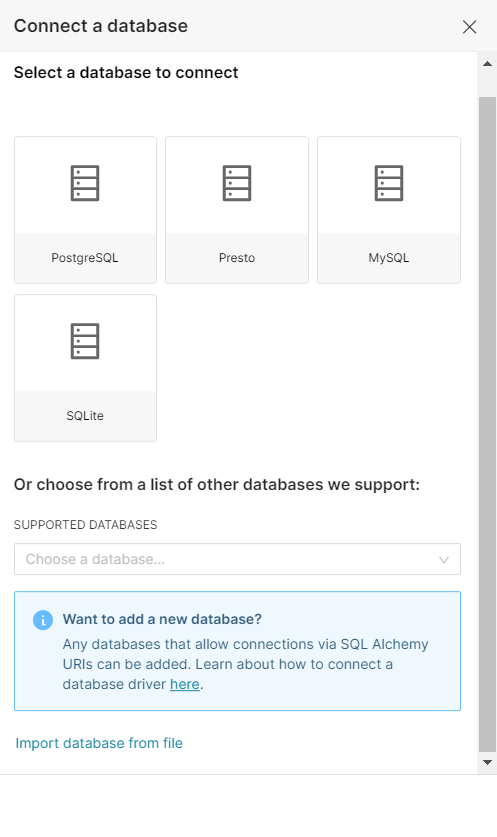
создание дэшборда

Создание соединения с СУБД:



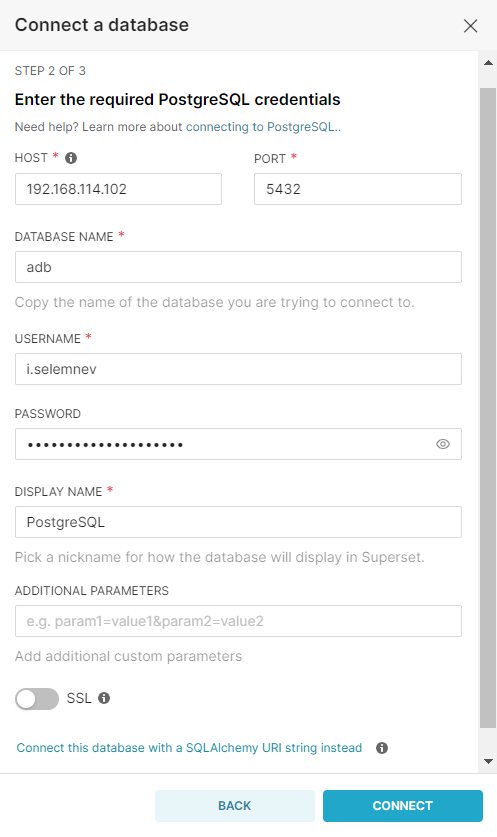
Создаётся соединение на указанной вкладке или в настройках в разделе Database Connections.

Сначала необходимо выбрать исходную систему, которая будет использоваться в качестве источника данных:

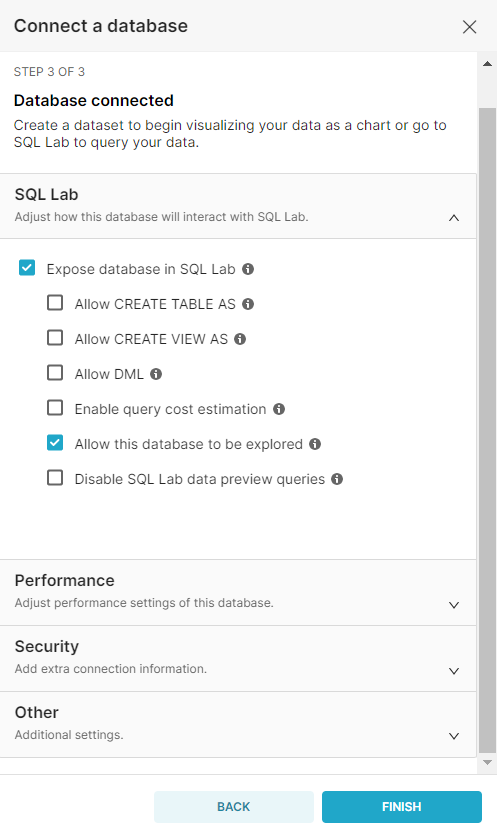


Можно подключить мн-во других систем, которых нет на рисунке выше, установив специальные драйверы.

Далее необходимо задать параметры подключения:



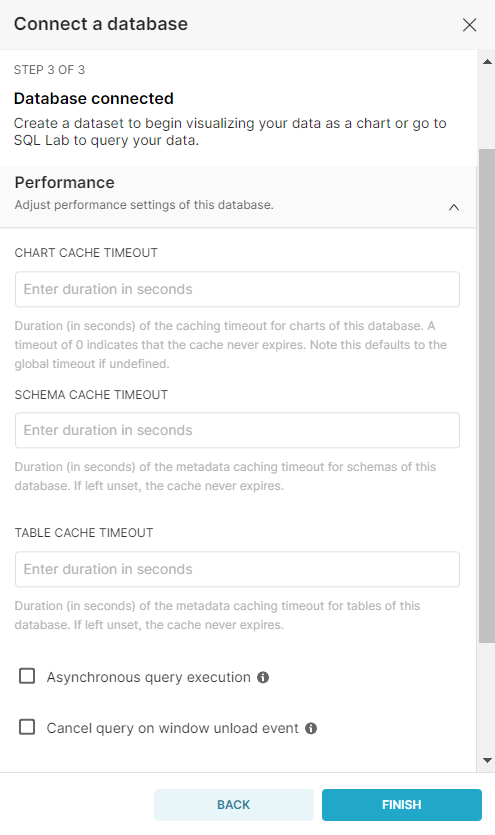
Также необходимо выставить некоторые настройки:



Можно предоставить:

* в целом возможность написания SQL-запросов в БД (Expose database in SQL Lab),
* позволить создавать таблицы и представления (Allow CREATE TABLE AS и Allow CREATE VIEW AS) посредством запросов,
* разрешить делать манипуляции с данными, помимо SELECT (Allow DML),
* предоставить возможность делать оценку стоимости запроса для некоторых БД в плане потребления ресурсов (Enable query cost estimation);
* дать возможность визуализировать SQL-запросы (Allow this database to be explored)
* отключить просмотр данных при получении метаданных таблицы в SQL Lab (Disable SQL Lab data preview queries).

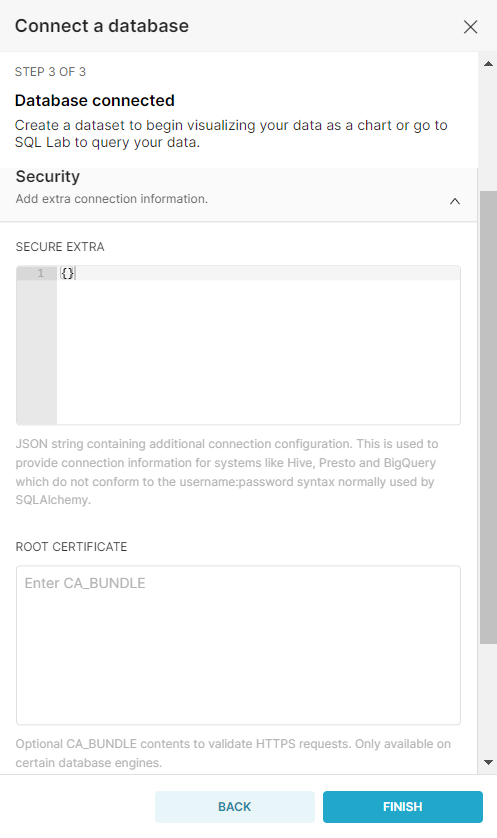
Параметр CHART CACHE TIMEOUT управляет временем хранения кэшированных данных для чартов:



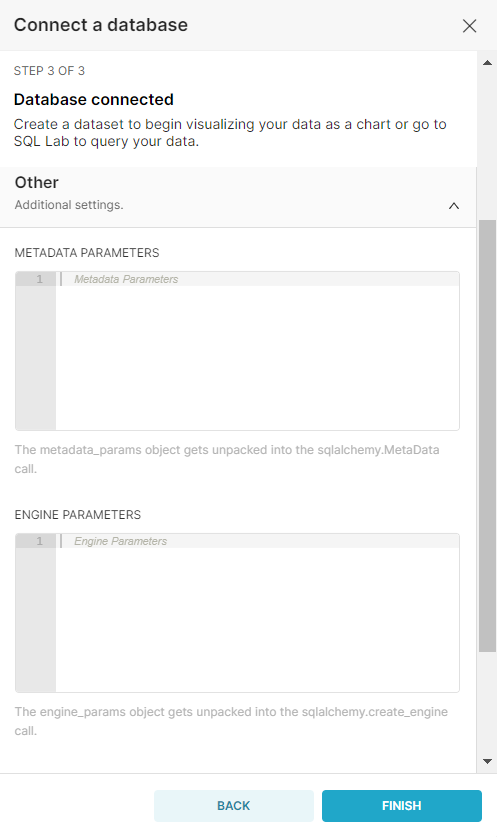
Если время хранения кэша превышает это значение, то кэшированные данные удаляются и затем пересчитываются снова при очередном запросе. Это полезно, чтобы удерживать актуальность данных и избежать отображения устаревших графиков.

Можно также установить таймауты для схем (SHEMA CACHE TIMEOUT) и таблиц (TABLE CACHE TIMEOUT).

Также для некоторых систем можно задать доп. конфигурацию подключения для обеспечения безопасности:



Также можно задать какие-либо metadata параметры или engine параметры:

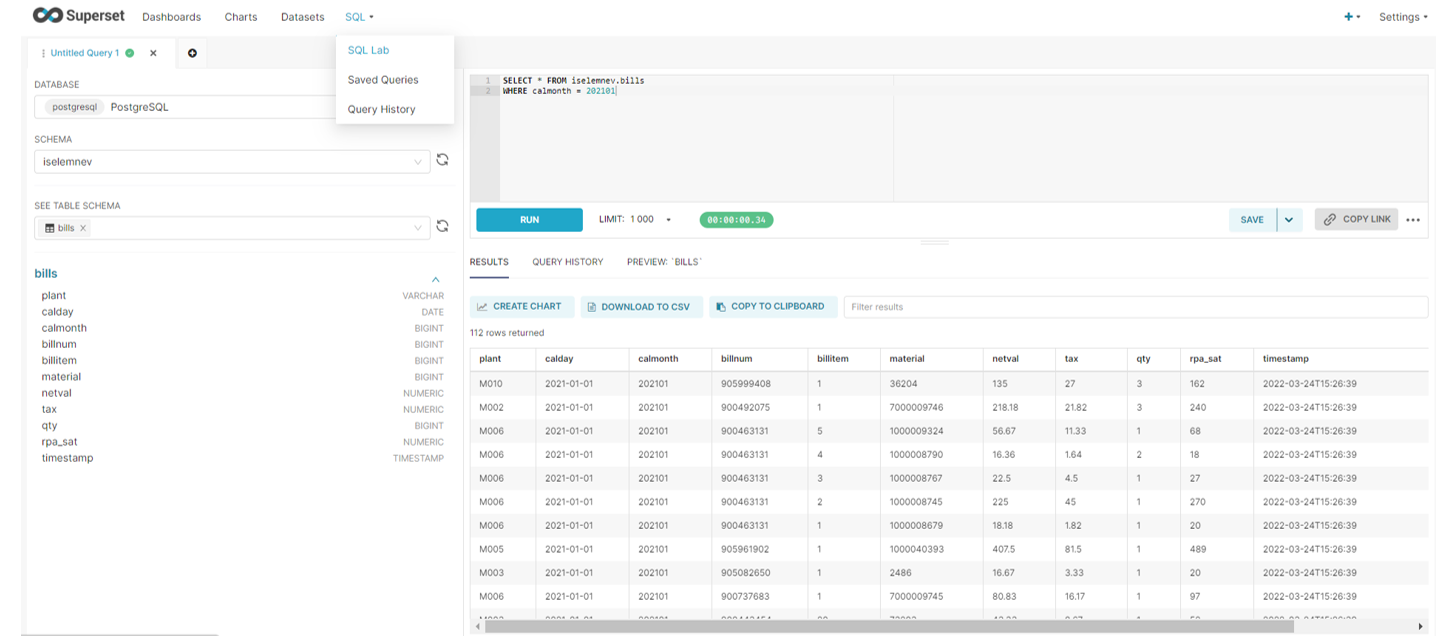


Но мы такие настройки не будем использовать.

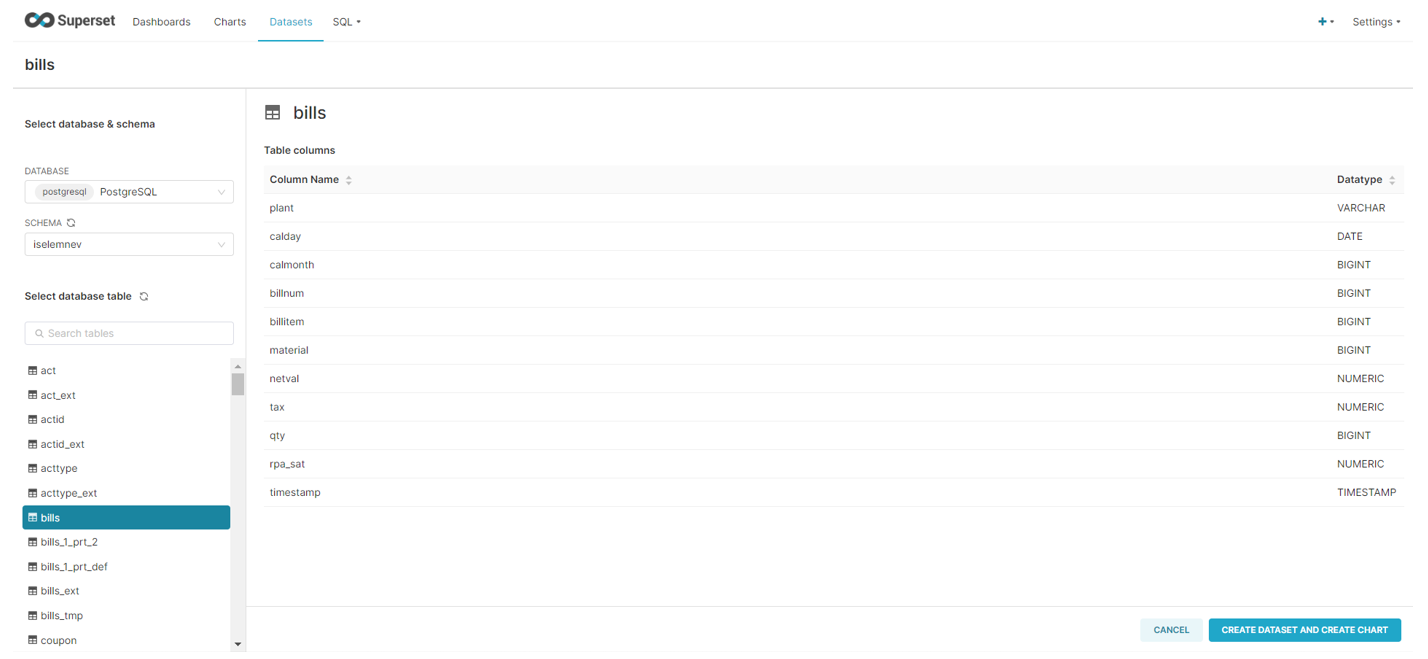
Создание датасета:

Датасет можно создать в SQL Lab посредством SQL-запроса к выбранным таблицам в схемах БД или же просто указав необходимые таблицы или представления БД.

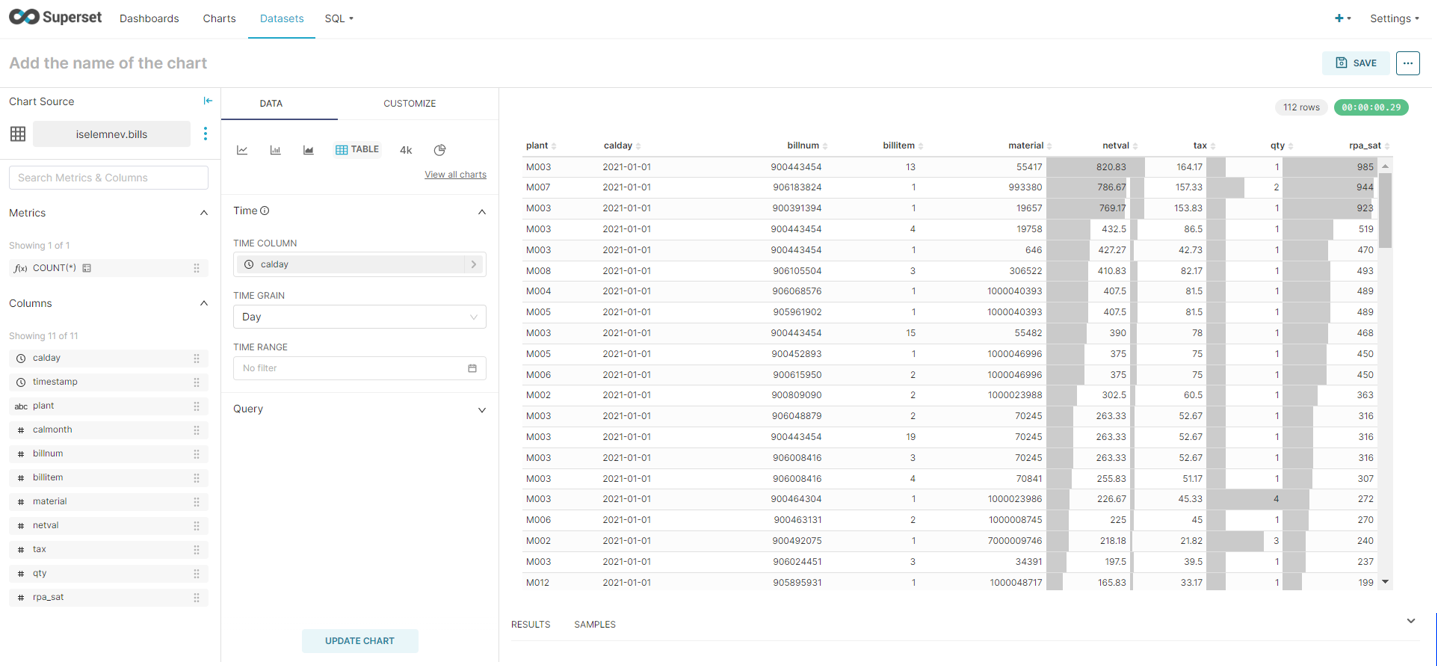
Создание датасета посредством SQL-запроса к выбранным таблицам в схемах БД:



Создание датасета посредством указания необходимых таблиц или представлений БД:

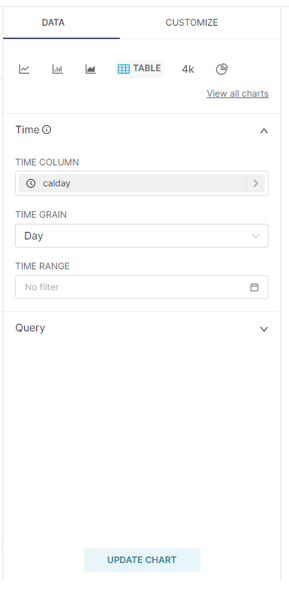


Создание чарта:

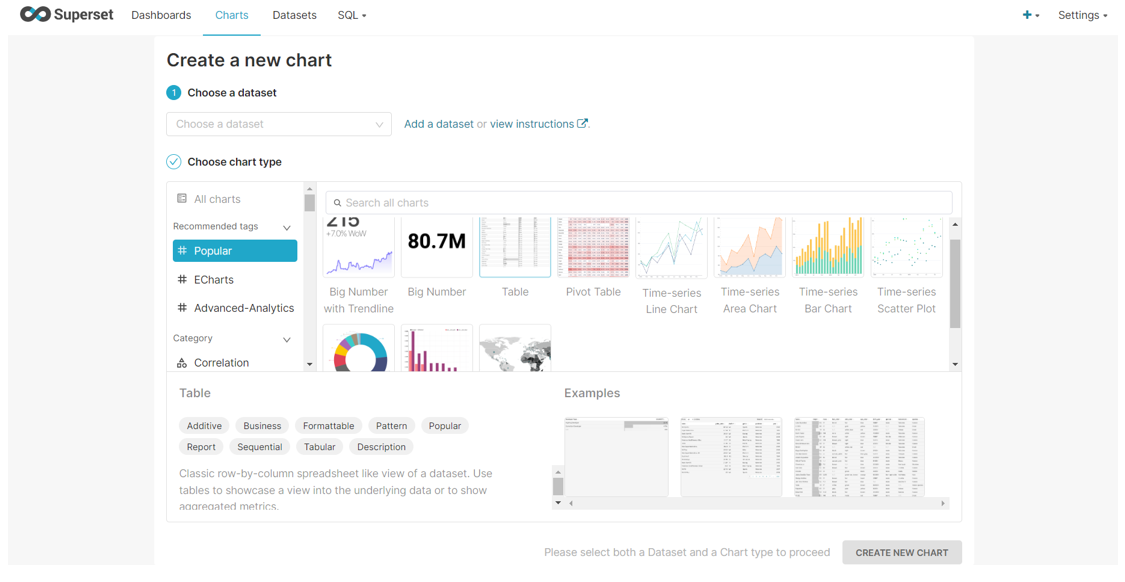


К его созданию можно перейти сразу после реализации датасета по кнопке Create dataset или Create chart.

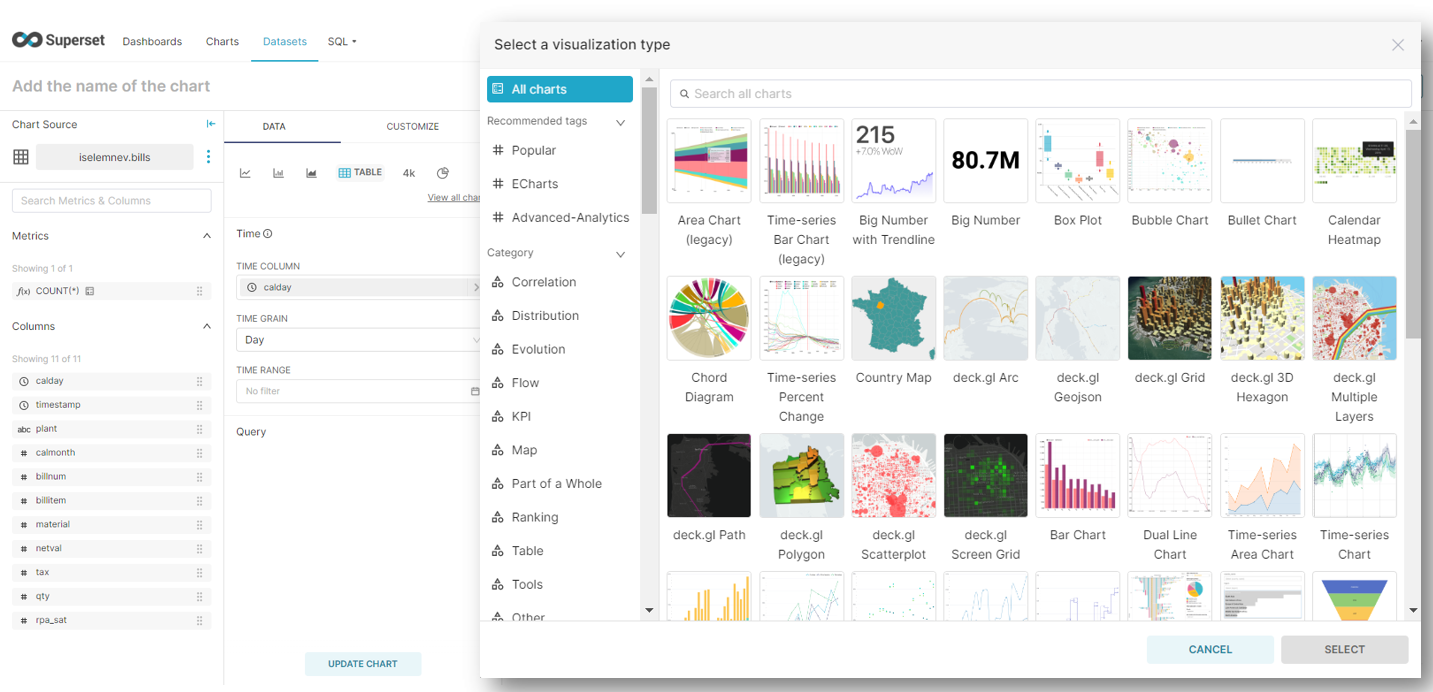
На вкладке Data в разделе Time можно выбрать столбец времени для визуализации. При выборе колонки Superset предложит написать собственный SQL-запрос, который должен будет возвращать datetime столбец, и поставить фильтр для нужного отображения времени.



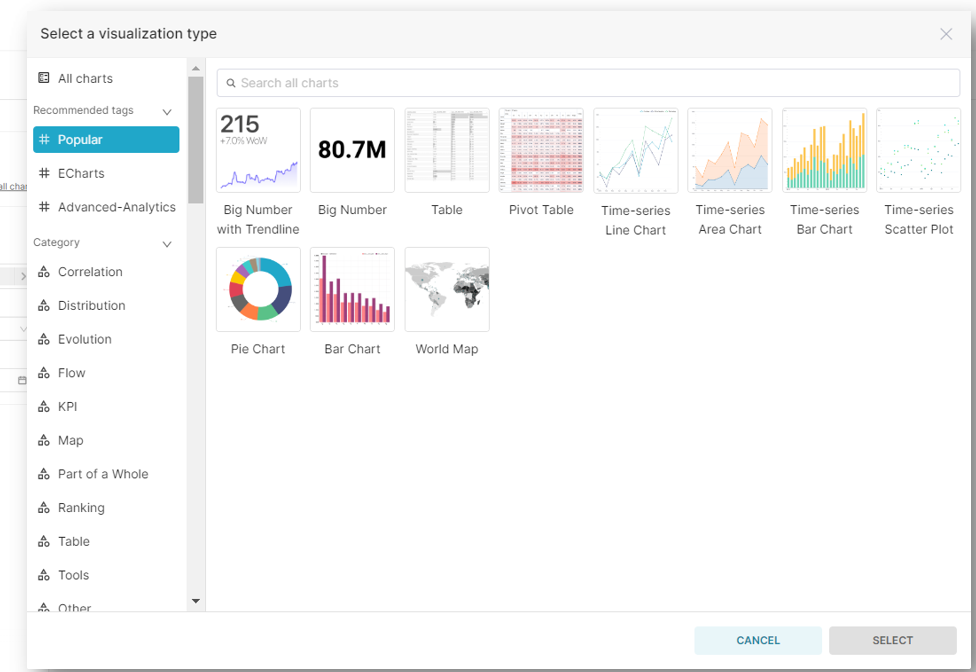
Также можно перейти на вкладку Charts и далее, при создании нового чарта, самим выбрать нужный датасет и тип визуализации:



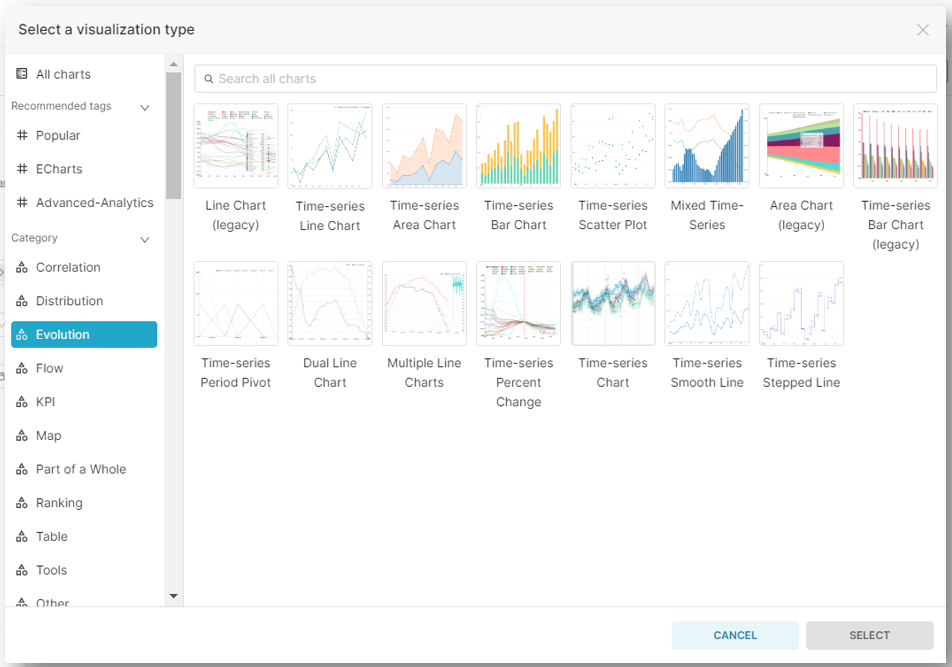
Через View all charts можно посмотреть доступные визуализации данных:

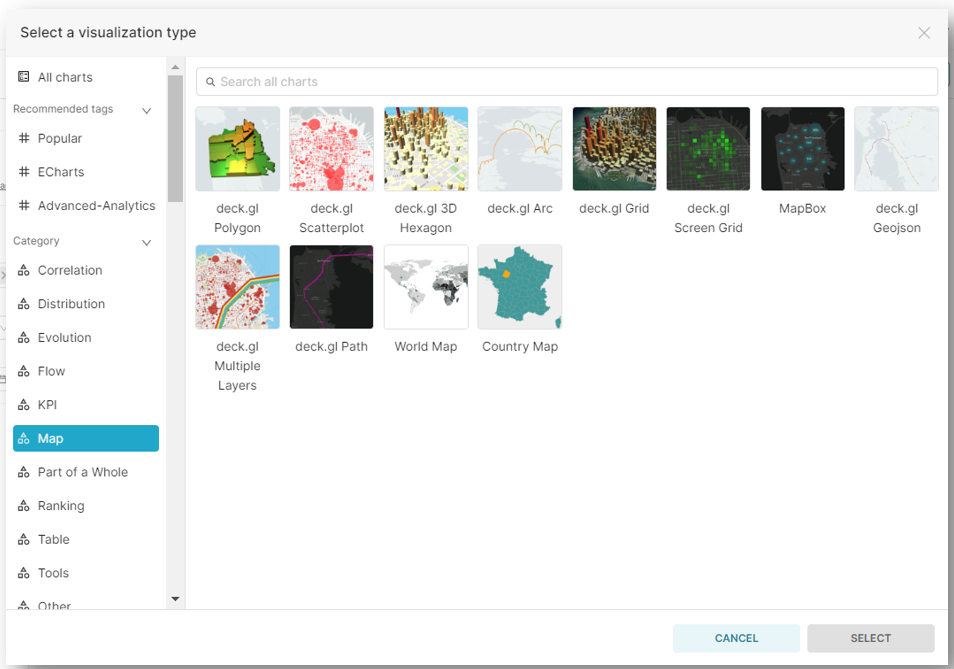


Этих визуализаций изначально достаточно большое кол-во в Superset. Рассмотрим некоторые из них.

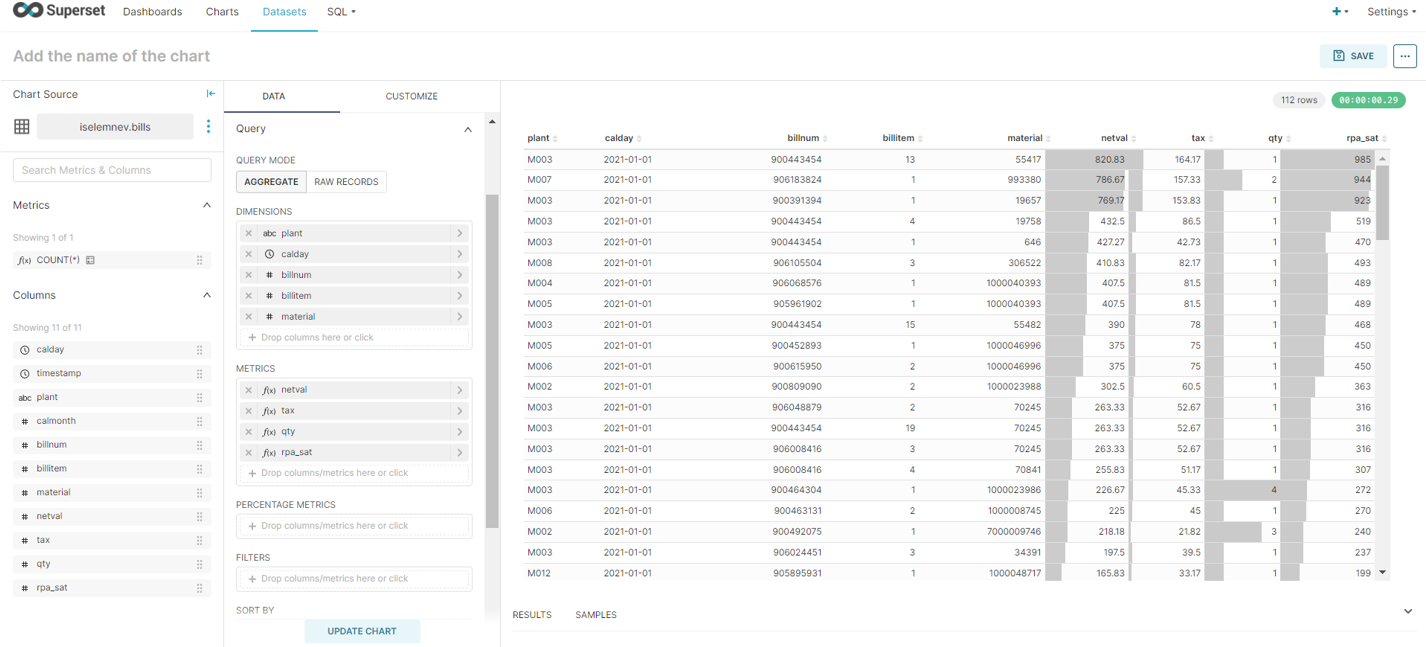




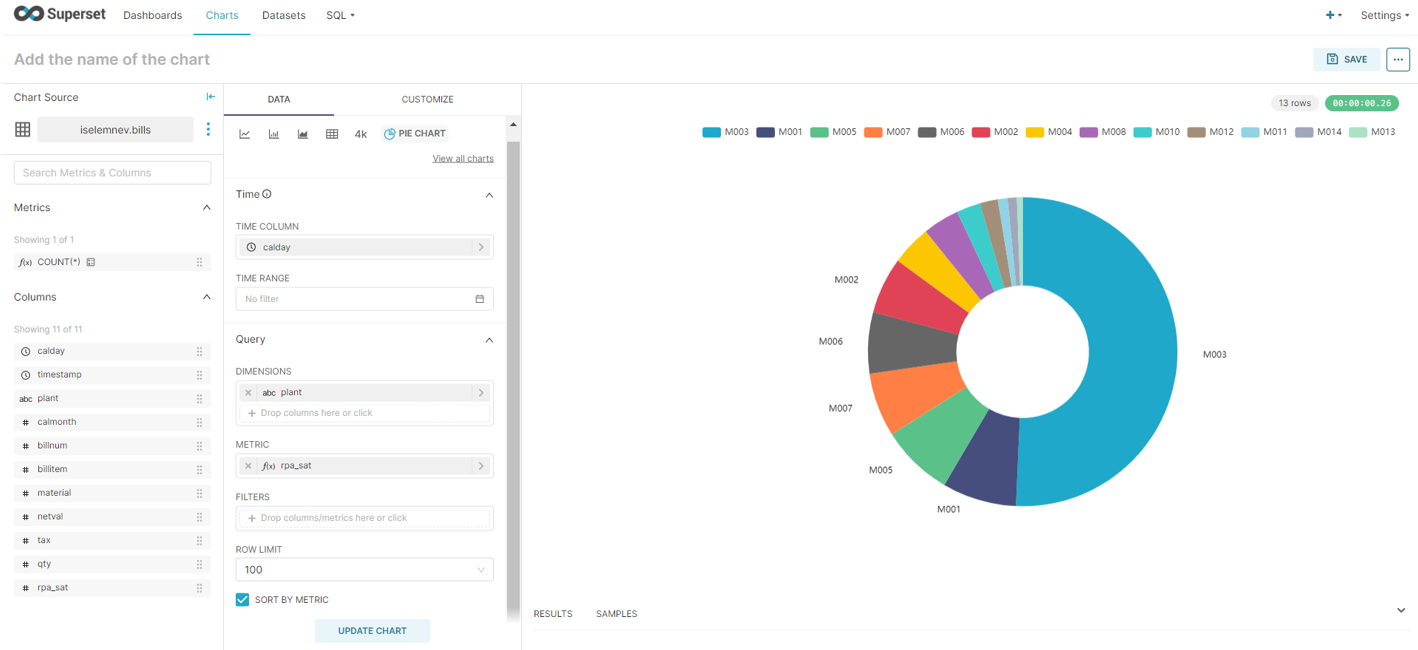




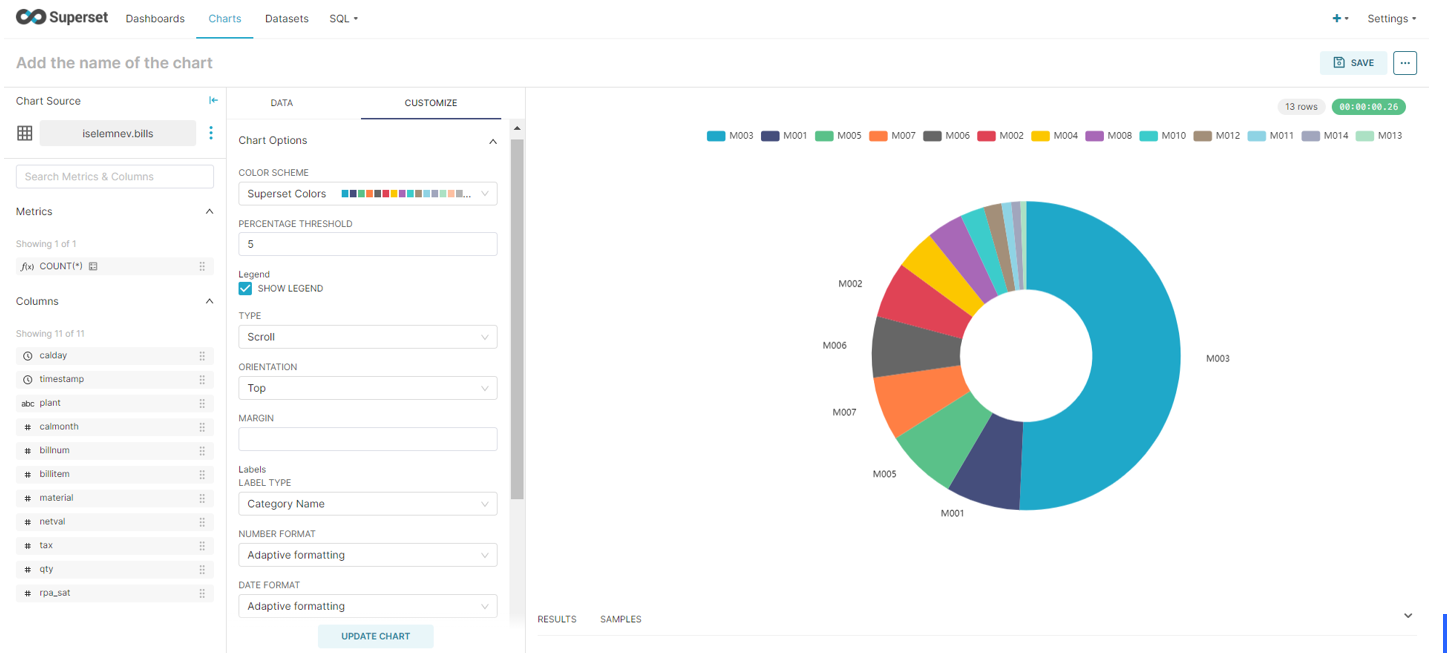
На вкладке Data в разделе Query можно выбрать режим запроса: агрегацию (AGGREGATE) или режим необработанных записей (RAW RECORDS), – определить признаки (DIMENSIONS) и показатели (METRICS), посчитать для показателей процент от общего числа (PERCENTAGE METRICS), добавить фильтр на колонки (FILTERS) и задать сортировку (SORT BY).



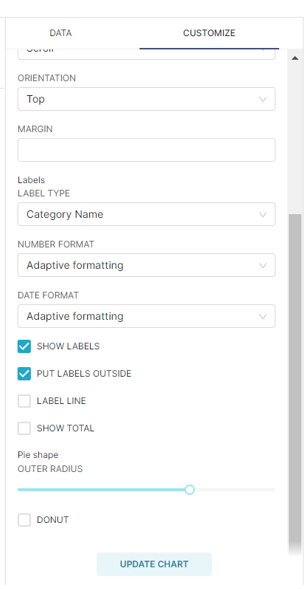
Создадим другой чарт с визуализацией Pie Chart на том же источнике, добавим поля «Магазин» и «Оборот».



На вкладке CUSTOMIZE можно выбрать наборы отображаемых цветов (COLOR SCHEME), определить порог процента, при котором будет отображаться подпись для магазина (PERCENTAGE THRESHOLD), включить отображение легенды для значений по цветам (Legend), выбрать тип легенды (TYPE) и её расположение (TYPE) и отступ (MARGIN). Можно также выбрать тип подписей (LABEL TYPE), их формат (NUMBER FORMAT) и формат для даты (DATE FORMAT), если она была бы визуализирована.

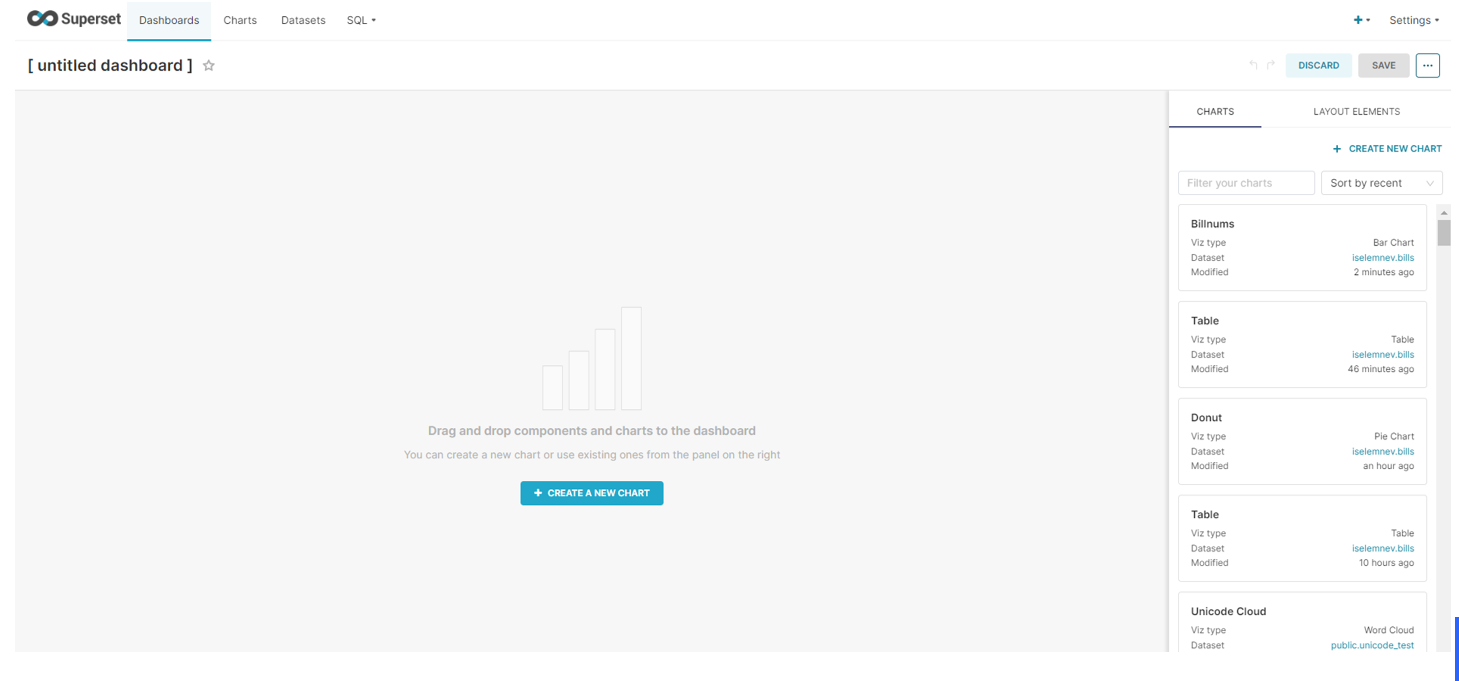


Также можно отключить отображение подписей (SHOW LABELS), поместить их снаружи или внутри (PUT LABELS OUTSIDE), сделать линии к соответствующим цветовым областям (LABEL LINE), показать total по показателю (SHOW TOTAL), можно задать радиус кружка, вырезать сердцевину и указать внутренний радиус.



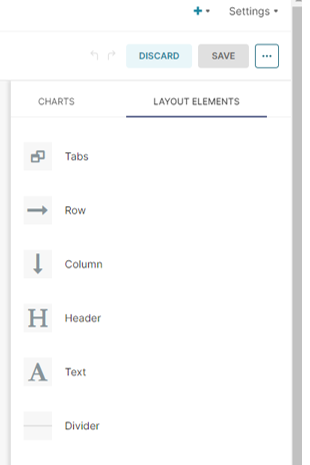
Для других чартов могут быть различные другие настройки.

На специальной вкладке можно создать свои дашборды:

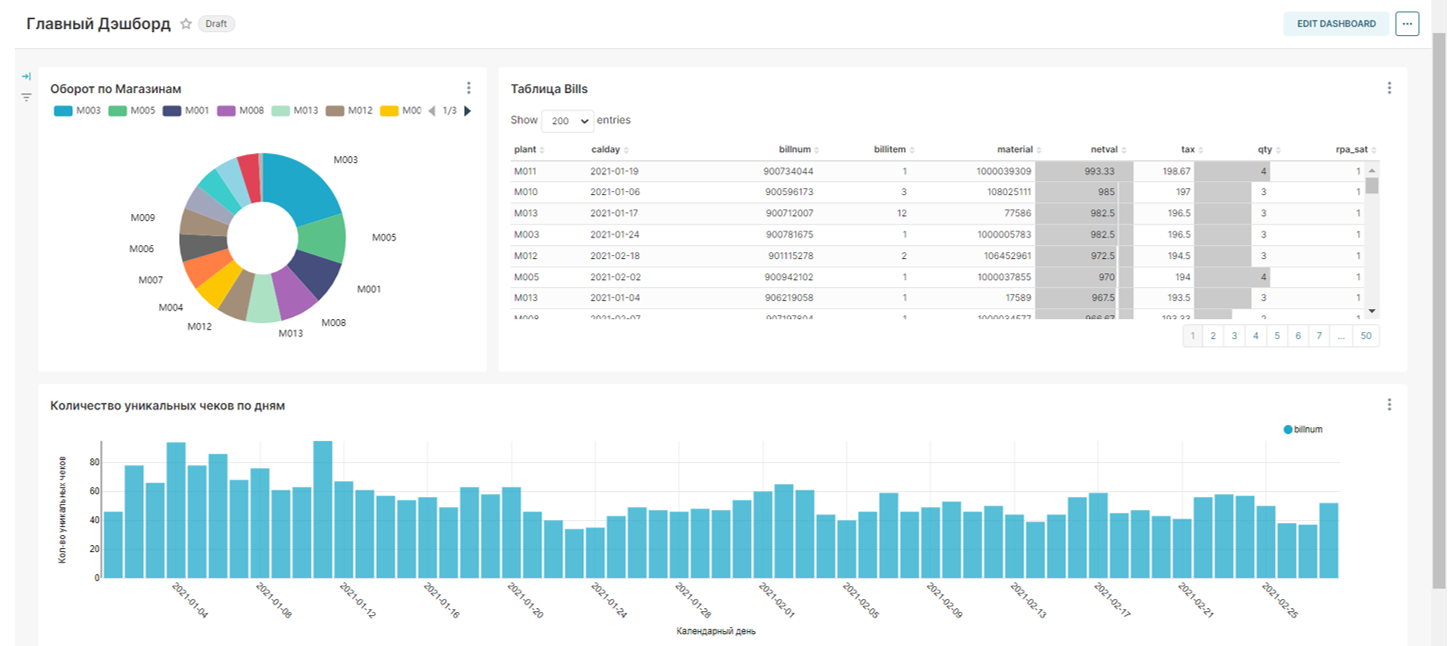


На панели справа отображаются существующие чарты, которые можно перетянуть в область дэшборда.

На вкладке LAYOUT ELEMENTS можно делать разметку области с помощью специальных элементов: создать несколько вкладок в дешборде, и добавлять заголовки или текст.



Пример созданного дэшборда на ранее созданных чартах:

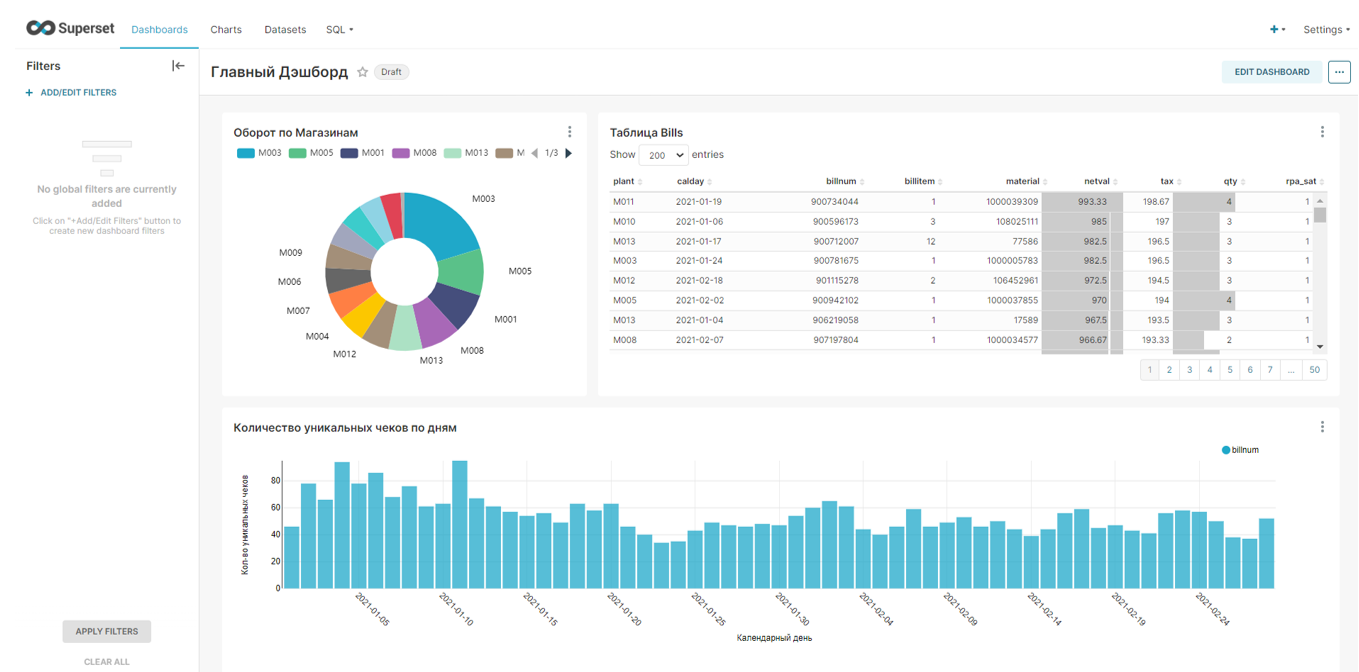


Также дополнительно добавлен новый, на котором отображается кол-во уникальных чеков по дням.

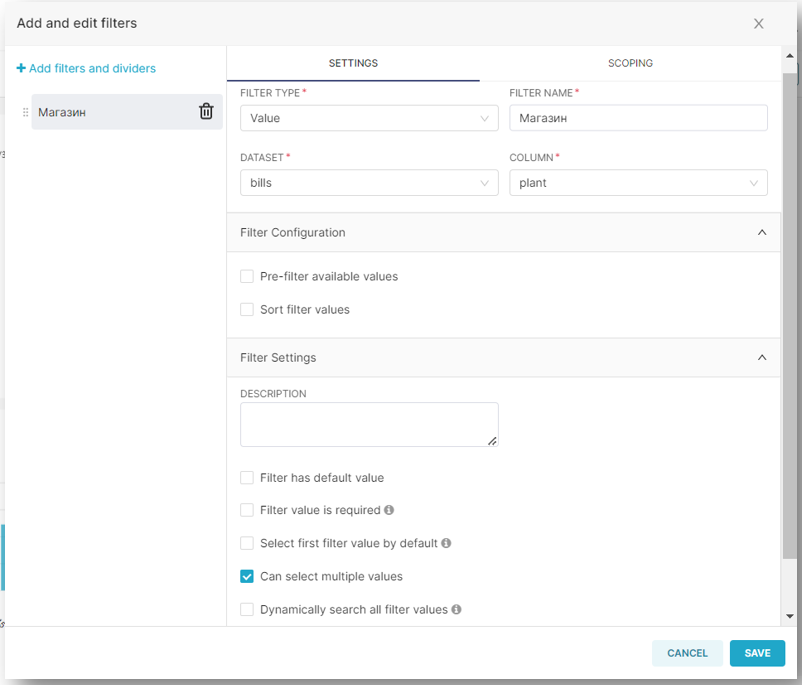
Другие примеры дэшбордов, которые уже имелись в Суперсете после его развёртывания см. на слайдах 103-106.

Фильтрация

На дэшборд можно установить определённые фильтры, открыв панель слева в режиме просмотра.

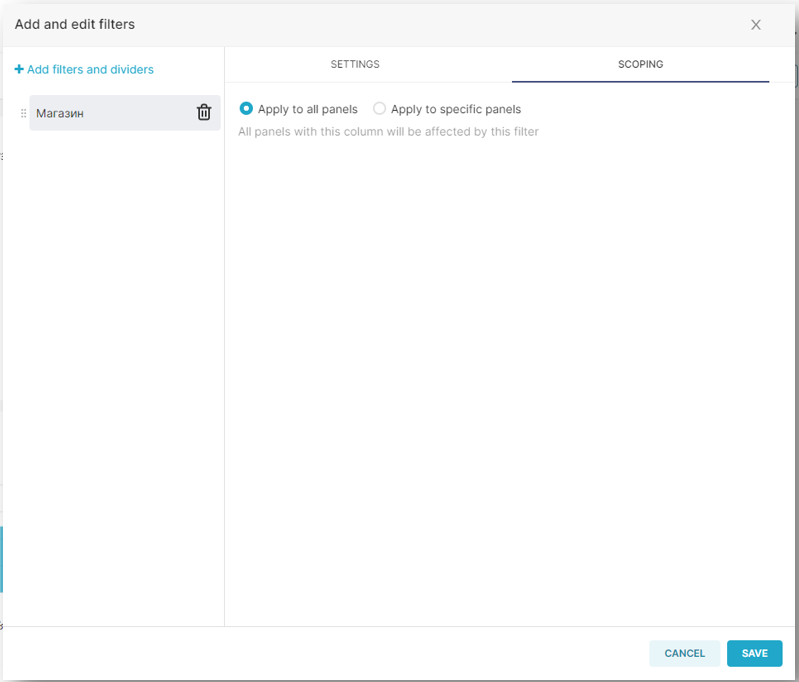


Можно указать тип фильтра (FILTER TYPE), его название (FILTER NAME), выбрать датасет (DATASET), и конкретную колонку из него (COLUMN).

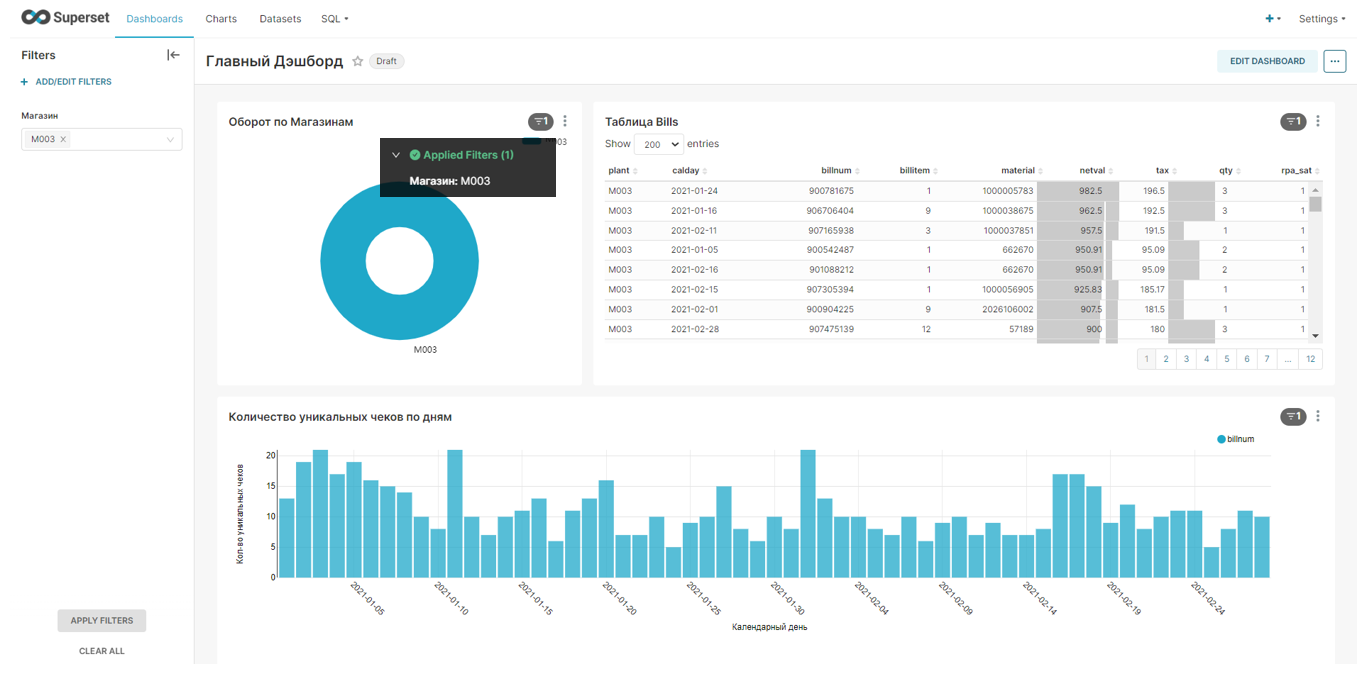


Также можно сделать предфильтр доступных значений основного фильтра (Pre-filter available values), отсортировать значения (Sort filter values). Можно добавить описание к фильтру (DESCRIPTION), указать значение по умолчанию (Filter has default value), сделать выбор значений для фильтра пользователем обязательным при его наложении (Filter value is required), сделать первое доступное значение фильтра значением по умолчанию (Select first filter value by default), предоставить возможность выбора нескольких значений (Can select multiple values), включить динамический поиск для фильтра, если он имеет слишком много значений (Dynamically search all filter values), и сделать фильтр исключающим по выбранным значениям.

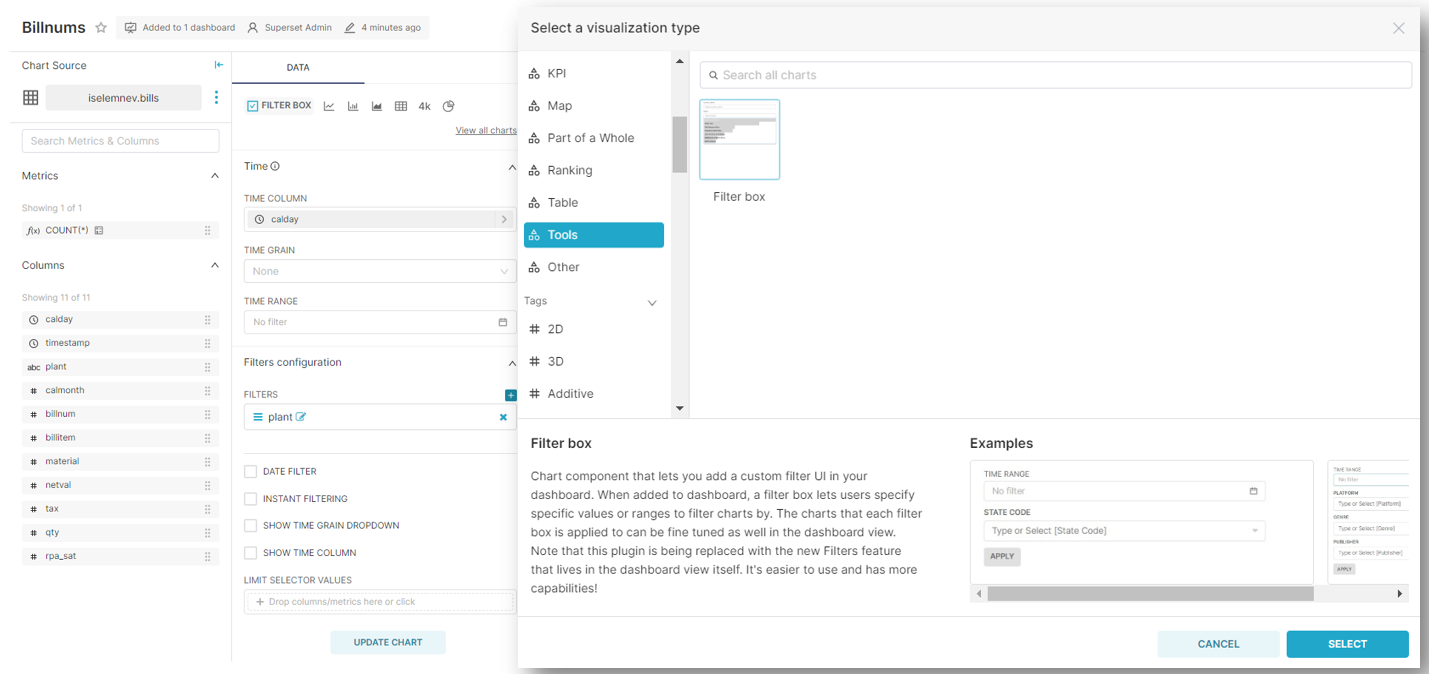
Настройками на вкладке SCOPING устанавливается применимость ко всем панелям или только к конкретным.



После применения фильтра на каждом чарте появляется специальный значок, который будет отображать в себе отфильтрованные значения:



Фильтры также можно создать отдельным чартом, выбрав тип визуализации Filter box:



И затем добавить его в дэшборд:

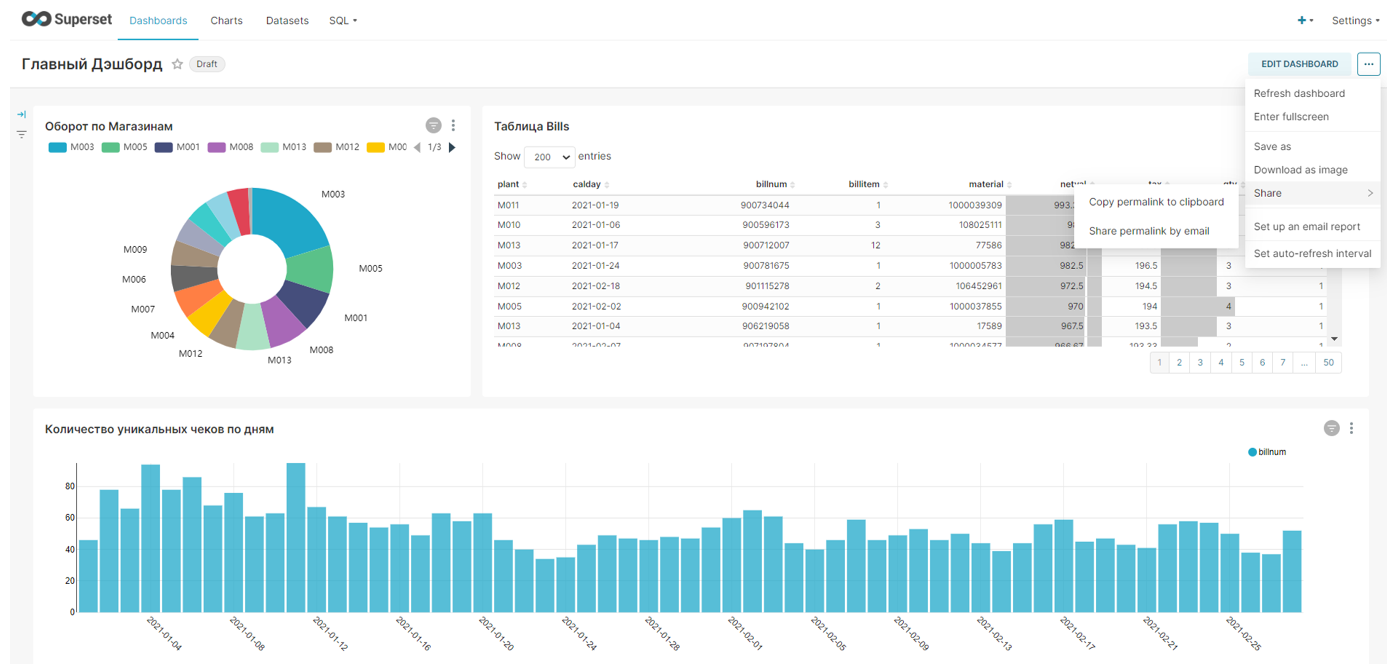


Также в Суперсете можно предоставить возможность делать кросс-фильтрацию при настройке их в чарт. Данная настройка должна быть включена для Суперсета. Кросс-фильтрация позволяет при выборе значения на одном чарте фильтровать по этому значению другие чарты.



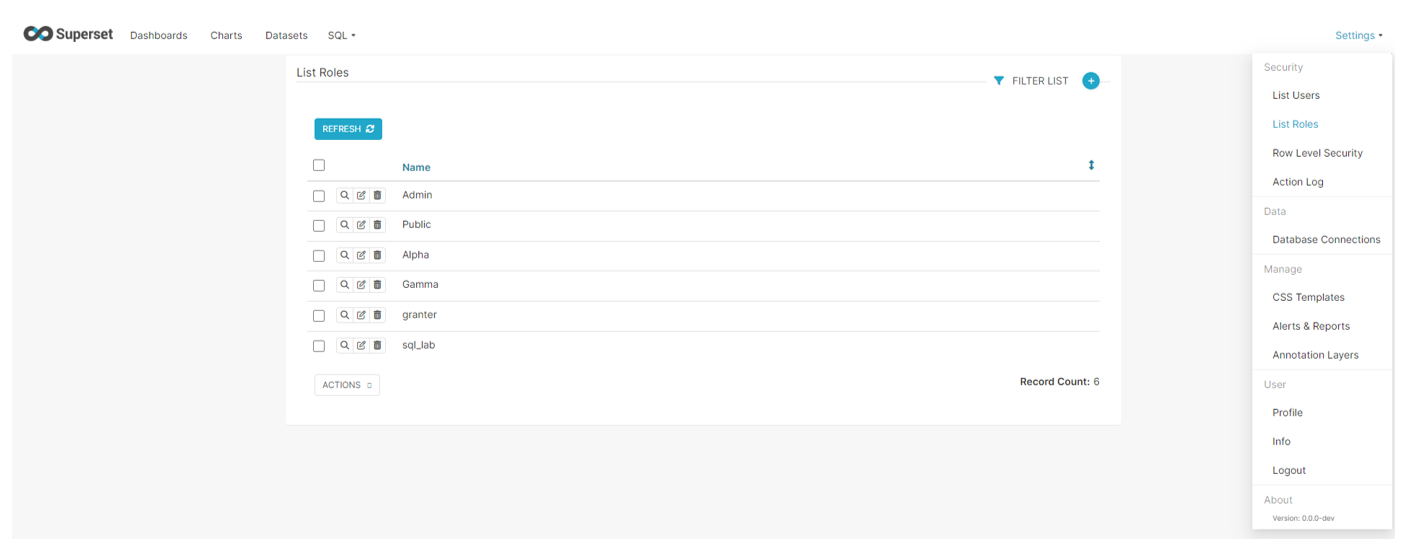
Распространение дэшборда

Мы можем скопировать ссылку на дэшборд и отправить её сразу по email.

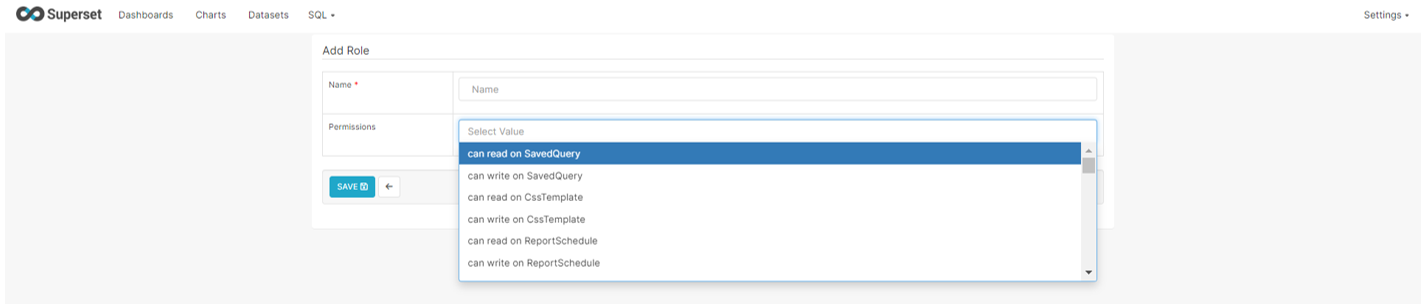


Роли

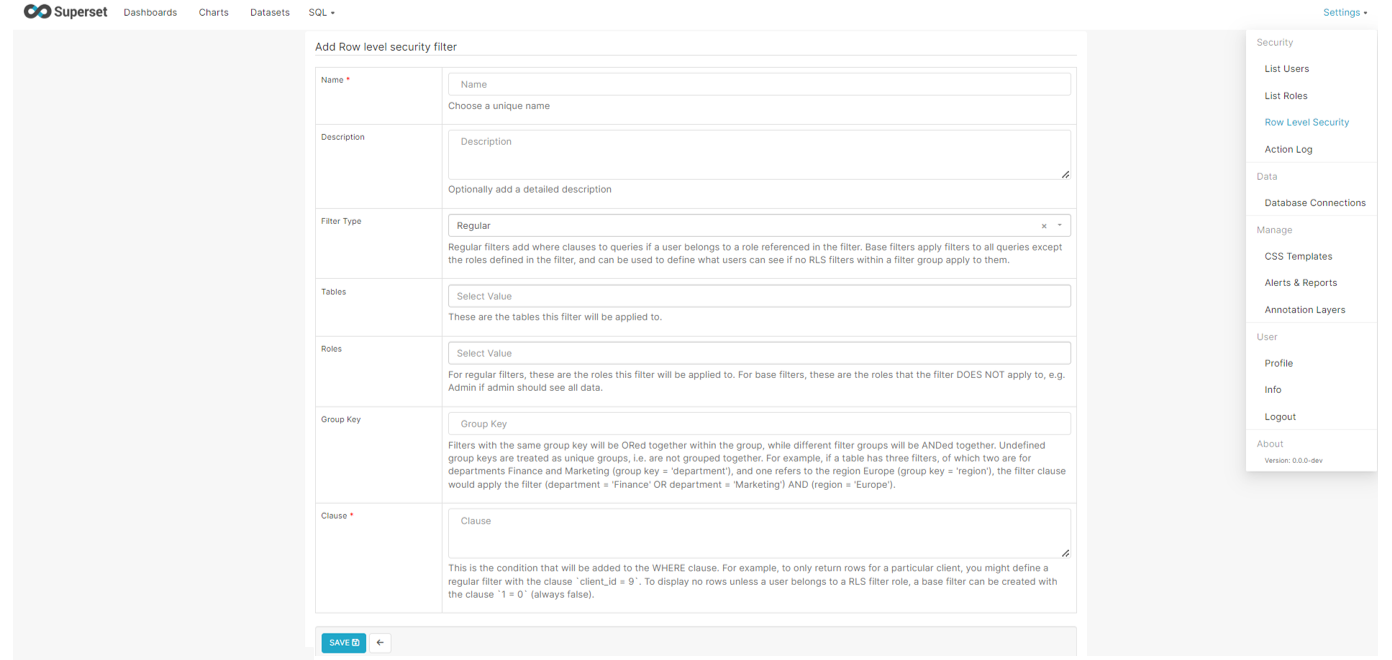
В Суперсете есть возможность создавать роли для пользователей:



настраивать их необходимым образом:



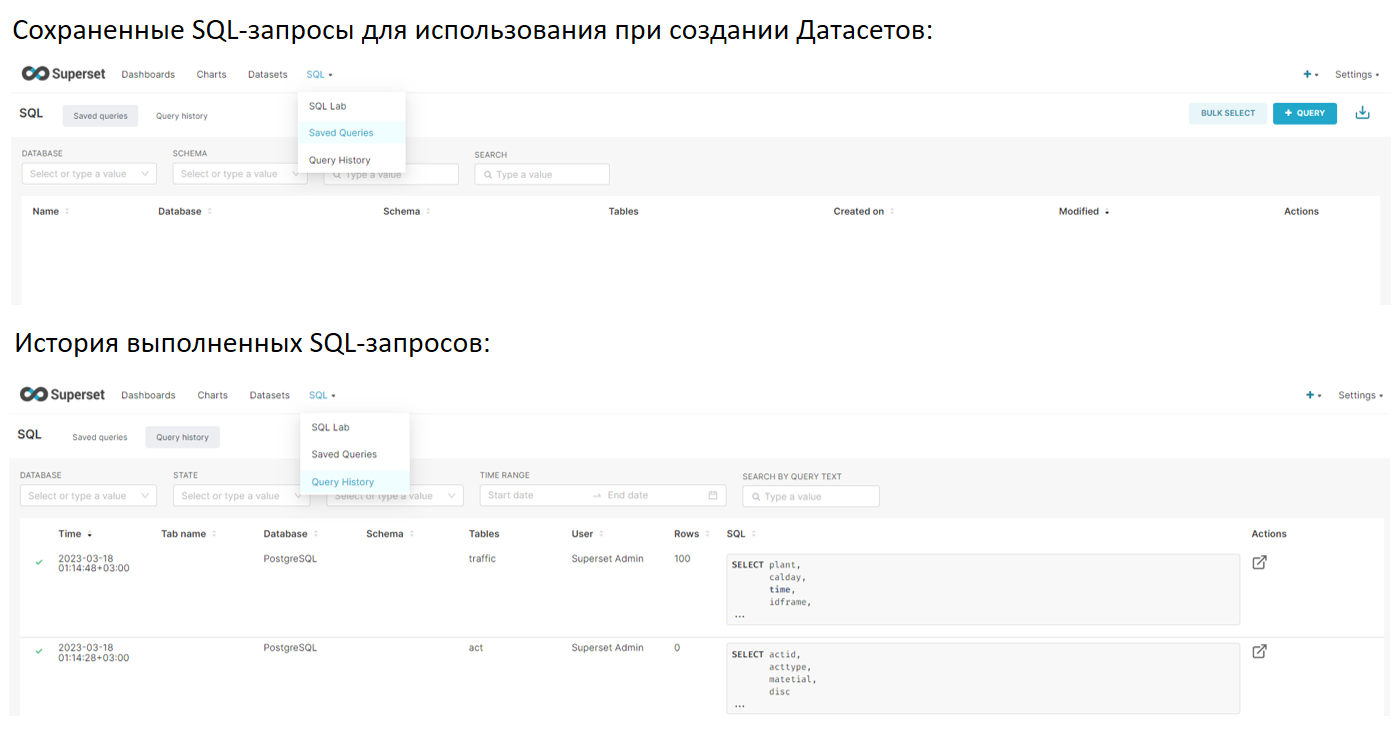
Также в Суперсет есть понятие Row Level Security:



Эта настройка позволяет ограничивать доступ к табличным данным для каких-либо ролей, прописав необходимые условия для этого в разделе Clause. Это условие WHERE будет добавляться ко всем запросам пользователя.

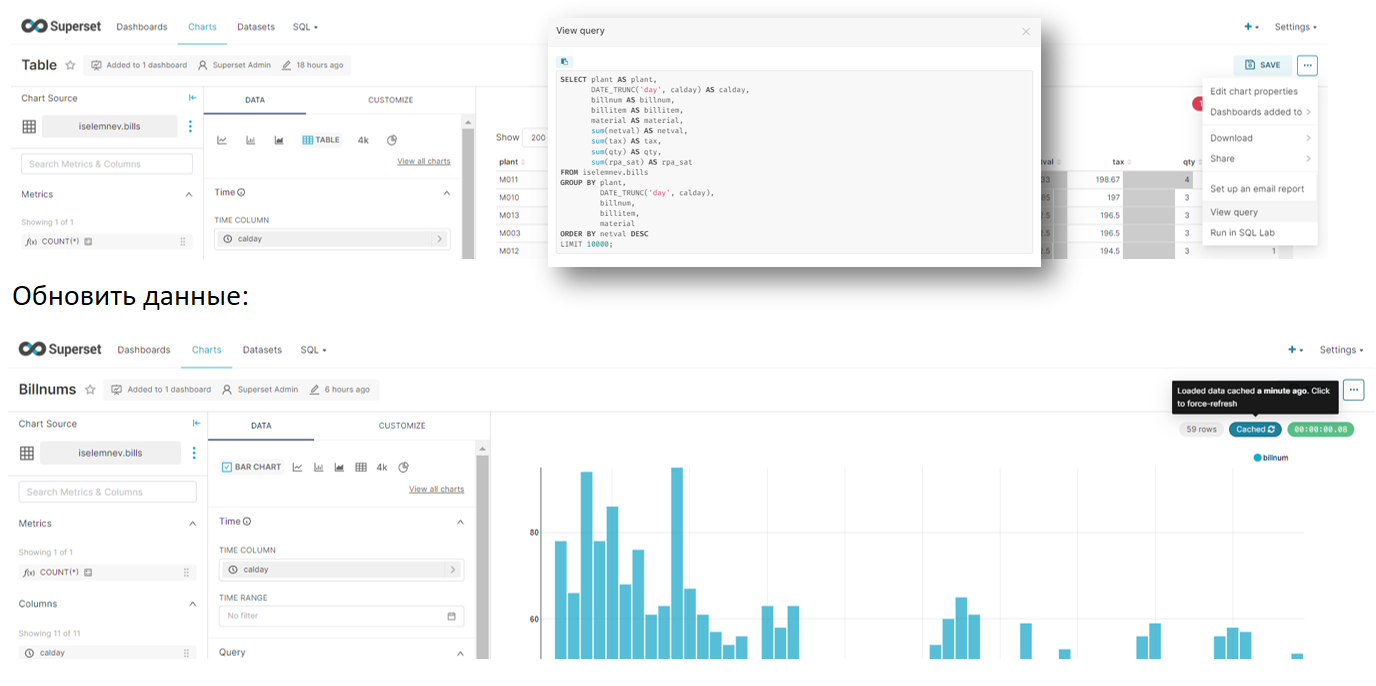
Сохранённые SQL запросы и история их выполнения

Также можем просмотреть на специальной вкладке сохранённые SQL-запросы из SQL Lab и историю их выполнения.



Отправляемые в БД SQL запросы и обновление данных

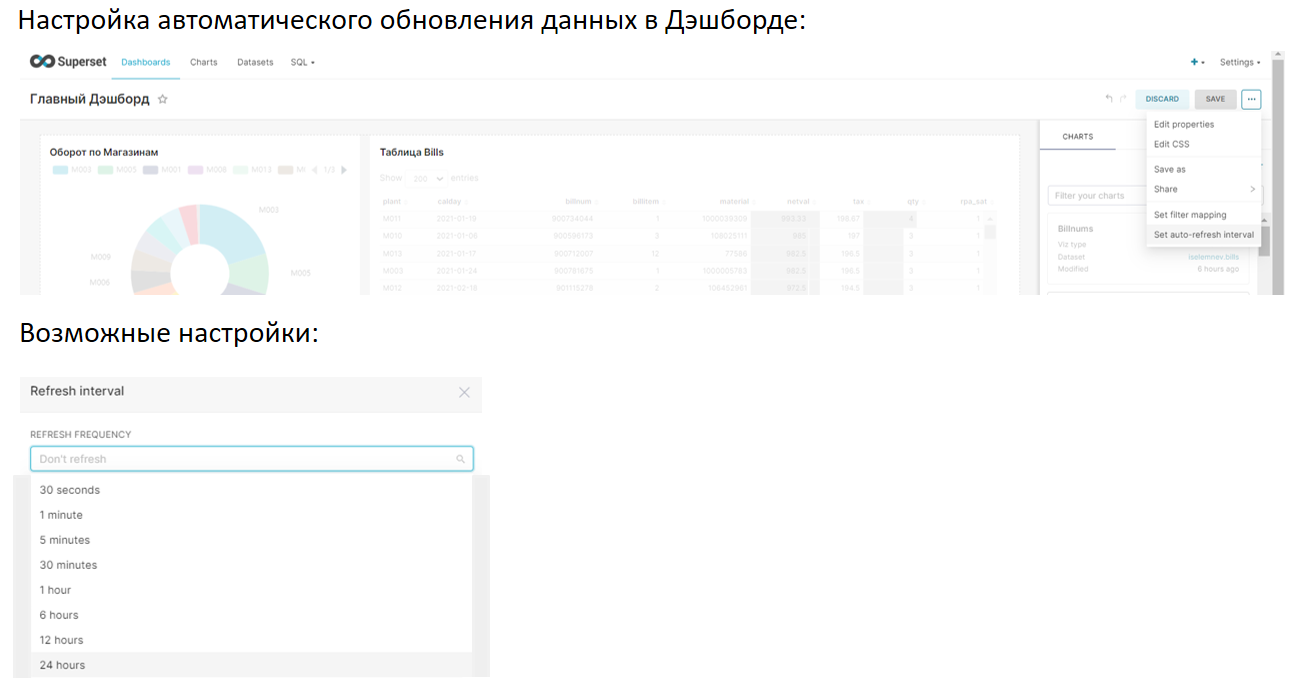
Каждый созданный объект отправляет запрос в подключенную БД посредством Python библиотеки SQL Alchamy. Сгенерированный запрос можно увидеть в следующем разделе меню:



По кнопке Cached есть возможность вручную обновить данные в чарте.

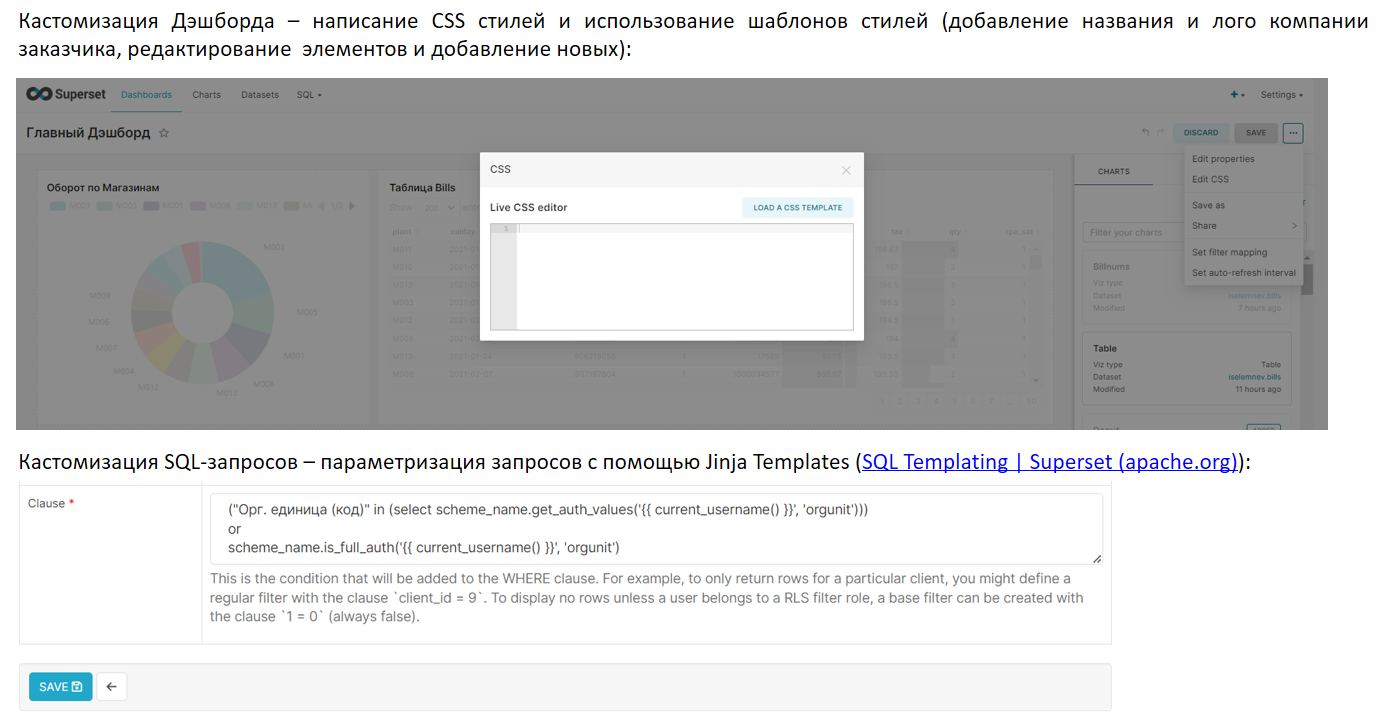
Автоматическое обновление данных

В дэшборде можно настроить автоматическое обновление данных. Возможные варианты настройки:

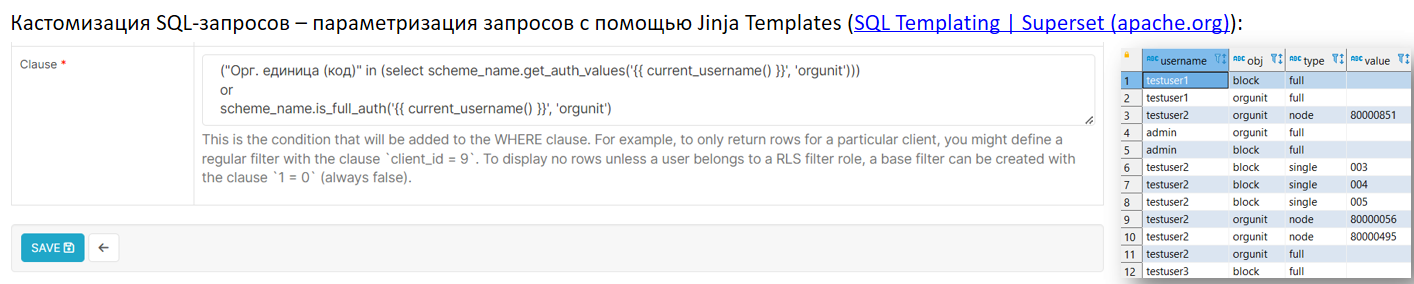


Кастомизация

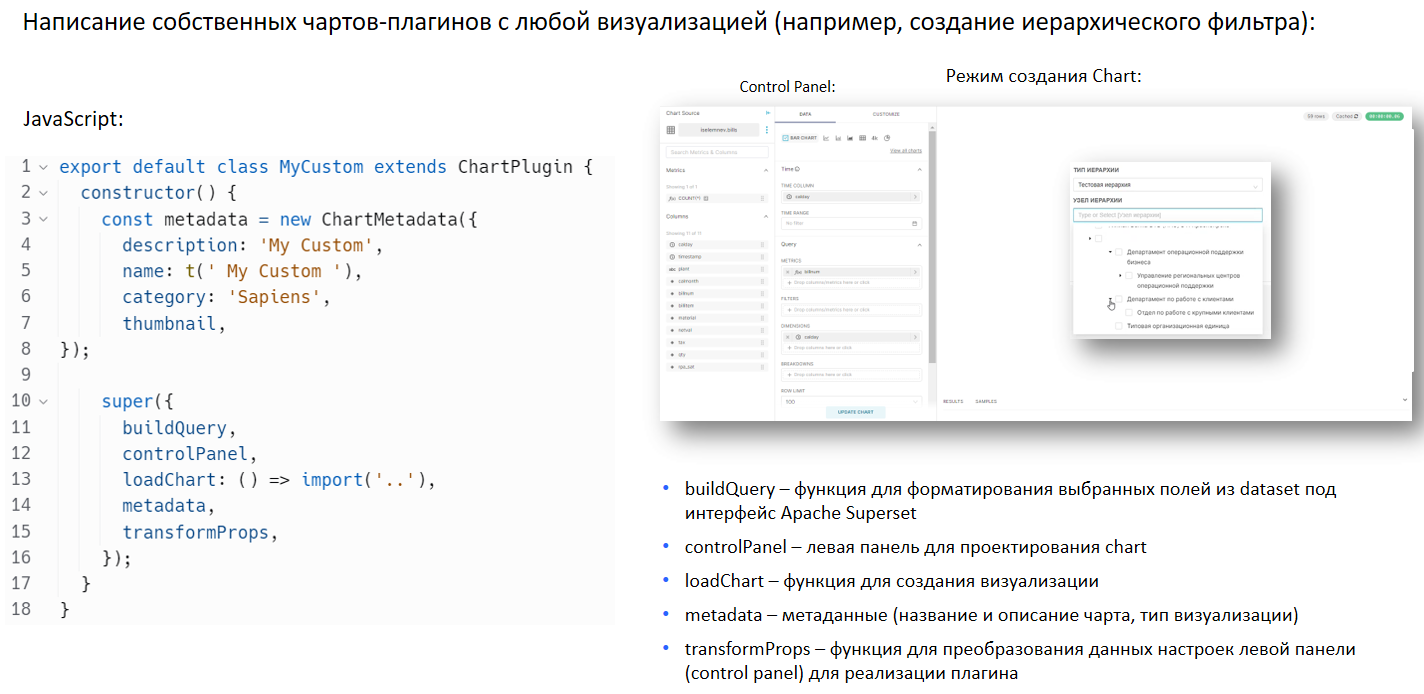
Т.к. Суперсет является Open Source решением и имеет открытый исходный код, то можно делать различного рода кастомизации, доработки и улучшения по своему усмотрению. Например, Суперсет сам предоставляет возможность прописывать свои CSS стили. Т.о. можно оформить дэшборд на свой вкус.



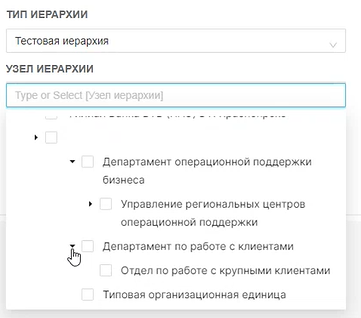
Для определения полномочий на просмотр данных в объектах мы можем параметризировать наш запрос, а именно условие WHERE. Будем передавать заранее написанную в Greenplum ф-цию и current\_username в качестве параметра. Вызываемая ф-ция будет доставать необходимые значения для конкретного пользователя из также ранее созданной таблицы. Данное условие будет накладываться при запросах к БД должным образом. Организационные единицы будут показываться только те, на которые пользователь имеет полномочия.



Для более продвинутых пользователей есть возможность писать собственные чарты с любой визуализацией, которые можно будет подключать как NPM пакеты.



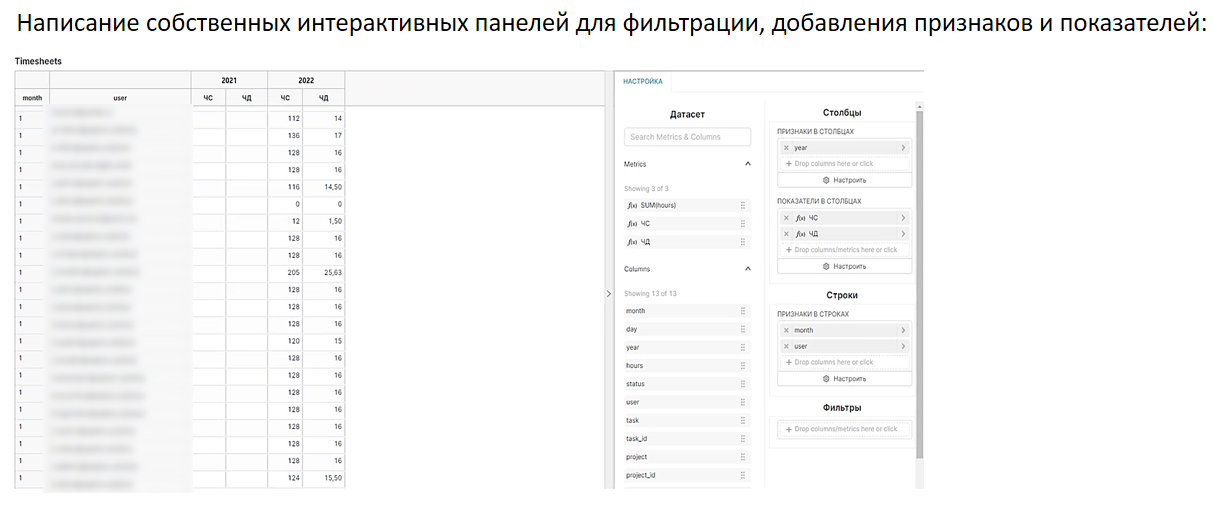
Увеличенная копия панели справа с предыдущего рисунка:



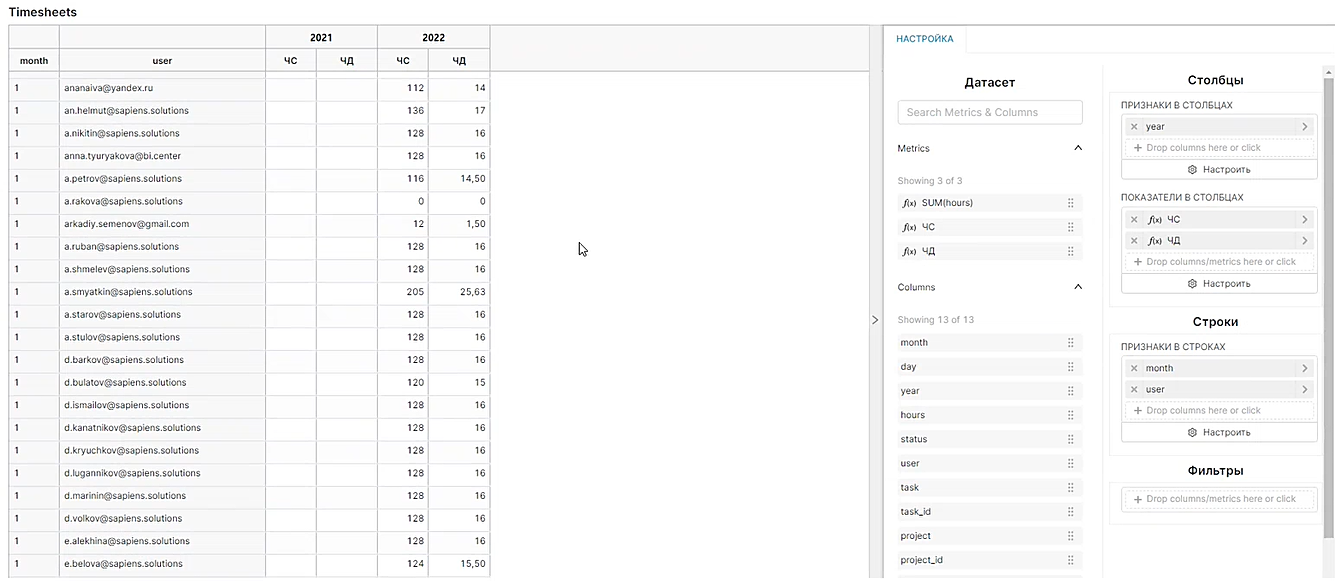
В рамках данного курса этот пункт подробно рассматриваться не будет.

Кастомизация: собственные интерактивные панели

Можно разрабатывать для дэшбордов собственные интерактивные панели в рамках одного чарта с таблицей, чтобы предоставить пользователям возможность гибко им управлять, не залезая в настройки чарта.







Кастомизация: расширение функционала и новые вкладки

А также можно дорабатывать код Суперсета, расширять его функционал и добавлять новые вкладки. Например, предоставить возможность загружать файлы или шаблоны:

