

Загрузка данных из BQ в S3

Garpun Feeds может выгружать файлы из Google BigQuery в S3.

S3 — сервис хранения неструктурированных данных. Внутри можно хранить все что захотим, но в основном используется как файловое хранилище. Каждый файл представлен в виде объекта, сами же объекты лежат в бакетах.

Про его настройки можно отдельно почитать в документации самого сервиса: <https://cloud.google.com/storage/docs/introduction>

Перед началом работы нужно добавить подключения к BQ и S3(если оно уже организовано, переходим к этапу [созданию потока](#)):

Авторизовавшись в системе Garpun переходим в раздел "Подключения"([ссылка](#)), выбираем S3, нажимаем "+подключение".

×

Добавление: Подключение

Хост хранилища *

Название подключения

Идентификатор ключа, который вы получили при генерации статического ключа *

Секретный ключ, который вы получили при генерации статического ключа *

❗ Не забудьте нажать кнопку сохранения, чтобы завершить настройку

Сохранить

- **Хост хранилища** - расположение хранилища в сети, его настраиваете при работе с хранилищем и хостингом
- **Название подключения** - то, как будет называться наше подключение.
- **Идентификатор ключа** - ID ключа шифрования в вашем хранилище. Не сам ключ а именно ID, система по этому ID будет отправлять запрос на работу с данными.
- **Секретный ключ** - Один из ключей шифрования, который генерируется при помощи статического ключа. Нужен для обращения к данным.

Авторизовавшись в системе Garpun переходим в раздел "Подключения"([ссылка](#)), Выбираем Google BigQuery, нажимаем "+подключение".

×

Добавление: Подключение

Google BigQuery *

🔗 Подключить

❗ Не забудьте нажать кнопку сохранения, чтобы завершить настройку

Сохранить

- Нажимаем кнопку подключить
- Выбираем необходимый аккаунт
- Нажимаем "Разрешить"

1. Источник данных > Приемник данных

В качестве источника выбираем Google BigQuery, в качестве приемника - хранилище S3, в качестве набора данных - "Стандартная выгрузка данных".


1

Источник данных > Приемник данных

?


Не нашли нужную систему?

Источник данных *

 Google BigQuery

>

Приемник данных *

 S3

Набор данных *

Стандартная выгрузка данных

+ Добавить

Далее

2. Настройка источника данных

2

Настройка источника данных

Выберите аккаунт для получения данных или добавьте новый *

Выберите вариант...

Q

+ Добавить

Редактировать подключения

Project ID в BigQuery *

Выберите вариант...

Standard SQL Query *

1

Используйте плейсхолдеры `$(feed.date_from)` и `$(feed.date_to)`, чтобы подставить в запрос период получения данных

Далее

- Выбираем аккаунт для получения данных
- Выбираем проект, из которого будем брать данные
- Прописываем SQL запрос

Используйте плейсхолдеры `$(feed.date_from)`, `$(feed.date_to)`, `$(feed.datetime_from)` и `$(feed.datetime_to)`, чтобы подставить в запрос период получения данных. Например, `WHERE date BETWEEN '$(feed.date_from)' AND '$(feed.date_to)'`

3. Настройка приемника данных

[illegible]

- Выбираем аккаунт для загрузки данных
- Название бакета. Инструкция по неймингу бакета для GCS: <https://cloud.google.com/storage/docs/buckets#naming>, для Yandex Object Storage - <https://cloud.yandex.ru/docs/storage/concepts/bucket#naming>
- Пресет для обработки и генерации файла. Если пресетом выбрано 'Нативная выгрузка из BigQuery (Parquet)', важно, чтобы юзер ВQ имел доступ к бакету
- Путь к файлу

- Создавать ли подпапки для партиций. При использовании опции в указанной папке будут созданы подпапки, разбитые по датам.
- "Использовать динамическое название файла". При включенной опции каждый запуск потока будет создавать в приемнике новый файл по выбранному шаблону

Выберите формат динамического названия файла *

Добавить к названию файла префикс с диапазоном сбора данных.[диапазон сбора]_[название]

Добавить к названию файла префикс с диапазоном сбора данных.[диапазон сбора]_[название]

Добавить к названию файл префикс с датой сбора данных.[дата запуска]_[название]

- Время хранения в днях.

4. Общие настройки

На этом этапе вам необходимо изменить название потока если необходимо. Название потока генерируется автоматически.

Выбрать период сбора при автоматическом перезапуске. По умолчанию устанавливается "на основе внутренних правил", что означает, что пересбор потока будет происходить за последние 30 дней.

Установить расписание

Нажать "Готово"