УДК 004.042

Шыхалиев Р.Г.
Институт Информационных Технологий НАНА, Баку, Азербайджан
ramiz@science.az

ОБ ИСПОЛЬЗОВАНИИ СИСТЕМ УПРАВЛЕНИЯ ПОТОКАМИ ДАННЫХ ДЛЯ МОНИТОРИНГА КОМПЬЮТЕРНЫХ СЕТЕЙ

В этой статье рассматриваются некоторые вопросы по обработке потока данных, включая модели потока данных, непрерывной обработки запросов и системы управления потоками данных и возможности использования их в мониторинге компьютерных сетей.

Ключевые слова: *анализ IP-трафика, потоки данных, модели потоков данных, непрерывный запрос, системы управления базами данных, системы управления потоками данных.*

1. Введение

Сегодня в компьютерных сетях (КС) очень быстрыми темпами растут число сетевых приложений, объем и скорость передачи сетевого трафика и во времени изменяются их характер и состав, а также поведение пользователей, что является особенно характерным для интернет-среды. Вследствие этого в КС появляется ряд проблем, связанных с оптимизацией сетевого трафика, сетевой инфраструктуры, обеспечением производительности сети, приемлемого уровня QoS (Quality of Service) показателей сервисов и приложений, безопасности сетевых приложений и пользователей и т.д. Решение этих и других проблем невозможно без наличия информации о сетевом трафике, для сбора и анализа которого используются различные средства мониторинга. Задачи этих средств состоят в сборе пакетов, анализе собранных данных и выводе результатов (отчетов) [1]. При этом для мониторинга КС используются комбинации различных аппаратных и программных средств, которые позволяют регистрировать различные статистические характеристики сетевого трафика и анализировать собранные данные.

В последние годы уровень развития аппаратных и программных средств позволяет регистрировать и сохранять большие потоки сетевых данных. Вместе с тем с ростом объема и скорости сетевого трафика при мониторинге КС требуется анализировать потоки большого объема сетевых данных в реальном масштабе времени, что оказывается ограниченным возможностями традиционных методов анализа данных, используемых в средствах мониторинга и анализа сетевого трафика. При этом объем данных мониторинга КС может изменяться от нескольких до сотен гигабайтов, но тенденция роста объема сетевого трафика дает основания сказать, что в ближайшем будущем объем данных будет измеряться терабайтами, а темп роста объема данных будет превосходить возможности средств мониторинга и технологий хранения данных.

Несмотря на то, что традиционные методы анализа данных себя хорошо зарекомендовали, при анализе потока данных появляются новые проблемы, которые с их помощью разрешить трудно. Это в основном связано с непрерывностью, последовательностью и быстротой поступления потоков данных, а также изменением их во времени и т.д. При этом быстрое и непрерывное поступление большего объема данных создает трудности в их хранении, вычислении и передаче по каналам вычислительных систем.

В последние годы для решения проблем анализа потоков данных, непрерывно поступающих в реальном масштабе времени, было разработано несколько новых технологий. В то же время некоторые существующие программные приложения, такие как

системы управления базами данных (СУБД) и генераторы правил [2], могут быть приспособлены для этой цели.

СУБД широко используются из-за надежности хранения больших объемов данных и возможности эффективно обрабатывать инициированные пользователем запросы. Поэтому существующие традиционные СУБД могут служить хорошей платформой для разработки средств анализа сетевого трафика. Однако у традиционных СУБД имеются некоторые серьезные недостатки, которые ограничивают их функциональные возможности по обработке потоков данных, непрерывно поступающих в реальном масштабе времени. Во-первых, традиционные СУБД являются пассивными хранилищами данных, в которых по поступлении запросов от пользователя осуществляются соответствующие транзакции. Во-вторых, предполагается, что текущее состояние данных является единственно возможным и важным. Следовательно, легко определить текущие значения элементов данных, в то время как получение предыдущих значений данных является очень сложной задачей и может быть восстановлено из лог-файлов СУБД. В-третьих, предполагается, что в СУБД элементы данных синхронизированы и на каждый запрос будет получен соответствующий точный ответ. Однако во многих приложениях, ориентированных на анализ потока данных, данные прибывают асинхронно и результаты запросов должны быть сформированы на основе имеющейся неполной информации. Наконец, традиционные СУБД не имеют возможности по обработке данных, поступающих в реальном масштабе времени [3].

Исходя из вышеизложенного, для сбора и анализа данных сетевого трафика, поступающих непрерывно, последовательно и в реальном масштабе времени, предлагается использовать системы управления потоками данных, СУПД [4, 5].

Основной задачей данной статьи является анализ основных моделей потоков данных и СУПД и возможностей использования их в мониторинге КС.

2. Модели потока данных

Понятие «потоки данных» появилось относительно недавно, и эта область развивается очень быстро. За последние несколько лет в этой области было издано множество работ. Эти работы имеют большое практическое значение в тех областях, где необходимо добыть знания из большого объема непрерывно растущих данных.

Поток данных – это последовательность упорядоченных (имеющих временные отметки или не имеющих) элементов, непрерывно поступающих в реальном масштабе времени. Обычно потоки данных поступают очень быстрыми темпами и невозможно упорядочить их поступление, а из-за ограниченности памяти становится трудным хранение потоков в целом.

Так как элементы могут поступать в отдельности, то поток данных может быть смоделирован как последовательность списков элементов [6]. Отдельные элементы потока могут принимать форму реляционных кортежей или спецификации объектов.

В модели потока данных к входным данным, которые необходимо обработать, нет возможности произвольного доступа из диска или памяти, и они поступают как последовательность непрерывных потоков данных. Обычно при произвольном способе доступа к дискам или памяти для чтения/записи произвольного блока данных не требуется последовательного просмотра блоков, начиная с самого первого, при этом для доступа к разным блокам данных тратится почти одинаковое время.

Модель потоков данных имеет некоторые отличия от обычной модели данных с реляционными отношениями:

- элементы данных в потоке поступают в реальном масштабе времени;
- система не имеет никакого контроля над упорядочением элементов потоков данных;

− потоки данных не имеют ограничения в размере;

− после обработки элемента потока данных он исключается из потока или архивируется, при этом для дальнейшего восстановления его нужно сохранить в памяти, которая имеет намного меньший размер, чем размер потока данных в целом.

Вместе с тем при обработке информации в модели потока данных не исключается наличие данных, имеющих обычные реляционные отношения.

У потоков данных имеется несколько моделей, и формально потоки данных можно описать следующим образом. Входной поток $s_1, s_2, ...$ поступает последовательно, элемент за элементом, и описывает основной сигнал S, который является одномерной функцией – $S:[1...N]\to R$. При этом вход может состоять из нескольких потоков или многомерных сигналов. В зависимости от того, как элементы s_i описывают S, модели потока данных отличаются:

− модель временных рядов. Каждый s_i-ый элемент равняется $S[i]$, и они поступают в порядке возрастания значения i. Эта модель подходит для временных рядов данных, например, при наблюдении за объемом сетевого трафика КС каждые 10 минут и т.д. В каждом таком периоде времени мы наблюдаем последующие новые данные;

− модель кассового аппарата. Здесь s_i-ые элементы возрастают до $S[j]$-х. Допустим, что $s_i=(j,I), I_i\geq 0$, это означает, что $S_i[j]=S_{i-1}[j]+I_i$, где S_i – это состояние сигнала после наблюдения i-го элемента в потоке. Как в кассовом аппарате, множество s_i-е может со временем возрастать до определенного $S[j]$. Эта модель является самой популярной моделью потока данных. Она подходит для мониторинга в КС IP-адресов, которые обращаются к веб-серверу и отправляют пакеты в сеть и т.д. Так как в КС одни и те же IP-адреса могут несколько раз получить доступ к веб-серверу или отправлять множество пакетов в сеть в течение долгого времени;

− модель турникета. Здесь s_i-ые элементы обновляются на $S[j]$-х. Допустим, что $s_i=(j,I_i)$, это означает, что $S_i[j]=S_{i-1}[j]+I_i$ где S_i – это сигнал после наблюдения *i*-го элемента в потоке и I_i может быть положительным или отрицательным. Эта модель является самой общей моделью потока данных, которая позволяет исследовать динамические ситуации, где элементы непрерывно поступают в систему и покидают ее.

Эти модели более подробно описаны в работе [7].

Как известно, приложения по обработке потоков данных требуют поддержки непрерывных запросов. Запросы по непрерывным потокам данных имеют много общего с запросами в традиционной СУБД. Однако имеется два важных различия, присущих модели потока данных, – это различие между одноразовыми и непрерывными запросами [8].

Одноразовые запросы (имеют место в традиционных СУБД) являются запросами, которые за один раз выполняются над определенным фиксированным набором данных, и результат возвращается пользователю (рис. 1).

Непрерывные запросы выполняются непрерывно, поскольку потоки данных поступают непрерывно. При этом ответы на непрерывные запросы производятся в течение времени, пока поступает поток данных и постоянно отражаются данные потока (рис. 1). Ответы на непрерывные запросы могут быть сохранены и обновлены вновь поступающими данными или они сами могут быть непосредственно произведены как потоки данных. Часто запросы по потокам данных могут объединить потоки данных и данные, имеющие реляционные отношения.

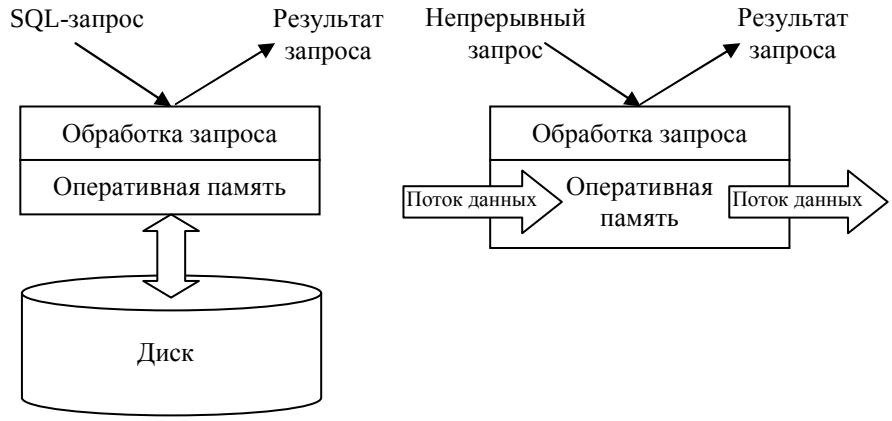

Рис.1. Различие между одноразовыми и непрерывными запросами

Семантика непрерывных запросов заключается в следующем. Для простоты допустим, что время представлено как последовательность целых чисел. Пусть $A(Q,t)$ является множеством ответов на непрерывный запрос Q, происходящий во времени t, для которого τ является текущим временем, а стартовым временем является 0. Если непрерывный запрос является монотонным, то достаточно рассмотреть запросы к вновь прибывающим элементам и результаты добавить в обучающие выборки. Таким образом, множество ответов монотонного непрерывного запроса Q, происходящего во времени τ, будет выражаться следующей формулой [9]:

$$A(Q,\tau) = \bigcup_{t=1}^{\tau} (A(Q,t) - A(Q,t-1)) \bigcup A(Q,0).$$

В отличие от монотонных непрерывных запросов, немонотонные непрерывные запросы должны быть повторно вычислены при каждой повторной оценке, что приводит к следующей семантике [9]:

$$A(Q,\tau) = \bigcup_{t=0}^{\tau} A(Q,t).$$

3. Системы управления потоками данных

СУПД должны управлять и обрабатывать элементы потока данных, прежде чем элементы данных будут заменены последующими входящими элементами данных. Однако эта задача является очень сложной в таких интенсивных и непредсказуемых средах, как КС, особенно в сети Интернет, в которых порядок поступления элементов данных не может управляться, а размер потока данных при этом неограничен.

В отличие от традиционных СУБД, в СУПД можно осуществлять непрерывные запросы по отношению к непрерывным потокам данных, которые в реальном масштабе времени поступают в систему и покидают ее, при этом на время обработки данные хранятся только в оперативной памяти и в систему могут поступать различные данные, например, данные фондовой биржи или сетевой трафик. Однако, как и любая СУБД, СУПД требует, чтобы схема, описывающая тип и структуру данных, была управляемой. Поэтому повторное использование и изменение СУПД приложений для анализа сетевого трафика не представит труда.

Общая архитектура предлагаемых сегодня СУПД для обработки потоков данных показана на рис.2. Входной монитор регулирует интенсивность входа, и при случае

неспособности системы обработать большие входящие потоки опускаются некоторые данные (например, при обработке сетевого трафика могут быть отброшены некоторые пакеты). Данные, как правило, хранятся в трех разделах памяти: во временной рабочей памяти (например, окна запросов); в памяти для хранения результатов, например, синопсисы потоков; в статической памяти, где хранятся метаданные (например, информация о физическом расположении источников данных). Непрерывные запросы регистрируются в хранилище запросов и помещаются в группы для совместной обработки, но вместе с тем и одноразовые запросы по предыдущему состоянию потока тоже могут быть выполнены. Процессор запросов связан с входным монитором и может оптимизировать планы запросов в зависимости от изменения интенсивности входных потоков. Или результаты передаются пользователям во временный буфер, где пользователи могут затем корректировать свои запросы.

Рис.2. Общая архитектура СУПД

Из-за уникальности особенностей потоков данных и непрерывности запросов к СУПД предъявляются следующие требования:

− модель данных и семантика запросов должны позволить выполнить упорядоченные и контролируемые по времени операции;

− невозможность сохранения полного потока предполагает использование приближенных структур [10, 11];

− планы потоковых запросов не могут использовать блокирование операторов, которые должны принимать весь вход, до получения результатов;

− из-за ограничения производительности и памяти невозможно выполнить откат по потокам данных. Алгоритмы обработки потоков данных в реальном режиме времени выполняются только один раз;

− приложения, которые проводят мониторинг потоков в режиме реального времени, должны быстро реагировать на необычные значения данных;

− на протяжении выполнения непрерывных запросов условия, в которых находится система, могут изменяться (например, изменяется скорость потока);

− совместное использование нескольких запросов необходимо для того, чтобы обеспечить масштабируемость системы.

Наиболее широко применяемой областью СУПД является анализ сетевого трафика КС. Для того чтобы СУПД можно было использовать для анализа сетевого трафика КС,

они должны обрабатывать данные с производительностью, пропорциональной загрузке КС. Например, для анализа заголовков протоколов IP и TCP в Gbps-сетях с полностью используемой полосой пропускания скорость, необходимая для обработки данных, должна быть больше, чем 42 Gbps (при условии, что фиксированный размер пакета является 1500 байтов, а размер заголовка 64 байта). Однако уменьшить потоки данных относительно текущей загрузки сети можно факторами 4.1–9.1 [12].

Применение СУПД для мониторинга КС позволит создать повторно используемые компоненты анализа сетевого трафика, которые могут быть применены как для анализа данных, поступающих в реальном масштабе времени, так и для анализа заранее собранных данных. При этом типичными задачами анализа сетевого трафика являются следующие:

− анализ загрузки системы, например, как часто определенные порты используются такими протоколами, как FTP или HTTP, при подключении к серверу, как используется полоса пропускания различными приложениями, как отдельные пользователи или группы пользователей используют полосу пропускания;

− анализ характеристик сетевого трафика, таких как среднее время жизни и размер пакетов, отношение между числом потерянных пакетов и временем жизни пакетов и т.д.;

− анализ характеристик сессий, таких как длительность взаимодействия клиентов с веб-серверами, время отклика веб-серверов на клиентские запросы, продолжительность использования пользователями интернет-сервисов и т.д.

Вышеперечисленные задачи предъявляют важные функциональные требования к средствам мониторинга и анализа, такие как способность взаимодействия протоколов всех уровней эталонной модели ISO OSI, включая прикладной уровень. При этом одним из важных требований, которые предъявляются к этим средствам, является возможность анализа сетевого трафика в реальном масштабе времени.

Сегодня существует множество СУПД, например, STREAM [13], GigaScope [14], TelegraphCQ [15], StatStream [16], Tribeca [17] и т.д., которые могут быть использованы для решения тех или иных задач мониторинга КС. Более обширную информацию о каждой СУПД можно получить соответственно в работах [13–17].

4. Заключение

Анализ сетевого трафика является одной из важных и сложных задач мониторинга КС. При этом СУБД могут быть использованы для хранения и анализа больших объемов данных сетевого трафика. Однако большой объем, характеристики наборов сетевых данных и особенности их анализа, а также характер поступления потока сетевого трафика не позволяют использовать традиционные СУБД непосредственно.

В последнее время одной из популярных идей в области обработки данных является анализ потока данных. Исходя из этого, в статье для сбора и анализа данных сетевого трафика, поступающих непрерывно, последовательно и в реальном масштабе времени, предлагается использовать СУПД, которая позволит справиться с таким большим объемом данных сетевого трафика в реальном масштабе времени.

Литература

1. Plagemann, T., Goebel, V., Bergamini, A., Tolu,G., Urvoy-Keller,G., Biersack,E.W.:Using Data Stream Management Systems for Traffic Analysis – A Case Study, Passive and Active Network Measurement, 5th International Workshop, PAM 2004, Antibes Juan-les-Pins, France, April 19-20, 2004, Proceedings, pp. 215–226.
2. L. Brownston, R. Farrell, E. Kant, and N. Martin, Programming expert systems in OPS5: an introduction to rule-based programming, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1985, 471 p.
3. Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S.,Seidman, G., Stonebraker, M., Tatbul, N., Zdonik, S.,:Monitiring Streams- A New Class of Data Management Applications, Proceedings of the 28th international conference on Very Large Data Bases (2002) (VLDB '02) Conference, Hong Kong, China, 2002, pp. 215–226.
4. S. Babu and J. Widom. Continuous queries over data streams. Technical report, Stanford University Database Group, March 2001. Available at http://ilpubs.stanford.edu:8090/527/
5. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and Issues in Data Stream Systems, Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA. ACM 2002, pp. 1–16.
6. P. Tucker, D. Maier, T. Sheard, L. Fegaras. Enhancing relational operators for querying over punctuated data streams. 2002. www.cse.ogi.edu/dot/niagara/pstream/punctuating.pdf
7. A. Gilbert, Y. Kotidis, S. Muthukrishnan and M. Strauss. Surfing wavelets on streams: One pass summaries for approximate aggregate queries. VLDB Journal, 2001, pp.79–88.
8. D. Terry, D. Goldberg, D. Nichols, and B. Oki. Continuous queries over append-only databases. In Proc. of the 1992 ACM SIGMOD Intl. Conf. on Management of Data, pp. 321–330.
9. A. Arasu, S. Babu, J. Widom. An Abstract Semantics and Concrete Language for Continuous Queries over Streams and Relations. Technical Report, Nov. 2002. http://ilpubs.stanford.edu:8090/563/.
10. Y. Zhu, D. Shasha. StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time. In Proc. Int. Conf. on Very Large Data Bases, 2002, pp. 358–369.
11. Micheel, J., Braun, H.-W., Graham, I.: Storage and Bandwidth Requirements for Passive Internet Header Traces, Workshop on Network-Related Data Management, in conjunction with ACM SIGMOD/PODS 2001, Santa Barbara, California, USA, May 21-23, 2001.
12. The STREAM Group. STREAM: The Stanford Stream Data Manager (short overview paper), IEEE Data Engineering Bulletin, Vol. 26 No. 1, March 2003, pp. 19–26.
13. Cranor, C., Johnson, T. Spatcheck, O., Shkapenyuk, V.: Gigascope: A Stream Database for Network Applications, ACM SIGMOD 2003, San Diego, California, USA, June 9–12, 2003, pp. 647-651
14. Sailesh Krishnamurthy, Sirish Chandrasekaran, Owen Cooper, Amol Deshpande, Michael J. Franklin, Joseph M. Hellerstein, Wei Hong, Samuel R. Madden, Vijayshankar Raman, Fred Reiss, and Mehul A. Shah. TelegraphCQ: An Architectural Status Report. IEEE Data Engineering Bulletin, Vol 26(1), March 2003, pp. 11–18.
15. Yunyue Zhu, Dennis Shasha “StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time” VLDB, August, 2002. pp. 358–369.
16. Mark Sullivan: Tribeca: A System for Managing Large Databases of Network Traffic, USENIX Annual Technical Conference (NO 98), June 15-19,1998, pp. 13–24.

UOT 004.042
Şıxəliyev Ramiz H.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
ramiz@science.az
Verilənlər axınını idarəetmə sistemlərinin kompyuter şəbəkələrinin monitorinqi üçün istifadəsi
Bu məqalədə verilənlər axınının emalı məsələlərinə, o cümlədən verilənlər axınının, fasiləsiz sorğuların emalı modellərinə və verilənlər axınının idarəetmə sistemlərinə və onların kompyuter şəbəkələrinin monitorinqi üçün istifadəsi imkanlarına baxılır.
***Açar sözləri*:** *IP-trafikin analizi, verilənlər axını, verilənlər axını modelləri, fasiləsiz sorğular, verilənlər bazasını idarəetmə sistemləri, verilənlər axınının idarəetmə sistemləri.*

Ramiz H. Shikhaliyev
Institute Information Technology ANAS, Baku, Azerbaijan
ramiz@science.az
Using the data stream management systems for computer networks monitoring
In this paper, some issues of the data stream processing, including data stream models, continuous query processing models and data stream management systems and their use computer networks monitoring are reviewed.
***Key words*:** *IP-traffic analysis are disscussed, data stream, data stream models, continuous query, data base management systems, data stream management systems.*