УДК 004.622+004.658.6

СОЗДАНИЕ МЕТОДА СРАВНЕНИЯ РЕЛЯЦИОННЫХ ТАБЛИЦ

А. Ш. Якупов[1], Д. А. Клинов[2]

[1]ООО «Пьяно Тек»; [2]ООО «Делион»

[1]asyakupov@kpfu.ru, [2]daniil.klinov@delion.ru

Аннотация

Статья посвящена созданию быстрого метода сравнения огромного количества данных таблиц в рамках реляционных систем управления базами данных. Проведено исследование существующих решений и показана востребованность создания эффективного метода сравнения реляционных отношений. Создан алгоритм с использованием вероятностной структуры данных «Исчисляемый фильтр Блума» и метода Монте-Карло. Предлагаемое решение уникально в своем направлении, так как использует наименьшее количество временных ресурсов. Построена вероятностная модель созданного алгоритма. В процессе написания статьи были выявлены пути развития алгоритма в сторону внедрения параллелизации процессов.

Ключевые слова: *мультимножество, сравнение реляционных таблиц, гетерогенная система, исчисляемый фильтр Блума, метод Монте-Карло, репликация, Oracle, PostgreSQL, вероятностная структура данных*

ВВЕДЕНИЕ

В современном мире наблюдается рост количества информации. Согласно статистике аналитической фирмы IDC «Эра данных 2025» [13], объем данных, которые человечество накопит уже меньше чем через 10 лет, составит 163 зеттабайт. Для сравнения: весь мировой объем интернет-трафика в 2016 году едва превысил 1 зеттабайт.

Для хранения огромного количества данных требуются мощные и современные системы управления базами данных (СУБД), примерами которых являются Oracle, PostgreSQL, MySQL, Microsoft SQL Server, MongoDB и другие [12]. На сегодняшний день наблюдается рост популярности использования

PostgreSQL [12]. В России это обосновано развитием сообщества благодаря мероприятиям PG Day и PG Conf и постоянным расширением функциональных возможностей PostgreSQL. Массовая миграция данных в рамках импортозамещения поднимает вопрос сравнения перенесенных данных.

При переносе большого количества данных путем замены одной СУБД на другую специалисты не могут гарантировать корректность и целостность переноса в связи с тем, что могут возникнуть внутренние, непредсказуемые ошибки из-за различий реализаций двух реляционных баз данных, также ошибками, связанными с различиями типов данных, внутренних функций, синтаксиса последовательностей и так далее [14]. Процесс миграции нужно рассматривать в комплексе с обеспечением мер по отказоустойчивости, резервированию и безопасности новой системы. Существующие на данный момент инструменты предоставляют возможность сравнения только построчно [6–10]. Данный подход является большим неудобством в мире, когда данные превышают несколько сотен гигабайт на одну реляционную таблицу из-за ограничений временных ресурсов.

СОПОСТАВЛЕНИЕ И АНАЛИЗ СУЩЕСТВУЮЩИХ ИНСТРУМЕНТОВ И РЕШЕНИЙ ДЛЯ СРАВНЕНИЯ ТАБЛИЦ РСУБД

Рассмотрим особенности существующих решений:

- **Red Gate SQL Data Compare**. Возможность копирования и переноса данных поиска из баз данных разработки в стадию производства. Генерация сценариев T-SQL для обновления одной базы данных с содержимым другой. Возможность вести точную историю всех предыдущих записей баз данных. Возможность сравнения и синхронизации данных, хранящихся в SQL Server Management Studio.
- **dbForge Data Compare**. Ручная настройка методов сравнения, позволяющая сократить время простоя системы, вызванное ошибками репликации. данных, и ускорение восстановления. Ускоренная разработка приложений, благодаря быстрому внедрению изменений данных. Индивидуальные сценарии синхронизации данных при низкой стоимости. Большая эффективность при сравнении больших баз данных. Широкая поддержка версий SQL Server.

- **EMS Data Comparer.** Отсутствие инструментов по сравнению структуры базы данных. Русификация от разработчика. Визуальное представление различий между данными в базах. Автоматическая и ручная выборки данных для сравнения. Возможность сохранять синхронизирующий сценарий. Большое разнообразие параметров для сравнения и синхронизации данных. Автоматическое сравнение/синхронизация данных. Возможность сохранения всех параметров, заданных в активной сессии. Гибкий графический интерфейс
- **SQLDelta**. Встроенное управление индексами, ключами и зависимостями между таблицами. Отчет в виде подробного html-файла. Загрузка баз данных происходит асинхронно.
- **SQL Comparison toolset from Idera**. Ведение историй сеансов сравнения. Обширный выбор фильтров, сортировок для системной аналитики после сравнения. Настройка сравнения, учитывающая ключи, столбцы, индексы. Не требует установки компонентов, динамических библиотек и ссылок.

Были изучены и проанализированы 5 инструментов сравнения таблиц РСУБД, которые суммарно имеют более 400 тысяч пользователей [6–10]. Все изученные инструменты используют построчный алгоритм сравнения таблиц. Некоторые инструменты используют параллелизацию алгоритма сравнения таблиц, но данный подход не является непосредственным преимуществом построчного сравнения таблиц перед сравнением таблиц с помощью сравнения их Исчисляемых фильтров Блума, так как параллелизацию можно использовать и при построении Исчисляемых фильтров Блума на таблицу.

После детального анализа существующих инструментов и решений для сравнения таблиц РСУБД не удалось найти инструмент, который бы использовал какую-либо иную технологию сравнения таблиц от сравнения таблиц построчно.

ОПИСАНИЕ АЛГОРИТМА СРАВНЕНИЯ ТАБЛИЦ И ЕГО ЭФФЕКТИВНОСТЬ

Целью конечного алгоритма является выявление простого факта несоответствия двух сравниваемых реляционных отношений с произвольным количеством записей и произвольным количеством атрибутов реляционных баз данных.

Задача состоит из нескольких этапов проверок, которые избегают использования построчного и последующего атрибутного сравнений данных.

В данный момент времени существующие на рынке инструменты сравнения используют обычный алгоритм перебора. Сложность алгоритма перебора составляет O(nm), где n – количество строк реляционного отношения, m – количество столбцов реляционного отношения.

Более того, существующие инструменты очень активно используют ресурсы компьютера/сервера, загружая части или весь объем данных двух сравниваемых отношений в оперативную память. Это связано не только с объемом данных, но и с производимой сортировкой атрибутов сравниваемых отношений. Алгоритм сортировки требует дополнительных затрат, и его сложность в худшем случае составляет O(mn^2), где n – количество строк реляционного отношения, m – количество столбцов реляционного отношения.

В свою очередь для выполнения сравнения Исчисляемых фильтров Блума достаточно их построить и сравнить между собой. Нет никакой сортировки и построчного сравнения каждого значения столбца в строке таблицы.

Сложность алгоритма в худшем случае линейна. Если требуется повторное сравнение таблиц после дополнительной миграции данных, то не требуется строить Исчисляемый фильтр Блума для таблиц, так как он может быть сохранен как заранее приготовленный вектор каждой из таблиц, что даст экономию в ресурсах вычислительной машины.

Этапы сравнения двух реляционных таблиц сводятся к последовательному выполнению шагов:

1. Сравнение типов таблиц;
2. Сравнение базовой статистики двух таблиц;
3. Построение и сравнение Исчисляемых фильтров Блума двух таблиц;
4. Сравнение таблиц с помощью метода Монте-Карло.

ВЕРОЯТНОСТНАЯ МОДЕЛЬ СТРУКТУРЫ ДАННЫХ «ИСЧИСЛЯЕМЫЙ ФИЛЬТР БЛУМА»

Необходимо сделать расчет вероятности равенства двух таблиц, у каждой из которых построен Исчисляемый фильтр Блума. Расчет вероятности равенства двух таблиц осуществляется с помощью сравнений двух заранее построенных Исчисляемых фильтров Блума.

Если при проверке равенства двух множеств с помощью сравнения их Исчисляемых фильтров Блума было установлено, что два множества не равны, то данных ответ является однозначным и не имеет погрешностей, так как Исчисляемые фильтры Блума, построенные на одинаковые множества элементов, не могут различаться.

Если при проверке равенства двух множеств с помощью сравнения их Исчисляемых фильтров Блума было установлено, что два множества равны: либо два множества действительно равны, либо был получен ложный положительный ответ, так как при построении Исчисляемых фильтров Блума двух неравных множеств с помощью случайного распределения данных были получены идентичные значения в каждом из соответствующих значений двух массивов Исчисляемых фильтров Блума.

Необходимо вычислить вероятность события, когда были построены два одинаковых Исчисляемых фильтра Блума для двух неравных множеств элементов.

Точный расчет данной вероятности не позволяет осуществить отсутствие информации о количестве неравных элементов двух множеств, при их наличии. На основании теории вероятностей и математической статистики можно наверняка вывести вероятностную формулу при наличии информации о количестве неравных элементов. Заранее неизвестно, какое количество элементов двух множеств не равны друг другу, но на основании переменной, которая будет определять количество неравных элементов, можно вычислить вероятность равенства двух множеств при сравнении соответствующих Исчисляемых фильтров Блума.

При наличии двух неравных множеств элементов вероятность события, когда были построены идентичные Исчисляемые фильтры Блума для этих

множеств, вычисляется по формуле

$$P = 1 / ((m + x * k - 1)! / ((x * k)! * (m - 1)!)),$$

где *P* – это вероятность события, *m* – количество ячеек в массиве Исчисляемого фильтра Блума, *x* – количество неравных элементов двух множеств элементов, *k* – количество объявленных хэш-функций при построении Исчисляемых фильтров Блума. Очевидно, что данная формула зависит от количества неравных элементов двух множеств. Другими словами, расчетная формула может дать ответ о том, что не установлено неравенство двух множеств с определенной вероятностью на основании сравнения двух Исчисляемых фильтров Блума.

Данная формула определяет количество вариаций выбора *kx* ссылок на определенные значения массива Исчисляемого фильтра Блума, которые образовали равное множество совместно со ссылками неравных элементов другого Исчисляемого фильтра Блума, инициировав тем самым ложноположительный ответ сравнения Исчисляемых фильтров Блума. Данное значение всегда равно 1, так как заранее можно установить, по какому набору ссылок неравных элементов сравниваемых неравных множеств образовались равные множества ссылок. Данное значение однозначное и неизменяемое, поэтому существует только один способ выбора *kx* ссылок на определенные значения массива Исчисляемого фильтра Блума.

Также данная формула определяет количество вариаций выбора *kx* элементов из множества с количеством элементов *m*.

На основании данных показателей вычисляется вероятность того, что произошло построение равных Исчисляемых фильтров Блума для неравных множеств элементов. Вычтя данную вероятность из единицы, получим вероятность того события, что произошло построение равных Исчисляемых фильтров Блума для равных множеств элементов.

Данная формула основывается на принципах теории вероятностей и комбинаторики, а именно, вычислении вероятности события и сочетания с повторениями.

На основании полученного значения по данной формуле определяется вероятность ложноположительного ответа, а на основании

ложноположительного ответа вычисляется вероятность верного положительного ответа, с учетом значения переменных *m* и *x* в самом худшем случае, а именно, когда *x* принимает значение 1.

ЗАКЛЮЧЕНИЕ

Проанализированы существующие инструменты сравнения реляционных таблиц и выявлены их недостатки. Разработано решение с использование вероятностной структуры данных «Исчисляемый фильтр Блума» и метода Монте-Карло для эффективного сравнения таблиц РСУБД.

Доказана эффективность данного алгоритма сравнения таблиц в разделе «Описание алгоритма сравнения таблиц и его эффективность» настоящей статьи и построена вероятностная модель алгоритма. Продолжить развитие приведенного алгоритма для последующей научной деятельности можно в сторону параллелизации процессов.

СПИСОК ЛИТЕРАТУРЫ

1. *Birialtsev E.* Intelligent search in Big Data [Text] // Approach to Data Integration. 2017. V. 46. No 19. P. 7–14.

2. *Chen G., Guo D., Luo L., Ren B.* Optimization of multicast source routing based on bloom filter // IEEE Communication Letters. 2018. No 4. P. 700–703.

3. *Kareev I.* Lower bounds for expected sample size of sequential procedures for the multinomial selection problems // Communications in Statistics. 2017. V. 913. No 1. P. 1–29.

4. *Wu K., Tan H., Liu Y., Zhang J., Zhang Q., Ni L.* Side channel: Bits over interference // IEEE Transactions on Mobile Computing. 2017. No 8. P. 1317–1330.

5. *Афанасьев Г.И., Марков А.Д.* База Данных NoSql и их сравнение с традиционными базами данных // Теория Инноваций. 2017. № 5-2. С. 4–10.

6. Официальная документация к инструменту сравнения таблиц РСУБД “Devart” [Электронный ресурс]. Режим доступа: https://www.devart.com (Дата обращения: 19.11.2018).

7. Официальная документация к инструменту сравнения таблиц РСУБД “Idera” [Электронный ресурс]. Режим доступа: https://www.idera.com (Дата обращения: 17.01.2019).

8. Официальная документация к инструменту сравнения таблиц РСУБД “Red Gate” [Электронный ресурс]. Режим доступа: https://www.red-gate.com (Дата обращения: 11.11.2018).

9. Официальная документация к инструменту сравнения таблиц РСУБД “SQL Delta” [Электронный ресурс]. Режим доступа: https://www.sqldelta.com (Дата обращения: 04.12.2018).

10. Официальная документация к инструменту сравнения таблиц РСУБД “SQL Manager” [Электронный ресурс]. Режим доступа: https://www.sqlmanager.net (Дата обращения: 03.12.2018).

11. Официальная документация РСУБД “Oracle Database” [Электронный ресурс]. Режим доступа: https://www.oracle.com/ru/database/ (Дата обращения: 10.02.2019).

12. Сайт DB-engines [Электронный ресурс]. Режим доступа: https://db-engines.com/en/ranking_trend (дата обращения 27.04.2019).

13. Сайт Seagate [Электронный ресурс]. Режим доступа: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (дата обращения: 25.04.2019).

14. Сайт Searchqlserver [Электронный ресурс]. Режим доступа: https://searchsqlserver.techtarget.com/definition/database (дата обращения: 23.05.2019)

15. Сайт W3techs. Trends in the usage of server-side languages for websites [Электронный ресурс]. Режим доступа: https://w3techs.com/ technologies/ history_overview/programming_language/m s/y (дата обращения 15.04.2019).

16. *Тишин А.О.* Разработка базы данных завершенных проектов // Евразийский научный журнал. 2017. № 5. С. 456–457.

CREATING A COMPARISON METHOD FOR RELATIONAL TABLES

A. S. Yakupov[1], D. A. Klinov[2]

[1]LLC “Piano”; [2]LLC “deLion”

[1]asyakupov@kpfu.ru, [2]daniil.klinov@delion.ru

Abstract

The article is devoted to creating a quick method of comparing a huge amount of data tables in relational database management systems. Creating an effective method for comparing relational systems is really relevant today. The study of existing solutions was conducted. The algorithm in this article was created using the probabilistic data structure «Countable Bloom filter» and the Monte Carlo Method. The proposed solution is unique in its direction, as it uses the least amount of temporary resources. A probabilistic model of the created algorithm is constructed, this algorithm can be used for parallelization.

Keywords: *multiset, comparison of relational tables, heterogeneous system, Countable Bloom filter, Monte Carlo method, replication, Oracle, PostgreSQL, Probabilistic data structure*

REFERENCES

1. *Birialtsev E.* Intelligent search in Big Data // Approach to Data Integration. 2017. V. 46. No 19. P. 7–14.

2. *Chen G., Guo D., Luo L., Ren B.* Optimization of multicast source routing based on bloom filter // IEEE Communication Letters. 2018. No 4. P. 700–703.

3. *Kareev I.* Lower bounds for expected sample size of sequential procedures for the multinomial selection problems // Communications in Statistics. 2017. V. 913. No 1. P. 1–29.

4. *Wu K. , Tan H., Liu Y., Zhang J., Zhang Q., Ni L.* Side channel: Bits over interference // IEEE Transactions on Mobile Computing. 2017. No 8. P. 1317–1330.

5. *Afanasev G.I., Markov A.D.* Baza Dannyh NoSql i ih sravnenie s tradicionnymi bazami dannyh // Teoriya innovazii. 2017. No 5-2. S. 4–10.

6. Official documentation for the table comparison tool for RDMS “Devart” [Internet resource]. Access mode: https://www.devart.com (Date of the application: 19.11.2018).

7. Official documentation for the table comparison tool for RDMS “Idera” [Internet resource]. Access mode: https://www.idera.com (Date of the application: 17.01.2019).

8. Official documentation for the table comparison tool for RDMS “Red Gate” [Internet resource]. Access mode: https://www.red-gate.com (Date of the application: 11.11.2018).

9. Official documentation for the table comparison tool for RDMS “SQL Delta” [Internet resource]. Access mode: https://www.sqldelta.com (Date of the application: 04.12.2018).

10. Official documentation for the table comparison tool for RDMS “SQL Manager” [Internet resource]. Access mode: https://www.sqlmanager.net (Date of the application: 03.12.2018).

11. Official documentation for RDMS “Oracle Database” [Электронный ресурс]. Access mode: https://www.oracle.com/ru/database/ (Date of the application: 10.02.2019).

12. Site DB-engines [Internet resource]. Access mode: https://db-engines. com/en/ranking_trend (Date of the application: 27.04.2019).

13. Site Seagate [Internet resource]. Access mode: https://www.seagate. com/ (Date of the application: 25.04.2019).

14. Site Searchqlserver [Internet resource]. Access mode: https://searchsqlserver. techtarget.com/definition/database (Date of the application: 23.05.2019)

15. Site W3techs. Trends in the usage of server-side languages for websites [Internet resource]. Access mode: https://w3techs.com/technologies/ (Date of the application: 15.04.2019).

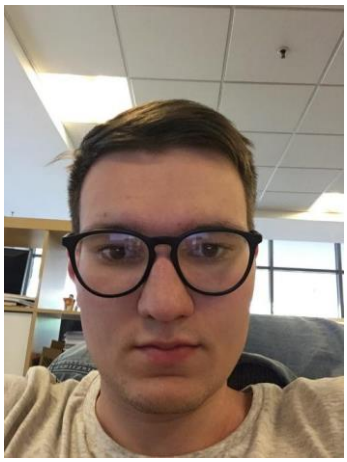
16. *Tishin A.O.* Razrabotka bazy dannyh zavershennyh proektov // Evraziiskii nauchnyi zhurnal. 2017. No 5. S. 456–457.

СВЕДЕНИЯ ОБ АВТОРАХ

ЯКУПОВ Азат Шавкатович – ассистент кафедры «Программная инженерия» Высшей школы информационных технологий и интеллектуальных систем, специалист в области баз данных.

Azat Shavkatovich YAKUPOV – Assistant at the department of “Software engineering” HS ITIS, senior of database developing.

asyakupov@kpfu.ru



КЛИНОВ Даниил Андреевич – бакалавр Высшей школы информационных технологий и интеллектуальных систем по направлению «Прикладная информатика».

Daniil Andreevich KLINOV – the bachelor of HS ITIS in the direction “Applied informatics”, junior of database developing.

daniil.klinov@delion.ru

Материал поступил в редакцию 29 июня 2019 года