Project Report

Bohdan Vey, Maksym Kuzushun

Big Data

Precomputed reports and streaming data

Для читтання стімінгу ми вирішили використовувати Kafka, оскілька це основне завдання Kafka, для якого вона сворювалась.
Дані ми вирішили зберігати в двох форматах:

1) Pandas Dataframe - дану табличку ми вирішили використовувати, для погодинних запитів, оскільки ми можемо зберігати необхідну інформацію про запити за останню годину локально на комп'ютері, через що наш сервер не буде витрачати час на те щоб звязатись з іншим сервером і дістати необхідні дані
   Ми також зберігаємо дані в одному форматі для всіх трьох запитів і використовуєм threads(тут також можна використовувати інші сервери), для того щоб обробити ці дані і дати відповідь на запитання, які нас цікавлять.
   Самі відповіді на запити ми зберігаємо в json форматі, так як цього потребує завдання, в папці data в файлах question_{1..3}.json

Ad-hoc querries

Кожен раз, коли ми отримуємо новий запис в кафці, ми відправляємо його в функцію, яка буде записувати новий запис у таблиці Cassandra. Ми маємо 5 різних таблиць, і 5 різних функцій для запису в ці таблиці, скрипт можна знайти тут. Схему таблиць можете бачити на діагнамі внизу. Посилання на cql файл для ініціалізації keyspace.

Для відповідей на запити у нас є Pyhton скрипт який вміє давати відповіді на будь-який з пятьох запитів. Посилання на папку з скриптом

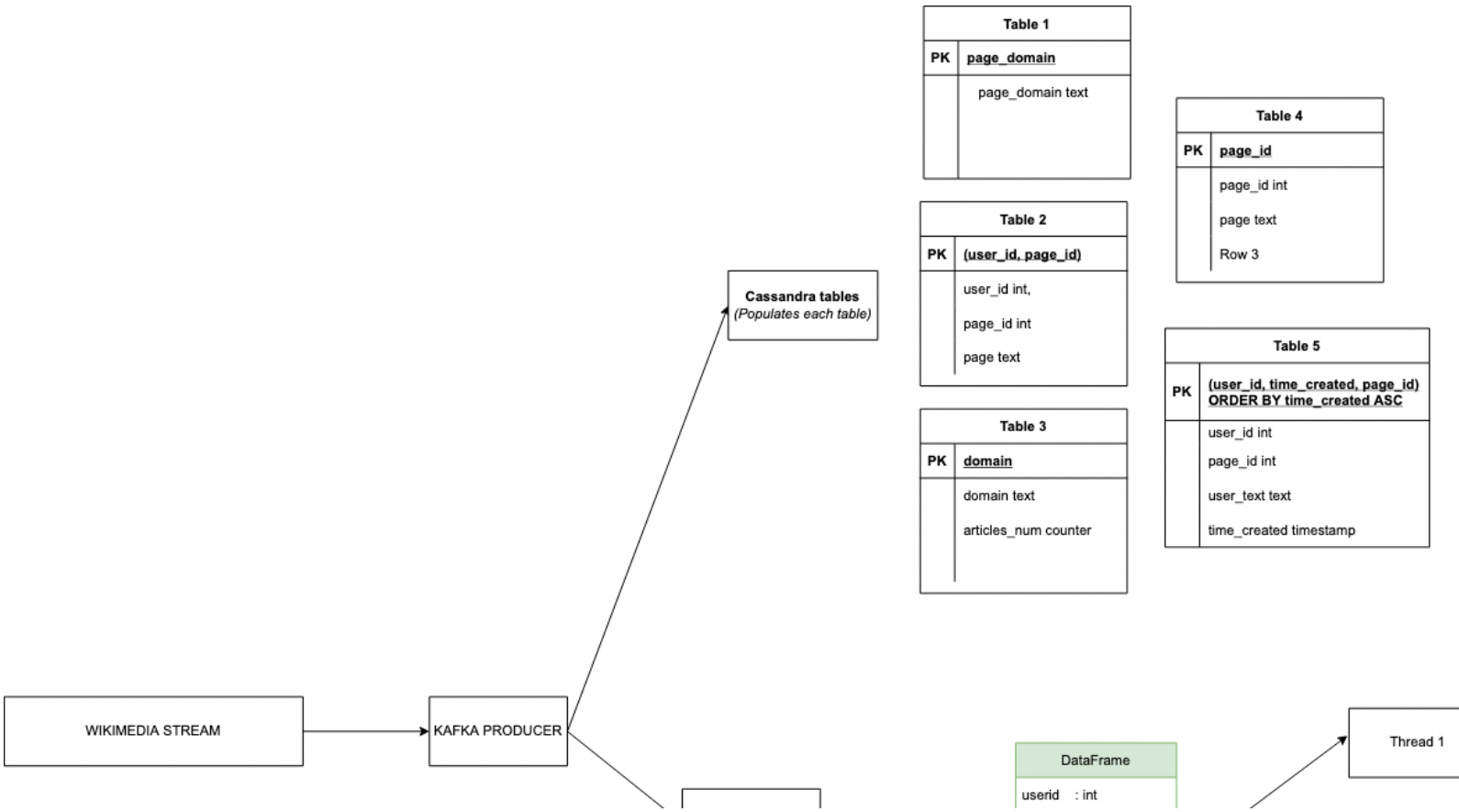
Table 1
PK
page_domain
page_domain text
Table 2
PK
(user_id, page_id)
user_id int,
page_id int
page text
Table 3
PK
domain
domain text
articles_num counter
Table 4
PK
page_id
page_id int
page text
Row 3
Table 5
PK
(user_id, time_created, page_id)
ORDER BY time_created ASC
user_id int
page_id int
user_text text
time_created timestamp
Cassandra tables
(Populates each table)
WIKIMEDIA STREAM
KAFKA PRODUCER
DataFrame
userid : int
Thread 1