Международный журнал прикладных наук и технологий "Integral"

Научная статья

Original article

УДК 004.42

DOI 10.55186/02357801_2022_7_2_7

ИМПОРТОЗАМЕЩЕНИЕ В ОБЛАСТИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ БИЗНЕС-АНАЛИТИКИ

IMPORT SUBSTITUTION IN BUSINESS INTELLIGENCE SOFTWARE

Шимановский Константин Викторович, кандидат экономических наук, доцент кафедрой информационных систем и математических методов в экономике, ФГАОУ ВО «Пермский государственный национальный исследовательский университет» (614068, г. Пермь, ул. Букирева, 15), тел. 8(342) 239-64-35, shimanovskiy@list.ru

Konstantin V. Shimanovskiy, Candidate of Sciences (Economics), Associate Professor of the Department of Information Systems and Mathematical Methods in Economics, Perm State National Research University (15 Bukareva st., Perm, 614068, Russia), tel. 8(342) 239-64-35, shimanovsky@list.ru

Аннотация. В статье рассматривается вопрос импортозамещения программного обеспечения для бизнес-анализа данных. Раскрываются основные архитектурные особенности российского ИТ-решения класса Business Intelligence – «Форсайт. Аналитическая платформа». Выделяются ее основные конкурентные преимущества (в части аналитической обработки информации) среди других российских аналогов. Определяются сходства и различия между OLTP и OLAP подходами при онлайн обработке данных, а также проводится анализ наиболее

эффективных импортонезависимых СУБД для хранения и обработки данных из BI платформы.

Abstract. The article discusses the issue of import substitution in the field of Business Intelligence software. The paper reveals the key architectural features of Russian BI solution called «Foresight. Analytical Platform». Its main competitive advantages in terms of analytical data processing compared with other Russian BI software are highlighted. The difference and connection between OLTP and OLAP approaches in online data processing are determined. The analysis of the most effective import-independent DBMS for storing and processing data from the BI platform is carried out.

Ключевые слова: платформа Business Intelligence, OLAP-кубы, гетерогенные источники данных, «Форсайт. Аналитическая платформа».

Keywords: Business Intelligence platform, OLAP cubes, heterogeneous data sources, «Foresight. Analytical Platform».

Введение

В современном мире анализ информации, или бизнес-анализ, занимает все более значимое место в сфере ИТ. С начала века скорость производства данных возросла уже на два порядка, и по оценкам компании IDC общий объем информации в мире вырастет к 2025 году до 175 зеттабайтов[1]. Для ее обработки нужны эффективные и современные средства бизнес-анализа, особый класс программных продуктов, который в мировой практике получил название Business Intelligent (сокращенно BI). Впервые этот термин появился еще до изобретения персонального компьютера [1], и дословно в переводе с английского означает «бизнес-интеллект» или, как определял его сам автор, «возможность понимания связей между представленными фактами». Но тогда, в далеком 1958 году, это было сугубо теоретическое понятие, не имеющее практического применения в ИТ-отрасли.

[1] один зеттабайт — это 10^{21} байтов

Российские BI платформы

Впервые ИТ-инструменты и задачи BI связали аналитики рейтингового агентства Gartner. С начала XXI века в формате ежегодных исследований [2] уже почти на протяжении 20 лет они публикуют «Magic Quadrant for Business Intelligence Platforms» (магический квадрант Гартнер). В этих исследованиях определяются позиции мировых компаний, занимающихся созданием программного обеспечения класса BI. Последние несколько в лет тройку лидеров стабильно входят ClikView, Tableau и Microsoft Power BI. Но в связи с геополитическими событиями в начале 2022 года большинство зарубежных вендоров приняли решение приостановить поставку лицензий на российском ИТ-рынке. По причине этого остро возникла потребность в поиске аналогичного open-source или отечественного проприетарного ПО. В настоящее время на рынке существует около 20 российских BI платформ [3]. Например, «Форсайт. Аналитическая платформа», Visiology, Luxms, Триофлай, Alpha BI и др. Каждая из них обладает своими преимуществами и ограничениями.

Среди всех этих отечественных программных продуктов в международном рейтинге магического квадранта Гартнер (по данному классу решений) участвовала только BI платформа «Prognoz Platform» (после ребрендинга в 2017 году – «Форсайт. Аналитическая платформа»). Эта платформа имеет почти 30-летнюю историю развития, а также обширную клиентскую базу в России и за рубежом [4]. Включение ее в столь авторитетный международный рейтинг наравне с мировыми гигантами BI-индустрии бесспорно подчеркивает ее уникальность и высокий уровень реализованного инструментария.

В рамках данной статьи автор рассматривает основные архитектурные аспекты технологии, которая применятся в «Форсайт. Аналитическая платформа» для аналитической обработки данных.

Многомерные OLAP-кубы

Онлайн режим обработки информации давно уже стал нормой для большинства бизнес-процессов в любой компании. Но важно различать два

формата такой обработки [5]. Первый – это работа с данными в режиме транзакций (Online Transaction Processing, OLTP), т. е. ввод, редактирование и удаление небольших «порций» информации. Тут решается задача накопления массива детальных/первичных сведений. Второй формат – это аналитическая обработка информации (Online Analytical Processing, OLAP). При анализе нам уже не нужно корректировать данные. Основная задача OLAP-анализа заключается в агрегировании первичной информации в разных разрезах. Например, первичные данные (транзакции) по заработной плате заполняются ежемесячно по каждому сотруднику, а проанализировать нужно агрегированные сведения за год в разрезе отделов компании.

На практике для такого бизнес-анализа используют многомерные OLAP-кубы. Они очень похожи на многомерные матрицы. Только вместо осей в многомерном кубе используют аналитические разрезы. Причем каждый аналитический OLAP-куб характеризуется своим набором детализирующих его аналитик.

Гетерогенность данных в многомерных BI кубах

Современная концепция «озёр данных» (Data Lakes) определяет для хранения информации новые, более сложные и масштабные условия. Появляются разные информационные слои: сырые данные, оперативные и консолидированные данные, витрины аналитических данных и т. п. Для каждого из этих слоев целесообразно использовать разные технологии СУБД. Среди российских проприетарных баз данных стоит выделить:

- линейку продуктов Аренадата;
- реляционную СУБД PostgrePRO;
- реляционную СУБД ЛИНТЕР;
- СУБД с повышенной защитой информации Jatoba;
- Колоночную in-memory СУБД Tarantool от mail.ru (с 2021 года VK).

Большинство этих СУБД разрабатываются как проприетарные форки[2], которые были сформированы из кодовой базы open-source, таких как PostgreSQL, Clickhouse, Greenplum и др.

Все эти СУБД ориентированы на обработку информации из разных слоёв «озера данных». Например, исторические данные с длительной динамикой нужно размещать в MPP-СУБД (что обеспечивает быстрое извлечение информации на больших объемах), транзакционная (часто изменяемая) исходная или расчетная информация хранится в реляционной СУБД и т. д.

В связи с этим для разных слоев данных в BI платформе приходится комбинировать разные СУБД. Они являются источником данных для BI, а итоговый аналитический срез данных объединяется уже на стороне BI-сервера в оперативной памяти. Таким образом, для работы многомерного OLAP-куба с информацией из «озера данных» необходимо обеспечить его взаимодействие с несколькими совершенно разными физическими таблицами из разных БД. Именно в такой архитектуре работает data engine в программном продукте «Форсайт. Аналитическая платформа» (далее платформа «Форсайт»).

Как BI «Форсайт» взаимодействует с реляционной и многомерной БД

В платформе «Форсайт» процесс получения («добычи») аналитических данных из исходных источников проходит несколько уровней абстракции – см. рисунок 1. Эти уровни тесно связаны друг с другом. Каждый последующий использует результаты предыдущего и добавляет к данным новые функции, свойства, признаки.

На нулевом уровне (Tier 0) все данные в платформе «Форсайт» хранятся во внешних источниках. Это могут быть реляционные СУБД, многомерные базы данных, веб-сервисы, файлы, языки программирования и др. С помощью различных универсальных или нативных драйверов платформа подключается к

[2] Форк (fork с англ. — «развилка, вилка») или ответвление — использование кодовой базы программного проекта в качестве старта для другого, при этом основной проект может как продолжать существование, так и прекратить его.

этим источникам и запросам и получает плоские данные (sql, mdx, http и другие виды запросов или прямое обращение через api).

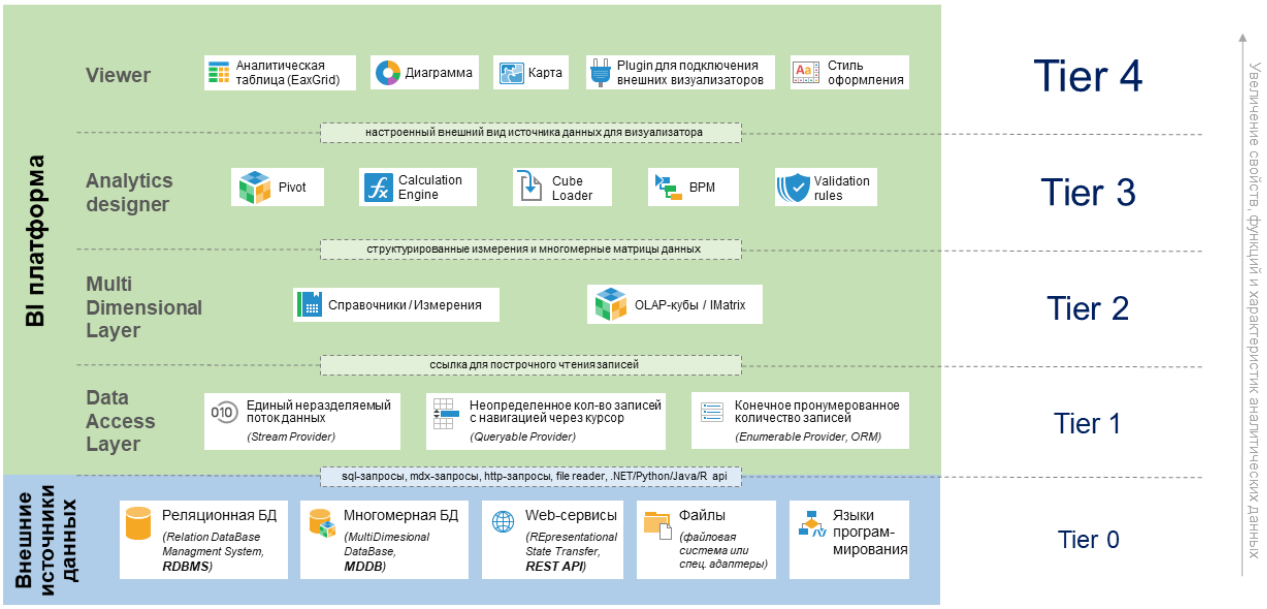

Рисунок 1. Уровни абстракции при создании аналитических данных в BI.

Как только BI-сервер получил с сервера СУБД (или иного источника хранения данных) плоские данные – это уже первый уровень абстракции (Tier 1). На этом уровне происходит конвертация типов данных. Важно понимать, что у разных внешних источников такие типы могут быть различными. Например, в Clickhouse для целого числа есть 6 разных типов [6]: int8, int16 и т. д., а в PostgreSQL их всего 3 [7]: smallint, integer и bigint. Или другой пример, финансовый тип currency поддерживают далеко не все потенциальные источники данных. И таких примеров достаточно много. Для BI платформы, которая поддерживает работу с гетерогенными источниками данных, важно и нужно приводить все эти типы данных к «общему знаменателю». Эту задачу решает специально реализованная в платформе «Форсайт» библиотека Data Access Layer [4].

Далее линейные и плоские источники данных структурируются в упорядоченные/иерархические измерения (Dimensions) или в наборы многомерных данных кубов (IMatrix[3]). Это уже второй уровень абстракции – «Tier 2». Он работает в многомерной парадигме. Именно здесь возникает гетерогенность

[3] IMatrix – это внутренняя система многомерного хранения данных в оперативной памяти BI сервера, с которой уже работают все инструменты платформы «Форсайт» (виртуальный куб, отчетность, bpm, экспорт/импорт данных и др.)

данных, когда для OLAP-кубов в оперативной памяти BI сервера совмещается информация из «плоских» источников разного происхождения.

На третьем уровне абстракции (Tier 3) многомерные данные начинают «обогащаться» аналитическими функциями BI инструментов. Например, в Pivot определяется расположение измерений для размещения crosstable. В многомерных расчетах появляются формулы и т. п. Все эти сведения добавляются к данным. И мы уже видим не просто плоские таблицы или многомерные матрицы, а аналитически осмысленные информационные представления.

На последнем, четвертом уровне (Tier 4), к аналитике добавляются свойства визуального отображения: разметка листа, оформление, размеры и т. п.

Механизмы кэширование для OLAP-кубов (in-memory)

Кэширование данных многомерного куба, пожалуй, самая сложная и затратная (с точки зрения аппаратных ресурсов сервера) операция. К ней нужно подходить крайне рационально. С одной стороны, кэш данных существенно увеличивает скорость реакции системы на информационные запросы пользователей. Не нужно каждый раз обращаться к исходным источникам данных, выполнять операцию их гетерогенного объединения и т. п. С другой стороны, совершенно бессмысленно загружать в оперативную память терабайты «холодных» или даже «теплых» данных. Целесообразно использовать комбинацию гибридной транзакционно-аналитической обработки информации [8], чередуя кэш BI платформы и прямой доступ к исходным источникам данных (live-connection).

Архитектура платформы «Форсайт» позволяет кэшировать многомерные данные при многопользовательской работе с данными. Этот механизм кэширования данных расположен на уровне «Tier 2». Связано это с тем, что переход информации с уровня «Tier 0» на уровень «Tier 2» занимает больше всего времени.

Во-первых, на уровне «Tier 0» приходит запрос с разными условиями фильтрации. Чем сложнее фильтры и объемнее общее количество записей во внешнем источнике, тем дольше будет обрабатываться такой запрос.

Во-вторых, на уровне «Tier 1» информация в большинстве случаев обрабатывается платформой построчно из курсора, который возвращают разные запросы. Такая построчная обработка:

- с одной стороны, длительная по времени – скорость чтения данных ограничена скоростью конкретного драйвера реляционной или многомерной БД;
- с другой стороны, обладает гибкостью и экономным расходом памяти – загрузить можно только те данные, которые необходимы сейчас и которые попадают под условия фильтрации в запросе.

Во-третьих, на уровне «Tier 2» каждая запись из полученных «плоских» данных должна пройти проверку на соответствие всем измерениям куба. Проверка заключается в поиске (LookUp) элементов измерений, соответствующих ключевым полям записи. Время этой операции напрямую зависит от размеров измерений (количества в них элементов). В финальную многомерную IMatrix куба попадают только те точки данных, все координаты которых прошли эту проверку.

По результатам тестирования, проведенного автором на бесплатно распространяемой демо-версии платформы «Форсайт», на уровнях «Tier 0, 1, 2» обычно уходит от 50% до 70% (а при сложных запросах и все 90%) общего времени прохождения всех этапов аналитической обработки данных. Поэтому, организовав на уровне «Tier 2» слой кэширования для IMatrix, можно существенно повысить скорость работы всей BI платформы.

Выводы

Платформа «Форсайт» является современным и эффективным средством аналитической обработки данных. Она включена в реестр отечественного ПО [9], соответствует всем критерия импотрозамещения и сможет заменить на российском ИТ-рынке зарубежные аналоги.

Литература

1. H. P. Luhn. A Business Intelligence System // IBM Journal of Research and Development (Volume: 2, Issue: 4, Oct. 1958, page(s): 314–319)

2. Kurt Schlegel, Bill Hostmann, Andreas Bitterer, Betsy Burton. Magic Quadrant for Business Intelligence Platforms, 1Q06 // Publication Date: 9 January 2006/ ID Number: G00136660
3. Громов С. Л. Исследование «BI круг Громова 2022», издание 3-е, апрель 2022 г.
4. Официальный сайт компании Форсайт. Веб-ссылка: https://www.fsight.ru.
5. Информационные системы в экономике: учебник для академического бакалавриата / В. Н. Волкова [и др.]; под редакцией В. Н. Волковой, В. Н. Юрьева. – М.: Издательство Юрайт, 2016. – 402 с. – Серия: Бакалавр. Академический курс.
6. Официальная онлайн документация по программному продукту Clickhouse. Веб-ссылка: https://clickhouse.com/docs/ru/sql-reference/data-types/int-uint.
7. Официальная онлайн документация по программному продукту PostgreSQL. Веб-ссылка: https://www.postgresql.org/docs/current/datatype-numeric.html#DATATYPE-INT.
8. Кузнецов С. Д., Велихов П. Е., Фу Ц. Аналитика в реальном времени, гибридная транзакционная/аналитическая обработка, управление данными в основной памяти и энергонезависимая память. Труды ИСП РАН, том 33, вып. 3, 2021 г., стр. 171–198.

References

1. H. P. Luhn. A Business Intelligence System // IBM Journal of Research and Development (Volume: 2, Issue: 4, Oct. 1958, page(s): 314-319)
2. Kurt Schlegel, Bill Hostmann, Andreas Bitterer, Betsy Burton. Magic Quadrant for Business Intelligence Platforms, 1Q06 // Publication Date: 9 January 2006/ ID Number: G00136660
3. Gromov S. L. Research "BI Circle of Gromov 2022", Issue 3, April 2022
4. Official website of the Foresight company. url: https://www.fsight.ru
5. Information systems in economics: a textbook for academic undergraduate studies / V. N. Volkova [and others]; edited by V. N. Volkova, V. N. Yuriev. – M.: Yurayt Publishing House, 2016. – 402 p. – Series: Bachelor. Academic course

6. Official documentation for the ClickHouse database. url: https://clickhouse.com/docs/en/sql-reference/data-types/int-uint
7. Official documentation for the PostgreSQL database. url: https://www.postgresql.org/docs/current/datatype-numeric.html#DATATYPE-INT
8. Kuznetsov S.D., Velikhov P.E., Fu Q. Real-time analytics, hybrid transactional/analytical processing, in-memory data management, and non volatile memory. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 3, 2021, pp. 171-198

Для цитирования: Шимановский К.В. ИМПОРТОЗАМЕЩЕНИЕ В ОБЛАСТИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ БИЗНЕС-АНАЛИТИКИ // Научно-образовательный журнал для студентов и преподавателей «StudNet» №2/2022.