УДК 004.65

ИССЛЕДОВАНИЕ АНАЛИТИЧЕСКИХ ВОЗМОЖНОСТЕЙ РАЗЛИЧНЫХ СУБД

Назаров А.А. (Университет ИТМО)
Научный руководитель – доцент, кандидат физико-математических наук, Графеева Н.Г.
(Университет ИТМО)

Введение. С увеличением количества информации возникает потребность в ее обработке. На текущий момент достаточно распространен способ, когда для анализа данных используется язык Python. Однако, для хранения больших объемов данных по-прежнему используются реляционные СУБД (Oracle, PostgreSQL, SQL Server, DB2 и другие). Следовательно, встает вопрос о значительных затратах по переносу данных из реляционной СУБД в программные средства, реализующие обработку информации, а также транспортировке результатов аналитики обратно, в реляционные таблицы. Также, при передаче данных по сети, необходимо учитывать требования к безопасности [1]. В связи с этим, набирают популярность решения, способствующие обработке данных внутри СУБД: с использованием параллелизма, индексирования и партиционирования, точек сохранения, курсоров и пользовательских табличных функций. То есть, классических преимуществ реляционного способа хранения и организации данных. Тем не менее, далеко не все средства управления базами данных предоставляют такую возможность: а если и предоставляют, то на разном уровне.

Основная часть. В настоящее время существует три основных способа интеграции между процессингом аналитики данных и реляционным хранилищем:

- **Слабое связывание** – реляционное хранилище рассматривается исключительно как источник данных.
- **Сильное связывание** – вся обработка данных происходит в реляционном хранилище. Подобный способ интеграции способствует масштабируемости, переносимости и высокой производительности [2].
- **Среднее связывание** – нечто среднее между первыми двумя: СУБД используется для возможной подготовки и предобработки данных, но сами алгоритмы реализованы средствами внешней системы.

Поскольку сильное связывание реализовать сложнее всего из-за скупости в структурах данных реляционных СУБД и кортеж-ориентированности, в основном исследования направлены на создание библиотек и модулей, которые можно подключить к реляционной базе данных с целью реализовать хотя бы среднее связывание. Противоположным подходом является реализация алгоритмов машинного обучения непосредственно на процедурном SQL (и многим авторам удалось реализовать единичные алгоритмы – широкий обзор на них представлен в работе [3], другие из них – представили работы, посвященные сравнению традиционной реализации алгоритмов машинного обучения и их разработкой с помощью средств СУБД [4]). Отдельно стоит отметить создание промежуточных расширений языка SQL, которые позволяют трансформировать возможности SQL и процедурного SQL и адаптировать их к структурам данных другого типа (например, деревьям, кучам и графам, очередям): среди них можно отметить Data Mining Query Language, Temporary Mining Query Language, Geo-Mining Query Language и другие.

Если рассматривать СУБД как целостный и самодостаточный программный объект, позволяющий заключить в себе как эффективное хранение, так и качественный анализ данных, стоит ввести критерии для сравнения: полнота и состав решений, интегрированных и адаптированных для решения задачи, язык реализации (позволяющий использовать или не использовать преимущества структур реляционных СУБД, уникальные реализации и оптимизации традиционных алгоритмов). У Oracle есть два достаточно мощных модуля, предназначенных для анализа данных: Oracle Data Mining [5] (на процедурном SQL –

единственное полноценное решение для реляционных СУБД, максимально адаптированное к особенностям структуры данных), а также Oracle Machine Learning (написанный на смеси SQL/R, но реализующий дополнительно алгоритм градиентного бустинга [6]); для PostgreSQL на текущий момент максимально конкурирующим решением выглядит только библиотека MADlib [7], реализованная на Python, но обладающая более скудным набором алгоритмов и проигрывающая в оптимизации. Что касается DB2, можно выделить наличие статистических критериев; у Microsoft Analysis Services интересен язык реализации – DMX, являющийся надстройкой традиционного SQL, адаптированный к работе с большими объемами данных и построению OLAP-систем.

Вывод. На текущий момент только СУБД Oracle дает возможности своим пользователям для эффективной аналитики данных средствами СУБД: это проявляется как в особенностях реализации модулей, так и в наполненности различными алгоритмами. Тем более важным выглядит задача по разработке аналитических пакетов для других современных СУБД.

Список литературы

1. Viloria A., Acuna G., Franco D.J.A., Hernandez-Palma H., Fuentes J.P., Rambal E.P. Integration of Data Mining Techniques to PostgreSQL Database Management System // Procedia Computer Science. – 2019. – №155 (5). – P. 575-580.

2. Han J., Kamber M. Data Mining: Concepts and Techniques. — Morgan Kaufmann, 2006. — 743 p.

3. Цымблер М. Обзор методов интеграции интеллектуального анализа данных в СУБД // Вестник Южно-Уральского Государственного Университета. Серия: вычислительная математика и информатика. — 2019. — Т. 8. — С. 32—62.

4. Калинина Е., Манохина Т., Ступаков С.Оптимизация процессов планирования запросов баз методами машинного обучения // Цифровая экономика. – 2022. – № S5(21).

5. Документация Oracle Data Mining. [Электронный ресурс] — Режим доступа: https://docs.oracle.com/en/ database/oracle/sql-developer/21.4/dmrrn (дата обращения: 07.11.2023).

6. Документация Oracle Machine Learning. [Электронный ресурс] — Режим доступа: https://docs.oracle.com/ en/database/oracle/machine-learning (дата обращения: 07.11.2023).

7. Документация библиотеки MADlib. [Электронный ресурс] — Режим доступа: https://madlib.apache.org/ docs/latest/index.html (дата обращения: 26.11.2023).

Автор _____ Назаров А.А
Научный руководитель _____ Графеева Н.Г