Колесник А. Б., магістрант
Державний університет «Житомирська політехніка»

АНАЛІЗ ВПОДОБАНЬ КОРИСТУВАЧІВ СТРІМІНГОВИХ СЕРВІСІВ НА ПРИКЛАДІ NETFLIX

В наш час, кіноіндустрія користується значною популярністю, за рахунок всесвітньої глобалізації та унізації. На культурне збагачення виділяються значні кошти з бюджетів країн, в свою чергу, приватні компанії також зацікавлені у випуску власних стрічок. Здавалось, ніяких проблем не може бути, стрічки виходять у кінопрокат, касети із записами продаються, але поява інтернету та незаконний перезапис касет приніс компаніям значні витрати. На допомогу виробникам прийшли стрімінгові сервіси, які беруть на себе роль кінотеатрів та магазинів. Вони безцінні дані про користувачів, що дало змогу популярним стрімінговому сервісу, досягти великого успіху. Не секрет, що Netflix, використовує усі можливі данні про користувача та його активність, щоб покращувати рекомендаційну систему, та зрозуміти вподобання користувача. У цій статті я розгляну, що таке прикладна аналітика великих даних на прикладі практичного використання ApacheKafka і Druid в Netflix.

Netflix представляє собою кінотеатр, де користувач отримує можливість переглядати фільми, серіалі або передачі. Щоб зрозуміти, який відеоматеріал більше до вподоби користувачу, або як сприймається якесь оновлення чи доповнення, компанія відстежує логи з усіх пристроїв, як Smart-TV, так і телефон чи комп'ютер. Таким чином можна визначити проблеми, які характерні для користувачів окремої платформи, чи взагалі. Для реалізації цих потреб використовується належна BigData технологія, яка здатна більш ніж 2 мільйона подій в секунду. Такою технологією є ApacheDruid.

ApacheDruid – це високопродуктивна аналітична база даних в реальному часі. Розроблена для workflow, в яких дійсно важливі швидкі запити і захоплення. Druid відрізняється миттєвою видимістю даних, спеціальними запитами, оперативної аналітикою і обробкою високого рівня паралелізму. Вона виконує потокову передачу даних з брокерів повідомлень, зокрема, ApacheKafka і AmazonKinesis, а також забезпечує одночасне завантажування файлів з озер даних, HDFS і Amazon S3. Druid підтримує найпопулярніші формати файлів для обробки структурованих і напівструктурованих даних. Завдяки сучасним підходам до зберігання, індексації, а також реалізації запитів, Druid з високим рівнем паралелізму дозволяє оперативно (менше ніж за секунду) отримати узгоджені результати для аналізу ризиків, виявлення шахрайства, аналітики потоку відвідувань, ланцюжків поставок, мережевий телеметрії, цифрового маркетингу, і багатьох інших кейсів обробки безлічі даних в реальному часі.[1]

Таким чином, конвеєр аналітики великих даних на базі ApacheKafka і Druid можна ввести в наступному вигляді :

– події, тобто метрики зчитуються прямо з ApacheKafka за принципом 1 топік для одного джерела даних;

– завдання індексації Kafka створюють кілька робочих процесів, які розподіляються між вузлами реального часу;

– кожен індексатор підписується на топік Kafka і зчитує свою частку подій з потоку;

– індексатори витягають значення з повідомлень про події у відповідності зі специфікацією прийому і накопичують створені рядки в пам'яті;

– як тільки рядок створено, його можна отримати за допомогою Druid SQL чи власного механізму запитів (nativequires) власні запити, які відправляються як JSON в кінцеву точку REST.

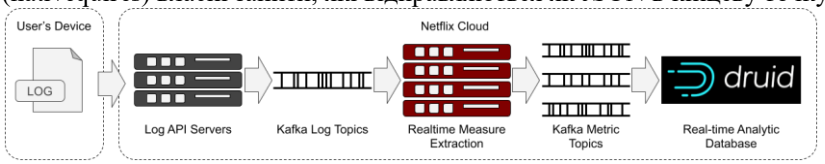

Рис. 1 BigDataPipeline на ApacheKafka и Druid в Netflix

Хоча Druid не є класичною реляційної СУБД, деякі терміни цієї концепції можна застосовуються у BigData. Наприклад, логічне угруповання однотипних даних у вигляді стовпців. Перед тим, як фільтрувати або групувати стовпці, слід переконатися, що вони включені в кожне джерело даних. Зазвичай в джерелі даних є стовпці трьох категорій: час, вимірювання і метрики. У Druidвсе залежить від часу, у кожного джерела даних є часовий відрізок. Прив'язавши дані до тимчасової помітки Druid оптимізує їх зберігання, розподіл і обробку запитів. Це дозволяє масштабувати джерело даних до трильйонів рядків, забезпечуючи час відповіді на запит за десятки мілісекунд. Для досягнення такого рівня масштабованості Druid ділить збережені дані на тимчасові відрізки (TimeChunk) з налаштованою тривалістю. Дані тимчасового відрізка зберігаються в одному або декількох сегментах. Кожен сегмент містить рядки даних, що потрапляють під часовий відрізок, який визначений ключовим стовпцем timestamp'a. Розмір сегментів може бути налаштований так, щоб існувала верхня межа кількості рядків або загального розміру файлу сегмента [2]

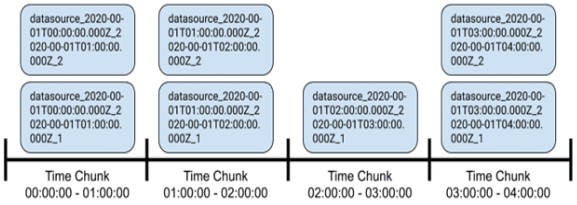

Рис. 2 Тимчасові відрізки

При запиті даних Druid відправляє запит всім вузлам в кластері, які містять сегменти для часових відрізків в межах діапазону запиту. Кожен вузол обробляє запит перед відправкою проміжних результатів назад брокеру, паралельно з обчисленнями над даними, які він зберігає. Брокер виконає остаточне злиття і агрегування перед відправкою набору результатів назад клієнту

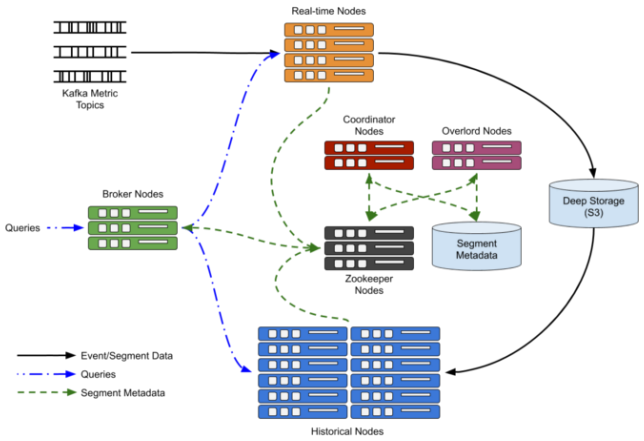

Рис 3. Архітектура системи аналітики BigData на ApacheKafka і Druid в Netflix

Повертаючись до Netflix, відзначимо, що дізнатися, коли були отримані всі події для конкретного відрізка часу, не так просто. Дані можуть надходити в Kafka із запізненням, або індексаторам може знадобитися час, щоб передати сегменти вузлам. Щоб обійти ці обмеження, дата-інженери відкидають дані, які надійшли занадто пізно і ущільнюють рядки для формування аналітичних запитів.