Тестовое задание Data Engineer

Описание задачи

Предположим данные о событиях пользователей(сессиях) поступают каждый день в виде json(например, из Kafka). Нам необходимо ежедневно загружать этот json в ClickHouse и там его обрабатывать для удобного хранения всей истории пользовательских сессий.

План выполнения

1. По инструкции с https://clickhouse.yandex/ установить в контейнере сервер ClickHouse
2. Посмотреть примеры работы с ClickHouse в документаци. Это поможет понять типовые подходы к загрузке данных.
3. Скачать тестовый датасет https://yadi.sk/d/ARJShvDUgazjMQ. В нем хранится пример потока данных для обработки, Нужно исходить из того, что каждый день будет приходить такой кусок данных.
4. Развернуть в контейнере Airflow и реализовать загрузку в ClickHouse в виде тасок в Airflow
5. Загрузку данных реализовать в два этапа (отдельные таски airlow)
    a. загрузка сырых данных json в том виде как они есть (временно храним пока не придет вторая таска и не обработает)
    b. преобразование (ETL) средствами ClickHouse “сырой” таблицы в таблицу оптимизированную для работы с большим объемом данных поступающих каждый день.

В качестве решение необходимо предоставить:

1) docker-compose
2) Скрипты по созданию таблиц для ClickHouse - с описанием структуры таблиц и того почему выбран тот или иной движок таблиц и его параметры
3) Скрипты airflow с тасками.
4) Результаты представить в виде pull request в github. Предоставить к своей репе доступ пользователю SKhakhulin и на него же сделать pull request