**Опциональная лабораторная работа №2**

**Олейников Михаил Николаевич (olejnikov)**

В текущей работе вам потребуется использовать продукт **RT.DataVision**, который развернут в рамках нашего курса. Если появятся вопросы, пожалуйста напишите в общий учебный чат или личным сообщением преподавателю.

**Цель задания**: попрактиковаться с инструментом Apache SuperSet.

1. Используйте интерфейс Apache SuperSet (http://vm-datavision.test.local:8090/superset/welcome/ ) и ваш логин / пароль, который вам был предоставлен на курсе.
2. Создайте три разных типов графиков (Charts) на основании вашего Dataset который определен в подключении к GreenPlum (указан как объект GP)
3. Разместите все три типа графиков в одном Дашборде (Dashboard) и сохраните его под своей фамилией, например yakupov\_lab11
4. Вы можете использовать любые ваши сырые данные (также можно взять данные в HDFS, которые мы использовали в первых работах). Загрузите данные в GreenPlum и трансформируйте их в информацию, используя подход виртуальных таблиц (View) или материализованных представлений (Materialized Table). На данном этапе можете применить Apache Airflow для создания расписания и механизма (кода) по трансформации данных

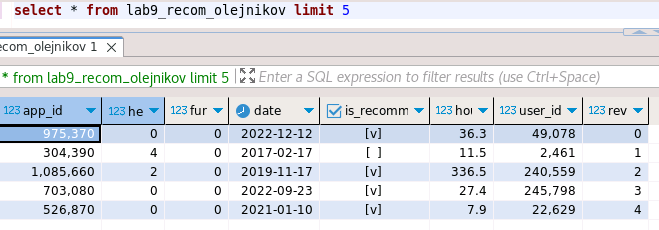
**Решение:**

Я сначала хотел попробовать через Import Data в Greenplum датасет импортировать. Так как доступа к yakupov нет, а у домашней папки ограничение 1 ГБ, ничего не оставалось, кроме как через Hadoop скачать сет ratings\_export.csv. Запустил импорт, оценочно пришлось бы ждать почти час. Остановил, опыт получен. Переходим к плану Б.

Используя HDFS файл **/user/yakupov/data/recommendations.csv** строим интеграцию Greenplum -> HDFS, используя PXF подход и EXTERNAL таблицу (беру датасет из ДЗ № 8):

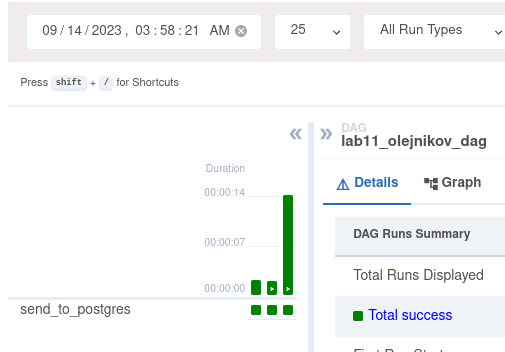
Импортируем HDFS файл **/user/yakupov/data/recommendations.csv** в Greenplum (код ДЗ № 8).

| create external table lab9\_recom\_olejnikov (  app\_id bigint,  helpful int,  funny int,  date date,  is\_recommended boolean,  hours numeric,  user\_id bigint,  review\_id bigint ) location ('pxf://user/yakupov/data/recommendations.csv?PROFILE=hdfs:csv&SERVER=hadoop') format 'csv' (delimiter ',' header); |
| --- |

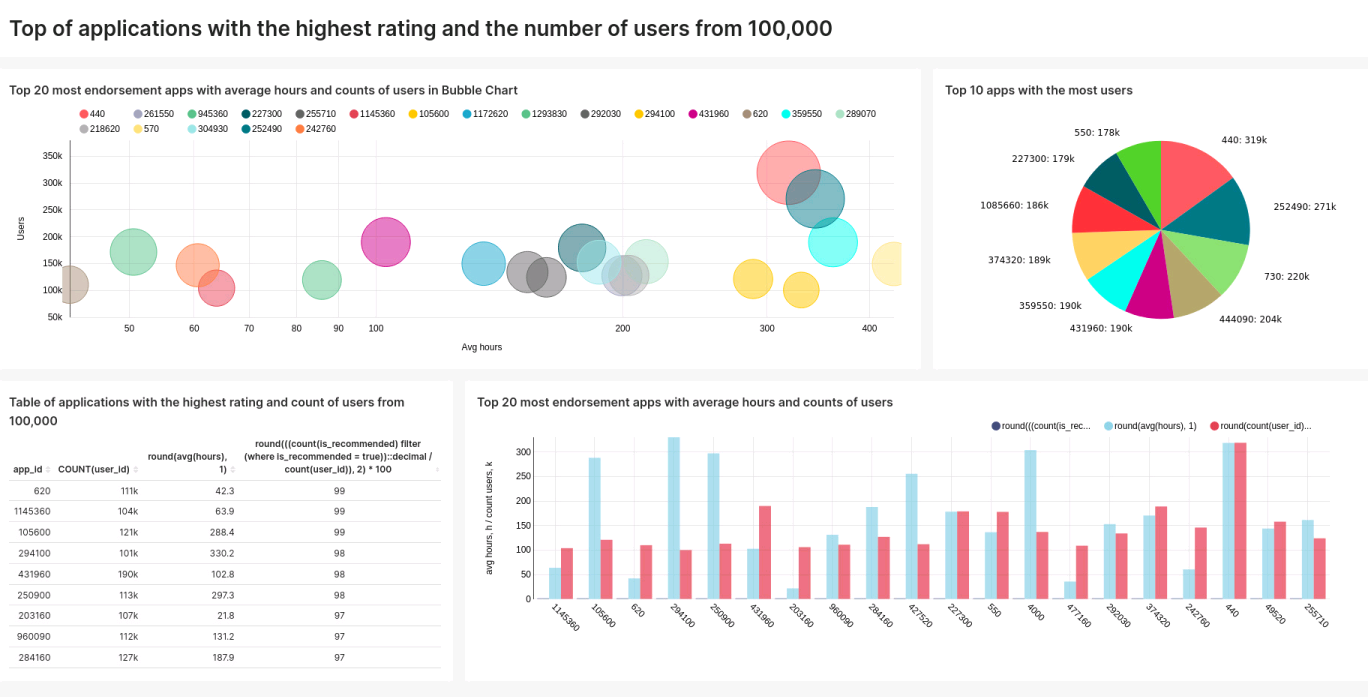


Трансформируем данные в информацию, используя подход материализованных представлений (Materialized Table) при помощи Apache Airflow. Ограничим датасет приложениями с рейтингом более 80% и количеством попробовавших его юзверей не менее 100 000. Тем самым уменьшим датасет с 38 млн до 6 млн строк.

| from datetime import datetime from airflow import DAG from airflow import configuration from airflow.operators.bash import BashOperator from airflow.utils.dates import days\_ago from airflow.operators.postgres\_operator import PostgresOperator from airflow.operators.python\_operator import PythonOperator import requests   DAG\_NAME = 'lab11\_olejnikov\_dag' GP\_CONN\_ID = 'olejnikov\_conn' LAB11\_SQL = ''' drop view if exists mv\_lab11\_olejnikov;  create materialized view mv\_lab11\_olejnikov as  select  \*  from  lab9\_recom\_olejnikov  where  app\_id in (  select app\_id  from lab9\_recom\_olejnikov  group by app\_id  having  count(user\_id) >= 100000 and  round(((count(is\_recommended) filter (where is\_recommended = true))::decimal / count(user\_id)), 2) > 0.8  ); '''  args = {'owner': 'olejnikov',  'start\_date': days\_ago(0),  'depends\_on\_past': False}   with DAG(DAG\_NAME, description='lab11\_olejnikov',  schedule\_interval='@once',  catchup=False,  max\_active\_runs=1,  default\_args = args,  params={'labels':{'env': 'prod', 'priority': 'high'}}) as dag:   send\_to\_postgres = PostgresOperator(  task\_id = 'send\_to\_postgres',  sql=LAB11\_SQL,  postgres\_conn\_id = GP\_CONN\_ID,  autocommit=True)  send\_to\_postgres |
| --- |



Итоговый результат: <http://vm-datavision.test.local:8090/superset/dashboard/p/KW2mdzaQZrn/>



Я к сожалению в этой версии 0.0.0-dev не понял как меня название колонок, прошу обратную связь. Или она просто сырая?

Айтишнику положено знать английский, но, простите меня за мой французкий, интерфейс тут непонятный, без русского жопаболь.

**(Предоставить преподавателю)** скопируйте, пожалуйста, ссылку на вашу витрину данных (Dashboard) в гугл документ и прикрепите ссылку в анкете. Не забудьте открыть доступ по ссылке.