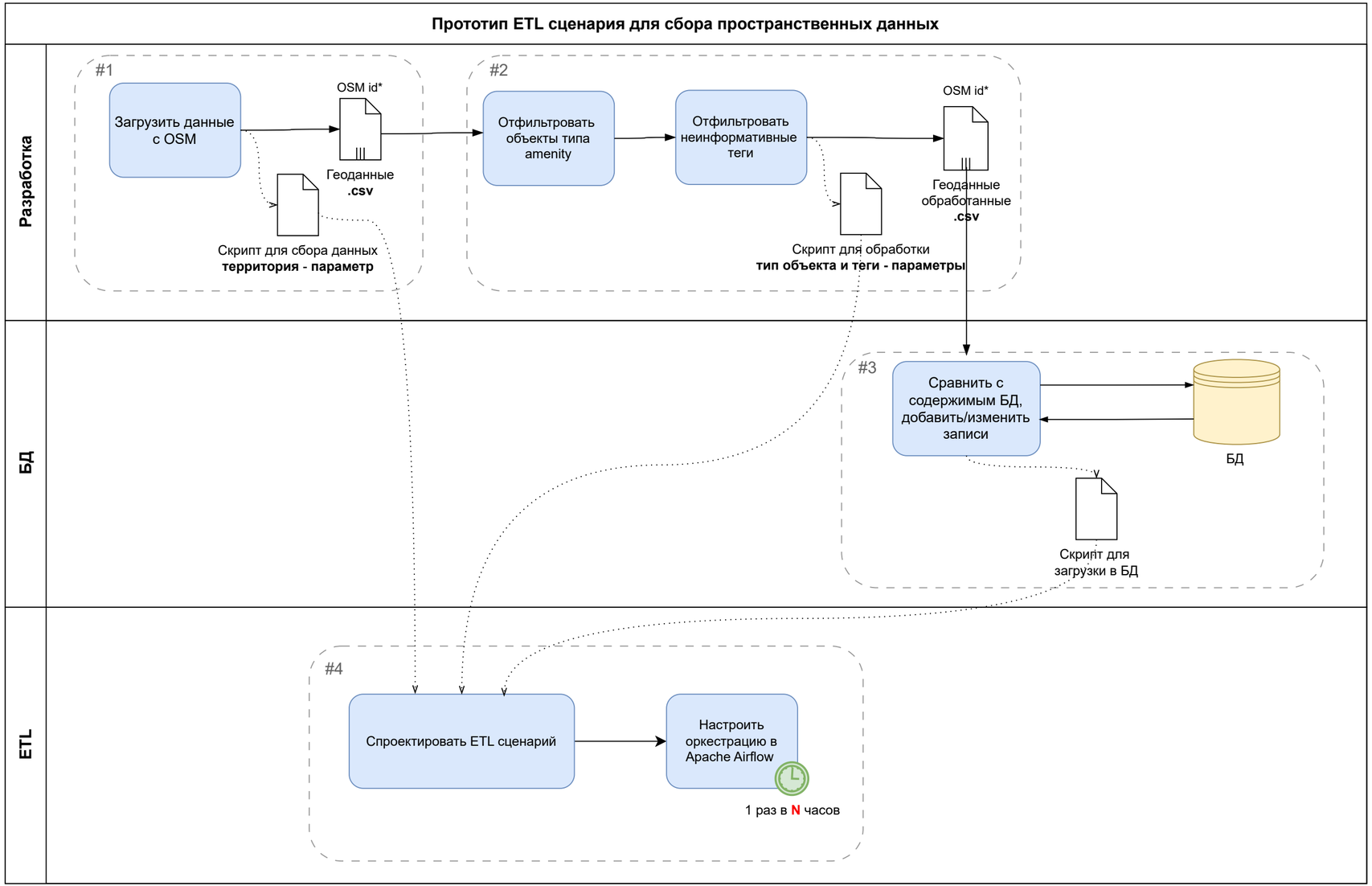

Задачи

#1 Прототип - Сбор данных.

Написать скрипт для сбора данных с OSM с помощью библиотеки osmnx.

На вход передается параметр - территория, с которой собирать данные.

На выходе - .csv файл со всеми полученными данными по объектам.

Важно, чтобы по каждому объекту было сохранено OSM id.

#2 Прототип - Обработка данных.

Написать скрипт для отбора данных типа amenity и фильтрации тегов.

На вход подается

- .csv файл со всеми полями полученными по объекту
- параметр types - тип(ы) объектов, которые хотим хранить

Необходимо:

1. Отобрать объекты types. Пока считаем types == amenity.
2. Оставить только "нужные" теги (Список нужных тегов - результат по задаче #12 https://github.com/citec-spbu/Spatial-Data-ETL/issues/12)

На выходе - .csv файл отфильтрованными объектами и тегами.

Важно, чтобы по каждому объекту было сохранено OSM id.

#3 Прототип - Загрузка данных в БД.

Написать скрипт для загрузки данных из .csv файла в Postgres.

При добавлении записей проверять наличие объекта в бд по OSM id. Если такой объект существует, внести изменения, если нет - добавить в базу

#4 Прототип - ETL процесс

Из полученных скриптов в задачах #1, #2, #3 собрать ETL процесс в Apache Airflow. Протестировать запуск процесса "по кнопке". Предусмотреть возможность запуска по таймеру.

#5 Установщик (идея на будущее)

Написать установщик, который поставит пользователю нужную версию Postgres и Docker для корректной работы ETL процесса при клонировании из нашего репозитория