Интеллектуальный анализ работы хранилища данных на основании обработки логов "Ростелеком"

команда DA44

25 сентября 2022

Skolkovo Hack 2022

Описание данных

Данные представляют схематичные логи запросов, выгруженные из базы Greenplum 5

rn	loguser	q
724484	etl_2048	from tbl_77830,from tbl_77829
884851	etl_1151	from tbl_336309
825173	etl_1151	into tbl_28425,JOIN tbl_28425,from tbl_97328
415920	etl_2048	from tbl_79634
469923	etl_1151	from tbl_19348

rn - это порядковый номер строки, он уникальный.

loguser - тип пользователя, выполняющий запрос. Где etl - группа разработчиков, которые загружают данные с помощью etl процессов, dev - обычные аналитики.

q - запрос query, упрощенный запрос, который пользователь отправлен в базу. Здесь через запятую склеены номера таблиц и оператор.

from, join означают, что данные извлекались,

into - запись данных

Задача

Задача - узнать, какие объекты в базе являются бесполезными.

Бесполезный объект - объект, который продолжает наполняться данными, но никто из разработчиков к ним не обращается,
то есть запрос into присутствует, но при этом нет ни одного селекта от обычного аналитика с именем, которое начинается на dev.

Решение нашей команды:

- проведен исследовательский анализ проблемы в jupiter notebook;
- разработана метрика "бесполезности": отношение числа записей в таблицу к числу обращений от разработчиков;
- реализовано веб-приложение Dash, в котором выведены:
  - топ-10 таблиц по нашей метрике;
  - топ-10 пользователей по кол-ву запросов в разре типа: etl и dev;
  - топ-10 пользователей по кол-ву операций в запросе в разрезе типа: etl и dev.

Используемые OpenSource Технологии:
Jupyter notebook, Python, Dash.

Определение метрики "бесполезности" таблицы

Метрика **useless**

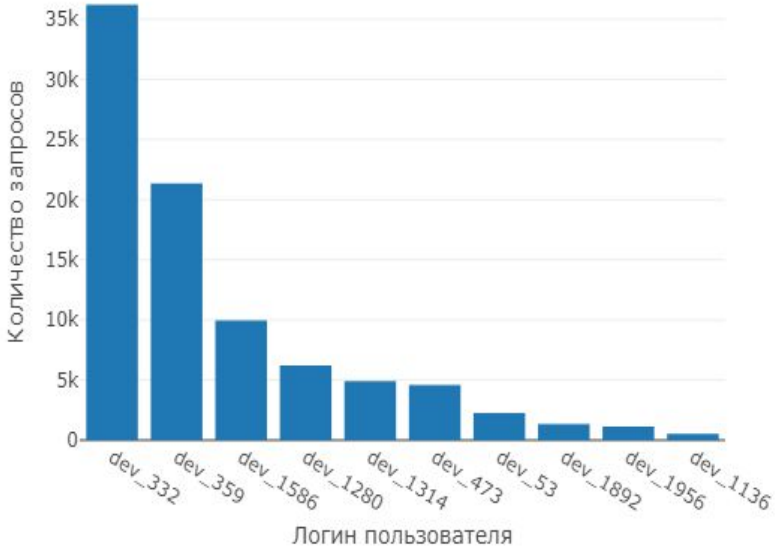
определена как отношение количества запросов into к количеству запросов from

ТОП-10 "бесполезных таблиц"

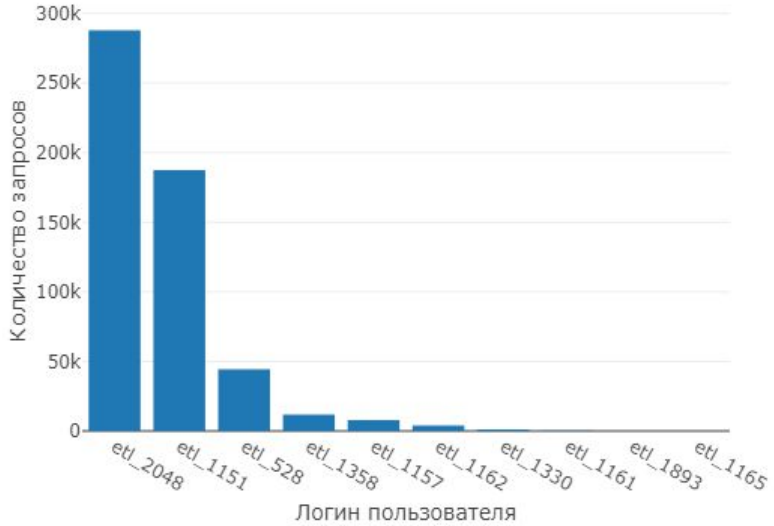
таблица	кол-во froms	кол-во joins	кол-во intos	метрика useless
tbl_211513	8	1	9594	1199.25
tbl_29455	1	1595	471	471.00
tbl_26776	1	3083	469	469.00
tbl_27182	1	8208	465	465.00
tbl_211512	21	0	3139	149.47
tbl_15647	2	0	203	101.50
tbl_44933	10	0	881	88.10
tbl_23594	1	0	54	54.00
tbl_211522	32	0	1171	36.59
tbl_23770	12	0	325	27.08

Распределение пользователей по кол-ву запросов

ТОП-10 dev пользователей



ТОП-10 etl пользователей

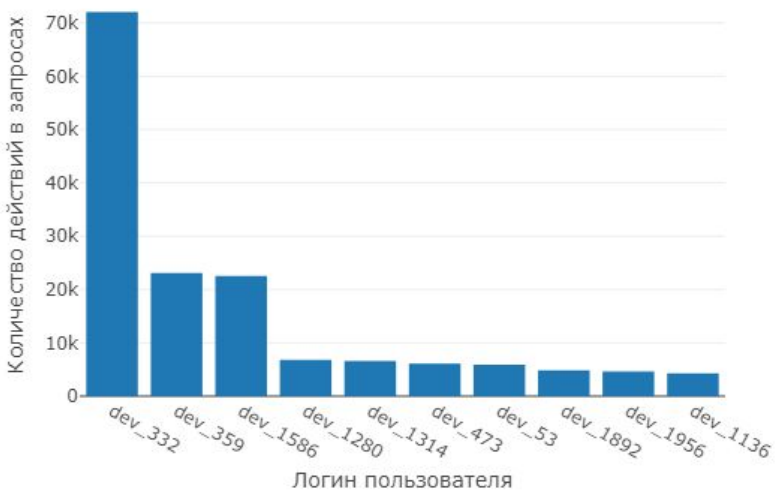


По графикам видно, что наибольшее количество запросов создают пользователи:

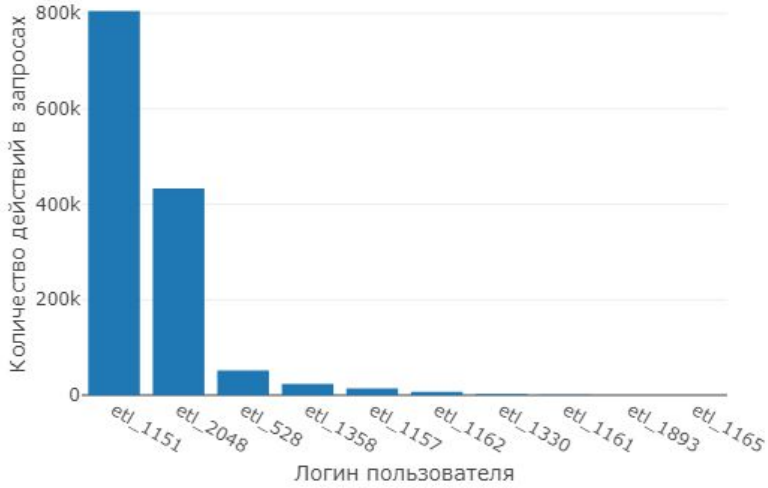
dev_332, dev_359, dev_1586, etl_2048, etl_1151, etl_528.

Распределение пользователей по кол-ву действий в запросах

ТОП-10 dev пользователей



ТОП-10 etl пользователей



По графикам видно, что запросы с наибольшим количеством действий создают те же пользователи, только меняются местами etl_2048 и etl_1151:

dev_332, dev_359, dev_1586, etl_1151, etl_2048, etl_528.

Рекомендации

По результатам ранжирования по метрике “бесполезности” нужно обратить внимание на таблицы: tbl_211513, tbl_29455, tbl_26776. Они очень часто записываются, но редко запрашиваются.

Также следует изучить действия пользователей dev_332, dev_359, dev_1586, etl_2048, etl_1151, etl_528, совершающих наибольшее количество запросов.

Как использовать

Необходимое ПО:

- Python 3.9+
- библиотеки Python:
  - dash
  - regex
  - plotly
  - pandas

Путь к csv-файлу с логами необходимо указать в файле config.py. Для запуска веб-приложения необходимо открыть терминал и выполнить команду:

- для Windows и Mac: python app.py
- для Linux: python3 app.py
- для запуска на виртуально машине: необходимо добавить host в конце app.py - app.run_server(debug=True, host='0.0.0.0')

Спасибо за внимание

Команда DA44





Github: den-dw, Morjella, Mike-solk, taisiiap, AlexeyK12