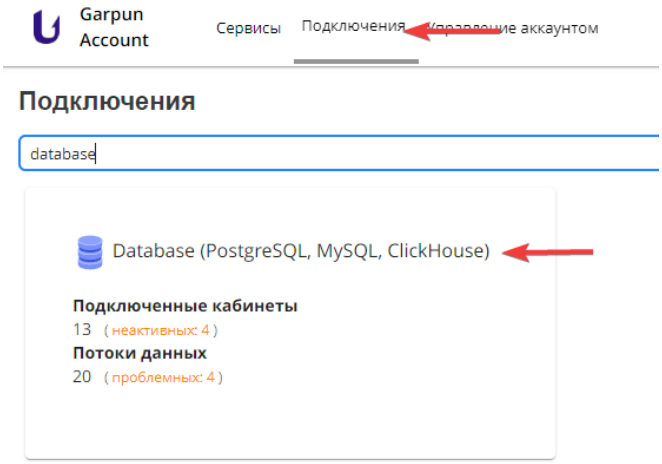
Импорт в базу данных(PostgreSQL, MySQL или ClickHouse)

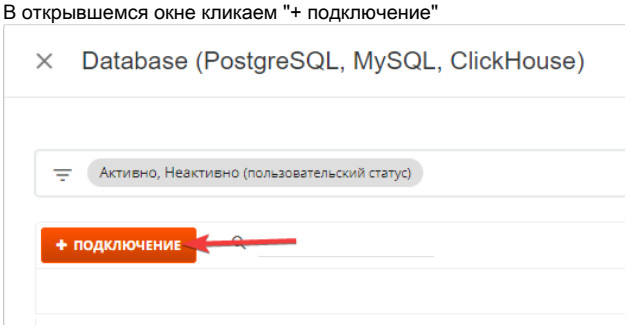
Базы данных как приемник данных являются альтернативой Google BigQuery.
Сейчас в базы данных можно провести загрузку из большинства систем, из которых доступен импорт в Google BigQuery.

Чтобы система могла забирать и загружать данные из базы, её необходимо корректно подключить.
Процесс подключения расписан отдельно, поскольку сама технология организации связи с базой универсальна. Но, поскольку есть множество разных видов баз, некоторые настройки указываются не в потоках (как например при работе с Google BigQuery) а в самом подключении.
Настройка подключения едина как на импорт так и на экспорт данных.

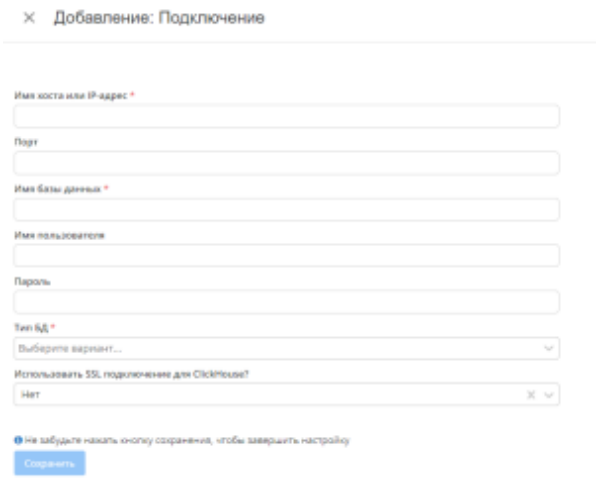
1. Переходим в раздел "Подключения" в вашем профиле Garpun. Можно так же воспользоваться ссылкой
2. Ищем в списке систем пункт Database (PostgreSQL, MySQL, ClickHouse), кликаем на него.



3. В открывшемся окне кликаем "+ подключение"



4. Далее приступаем к вводу необходимых параметров:



а. Имя хоста или IP-адрес - адрес обращения к базе, обязательный параметр. Эту информацию можно уточнить у администратора вашей базы, если вы им не являетесь.
б. Порт - необязательный параметр. В зависимости от того, как организована база, для подключения к ней может быть необходим определенный порт
в. Имя базы данных - указывать обязательно. Без этого параметра система не будет знать к какому объекту обращаться при загрузке

/выгрузке данных
г. Имя пользователя и пароль - необходимо указывать если они нужны для доступа к данным и работе с базой
д. Тип БД - указывать обязательно. У каждого типа БД свои особенности подключения и отправки запросов. В данный момент на сервисе есть поддержка PostgreSQL, MySQL и ClickHouse
е. SSL подключения для ClickHouse используется только для ClickHouse если в этом есть необходимость
5. Нажимаем "Сохранить"

Отличия облачного хранения и сервера во внутреннем контуре:

1. Облачное хранение позволяет использовать ClickHouse как сервис в облаке, что значительно упрощает управление и масштабирование инфраструктуры. Облачные поставщики предоставляют готовые образы ClickHouse, которые можно развернуть в несколько кликов.
2. Сервер во внутреннем контуре предполагает использование собственных вычислительных ресурсов для установки и настройки ClickHouse. Это требует больше времени и усилий, но также позволяет настроить систему под конкретные потребности.
3. Облачное хранение может быть более экономичным в плане затрат на оборудование и поддержку инфраструктуры. С другой стороны, сервер во внутреннем контуре обеспечивает большую гибкость и контроль над системой.

Если у вас защищенная база и доступ осуществляется только с разрешенных IP-адресов, просьба обратиться в поддержку Garpun за получением актуального списка наших адресов, с которых ведется подключение к базам.

При выгрузке большого объема данных из BQ в ClickHouse(>10 гб в одной партиции) используется мультипоток:

1. BQ -> S3 GCS (Google Cloud Storage). Тут мы сохраняем данные из BQ в Google Cloud Storage в формате Parquet
2. S3 GCS -> S3 (Yandex Object Storage). Передаем данные между S3 хранилищами
3. S3 (Object Storage) -> Clickhouse. Финальная передача готовых данных из S3 Yandex в ClickHouse.

Рассмотрим создание потока по передаче данных в базу на примере передачи статистики из Google BigQuery в ClickHouse

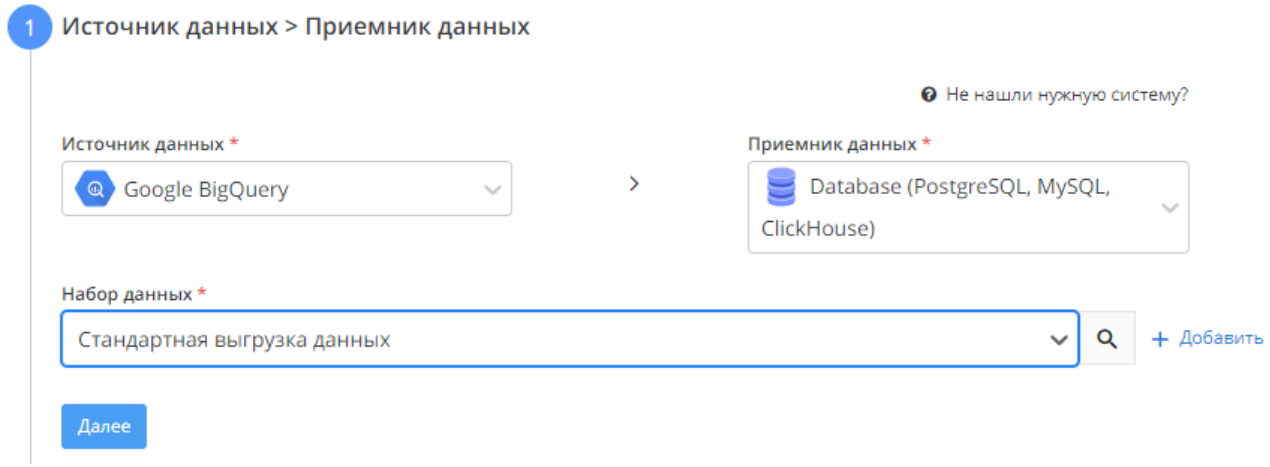
Первоначально нам необходимо войти в систему https://feeds.garpun.com/ и нажать на

1) Источник данных > Приемник данных

В качестве источника данных выбираем **Google BigQuery,** в качестве приемника - **Database (PostgreSQL, MySQL, ClickHouse)**.

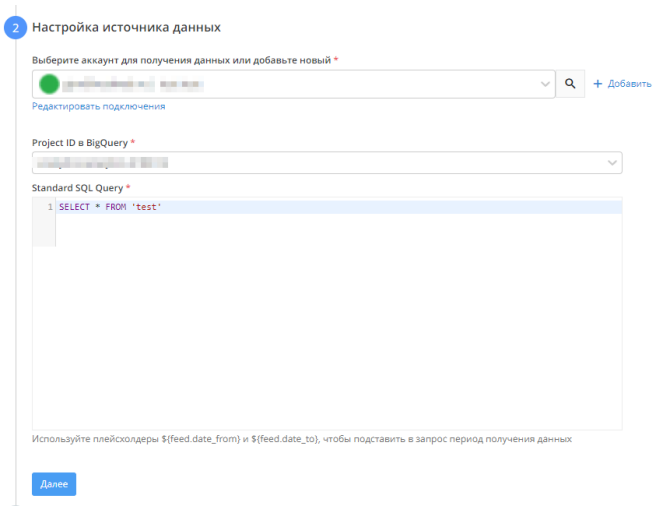
После выбора источника и приемника появится выпадающий список с возможными наборами данных. Для каждой системы они могут отличаться в зависимости от метрик, которые передаются.

Чтобы посмотреть, какие параметры будут передаваться, необходимо нажать на значок лупы справа от набора данных.



2) Настройка источника данных

На втором этапе выбираем аккаунт, или нажимаем кнопку + Добавить, для того, чтоб добавить новое подключение, указываем Project ID и используем SQL-запрос, чтоб определить данные, которые будут выгружены

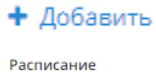


3) Настройка приемника данных

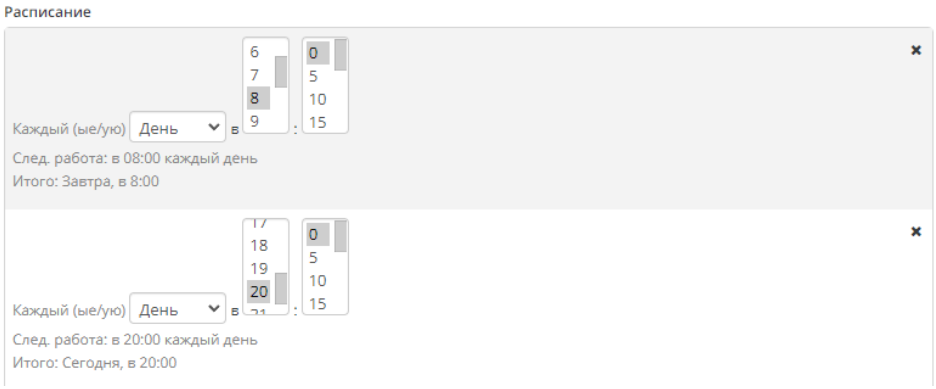
- Выбираем существующее подключение в списке, либо добавляем новое с помощью соответствующей кнопки
- Выбираем название базы данных. Это необходимо для корректной передачи информации и создания таблиц.
- Название схемы базы данных. Необходимо заполнять только если ваша база работает на PostgreSQL
- Указываем способ записи данных в таблицу. По умолчанию установлен способ "обновить".
- При выгрузке в ClickHouse можно включить доп.опцию, которая будет создавать реплицированные таблицы в разных узлах кластера для обеспечения сохранности полученных данных

4) Общие настройки

- В графе “Название потока” ввести название либо оставить сгенерированное автоматически
- В графе “Период сбора при автоматическом запуске” можно выбрать за какой период фид будет осуществлять пересбор статистики.
- В графе “Расписание” - выбрать например 8:00 утра, в это время фид будет запускаться ежедневно. При нажатии на кнопку



можно добавить дополнительную строку, таким образом фид будет отрабатывать по более гибкому график



- Нажать “Готово”

Результат

В результате отработки потока будет создана таблица с указанной схемой.

Параметры движка, а именно сам движок, поле для партицирования, группировка и index_granularity задаются автоматически:

MergeTree – движок по умолчанию

PARTITION BY и ORDER BY – зависят от выбранного в потоке набора данных

index_granularity – 8192

allow_nullable_key – 0 по умолчанию, может быть 1, если на третьем шаге настроек потока поставить чекбокс у настройки "Включить nullable поля для ключевых параметров"

ON CLUSTER `{cluster}` – по умолчанию не используется, можно включить, если на третьем шаге настроек потока поставить чекбокс у настройки "Создавать реплицированные таблицы для Clickhouse"

TTL - может использоваться в некоторых наборах данных, которые подразумевают создание временных таблиц.

В описание таблицы автоматически будет добавлена подобная информация:

```
    Garpun Feeds
: Yandex Direct -> Database (PostgreSQL, MySQL, ClickHouse) (ID: 11caef1c-ccfb-42de-add8-972c862da48d)
: Yandex Direct Connector
:   CH
```

ID из описания таблицы можно использовать для того, чтоб найти поток данных, который записывает в конкретную таблицу:

https://feeds.garpun.com/card?e=2770&o=11caef1c-ccfb-42de-add8-972c862da48d&a=garpun_feeds