

7. Mehdiyeva G. Y., Ibrahimov V. R., Imanova M. N. Some refinement of the notion of symmetry for the Volterra integral equations and the construction of symmetrical methods to solve them // J. Computational and Appl. Mathematics. 2016. V. 306. P. 1—9.

8. Aliev T. A. Robust Technology with Analysis of Interference in Signal Processing. N. Y.: Kluwer Acad./Plen. Publishers, 2003.

9. Aliev T. A. Digital Noise Monitoring of Defect Origin. London: Springer, 2007.

10. Петров В. В. Помехоустойчивость измерений уровня излучения движущегося акустического источника на фоне изотропного шума среды при использовании комбинированного приёмника // Измерительная техника. 2015. № 1. С. 49—52.

11. Глинченко А. С., Алешечкин А. М., Комаров В. А. Повышение достоверности спектральных измерений параметров сигналов при малых отношениях сигнал—шум // Метрология. 2015. № 3. С. 46—57.

12. Сандомирский С. Г. Влияние точности измерения и диапазона изменения физической величины на коэффициент корреляции между ее истинными значениями и результатами измерения // Измерительная техника. 2014. № 10. С. 13—17.

13. Aliev T. A., Guluyev G. A., Pashayev F. H., Sadygov A. B. Noise monitoring technology for objects in transition to the emergency state // Mechanical Systems and Signal Proc. 2012. V. 27. P. 755—762.

14. Aliev T. A., Abbasov A. M., Guluyev Q. A., Pashaev F. H., Sattarova U. E. System of robust noise monitoring of anomalous seismic processes // Soil Dynamics and Earthquake Eng. 2013. V. 32. No. 1. P. 11—25.

Дата принятия: 21.07.2016.

ИЗМЕРЕНИЯ В ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЯХ

52.08:004.62:681.518.3

Анализ существующих технологий обработки потоков пространственно-временных данных для современных информационно-измерительных систем

А. А. МАЙОРОВ, А. В. МАТЕРУХИН

Московский государственный университет геодезии и картографии,
Москва, Россия, e-mail: maiorov@miigaik.ru

Исследованы существующие технологии обработки потоков пространственно-временных данных от системы датчиков (сенсоров) для выбора наиболее подходящих технологий для практической реализации различных информационно-измерительных систем. Проанализированы типовые технологии обработки потоков пространственно-временных данных от датчиков. Предложены критерии выбора необходимой технологии для реализации в конкретной информационно-измерительной системе.

Ключевые слова: потоки пространственно-временных данных, информационно-измерительные системы, технологии обработки данных.

The existing technologies of processing of space-time data flow from the sensor system (sensors) to select the most suitable for the practical implementation of various information-measuring systems were investigated. The model technologies of processing of space-time data flow from the sensors were analyzed. Criteria of selecting appropriate technologies for implementation in specific information-measuring system are proposed.

Key words: spatio-temporal data streams, information-measuring systems, data processing technology.

Информационно-измерительные системы (ИИС), как новый класс измерительной техники, появились примерно в конце 50-х гг. прошлого века. Тогда они представляли собой системы централизованного циклического получения измерительной информации с элементами вычислительных устройств на базе дискретной полупроводниковой техники. Вместе с развитием вычислительной техники и информационных технологий изменялось понимание концепции ИИС. В

настоящее время под ИИС понимают совокупность функционально объединённых измерительных, вычислительных и других вспомогательных технических средств для получения измерительной информации, её преобразования, обработки в целях представления потребителю (в том числе ввода в автоматизированные системы управления) в требуемом виде либо автоматического осуществления логических функций измерения, контроля, диагностирования, идентификации

(распознавания образов) [1]. В англоязычной научной литературе выражение «system of information and measurement» практически не используется, а полным аналогом русского термина «информационно-измерительная система» является «data acquisition system». Это словосочетание иногда ошибочно переводят как «система сбора данных». В российской научной традиции под системой сбора данных понимают только часть ИИС, а именно — комплекс средств, осуществляющий автоматизированный сбор информации о физических параметрах в заданных точках объекта исследования с аналоговых и (или) цифровых источников сигнала, а также первичную обработку, накопление и передачу данных [2]. Появление ИИС связано с переходом к получению и использованию результатов не отдельных измерений, а потоков измерительной информации [3]. Для разных классов измерительной информации проблемы обработки этих потоков отличаются друг от друга и требуют решения различных задач. Для многих практически значимых классов ИИС большое количество измерительной информации — это потоки пространственно-временных данных от распределённых сетей датчиков (сенсоров). В настоящее время сенсорные сети в режиме реального времени генерируют огромное количество пространственно-временных данных. Примерами таких систем являются:

- подсистемы периметрального контроля охраняемых объектов, предназначенные для измерений координат, скоростей и ускорений нарушителей. Подсистемы контроля основаны на сейсмическом, акустическом, магнитометрическом, трибоэлектрическом, тепловом (инфракрасном) и других физических принципах регистрации воздействий нарушителей на чувствительный элемент;

- ИИС мониторинга окружающей среды;

- подсистемы получения измерительной информации об окружающей среде от самоуправляемых групп роботов.

Во всех приведённых выше примерах ИИС с помощью датчиков определяют некоторые пространственно-распределённые величины, причём каждый отдельный датчик выдаёт поток информации, содержащий как временные отметки выполнения измерений, так и указание на местоположение

датчика в пространстве в момент соответствующего измерения. В современных ИИС такой поток данных можно представить в виде последовательности кортежей, состоящих из фиксированного набора атрибутов.

В настоящей статье авторы проанализировали существующие технологии обработки потоков пространственно-временных данных от системы датчиков (сенсоров) и разделили их на три группы. Цель проведённого анализа — показать, какие технологии являются наиболее подходящими для практических реализаций различных ИИС.

Первая выделенная группа технологий — системы обработки потоков пространственно-временных данных от системы датчиков (сенсоров), основанные на парадигме Map—Reduce. Map—Reduce — это модель распределённой обработки данных, предложенная компанией Google для параллельной обработки больших объёмов данных на компьютерных кластерах. Общая схема этой модели изображена на рис. 1. Входные данные D разбиваются на отдельные фрагменты $D_1, D_2, \dots, D_n$. К каждому фрагменту применяется один и тот же алгоритм обработки M , который на выходе выдаёт набор пар (ключ, значение). Таким образом, выполняется фаза *Map*. Затем отсортированные по ключам пары подаются на вход обработчиков *Reduce* ($$), эта обработка составляет фазу *Reduce*. Полученные в результате выполнения последней фазы результаты $O_1, \dots, O_k$ образуют выходные данные — результат обработки O .

Рассмотренная технология включает не только модель вычислений, но и поддержку этой модели, а именно автоматическую нарезку входного файла с данными для последующего распараллеливания обработки входных данных, эффективную балансировку загрузки вычислительных узлов кластера и алгоритмы обеспечения отказоустойчивой работы кластера.

Наиболее известная реализация описываемой технологии — это Hadoop. Изначально указанная технология не предполагала обработку пространственных данных, но в последние годы были разработаны и реализованы соответствующие алгоритмы, например ESRI tools for Hadoop [4]. Описанная выше технология хорошо подходит для обработки уже собранных и сохранённых в файл пространственно-временных данных [5, 6]. Во-первых, файл с данными может иметь упрощённую структуру. В самом простом варианте — это полученные от датчиков кортежи данных с разделителями. Во-вторых, такой файл практически не имеет ограничений по размеру. Более того, чем больше размер файла, тем эффективнее использование рассмотренного вида обработки данных по сравнению с другими.

Вторая выделенная авторами группа — технологии обработки потоков пространственно-временных данных от системы датчиков (сенсоров), основанные на использовании систем управления реляционными базами данных. В настоящее время технологии, реализуемые в системах управления реляционными базами данных, являются наиболее изученными и широко используются при создании различных ИИС. Подход основан на идеях Э. Ф. Кодда, сформулированных в 1969—1970 гг. [7]. Про-

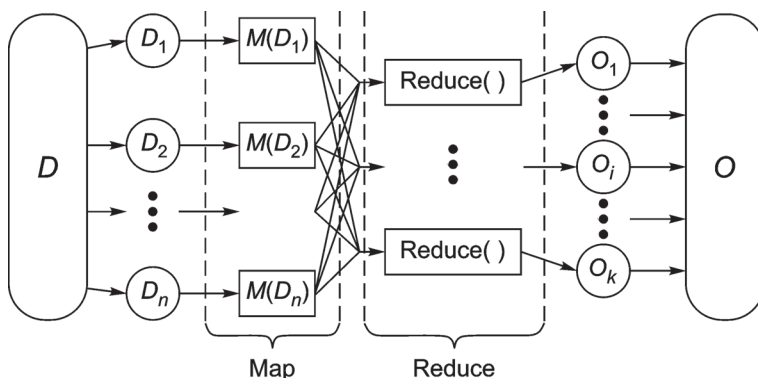


Рис. 1. Модель обработки данных, основанная на парадигме Map—Reduce: D — входные данные; *Map* и *Reduce* — фазы обработки входных данных; $M()$ — алгоритм обработки фрагментов данных в фазе *Map*; $Reduce()$ — алгоритм обработки фрагментов данных в фазе *Reduce*; O — выходные данные

мысленные реализации этой идеи практически до конца прошлого века плохо справлялись с поддержкой как пространственных данных, так и пространственной аналитики. Однако для современных систем управления реляционными базами данных созданы расширения, поддерживающие такие пространственные данные как точки, линии, полигоны, покрытия и сети, и позволяющие эффективно выполнять пространственные запросы. В качестве примера можно привести системы управления реляционными базами данных: Oracle Spatial, IBM DB2 или PostGIS.

Схема обработки данных с использованием системы управления реляционной базой данных изображена на рис. 2. Поступающие от датчиков (сенсоров) данные D передаются для записи в реляционную базу данных RDB и становятся доступны для запросов и анализа с помощью запросов Q . Для увеличения скорости записи данных в базу используется буферизация в оперативной памяти MM . Выходные данные, т. е. результаты обработки обозначены O .

Данные, поступающие в режиме реального времени, бесшовно интегрируются с более ранними, что делает возможным выполнение аналитических запросов с использованием информации в реальном времени и исторических данных. Такой вариант актуален для многих видов ИИС. Технология обеспечивает выполнение полного набора требований ACID (Atomicity, Consistency, Isolation, Durability), т. е. атомарность, непротиворечивость, изолированность, сохранность данных по отношению к записываемой измерительной информации, что достаточно важно для некоторых областей применения ИИС.

Проблема, связанная с применением этой технологии, заключается в физических ограничениях на интенсивность потока измерительной информации, которую может обработать ИИС, использующая систему управления реляционной базой данных. Поскольку данные, поступающие от датчиков (сенсоров), передаются для записи в базу данных, хранящуюся на жёстком диске, то ограничителем выступает такая характеристика диска как время произвольного доступа. Применение твёрдотельных накопителей позволяет улучшить эту характеристику, но не изменить это ограничение существенным образом. На текущий момент максимальная интенсивность потока измерительной информации, которая может быть обработана с помощью технологии использования систем управления реляционными базами данных, составляет несколько тысяч обновлений от системы датчиков в секунду.

Третья выделенная авторами группа — технологии обработки потоков пространственно-временных данных от системы датчиков (сенсоров), основанные на использовании систем управления потоками данных (data stream management systems или stream processing engines). Разработка указанных технологий началась в начале XX в. в американских исследовательских университетах. Это проекты: TelegraphCQ, выполненный в Калифорнийском университете в Беркли [8]; Aurora, реализованный совместно Массачусетским технологическим институтом MIT, Брауновским университетом и университетом Брендаиса [9]; STREAM Стэнфордского университета [10].

Схема обработки информации системой управления потоками данных изображена на рис. 3. Базовая идея представленной технологии заключается в том, что с помощью оконных операторов (window operators) потоки в оперативной памяти MM преобразовываются в таблицы W , к

которым применяются запросы Q . Данные, выходящие за пределы окна, можно архивировать в базе данных DB , которая не обязательно должна быть реляционной. Выходные данные (результаты обработки) обозначены O . Ограничение, следующее из такого подхода, состоит в том, что запросами Q обрабатываются только данные, попавшие в соответствующее временное окно. Другим отличием от технологий обработки в реляционных базах данных является непрерывное выполнение заданного алгоритма обработки. В системах баз данных информация сначала поступает в базу данных и сохраняется там, а потом система направляет к этим данным запросы. В систему управления потоками данных запросы поступают раньше, чем данные. Данные проходят через несколько постоянно выполняемых запросов и преобразованными поступают в соответствующие приложения. Примеры современных систем, основанных на этой группе технологий, — TIBCO Streambase, Apache/Storm, Microsoft Streaminsight, IBM Infosphere.

Характерными и специфичными для процессов обработки пространственно-временных данных в таких системах являются проблемы выполнения пространственных запросов. Это связано с тем, что существующие алгоритмы пространственной аналитики разрабатывались для реляционных баз данных с предварительно построенными пространственными индексами. Выполнение в режиме реального времени запросов на обработку потоков пространственно-временных данных, например, таких как построение пересечения, проверки вложенности, расчёта ближайших соседей или кратчайшего пути, требует разработки и программной реализации новых алгоритмов. В этом направлении ведутся исследования, но работа ещё завершена не полностью [11–13].

Результаты проведённого анализа позволяют сделать следующие выводы. Если в разрабатываемой ИИС не требуется анализировать поступаю-

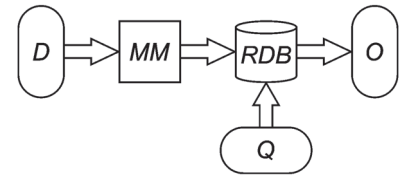


Рис. 2. Модель обработки данных, основанная на технологии использования систем управления реляционными базами данных:

D , O — то же, что и на рис. 1; MM — оперативная память; RDB — реляционная база данных; Q — запросы

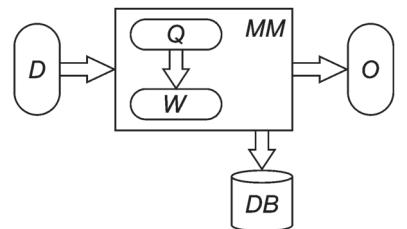


Рис. 3. Модель обработки данных, основанная на технологии использования систем управления потоками данных:

D , O — то же, что и на рис. 1, 2; Q , MM — то же, что и на рис. 2; W — таблицы данных; DB — база данных

щие от сенсоров пространственно-временные данные в режиме реального времени, но при этом необходимо их проанализировать на любом заданном временном интервале, и объём поступающих данных потенциально не ограничен (хранятся все старые данные), то наилучшим выбором технологии обработки будет эффективная пакетная обработка на основе парадигмы Map—Reduce. Если же от ИИС требуется не только обработка сохранённых данных, но и оценка данных в режиме реального времени, то следует выбирать между технологией использования систем управления реляционными базами данных и технологией использования систем управления потоками данных. Критерием выбора будет интенсивность потока пространственно-временных данных от системы сенсоров. При высокой интенсивности потока единственным возможным вариантом будет технология использования систем управления потоками данных, что приведёт к ограничениям по использованию пространственной аналитики в разрабатываемой ИИС. Когда интенсивность потока не слишком высока, то предпочтительнее использовать технологию системы управления реляционными базами данных, для которой существует большое количество готовых библиотек для поддержки различных пространственных запросов.

С точки зрения перспективы, наиболее многообещающей технологией обработки пространственно-временных данных в ИИС авторам представляется использование систем управления потоками данных. Существующий недостаток разработанных и реализованных алгоритмов пространственной аналитики для этих систем открывает новую область научных исследований.

Л и т е р а т у р а

1. **Муха Ю. П., Королёва И. Ю.** Информационно-измерительные системы. Волгоград: ВолгГТУ, 2015.
2. **Крюков В. В.** Информационно-измерительные системы. Владивосток: ВГУЭС, 2000.
3. **Цапенко М. П.** Измерительные информационные системы. М.: Энергия, 1974.

4. **The GIS Tools** for Hadoop are a collection of GIS tools for spatial analysis of big data. [Электрон. ресурс]: <https://github.com/Esri/gis-tools-for-hadoop> (дата обращения: 01.09.2016).

5. **Aji A., Wang F., Vo H., Lee R., Liu Q., Zhang X., Saltz J.** Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce: Proc. VLDB Endow. 2013. [Электрон. ресурс]: <http://www.doi.org/10.14778/2536222.25362271009-1020> (дата обращения: 01.09.2016).

6. **Fox A., Eichelberger C., Hughes J., Lyon S.** Spatio-temporal indexing in non-relational distributed databases: IEEE Int. Conf. 2013. P. 291—299. DOI: 10.1109/BigData.2013.6691586.

7. **Codd E. F.** A relational model of data for large shared data banks // Commun. ACM. 1970. V. 13. No. 6. P. 377—387.

8. **Chandrasekaran S., Cooper O., Deshpande A., Franklin M. J., Hellerstein J. M., Hong W., Krishnamurthy S., Madden S. R., Reiss F., Shah M. A.** TelegraphCQ: continuous dataflow processing 2003 ACM SIGMOD: Proc. Int. Conf. Management of data (SIGMOD '03). N. Y., USA. P. 668—669.

9. **Abadi D. J., Carney D., Cetintemel U., Cherniack M., Convey Ch., Lee S., Stonebraker M., Tatbul N., Zdonik S.** Aurora: a new model and architecture for data stream management // J. The VLDB, 2003. V. 12. No. 2. P. 120—139.

10. **Arasu A., Babcock B., Babu Sh., Datar M., Ito K., Nishizawa I., Rosenstein J., Widom J.** STREAM: the stanford stream data manager (demonstration description): Proc. of the 2003 ACM SIGMOD Int. conf. on Management of data (SIGMOD '03). N. Y., USA. P. 665—666.

11. **Kazemitabar S. J., Demiryurek U., Ali M., Akdogan A., Shahabi C.** Geospatial stream query processing using Microsoft SQL Server StreamInsight: Proc. VLDB Endow. 2010. P. 1537—1540. DOI:10.14778/1920841.1921032.

12. **Pandey V., Kipf A., Vorona D., Mühlbauer T., Neumann T., Kemper A.** High-Performance Geospatial Analytics in HyPerSpace: Proc. of the 2016 Int. Conf. Management of Data (SIGMOD '16). ACM, N. Y., USA. P. 2145—2148.

13. **Hendawi A. M., Shi Y., Fattah H., Karwa J., Ali M.** RxSpatial: a framework for real-time spatio-temporal operations: demo: Proc. 10th ACM Int. Conf. on Distributed and Event-based Systems (DEBS '16). 2016. N. Y., USA. P. 366—367.

Дата принятия: 12.10.2016.