ПОТОКОВАЯ АСИНХРОННАЯ ПЕРЕДАЧА ДАННЫХ МЕЖДУ МОДУЛЯМИ В ETL-КОНВЕЙЕРЕ ПОД УПРАВЛЕНИЕМ РАСПРЕДЕЛЕННОГО БРОКЕРА СООБЩЕНИЙ APACHE KAFKA

Кузьменко Д.Е.[1], Кайда А.Ю.[2]
[1] Место учебы, ТПУ, ИШИТР, 8ПМ21, e-mail: dek29@tpu.ru
[2] Место работы, ТПУ, ИШИТР, ст. преп. ОИТ, e-mail: ayk13@tpu.ru

Аннотация

В статье рассматривается проблема, препятствующая эффективной асинхронной передаче данных, а также предложено решение, позволяющее избежать поэтапного режима передачи. Описано использование асинхронной передачи данных между модулями в ETL-конвейере с применением распределенного брокера сообщений Apache Kafka. Представлена схема модулей ETL-конвейера и описаны полученные результаты их работы.

Ключевые слова: асинхронная передача данных; ETL-конвейер; потоковая обработка.

Введение

С ростом популярности онлайн-сервисов возникают сложности с передачей данных, поскольку объем данных, который необходимо передавать, значительно увеличился. Из-за этого возрастает интерес к методам обработки и хранения данных в децентрализованных системах. Для эффективной работы с информацией, собранной из различных источников, требуются современные аналитические инструменты. [1].

Одним из таких решений является промежуточное программное обеспечение, так называемые брокеры сообщений, которые обеспечивают обмен данными между приложениями или модулями в реальном времени. Они позволяют передавать различные виды информации, будь то банковская транзакция или же целый словарь [1].

Конвейер, в частности ETL конвейер, состоит из модулей, обменивающихся между собой данными. Проходящие через конвейер данные обрабатываются в потоковом режиме. Поток данных – это упорядоченная последовательность данных, которой соответствует определенный источник или получатель. ETL-процесс присутствует в каждом из модулей конвейера, он включает в себя: extract – извлечение данных из внешних источников; transform – их преобразование для того, чтобы они соответствовали заданным условиям; и load – загрузку их в последующий модуль или в хранилище данных. Схематичное представление ETL-процесса приведено на рисунке 1 [1].

Рис. 1. Схематичное представление ETL-процесса

Разработка и описание ETL-конвейера под управлением Apache Kafka

В данном пункте подробно описан этап разработки ETL-конвейера с использованием распределенного брокера сообщений Apache Kafka. Модули в ETL-конвейере под управлением Apache Kafka написаны на языке программирования Python.

В разработанном конвейере существует два вида модулей: первый – это модули обработки; второй – модули приема и передачи. Модули обработки преобразуют входные данные в соответствии с заданной логикой перед отправкой их в следующий модуль. Модули передачи данных обеспечивают передачу данных между модулями обработки через сервер Apache Kafka.

Запуск ETL-конвейера осуществляется с помощью bash-скрипта [2]. Для передачи информации между модулями используются стандартные потоки ввода и вывода (stdin() и stdout()) с помощью библиотеки sys [3, 4]. Схематичное представление ETL-конвейера под управлением распределённого брокера сообщений Apache Kafka приведено на рисунке 2 [1].

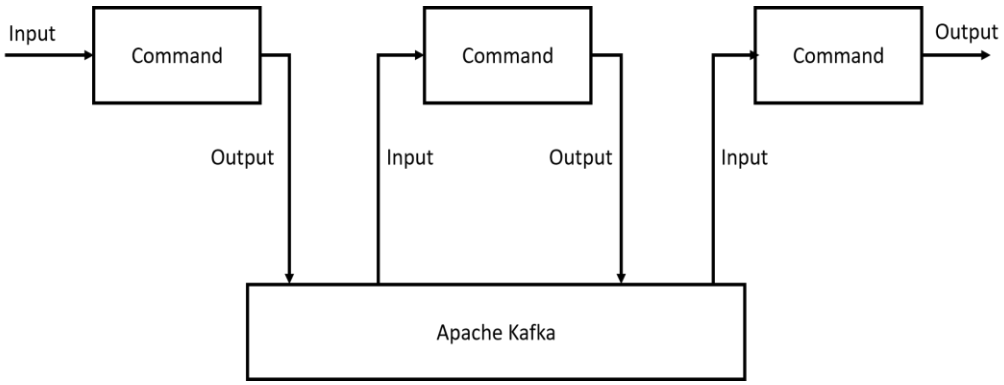

Рис. 2. Схематичное представление ETL-конвейера под управлением распределённого брокера сообщений Apache Kafka [1]

Конвейер предназначен для потоковой обработки данных: массив данных проходит через каждый модуль, где он трансформируется и передается дальше. Apache Kafka используется для передачи данных между модулями обработки. Трансформированные данные из модулей обработки отправляются на сервер Apache Kafka через модули передачи, далее модули приема получают их для дальнейшей обработки. Передаваемые сообщения могут содержать любые данные.

Проблематика поэтапного режима передачи данных между модулями

В текущем пункте описана проблема, возникающая при поэтапной передаче данных между модулями при работе вышеописанного конвейера. Стандартный вывод Python (sys.stdout.write()) буферизуется, он собирает некоторые данные перед их записью в стандартный вывод, после заполнения буфера данные записываются в выходной поток. Данные сохраняются в буфере при условиях: 1) получения данных с устройства ввода; 2) непосредственно перед отправкой на устройство вывода; 3) перемещения данных между процессами на узле. В отличие от команды print(), команда sys.stdout.write() не переключается на новую строку после вывода одного сообщения. Для этого необходимо использовать символ, обозначающий переход к новой строке («\n») [4]. Каждый модуль конвейера обработки накапливает данные в своем выходном потоке до завершения трансформации данных. После заполнения буфера памяти трансформированными данными они передаются на вход следующего модуля.

Время обработки подготовленного датасета из 150000 записей, в ETL-конвейере, составляет 18,5 часов. Следовательно, приблизительное время обработки одной записи – около 0,44 секунды. Из этого следует, что при использовании конвейера данные будут доступны только после завершения прохождения всего массива данных через все этапы конвейера. Это связано с тем, что каждый модуль накапливает данные в своем выходном потоке (буферизация данных). Например, при обработке большого объема данных, такого как 1 миллион записей, время работы конвейера может занять несколько суток. Таким образом, работу с выходными данными можно начать только после завершения всего процесса обработки. Эта проблема может затруднить последующую работу с полученными данными.

Решение проблемы, связанное с поэтапным режимом передачи данных между модулями

В данном пункте представлено решение, позволяющее избежать поэтапного режима работы ETL-конвейера – асинхронный режим передачи данных.

Для каждого модуля приема и передачи данных необходимо использовать разные топики, располагающиеся в Apache Kafka. Данное решение позволяет разделить однородные выходные данные из модулей обработки и предотвращает дублирование данных.

Использование отдельных топиков для каждого этапа работы помогает избежать неправильной интерпретации отправленных сообщений модулями обработки, для которых они не предназначены. В противном случае данные на выходе конвейера могут быть хаотичными и не иметь структуры.

При отправке или получении сообщений Apache Kafka создает записи в журнале действий. Логи содержат информацию об отправленных или полученных сообщениях, а также указывают на то, какой модуль прочитал или отправил сообщение. При возникновении сбоя можно обратиться к журналу действий и найти момент, который привел к сбою для дальнейшего анализа.

Изменение конфигурационного файла сервера Apache Kafka позволяет получать логи за любой период времени. Установка параметра на значение минус один для «возраст файла с логами, который необходимо удалить при превышении» предотвращает удаление логов. После внесения изменений в конфигурацию можно обращаться к логам с момента изменения параметра.

Например, если в процессе обработки 1 миллиона записей произошел сбой на полумиллионной записи, можно найти момент ошибки в журнале логов и запустить конвейер с нужного места.

Для асинхронной передачи данных между модулями в ETL-конвейере необходимо очищать буфер памяти в модулях обработки на каждом этапе отправки данных. После отправки данных следует отправлять сигнал о новой строке. Чтобы избежать дублирования данных, рекомендуется распределять однородные выходные данные из модулей обработки по разным топикам.

На рисунке 3 приведен график зависимости времени обработки от количества поданных на вход текстовых документов.

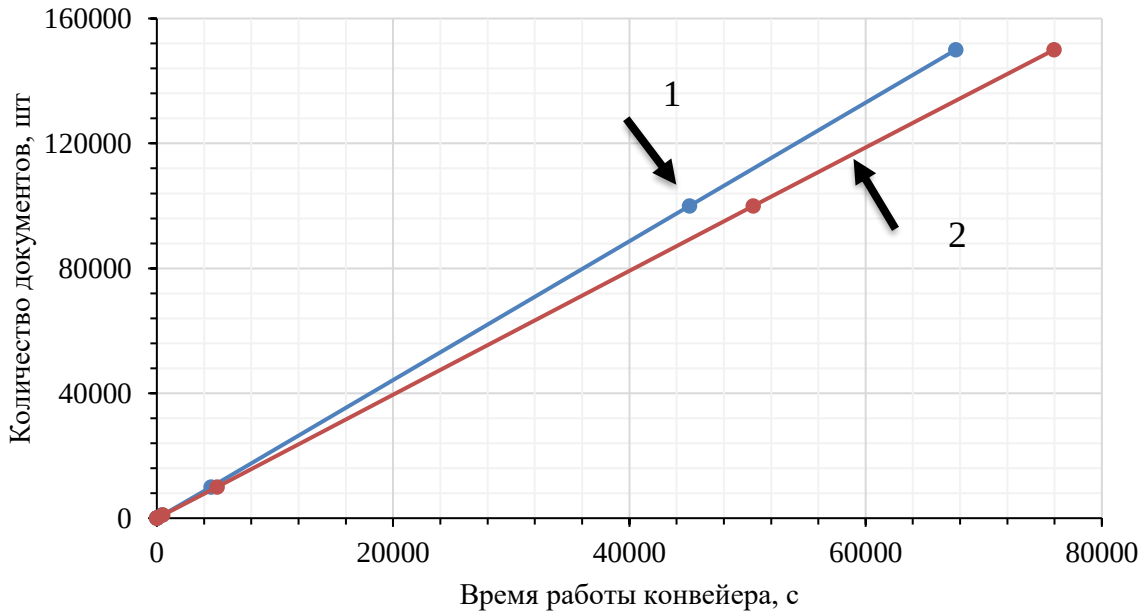

Рис. 3. Зависимость времени обработки от количества поданных на вход текстовых документов, при работе ETL-конвейера:
1 – ETL-конвейер с асинхронной передачей данных между модулями;
2 – ETL-конвейер с поэтапной передачей данных между модулями

Достоинство данного решения заключается в увеличении скорости обработки данных по сравнению с поэтапной передачей в среднем на 12 % (рисунок 3).

Заключение

В статье представлено решение проблемы, препятствующее эффективной асинхронной передаче данных между модулями в ETL-конвейере. Проблема накопления данных в выходном буфере рассмотрена в пункте, описывающем проблематику поэтапного режима передачи данных между модулями. В пункте, связанном с решением проблемы, поэтапного режима передачи данных между модулями предложено решение по предотвращению буферизации данных.

На базе рассмотренного ETL-конвейера сделан вывод о том, что асинхронный режим передает данные мгновенно, не ожидая завершения обработки на предыдущих этапах, в отличие от поэтапного режима работы. Асинхронная передача данных между модулями оказалась быстрее поэтапной обработки в среднем на 12 %, на датасете до 150000 записей.

Список использованных источников

1. Кузьменко Д.Е. ETL-конвейер для потоковой обработки текстовых данных под управлением распределенного брокера сообщений Apache Kafka // сборник трудов XX Международной научно-практической конференции студентов, аспирантов и молодых ученых («Молодежь и современные информационные технологии». – Томск: Изд-во ТПУ, 2023. – С. 237-239.

2. Кайда А.Ю. Магистерская диссертация: Разработка протокола. передачи данных для системы управления потоками данных. – ТПУ. –2019. – 108с. – URL: https://earchive.tpu.ru/handle/11683/53908 (дата обращения 21.01.2024).

3. A generic and customizable framework for the design of ETL scenarios // ScienceDirect: сайт. URL: https://www.sciencedirect.com/science/article/abs/pii/S0306437904000985 (дата обращения 21.01.2024)

4.Sys – System-specific parameters and functions // Python: сайт. – URL: https://docs.python.org/3/library/sys.html (дата обращения 06.02.2024).