УДК 004.62

РАЗРАБОТКА ИНСТРУМЕНТА ПОСЛЕДОВАТЕЛЬНОГО СНЯТИЯ СНИМКОВ АГРЕГИРОВАННЫХ ДАННЫХ ИЗ ПОТОКОВЫХ ДАННЫХ

Гурьянов А. И.

студент,

Казанский (Приволжский) федеральный университет,

Казань, Россия

Якупов А. Ш.

старший преподаватель,

Казанский (Приволжский) федеральный университет,

Казань, Россия

Аннотация

В настоящее время во многих предметных областях широко распространены потоковые данные. При этом высокую востребованность имеет потоковая обработка данных с минимальной задержкой, в режиме реального времени. В значительном количестве случаев при потоковой обработке данных производится агрегация данных для вычисления значений разнообразных бизнес-метрик.

Таким образом, на рынке существует потребность в инструменте последовательного снятия снимков агрегированных данных из потоковых данных. Статья посвящена анализу актуальности и разработке такого инструмента.

Ключевые слова: потоковые данные, потоковая обработка данных, материализованные представления, потоковые алгоритмы, экстраполяция.

DEVELOPMENT OF A TOOL FOR SEQUENTIAL SNAPSHOTTING OF AGGREGATED DATA FROM STREAMING DATA

Gurianov A. I.
student,
Kazan (Volga region) Federal University,
Kazan, Russia

Yakupov A. S.
Senior Lecturer,
Kazan (Volga region) Federal University,
Kazan, Russia

Abstract

Currently, streaming data is widespread in many subject areas. At the same time, stream processing with minimal delay, in real time, is in high demand. In a significant number of cases, data aggregation is performed during stream processing to calculate the values of various business metrics.

Thus, there is a need in the market for a tool for sequential snapshotting of aggregated data from streaming data. The article is devoted to the analysis of the relevance and development of such a tool.

Keywords: streaming data, stream processing, materialized views, streaming algorithms, extrapolation.

Потоковые данные – это данные, которые непрерывно генерируются различными источниками [7].

В настоящее время информация является важнейшим элементом выживаемости и конкурентоспособности организаций [2]. Потоковые данные являются важным стратегическим ресурсом организации, их можно

использовать для решения широкого спектра прикладных задач. Однако для этого необходимо не только собирать данные, но и обрабатывать и анализировать их, чтобы извлечь из данных необходимую информацию [5].

В современном мире очень часто возникает необходимость обрабатывать потоковые данные в реальном времени, с минимальной задержкой [3]. Такую обработку данных называют потоковой.

Потоковая обработка данных активно применяется в большом количестве предметных областей [1]. В частности, потоковая обработка данных является необходимым условием существования таких сфер, как интернет вещей и социальные сети [6].

Кроме того, обработка потоковых данных в реальном времени играет крайне важную роль в сфере информационной безопасности. В потоковых данных журналов информационных систем очень важно оперативно находить аномальную активность, которая может являться признаком атаки. Также потоковый анализ данных применяется для выявления мошеннических финансовых транзакций.

Во многих случаях обработка данных в реальном времени может предоставить значительное конкурентное преимущество [9]. В частности, применение потоковой обработки данных дает возможность повысить актуальность информации, используемой менеджерами компании для принятия решений [4]. Это дает возможность оперативно реагировать на изменения внешней и внутренней среды компании.

Потоковая обработка данных часто противопоставляется пакетной. При пакетной обработке потоковых данных они сначала сохраняются в некоторое хранилище данных, и далее обрабатываются большими пакетами.

Потоковые данные тесно связаны с big data, так как суммарный объем данных потока часто бывает огромен. Хранение потоковых данных для последующего анализа традиционными методами пакетной обработки данных во многих случаях является нерентабельным. Кроме этого, чем больше объем

данных, тем выше задержка до получения результата при использовании методов пакетной обработки данных. Эта задержка может достигать многих часов, что во многих случаях неприемлемо, так как полученный с такой задержкой результат в значительной мере теряет актуальность.

В то же время, при использовании методов потоковой обработки данных можно в каждый момент времени хранить лишь малую долю данных потока, на порядки меньшую, чем общий объем данных потока. По этой причине важным фактором востребованности потоковой обработки данных является происходящий в настоящий момент стремительный рост объема данных.

При обработке данных часто возникает потребность вычисления значения какой-либо функции на основе группы входных значений. Подобные функции называются агрегатными, а процесс вычисления их значений – агрегацией. Примерами агрегатных функций являются: сумма, максимум, минимум, среднее арифметическое, мода, медиана, количество уникальных элементов.

При потоковой обработке данных часто применяются различные приближенные алгоритмы. Они имеют гораздо более высокую эффективность по времени и по памяти, чем аналогичные точные алгоритмы. При этом точность результатов этих алгоритмов является достаточно высокой. Приближенные алгоритмы имеют очень важное значение для потоковой обработки данных, так как при очень большом суммарном объеме данных потока даже алгоритмы, имеющие линейную сложность по времени или памяти часто оказываются неприемлемыми для потоковой обработки данных.

Приближенные алгоритмы часто используются для таких задач, как подсчет количества уникальных элементов, проверка принадлежности элементов к множеству, поиск преобладающего элемента, нахождение частот элементов потока, поиск медианы [8].

Кроме того, актуальной задачей является выявление тенденций, присутствующих в потоковых данных, с целью краткосрочного прогнозирования состояния потока. Это, в частности, может быть использовано

для раннего обнаружения и предотвращения нештатных ситуаций в таких сферах, как транспорт [11]. Для решения этой задачи можно применить методы экстраполяции.

В базах данных для хранения результатов запросов используются материализованные представления (Materialized View). Однако в большинстве существующих реализаций материализованные представления имеют значительный недостаток – единственным способом их обновления является полное повторное выполнение запроса. Таким образом, обновление материализованного представления – длительная операция, требующая значительных ресурсов. Это значит, что материализованные представления в большинстве реализаций непригодны для потоковой обработки данных.

Исходя из вышесказанного, на рынке существует потребность в инструменте анализа потоковых данных, строящем материализованные представления потоковых данных, включающие в себя агрегации, и инкрементально обновляющем эти материализованные представления по мере поступления новых данных [12; 13].

В настоящий момент на рынке представлено несколько таких инструментов, однако они имеют существенные недостатки.

Ни один из присутствующих на рынке инструментов такого рода не поддерживает приближенные алгоритмы обработки потоковых данных и прогнозирование состояния потока, несмотря на их высокую востребованность.

Практически все такие инструменты рассматривают потоковые данные как append-only, то есть обновление и удаление пришедших ранее записей не поддерживается. Это существенно ограничивает набор возможных сценариев их использования.

Единственным инструментом, поддерживающим операции обновления и удаления, является Materialize, позиционируемый как как Data Warehouse для потоковых данных. Однако данный инструмент, хотя и имеет открытый исходный код, выпускается под лицензией Business Source License 1.1, которая

значительно ограничивает его бесплатное использование. Эти ограничения делают невозможным бесплатное использование Materialize в большинстве практических сценариев.

При этом лицензия каждой версии Materialize ровно через 4 года после выпуска версии автоматически изменяется на Apache License 2.0, что означает снятие вышеописанных ограничений. При этом первая версия Materialize была выпущена 14 февраля 2020 года, и перейдет под свободную лицензию только 14 февраля 2024 года.

Использование версии Materialize четырехлетней давности во многих случаях будет являться неприемлемым из-за ее значительного устаревания, а также потенциально из-за наличия известных уязвимостей, которые могут быть использованы злоумышленниками для проведения атаки.

Для использования Materialize без ограничений необходимо приобрести коммерческую лицензию. При этом, в текущей обстановке приобретение коммерческой лицензии является рискованным, так как присутствует значительная вероятность политически мотивированных недобросовестных действий лицензиара.

На основе вышесказанного нами был разработан инструмент, обладающий следующими свойствами [10]:

1. Дает возможность построения на потоковых данных инкрементально обновляемых материализованных представлений. Материализованные представления поддерживают операции обновления и удаления.
2. Поддерживает приближенные алгоритмы обработки потоковых данных.
3. Дает возможность прогнозирования состояния потока с помощью экстраполяции.
4. Имеет открытый исходный код.
5. Является отечественным.

6. Выпускается под свободной лицензией MIT. Возможно бесплатное использование инструмента, в том числе коммерческое, без каких-либо ограничений.
7. Является расширяемым, с возможностью добавления новых агрегатных функций и коннекторов к источникам данных.

Таким образом, в настоящее время существует потребность в инструменте последовательного снятия снимков агрегированных данных из потоковых данных, дающий возможность применения приближенных алгоритмов анализа потоковых данных и прогнозирования состояния потока. Авторами статьи такой инструмент был разработан.

Библиографический список:

1. Апатова Н. В. Управление в экосистеме бизнеса в период цифровой трансформации // Эффективное управление экономикой: проблемы и перспективы. 2022. С. 238–241.

2. Гурьянова Э. А., Гурьянов А. И. Анализ и перспективы рынка SaaS в Российской Федерации // Вестник экономики, права и социологии. 2022. №1. С. 182–185.

3. Ельченков Р. А., Дунаев М. Е., Зайцев К. С. Прогнозирование временных рядов при обработке потоковых данных в реальном времени // International Journal of Open Information Technologies. 2022. Т. 10, №6. С. 62–69.

4. Логиновский О. В., Шестаков А. Л., Шинкарев А. А. Построение современных корпоративных информационных систем // Управление большими системами: сборник трудов. 2019. №81. С. 113–146. doi: 10.25728/ubs.2019.81.5 (новая статья)

5. Маркова В. Д., Кузнецова С. А. Развитие менеджмента в цифровой экономике: аналитический обзор исследований // Мир экономики и управления. 2020. Т. 20, №3. С. 166–183. doi: 10.25205/2542-0429-2020-20-3-166-183

6. Маркова В. Д., Кузнецова С. А. Развитие стратегического менеджмента в цифровой экономике // Вестник Томского государственного университета. Экономика. 2019. №48. С. 217–232. doi: 10.17223/19988648/48/15

7. Определение потоковой передачи данных // Amazon Web Services (AWS). – URL: https://aws.amazon.com/ru/streaming-data/ (дата обращения 26.03.2023)

8. Толпинская Н. Б., Сычев А. Алгоритмы обработки потоковых данных // Проблемы автоматики и управления. 2019. №1 (36). С. 110–117. doi: 10.5281/zenodo.3253017 (новая статья)

9. Трофимов В. В., Трофимова Л. А. О концепции управления на основе данных в условиях цифровой трансформации // Петербургский экономический журнал. 2021. №4. С. 149–155. doi: 10.24412/2307-5368-2021-4-149-155 (новая статья)

10. artemgur/Diplom // GitHub. – URL: https://github.com/artemgur/diplom (дата обращения 26.03.2023)

11. Brandt T. L., Grawunder M. Moving Object Stream Processing With Short-Time Prediction // Proceedings of the 8th ACM SIGSPATIAL Workshop on GeoStreaming. 2017. doi: 10.1145/3148160.3148168

12. Incremental Computation in the Database // Materialize. – URL: https://materialize.com/guides/incremental-computation/ (дата обращения 26.03.2023)

13. McSherry F. View Maintenance: A New Approach to Data Processing // Materialize Blog. 2020. – URL: https://materialize.com/blog/olvm/ (дата обращения 26.03.2023)

Оригинальность 89%