УДК 004.65
DOI: 10.26102/2310-6018/2022.38.3.027

Разработка алгоритма приближенной обработки конвейера запросов в реляционной системе управления базами данных

А.В. Филимонов✉

Академия Федеральной службы охраны Российской Федерации,
Орёл, Российская Федерация
fay0@yandex.ru✉

Резюме. В статье рассматривается алгоритм приближенной обработки запросов в системах управления базами данных реляционного типа. Описываемый алгоритм позволяет получить приближенные результаты запросов с агрегированием и группированием, что позволяет применить его в задачах аналитической обработки запросов с целью снижения времени отклика при обработке запросов. Представленные алгоритмы реализуют метод случайной кластерной выборки и используют математическое обеспечение, позволяющее получить оптимизированное распределение пространства выборки с применением метрики качества выборки. В качестве такой метрики выбран коэффициент вариации. Также в статье продемонстрирована модель конвейера аналитических запросов, представленная в форме направленного ациклического графа. Алгоритм приближенной обработки запросов расширен для условий применения его в потоке запросов, что позволяет оценить доверительный интервал вместе с результатом обработки конвейера запросов. Данный алгоритм может быть применен при разработке специального программного обеспечения процессора базы данных, реализующего архитектуру приближенной обработки запросов в реляционных базах данных. Такой подход находит место в поле исследований синтеза структуры гибридных хранилищ данных, реализующих транзакционно-аналитическую обработку данных. В дальнейшем исследовании предполагается получение экспериментальной оценки представленного подхода.

Ключевые слова: приближенная обработка запросов, алгоритмы обработки запросов, конвейер запросов, кластерная выборка, хранилище данных, гибридная транзакционно-аналитическая обработка данных.

Для цитирования: Филимонов А.В. Разработка алгоритма приближенной обработки конвейера запросов в реляционной системе управления базами данных. *Моделирование, оптимизация и информационные технологии.* 2022;10(3). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1242
DOI: 10.26102/2310-6018/2022.38.3.027

Developing an algorithm for approximation query pipeline processing in a relational database management system

A.V. Filimonov✉

Russian Federation Security Guard Service Federal Academy,
Orel, Russian Federation
fay0@yandex.ru✉

Abstract. The article considers an algorithm for approximate query processing in relational database management systems. The described algorithm makes it possible to obtain approximate results of queries with aggregation and grouping, which helps to apply it for the purposes of analytical query processing in order to reduce the response time when processing queries. The presented algorithms implement the method of random cluster sampling and employ software that provides means for obtaining an optimized distribution of the sample space using a sample quality metric. The coefficient of variation is chosen as such metric. The article also proposes a model of the analytical query pipeline

given in the form of a directed acyclic graph. The approximate query processing algorithm is extended for the conditions of its application in a query flow, which enables the estimation of the confidence interval along with the result of processing the query pipeline. This algorithm can be utilized in the development of special database processor software that implements the architecture of approximate query processing in relational databases. This approach finds a place in the field of research on the synthesis of the structure of hybrid data warehouses that implement transactional-analytical data processing. Further research is expected to obtain an experimental evaluation of the presented approach.

Keywords: approximate query processing, query processing algorithms, query pipeline, cluster sampling, data warehouse, hybrid transactional-analytical data processing.

For citation: Filimonov A.V. Developing an algorithm for approximation query pipeline processing in a relational database management system. *Modeling, Optimization and Information Technology.* 2022;10(3). Available from: https://moitvivt.ru/ru/journal/pdf?id=1242 DOI: 10.26102/2310-6018/2022.38.3.027 (In Russ.).

Введение

Приблизительная обработка запросов является важным подходом в задачах обработки больших данных. Различные методы аппроксимации обеспечивают интерактивное время отклика при анализе массивных наборов данных, а также находят свое применение в обработке высокоскоростных потоков данных. Методы приближенной обработки подразделяются на два класса: методы на основе сжимающих преобразований без потерь, методы, использующие сжатие данных с потерями. Но методы обоих классов базируются на одном высокоуровневом подходе – применении сжатого отображения данных вместо полного набора данных. Сжатое отображение может быть получено различными способами, включая, помимо прочих, методы случайной выборки [1, 2], гистограммные методы, вейвлет-преобразования [3], скетчинг (sketching) [4], применение материализованных представлений. Полученное в результате сжатое представление должно сохранять необходимые свойства исходного массива данных, при этом занимая гораздо меньше места.

Случайная выборка наиболее широко используется для приблизительной обработки запросов в больших базах данных [5]. Такие методы имеют потенциал для значительного сокращения используемых ресурсов и времени отклика, обеспечивая при этом небольшую ошибку аппроксимации. Как показано в [6], для многих приложений нет необходимости точного ответа на запрос. Таким образом, небольшая ошибка допустима взамен экономии ресурсов, потребляемых в процессе обработки запроса, а также для снижения задержки. Например, это подходит для запросов, результаты которого отображаются в виде графиков и диаграмм в приложениях визуализации данных. Ошибки в результате запроса приводят к неточности визуального представления, но в большинстве случаев небольшие различия допустимы. Особенно в случаях, когда ошибки достаточно малы, разница между приблизительными и точными цифрами практически незаметна. Несмотря на то, что выборка позволяет гибко обрабатывать различные запросы, ее использование не требует дополнительного места для хранения. Эта черта отличает выборку от других подходов к ответам на большие запросы.

Поскольку приближенная обработка запросов широко применяется на практике, существует тренд исследований, направленный на ее изучение в таких аспектах, как обеспечение точности результата запроса, эффективности использования пространства и времени, определение границ ошибок результатов запросов и так далее. Настоящее исследование сосредоточено на вопросе оптимизации пространства выборки, которая позволит получить результаты выполнения запросов с определенной точностью за

наименьшее время, тем самым снизив задержку и ресурсоемкость системы хранилищ данных.

Для достижения общей цели исследования по разработке математического и специального программного обеспечения, приближенной обработки запросов, обеспечивающих эффективную работу с конвейером запросов в хранилищах данных, необходимо получить алгоритм, который будет обеспечивать приближенную обработку за счет применения разработанного в [7] математического обеспечения.

Алгоритм определения объема случайной выборки по стратам в запросах с единственным атрибутом агрегирования и группирования

Как представлено в [7], точность результатов приближенного выполнения запросов возможно повысить за счет применения стратифицированной выборки и метрики качества выборки на основе коэффициента вариации. Такой подход позволяет учесть отличие статистических характеристик случайных величин в различных стратах, что часто присутствует на практике. Применение коэффициента вариации в качестве метрики обеспечит эффективное выравнивание точности оценок различных случайных величин между стратами. Такая метрика позволит получить лучшее распределение ошибки в оценке агрегированного значения из разных страт для случаев, когда оцениваемые величины различаются по модулю.

На основе описанной выше идеи было разработано математическое обеспечение, решающее задачу оптимизации распределения выборки в запросах. Многопараметричность рассматриваемых запросов не дает возможности сформировать некоторую обобщенную модель. Таким образом, решение задачи оптимизации представляется системой частных решений для классов запросов с различным значением числа функций агрегирования и операторов группирования. Ниже представлено решение задачи для запроса с одной функцией агрегирования и одиночным группированием:

$$s_i = M \frac{\sqrt{w_i}\sigma_i / \mu_i}{\sum\limits_{j=1}^{r} \frac{\sqrt{w_j}\sigma_j}{\mu_j}}, \quad (1)$$

где $s = \left[s_1, s_2, \ldots, s_r \right]$ – r-местный вектор, каждый элемент которого содержит значения объема выборки для каждой из r страт;

M – бюджет выборки;

$w = \{w_i\}$ – r-местный вектор весовых коэффициентов, который позволяет дополнительно подстраивать точность для необходимых страт, определяемых в зависимости от прикладных задач. При равном приоритете точности всех оценок в каждой страте элементы этого вектора принимают единичные значения;

$\mu = \left[\mu_1, \mu_2, \ldots, \mu_r \right]$ – вектор средних значений для страт;

$\sigma = \left[\sigma_1, \sigma_2, \ldots, \sigma_r \right]$ – вектор стандартных отклонений для страт.

Разработанное математическое обеспечение может быть применено при синтезе алгоритма приближенной обработки запросов функционирующего в пространстве процессора СУБД. Далее представлена схема алгоритма (Рисунок 1) для частного случая запросов с единственной функцией агрегирования, решение которого определяется выражением (1).

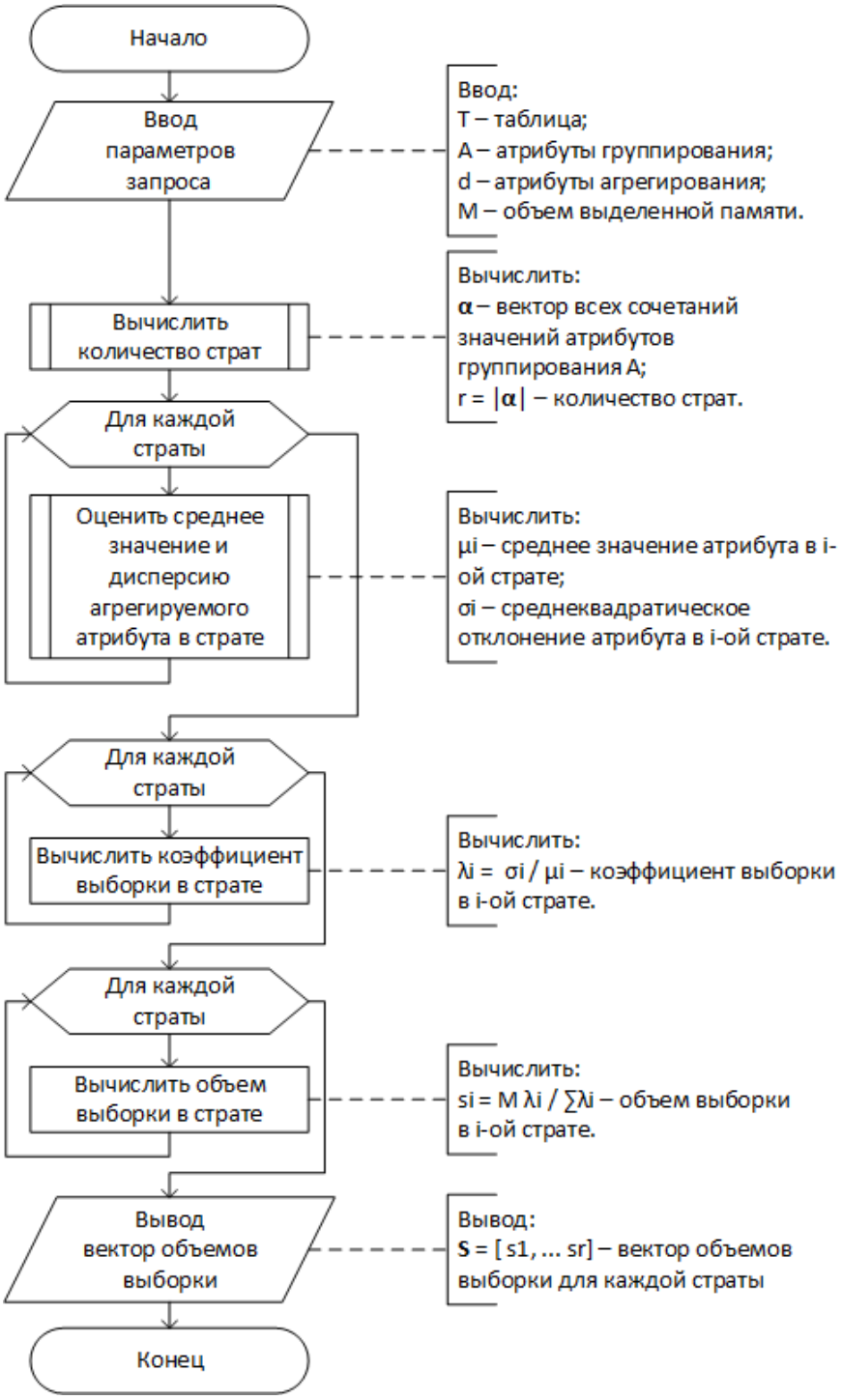

Рисунок 1 – Схема алгоритма приближенной обработки запроса
Figure 1 – Flow chart of query approximate processing

Алгоритм приближенной обработки конвейера запросов

В современных хранилищах данных существует тенденция не только к увеличению объема данных, но и повышения сложности задач аналитической обработки этих данных. Для решения таких задач анализа существует устоявшийся подход, который основан на переходе от одного комплексного запроса к последовательность более простых локальных запросов. Такая последовательность может быть описана моделью конвейера запросов (query pipeline) [8].

Конвейер запросов – это набор запросов и таблиц, расположенных в топологическом порядке выполнения. Модель конвейера представляется ориентированным ациклическим графом (ОАГ), в котором таблицы представлены узлами, а запросы – ребрами. Результаты каждого частного запроса материализуются в виде таблицы, которая может обслуживать другие запросы в потоке. Таким образом массивные вычисления разбиваются на множество подзадач с промежуточными результатами. Такая модель предоставляет следующие преимущества:

- *Модульность*. Сложная задача разбита на несколько компонентов с четко определенным предназначением. Такие компоненты обладают простотой поддержки и отладки, что позволяет эффективно контролировать работу всего конвейера. Также компоненты дают возможность повторного переиспользования, что позволяет повысить технологичность процесса анализа данных;

- *Оптимизация обработки полного потока запроса*. Применение продвинутых инструментов оптимизации запросов не всегда дает возможность полного устранения вычислительной избыточности при обработке сложных аналитических запросов. Модель конвейера материализует промежуточные результаты выполнения подзапросов. Таким образом, их возможно повторно использовать вместо дублирования в вычислениях. Более того, промежуточные результаты также могут быть разделены между конвейерами. Это хорошо совмещается с практикой применения хранилищ данных, в которых существуют таблицы, одновременно участвующие в нескольких конвейерах обработки;

- *Масштабируемость*. Конвейерная модель позволяет масштабировать систему как по объему данных, так и по количеству задач. Вспомогательная инфраструктура представляет собой комбинацию нескольких обработчиков. Вместо того, чтобы один обработчик выполнял сложную задачу, конвейер может обрабатываться несколькими обработчиками параллельно.

При внедрении методов приближенной обработки конвейерную модель, случайная выборки из части таблиц дает приближенные результаты, которые будут использованы следующими в конвейере запросами. Кроме того, может быть задействовано несколько входных таблиц. Обработка запроса к таблице, к которой применяется выборка отличается от обработки запроса к приближенной таблице. Таблица с выборкой содержит подмножество данных, в котором каждая запись является фактическим наблюдением исходных данных. Случайность оценки в этих таблицах зависит от характеристик выборки. С другой стороны, в приближенной таблице данные оцениваются. Как следствие, неопределенность содержится не только в подмножестве исходных результатов, но и в самих наблюдаемых данных. Таким образом, применение выборки в модели конвейера запросов является сложной и нетривиальной задачей. Ниже представлен алгоритм (Рисунок 2), позволяющий применить приближенную обработку конвейера запросов с учетом, изложенных выше особенностей.

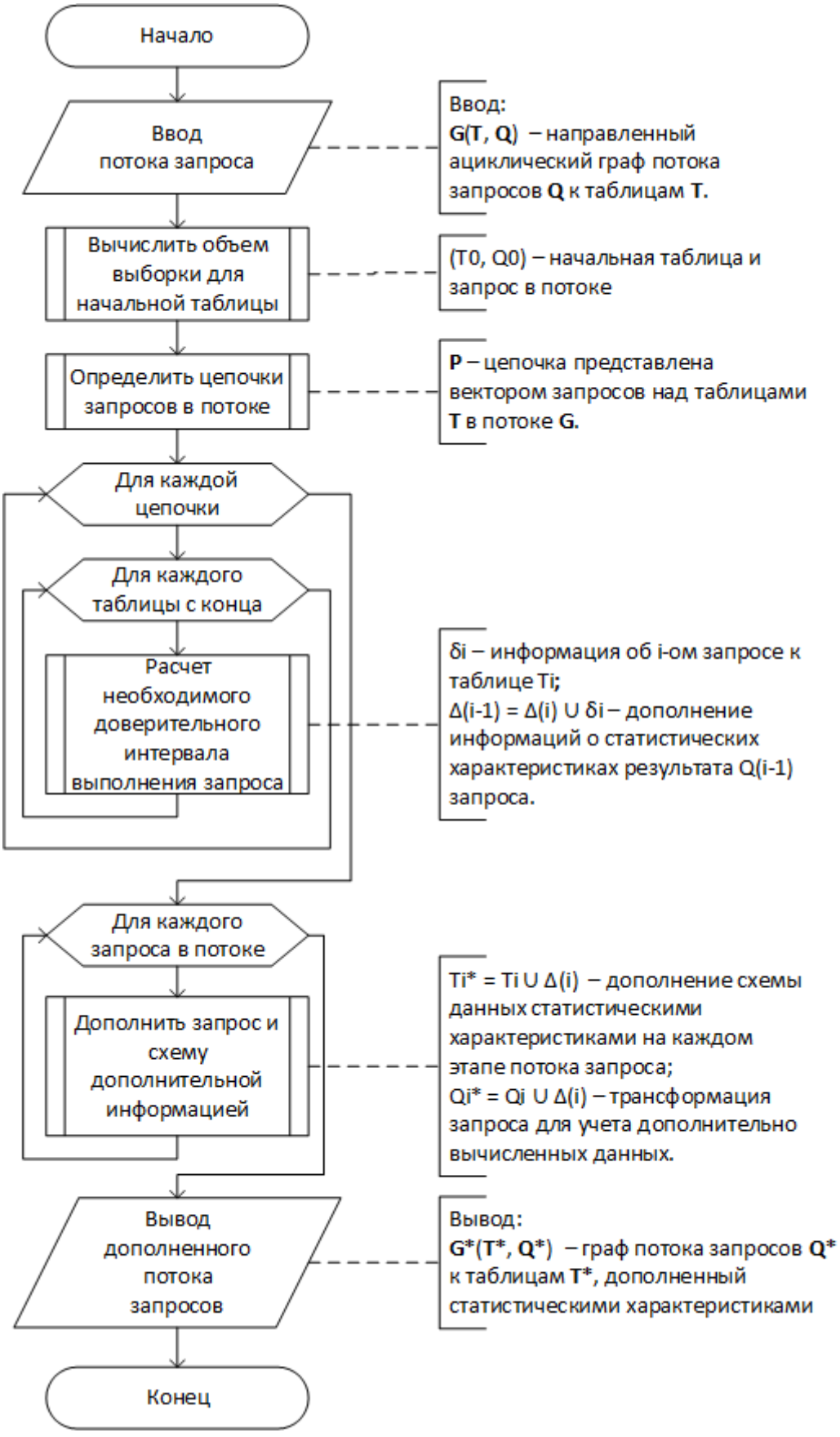

Рисунок 2 – Схема алгоритма приближенной обработки конвейера запросов
Figure 2 – Flow chart of the approximation query pipeline processing

Приближенная обработка конвейера запросов предполагает получение результата с доверительным интервалом. В таком случае процесс необходимо адаптировать так, чтобы необходимые данные для вычисления доверительного интервала проходили через весь конвейер. Для этой цели представлен двухпроходный алгоритм. Первая фаза алгоритма выполняет обратный проход по конвейеру – запрос (от приемника к источнику). На каждой итерации этой фазы алгоритм вычисляет информацию, необходимую для получения всех последующих оценок и их доверительных интервалов. На этапе прямого прохода полученная дополнительная информация учитывается посредством дополнения самих запросов в конвейере и схем таблиц с промежуточными результатами. Результатом этого алгоритма является ОАГ конвейера запросов, топологически эквивалентный исходному, но узлы и ребра этого графа дополнены информацией для выполнения приближенной обработки каждого подзапроса, что позволит обеспечить результат выполнения требуемыми статистическими гарантиями, применяя алгоритм приближенной обработки для каждого частного запроса в конвейере.

Заключение

На основе приведенного в [7] математического обеспечения были представлены разработанные алгоритмы приближенной обработки данных в реляционных СУБД в расширении работы с потоком запросов к хранилищам данных. Алгоритм приближенной обработки конвейера запросов позволяет адаптировать ОАГ конвейера для применения к его этапам алгоритма приближенной обработки частных запросов.

В дальнейшем представленные алгоритмы могут найти свое применение в синтезе архитектуры специального программного обеспечения процессора запросов системы управления базами данных, тем самым обеспечивая функцию приближенной обработки запросов в хранилищах данных, построенных на базе реляционных СУБД. Такой подход находится в плоскости существующей задачи разработки системы гибридной транзакционно-аналитической обработки запросов [**Ошибка! Источник ссылки не найден.**].

Как представлено в [10], такие системы потенциально дают ряд преимуществ, среди которых: отсутствие необходимости переноса данных при помощи инструментов ETL из OLTP в OLAP хранилища; данные доступны для аналитической обработки непосредственно с момента их создания транзакцией; устраняется или по меньшей мере уменьшается потребность в размножении копий одних и тех же данных в различных системах.

СПИСОК ИСТОЧНИКОВ

1. Babcock B., Chaudhuri S., Das G. Dynamic sample selection for approximate query processing. *Proceedings of International Conference on Management of Data, SIGMOD '03*. 2003;539–550. DOI: 10.1145/872819.872822.
2. Ganti V., Lee M., Ramakrishnan R. (2000). ICICLES: self-tuning samples for approximate query answering. *VLDB.* 2000;176–187.
3. Cormode G., Garofalakis M., Haas P.J., Jermaine C. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*. 2012;4(1–3):1–294. DOI:10.1561/1900000004.
4. Xu B., Tirthapura S., Busch C. Sketching asynchronous data streams over sliding windows. *Distributed Computing*. 2008;20(5):359–374. DOI:10.1007/s00446-007-0048-7.

5. Chaudhuri S., Ding B., Kandula S. Approximate query processing: No silver bullet. *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data*. 2017;511–519. DOI: 10.1145/3035918.3056097.
6. Григорьев Ю.А., Ухаров А.О., Плутенко А.Д. Использование вейвлет-преобразования для приближенной обработки многомерных данных. *Информатика и системы управления*. 2008;15(1):3–13.
7. Громей Д.Д., Козлов С.В., Филимонов А.В. Оптимизация распределения пространства выборки для запросов с группированием в процессе их приближенной обработки. *Системы управления и информационные технологии.* 2022;89(3):48–54. DOI: 10.36622/VSTU.2022.89.3.011.
8. Cao Y., Fan W. Data driven approximation with bounded resources. *Proceedings of the VLDB Endowment*. 2017;10(9):973–984. DOI: 10.14778/3099622.3099628.
9. Al-wesabi O.A., Abdullah N., Sumari P. (2020). Hybrid Storage Management Method for Video-on-Demand Server. *Emerging Trends in Intelligent Computing and Informatics*. 2020;1073:695–704. DOI: 10.1007/978-3-030-33582-3_65.
10. Козлов С.В., Невров А.А., Латышев И.П., Филимонов А.В. Подходы к приближенной обработке аналитических запросов в реляционных системах управления базами данных. *I-methods.* (2021);13(4). Доступно по: http://intech-spc.com/wp-content/uploads/archive/2021/4/7-kozlov.pdf (дата обращения: 30.09.2022).

REFERENCES

1. Babcock B., Chaudhuri S., Das G. Dynamic sample selection for approximate query processing. *Proceedings of International Conference on Management of Data, SIGMOD '03*. 2003;539–550. DOI: 10.1145/872819.872822.
2. Ganti V., Lee M., Ramakrishnan R. (2000). ICICLES: self-tuning samples for approximate query answering. *VLDB.* 2000;176–187.
3. Cormode G., Garofalakis M., Haas P.J., Jermaine C. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*. 2012;4(1–3):1–294. DOI:10.1561/1900000004.
4. Xu B., Tirthapura S., Busch C. Sketching asynchronous data streams over sliding windows. *Distributed Computing*. 2008;20(5):359–374. DOI:10.1007/s00446-007-0048-7.
5. Chaudhuri S., Ding B., Kandula S. Approximate query processing: No silver bullet. *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data*. 2017;511–519. DOI: 10.1145/3035918.3056097.
6. Grigor'ev Yu.A., Ukharov A.O., Plutenko A.D. Ispol'zovanie veivlet-preobrazovaniya dlya priblizhennoi obrabotki mnogomernykh dannykh. *Informatika i sistemy upravleniya*. 2008;15(1):3–13. (In Russ.).
7. Gromei D.D., Kozlov S.V., Filimonov A.V. Optimization of sample space distribution for questions with grouping in the process of their approximate processing. *Sistemy upravleniya i informatsionnye tekhnologii*. 2022;89(3):48–54. DOI: 10.36622/VSTU.2022.89.3.011. (In Russ.).
8. Cao Y., Fan W. Data driven approximation with bounded resources. *Proceedings of the VLDB Endowment*. 2017;10(9):973–984. DOI: 10.14778/3099622.3099628.

9. Al-wesabi O.A., Abdullah N., Sumari P. (2020). Hybrid Storage Management Method for Video-on-Demand Server. *Emerging Trends in Intelligent Computing and Informatics.* 2020;1073:695–704. DOI: 10.1007/978-3-030-33582-3_65.
10. Kozlov S.V., Nevrov A.A., Latyshev I.P., Filimonov A.V. Approaches to approximate processing of analytical queries in relational database management systems. *I-methods.* (2021);13(4). Available from: http://intech-spc.com/wp-content/uploads/archive/2021/4/7-kozlov.pdf (accessed on: 30.09.2022) (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Филимонов Алексей Валерьевич, Академия Федеральной службы охраны Российской Федерации, Орел, Российская Федерация.
e-mail: fay0@yandex.ru

Alexey Valerievich Filimonov, Russian Federation Security Guard Service Federal Academy, Orel, Russian Federation.

Статья поступила в редакцию 19.09.2022; одобрена после рецензирования 27.09.2022; принята к публикации 30.09.2022.

The article was submitted 19.09.2022; approved after reviewing 27.09.2022; accepted for publication 30.09.2022.