**Задание 1:**

Для выполнения первого задания был выбран следующий датасет: <https://www.kaggle.com/datasets/asaniczka/public-opinion-on-climate-change-updated-daily>.

Перед загрузкой данных из него в YDB, в ней была создана таблица (**ydb\_scripts.sql**). Для загрузки данных в таблицу YDB была использована команда ***ydb import file csv -p climate\_change\_comments -i file.csv***. После загрузки данных (Рисунок 1) были созданы эндпоинты трансфера (Рисунок 2) и после был выполнен сам трансфер (Рисунок 3-4).

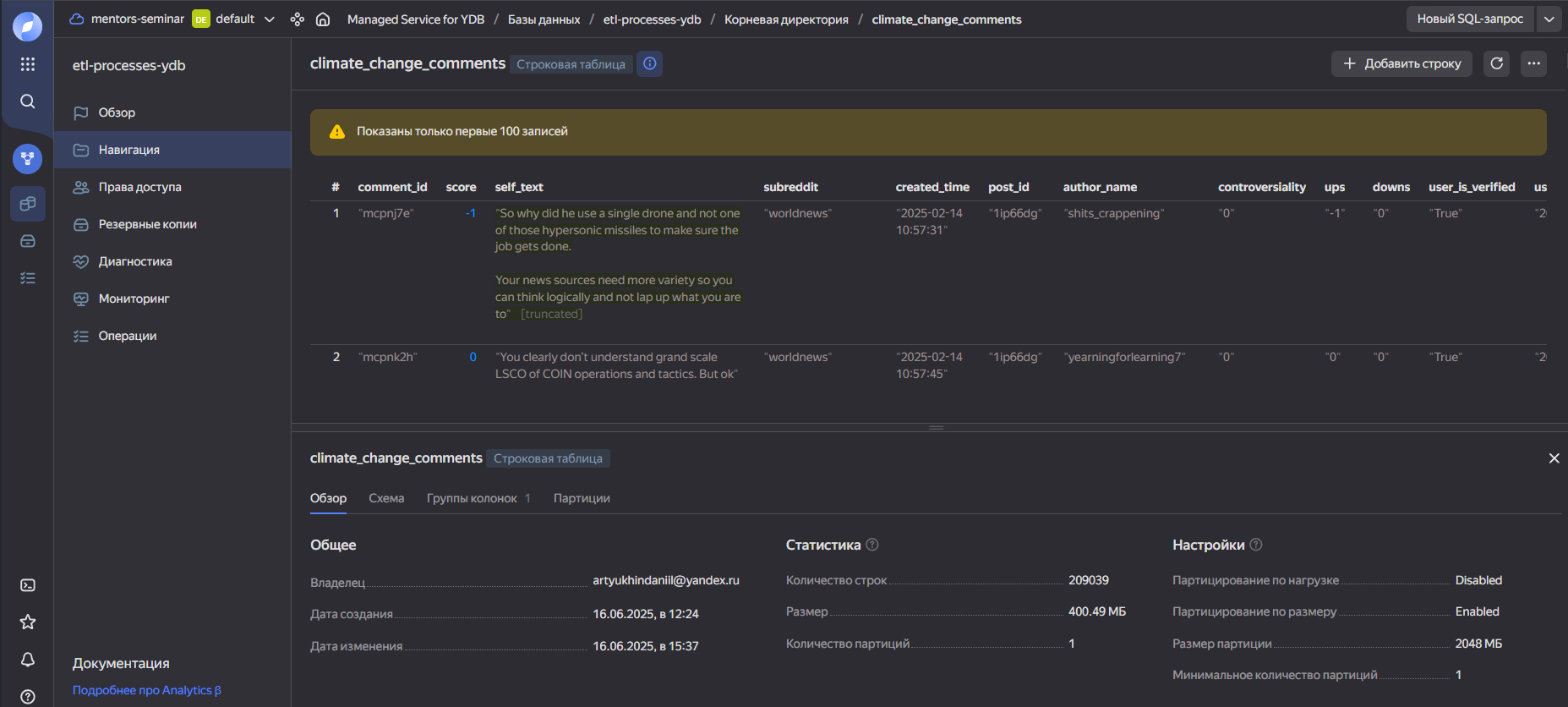


Рисунок 1

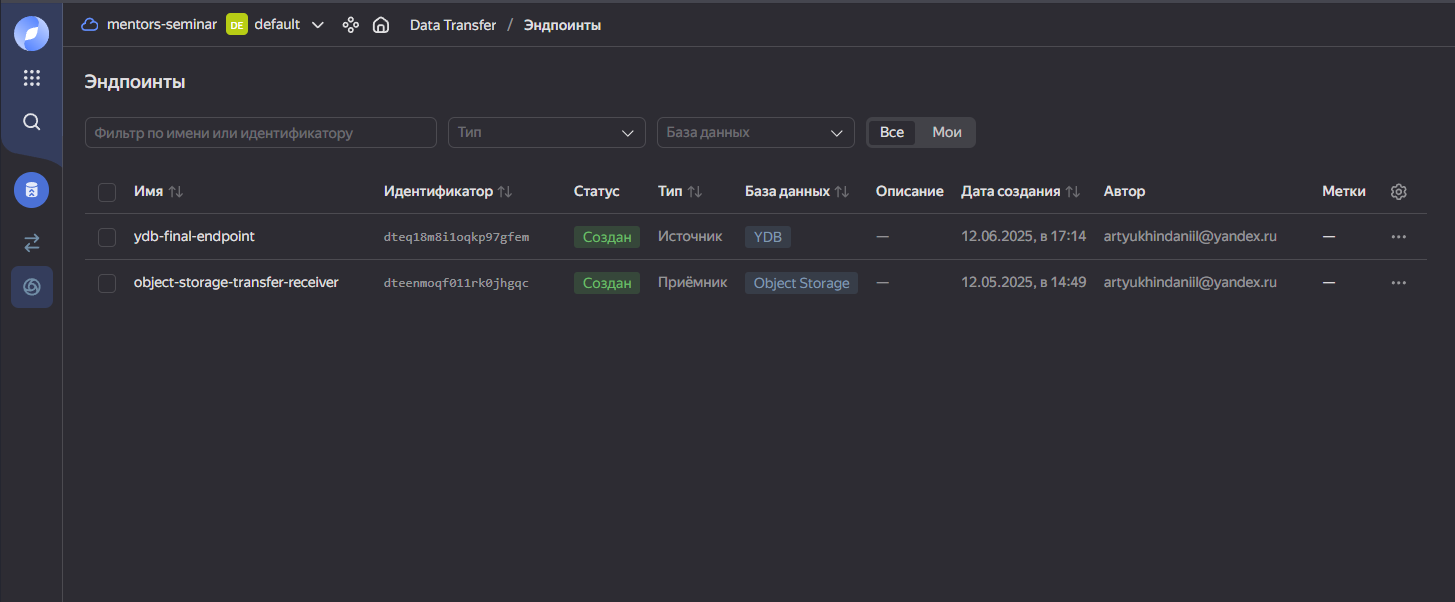


Рисунок 2

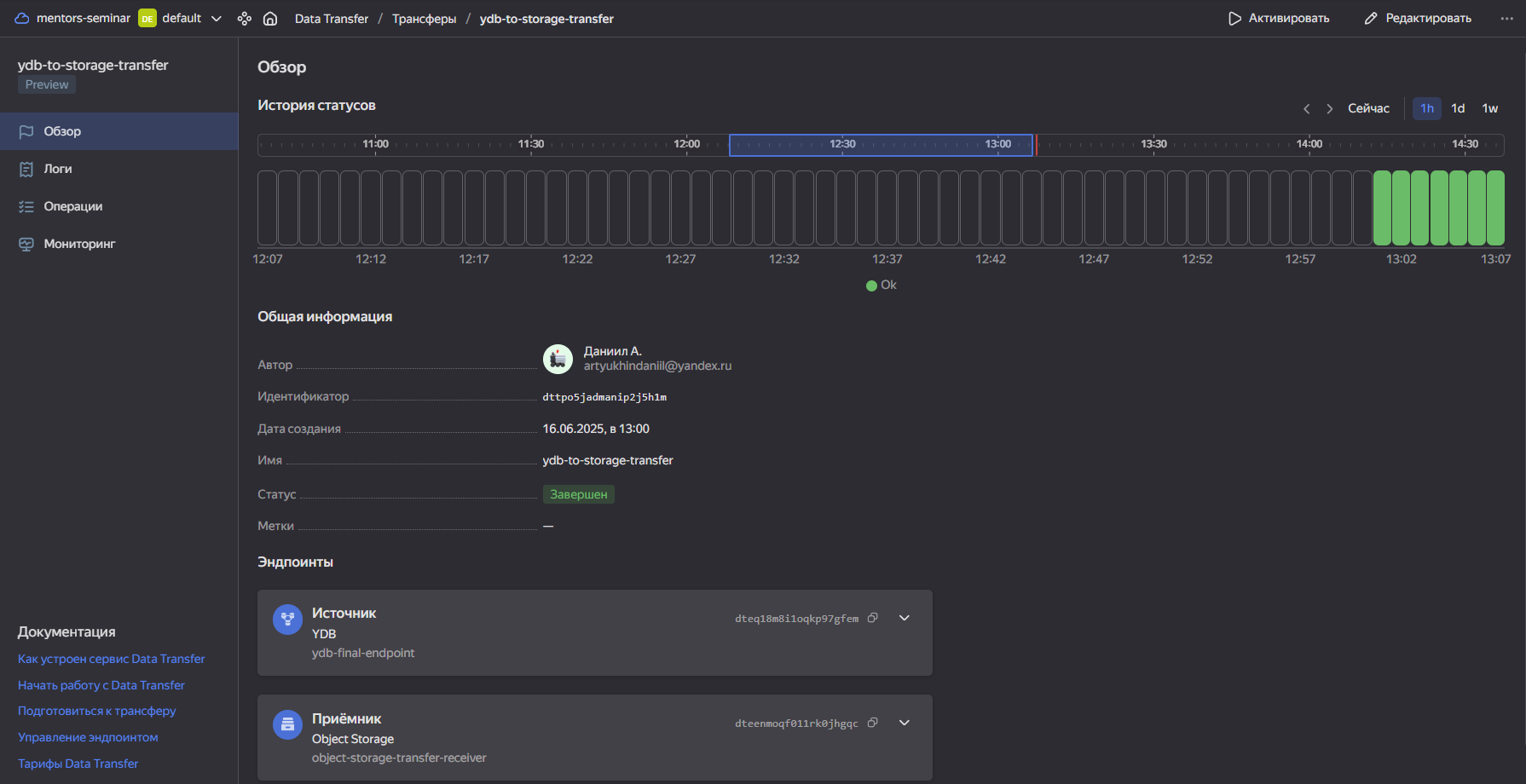


Рисунок 3

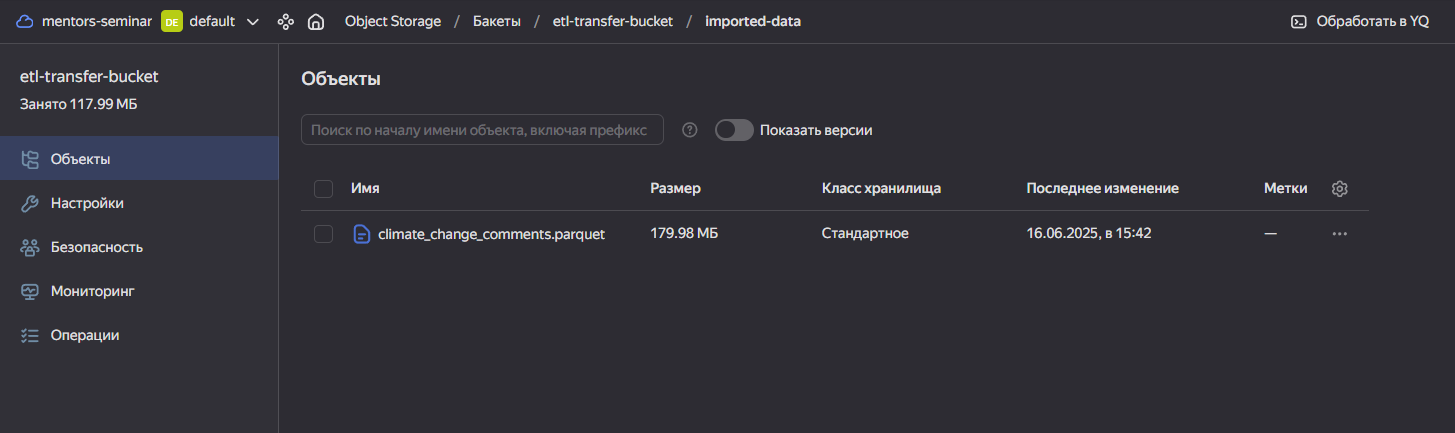


Рисунок 4

**Задание 2:**

Для выполнения данного задания нам необходимо создать Airflow DAG (**dataproc-cloud-dag.py**), который будет создавать временный кластер Yandex Cloud Dataproc, создавать в нем задание PySpark (**spark-task-script.py**) с написанным нами скриптом, и удалять кластер после выполнения этого задания Pyspark. После, мы запускаем данный DAG в Managed Service For Airflow и проверяем результат (Рисунок 5-6).

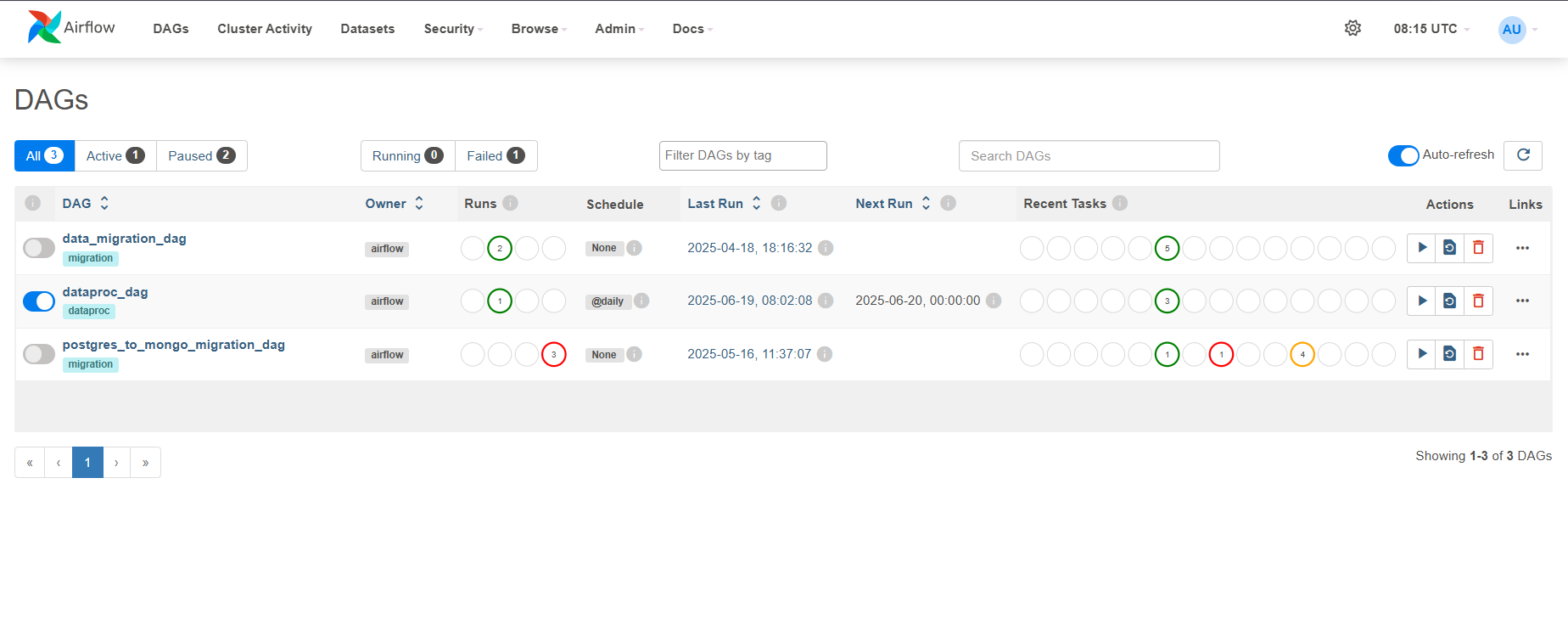


Рисунок 5

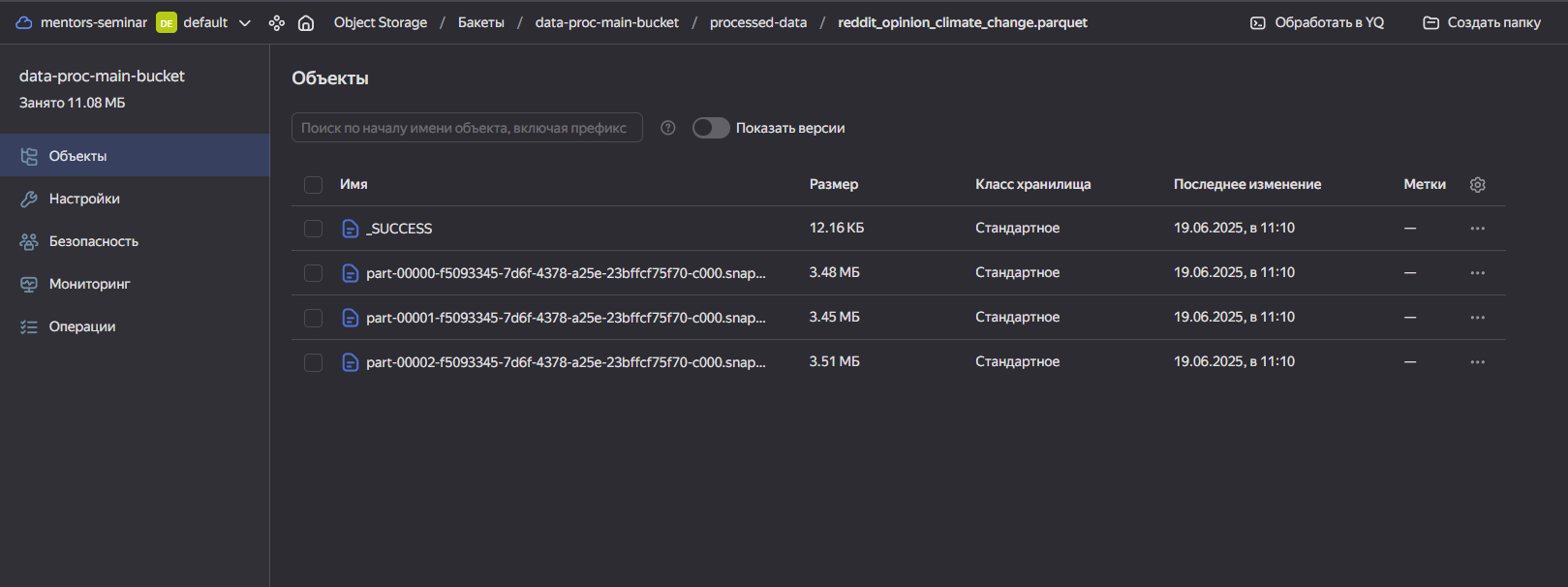


Рисунок 6

**Задание 3**