УДК 004.67 DOI: 10.26102/2310-6018/2019.27.4.002

РАЗРАБОТКА КОНЦЕПТУАЛЬНОЙ МОДЕЛИ ОПЕРАТИВНО-АНАЛИТИЧЕСКИХ ВИТРИН ДАННЫХ

А.П. Раевич[1], Б.С. Добронец[2]
Федеральное государственное автономное образовательное учреждение высшего образования «Сибирский федеральный университет», Красноярск, Российская Федерация
[1]*e-mail: raevich.ap@yandex.ru*
[2]*e-mail: bdobronets@yandex.ru*

Резюме: Интегрированные с аналитическими системами хранилища ориентированы на многомерное моделирование данных, которое обеспечивает быстроту выполнения аналитических запросов, но обладает существенными недостатками при работе с большими данными. В статье предложен подход к построению концептуальной модели оперативно-аналитических витрин данных, позволяющий совмещать концепции оперативных витрин и аналитических витрин данных. Оперативные витрины данных, представляющие собой информационные срезы узконаправленной, тематической информации, призваны решать проблему оперативного доступа к источникам больших данных за счет консолидации и ранжирования информационных ресурсов по уровню востребованности. В отличие оперативных витрин данных, являющихся зависимыми от источников, аналитические витрины данных рассматриваются как независимые источники данных, создаваемые пользователями с целью обеспечения структуризации данных для решаемых задач. В работе приводится сравнение подходов к построению аналитических запросов на основе линейных запросов и ассоциативных связей. Полученные в работе результаты используются в построении BI кластера, на основе которого выполняется быстрое проектирование, аналитика, разработка и внедрение моделей бизнес-процессов с использованием предподготовленных оперативных и аналитических витрин данных.

Ключевые слова: системы бизнес-аналитики, оперативно-аналитические витрины данных, ассоциативная модель данных.

Для цитирования: Раевич А.П., Добронец Б.С. Разработка концептуальной модели оперативно-аналитических витрин данных. *Моделирование, оптимизация и информационные технологии.* 2019;7(4). Доступно по: https://moit.vivt.ru/wp-content/uploads/2019/11/Raevich_4_19_1.pdf DOI: 10.26102/2310-6018/2019.27.4.002

DEVELOPMENT OF A CONCEPTUAL MODEL OF OPERATIONAL - ANALYTICAL DATA MARTS

A.P. Raevich, B.S. Dobronets
Federal State Autonomous Educational Institution of Higher Education "Siberian Federal University", Krasnoyarsk, Russian Federation

Abstract: The storages integrated with analytical systems are focused on dimensional data modeling, which provides quick execution of analytical queries, but has significant drawbacks when working with big data. The article proposes an approach to constructing a conceptual model of operational-analytical data marts, which allows combining the concepts of operational data marts and analytical data marts. Operational data marts are information slices of narrowly focused, thematic information, which designed to solve the problem of operational access to big data sources through the consolidation and ranking of information resources in terms of demand. In contrast to operational data marts that are dependent from

sources, analytical data marts are considered as independent data sources created by users in order to provide data structuring for the tasks being solved. The paper provides a comparison of approaches to the construction of analytical queries based on linear queries and associative relationships. The results obtained in this work are used in building a BI cluster on the basis of fast design, analytics, development and implementation of business process models which are performed with using prepared operational-analytical data marts.

Keywords: business intelligence systems, operational-analytical data marts, associative data model.

For citation: Raevich A.P., Dobronets B.S. Development of a conceptual model of operational - analytical data marts. *Modeling, Optimization and Information Technology*. 2019;7(4). Available from: https://moit.vivt.ru/wp-content/uploads/2019/11/Raevich_4_19_1.pdf DOI: 10.26102/2310-6018/2019.27.4.002 (In Russ).

Введение

Информация, данные и знания в современном, динамически меняющемся мире, являются неоспоримой ценностью и решающим фактором развития как для малого предприятия, так и для целой страны.

В результате развития вычислительной техники, а также за счет повсеместного внедрения средств мониторинга (датчики, программно-аппаратные автоматизированные системы) и учета поведенческой деятельности человека генерируется значительный поток данных, которые необходимо хранить и в последующем анализировать. Кроме того, поток данных ускорился благодаря использованию современных устройств связи для мобильных абонентов и всемирному успеху социальных платформ.

Это приводит к необходимости сбора огромного количества данных (объемы собираемых данных в хранилища, зачастую, превышают 10^{15} байт информации) как для решения стратегических задач анализа и принятия решений, так и для коммерческих целей по ведению учета и составлению отчетности [1].

Однако процесс принятия решений, основанный на обнаружении практически полезных и доступных интерпретации знаний, а также поиске скрытых закономерностей, становится эффективнее при использовании средств визуального моделирования данных, позволяющих эксперту-аналитику проверять гипотезы и выдвигать требования к данным, анализировать полученные результаты [2,3]. Для этого все чаще применяются системы бизнес-аналитики (Business Intelligence – BI), дающие возможность оперативной трансформации данных и предоставления этих данных в удобной для восприятия человеком форме посредством интерактивных визуализаций.

Системы BI, как класс систем реализующих принципы систем поддержки принятия решений (СППР) акцентируются на интеллектуальном управлении данными и подразумевают применение совокупности технологий, программного обеспечения и практик, направленных на достижение целей бизнеса.

Исходные детальные данные, которые необходимо анализировать, как правило, хранятся в транзакционных OLTP (On-Line Transaction Processing) системах. Для транзакционных систем, рассчитанных на выполнение операций в режиме реального времени, основным приоритетом выполнения операций является обеспечение минимального времени отклика при максимальной загрузке системы.

Кроме того, исходные данные для анализа могут содержаться в файлах и во внешних системах, быть слабо структурированными (для обработки такие данные должны проходить через специальные процедуры структуризации). Как правило при

использовании в компании разнородных источников для их объединения требуется ведение мастер данных: справочников, классификаторов и другой нормативно-справочной информации [4].

В следствие этого в современных системах BI применяется концепция использования хранилищ данных (ХД), которые решают задачи структуризации и объединения исходных данных с учетом их возможной недостоверности, противоречивости и быстроты изменений посредством выполнения ряда процедур: консолидации данных, трансформации, очистки и предобработки данных [4,5].

Широко применяемые в корпоративных хранилищах модели данных опираются на многомерное представление данных. Такие модели и схемы представления данных просты для понимания и построения запросов, оптимизированы для анализа и эффективны с точки зрения описания бизнес-процессов, но с точки зрения физического представления «больших данных» имеют ряд существенных недостатков. К недостаткам можно отнести невозможность загрузки всех данных в многомерную модель из-за их размера, сложность построения моделей для распределенных по узлам кластера данных, необходимость денормализации таблиц фактов, что может значительно увеличить объем обрабатываемых данных.

Немаловажным считается тот факт, что детальный и быстрый анализ исходных данных для извлечения скрытых знаний невозможен без «понимания» контекста данных. При сравнении различных моделей данных указывается, что отдельные единицы информации могут быть малозначимыми, с точки зрения восприятия конечным пользователем [5].

Информация может быть собрана в бесконечное количество наборов и информационных элементов, которые существуют одновременно в любом количестве разных информационных наборов. И чтобы довести информацию до уровня интеллектуальных знаний, эта информация должна быть контекстуализирована.

Таким образом при проектировании информационных структур для представления данных, призванных решать задачи для систем бизнес-аналитики необходимо учитывать объемы обрабатываемых данных, а также время, необходимое время на изменение аналитической модели при изменениях бизнес-задач, в том числе по причине изменений структуры данных на источнике.

Поэтому, в отличие от традиционного подхода использования хранилищ данных совместно с системами бизнес-аналитики, активно применяются: различные подходы к построению витрин данных, вычисления в памяти (In-Memory Computing) в системах бизнес-аналитики и использование ассоциативных моделей данных.

Предлагаемый в работе подход построения оперативно-аналитических витрин данных призван обеспечивать доступность источников, быстроту и структуризацию данных для решаемых задач.

Концепция оперативных витрин данных

Чтобы довести информацию до уровня интеллектуальных знаний посредством инструментария BI пользователям требуются в работе не только агрегированные данные, но и детальные: для проверки гипотез, поиска скрытых закономерностей, построения трендов и многих других аналитических задач.

Концепция оперативных витрин, как часть общей концепции оперативно-аналитических витрин, призвана решать задачи доступности большого количества источников, а также доступности источников больших данных, когда количество источников и размерности данных сводятся к минимально необходимым наборам, что

позволяет строить выборки данных и получать результат в режиме близкому к реальному времени.

Отметим, что витрины данных (DataMarts) в концепции оперативных витрин данных рассматриваются как логически и физически разделенные подмножества данных, представленные в виде срезовых массивов тематической, узконаправленной информации, ориентированной на потребности определенной группы пользователей.

В стандартном моделировании каждая сущность или объект реального мира может быть представлен в виде отдельной таблицы с множеством аргументов. Будем называть совокупность хранящихся вместе, взаимосвязанных таблиц минимальной избыточности, источником данных.

Обозначим таблицу источника данных как элемент информационного ресурса $IR = \{ir_1, ir_2, \dots, ir_N\}$, $N = |IR|$, к которому требуется оперативный доступ от группы потребителей. Тогда информационный каталог ресурсов IK будет состоять из кортежа информационных ресурсов:

$$IK = \langle IR_i \rangle, \forall ir_{ij} \in IR_i, j = 1 \dots N, N = |IR_i| \quad (1)$$

где IK – информационный каталог, в котором могут содержаться взаимосвязанные и не взаимосвязанные информационные ресурсы;

IR_i – i-й информационный ресурс, каждый j-й элемент которого ir_{ij} принадлежит одному из двух классов: классу нормативно-справочной информации NSI или классу детальных данных DD.

Очевидно, что востребованность различных элементов информационных ресурсов через запросы на выборку данных за длительный период времени T у разных групп пользователей будет отличаться.

Обозначим текущую частоту обращений к элементам информационного каталога через $f_c(ir_{ij})$. Тогда для предопределенного временного периода $T = \langle t_1, \dots, t_l, \dots, t_c \rangle$ может быть вычислена максимальная частота запросов к каждому элементу ir_{ij} как:

$$f_{max} = \max_T(\langle f_{t_1}(ir_{ij}), \dots, f_{t_l}(ir_{ij}) \rangle), t_l \in T, l = 1 \dots c \quad (2)$$

где f_{max} – максимальное значение агрегатной функции для кортежа значений частот востребованности элементов информационных ресурсов;

f_{t_l} – частота пользовательских обращений к ресурсу ir_{ij} в период t_l.

Для ранжирования информационных ресурсов по уровню востребованности вводится понятие нормализованного индекса:

$$w_c = \frac{f_c}{f_{max}}, \quad (3)$$

Таким образом, текущая частота w_c востребованности каждого элемента информационного ресурса для заданного периода времени T может определяться на множестве значений $\{x \in \mathbb{R} | 0 \le x \le 1\}$.

В качестве источников информации для хранилища оперативных витрин данных могут выступать транзакционные системы или другие хранилища, предназначенные для обработки «больших данных», такие как Hadoop (Рисунок 1).

Загрузка наиболее востребованных таблиц в оперативное хранилище данных выполняется посредством фильтрации элементов информационного каталога с помощью нормализованного индекса w_c.

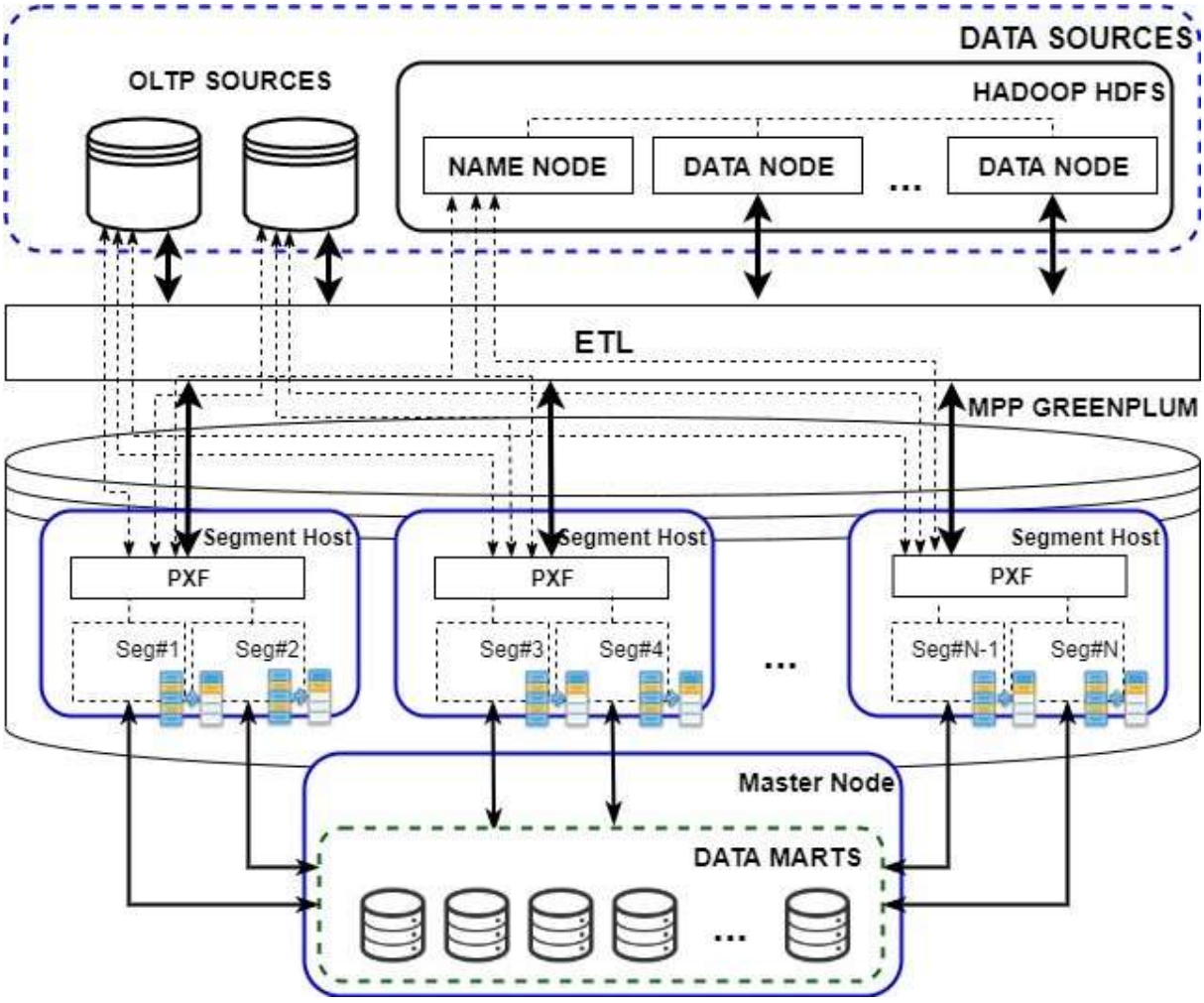

Рисунок 1. Хранилище оперативных витрин данных

Если все множество востребованности элементов информационного каталога $\{w_{c_i}\}$ рассматривать как базовую шкалу, то на ней можно построить следующие множества:

$$w_c \in \begin{cases} \text{"H"};\ ir_{ij} \in \{NSI, DD\};\ w_c \in [p_1 \dots p_2) \\ \text{"}M\text{"};\ ir_{ij} \in \{DD\};\ w_c \in [p_2 \dots p_3) \\ \text{"}L\text{"};\ ir_{ij} \in \{DD\};\ w_c \in [p_3 \dots p_5] \end{cases} \quad (4)$$

где "H" – класс наиболее часто востребованных элементов информационных ресурсов, которые должны быть включены в оперативное хранилище. Элементы информационных ресурсов данного класса могут относиться к типам детальных данных $\{DD\}$ и нормативно-справочной информации $\{NSI\}$ (не учитывается частота востребованности ресурса);

"M" – класс со средней востребованностью информационного ресурса. Элементы информационных ресурсов данного класса, относящиеся к типам детальных данных $\{DD\}$, могут быть включены в оперативное хранилище, только при наличии свободных ресурсов хранилища;

"L" – класс элементов информационных ресурсов с низкой востребованностью со

стороны потребителей детальных данных $\{DD\}$, не включаются в оперативное хранилище;

$\langle p_i \rangle$ - кортеж параметрических переменных, задающих разбиения всего интервала частот w_c на меньшие интервалы $p_1 < p_2 < p_3 < p_4$, которые не пересекаются.

Загрузка данных из источников выполняется начиная с класса "Н", включающего в себя детальные данные и нормативно-справочную информацию, используемую для контекстуализации детальных данных.

Применение ассоциативных моделей

Ключевым элементом любой системы бизнес-аналитики является заложенная в нее модель данных, отвечающая за быстроту, гибкость и функциональность системы. Для формализации описания информационных моделей представления данных в системах бизнес-аналитики используется математическая модель, основанная на теоретико-множественном подходе к описанию информационных процессов [6].

Традиционный подход к вычислениям в системах бизнес-аналитики основан на применении реляционной модели данных, предложенной Эдгаром Коддом в 1970 году [7]. Фактически эта модель стала стандартом, на который ориентируются современные системы управления базами данных (СУБД).

В отличие от традиционного подхода BI к вычислениям, когда обрабатываемые данные хранятся в реляционной базе данных на внешнем оборудовании, подход in-memory к вычислениям позволяет хранить обрабатываемые данные непосредственно в оперативной памяти вычислителя [8].

За счет этого достигаются следующие основные преимущества:

- вычисления в памяти предполагают упрощение процесса анализа данных из-за сокращения уровней агрегации и консолидации данных;
- адаптация модели данных может быть выполнена согласно меняющихся потребностей бизнеса путем подключения новых источников данных в качестве дополнительных источников информации на «лету»;
- аналитика в памяти уменьшает фрагментацию данных и повышает их точность;
- in-memory системы могут хранить активные структуры данных и выполняемый код пользовательских запросов непосредственно в памяти, тем самым выполняя расчеты для построения модели данных, фильтрации данных и построения агрегатов без необходимости сохранения промежуточных результатов на диск.

В основе большинства существующих BI-систем лежат технологии, основанные на линейных запросах к реляционным структурам данных. Поэтому такие инструменты бизнес-анализа, фактически, стали стандартными для систем поддержки принятия решений. Компанией QlikTech при создании системы бизнес-аналитики QlikSense была использована комбинация in-memory технологии и ассоциативной архитектуры модели данных в BI [8].

Для линейных запросов итоговый набор данных может быть представлен в виде выборки:

$$R_n = \sigma_{\varphi}(R), \tag{5}$$

где R_n – выборка (экземпляр отношения) из транзакционной базы представляет собой реляционную таблицу. Таким образом отношение R_n является результатом применения выборки к транзакционной БД и представляет собой множество n- нарных кортежей $\{\langle a_1, a_2, \dots, a_N \rangle\}$, $N = |R_n|$;

σ_φ – определяет формулу, по которой строится результирующее отношение и накладываются ограничения, задающие мощность отношения;

R – это схема отношения для заданного набора множеств атрибутов $S_1, S_2, S_3, \dots, S_n$, где каждый элемент $a_i \in R$ взят из определенного набора S, согласно заданных операций соединения.

Ассоциативным правилом, согласно [9], называется импликация $X \longrightarrow Y$, в которой наборы X и Y не пересекаются:

$$X \longrightarrow Y: X \subset I, \text{Y} \subset I, X \cap Y = \emptyset \quad (6)$$

Таким образом задача поиска ассоциативных правил заключается в поиске закономерностей между событиями из множества кортежей $\{\langle a_1, a_2, \dots, a_N \rangle\}$. При построении ассоциативных связей, связи устанавливаются двунаправленные между атрибутами таблиц (Рисунок 2):

$$S_i: \langle a_{ik} \rangle \longrightarrow S_j: \langle a_{jm} \rangle, S_i: \langle a_{ik} \rangle \longleftarrow S_j: \langle a_{jm} \rangle, i \neq j \quad (7)$$

где a_{ik} – ik-й атрибут отношения S_i для которого установлена связь с jm-м атрибутом a_{jm} отношения S_j, для которой экземпляры кортежей равны.

Организация данных на уровне ассоциативных связей позволяет видеть связи или их отсутствие между конкретными данными без построения линейных запросов, постепенно продвигаясь по зависимостям в модели. В сою очередь, механизм ассоциативной связи обеспечивает полное объединение связных таблиц, что обеспечивает доступ ко всем данным единовременно. Связывание данных в ассоциации выполняется на основе выбранных значений.

Выбор сущностей, атрибутов и фиксация взаимосвязей между сущностями зависит от семантики предметной области и выполняется системным аналитиком субъективно в соответствии с его личным пониманием специфики прикладной задачи. Таким образом использование комбинации технологий выполнения вычислений в памяти и построения ассоциативной архитектуры данных позволяет строить многомерную модель для решения бизнес-задач непосредственно в памяти вычислителя.

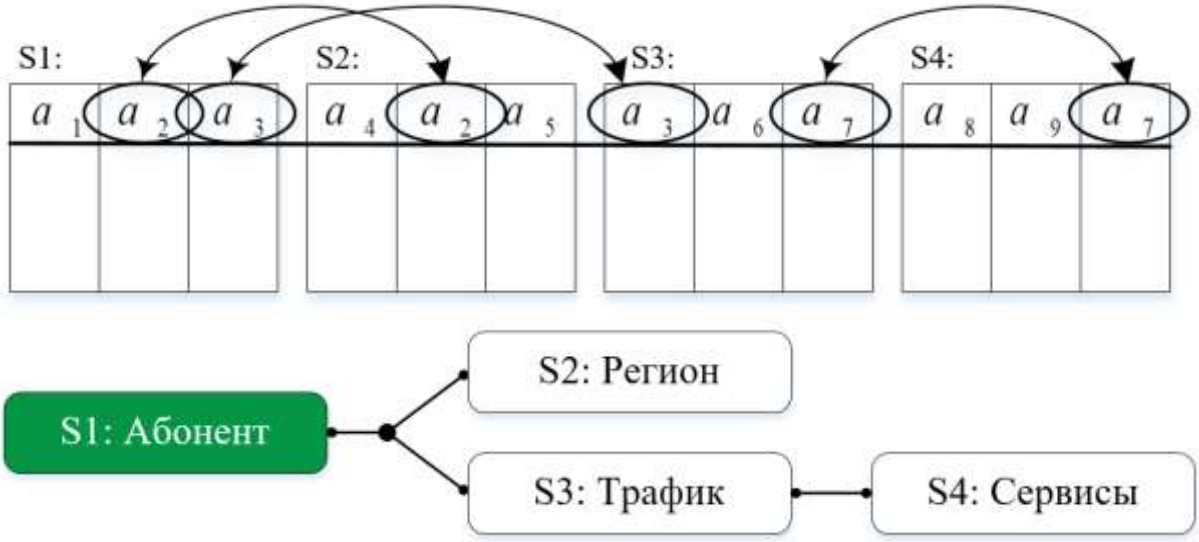

Рисунок 2. Структура запроса в ассоциативной модели

Набор данных, который предназначен для решения некоторого класса пользовательских задач, может быть описан в виде многомерного представления данных

H_n, имеющего n измерений. Тогда решение пользовательской задачи $F(H_n)$ в общем виде имеет многомерное представление H_n и записывается в виде кортежа:

$$H_n = \langle D_n, F_n \rangle \quad (8)$$

где $D_n = \{D_i\}$, $i = 1 \dots N$ - множество измерений куба, задающих систему координат для пространства данных. С каждым измерением связано конечное множество его значений, образующих грань куба. $\forall \mathrm{D}_i = \{d_{ij_i}\}$, $j_i = 1 \dots M_i$ – множество значений измерения, где d_{ij_i} - j_i значение измерения D_i.

F_n – это фактовые данные, где каждой комбинации из значений измерений $\forall\left(d_{1j_1}, d_{2j_2}, \dots, d_{nj_n}\right)$ соответствует определенный набор значений фактов $\exists F_{d_{1j_1}d_{2j_2}\dots d_{nj_n}} = \left\{f_{kd_{1j_1}d_{2j_2}\dots d_{nj_n}}\right\}$, $k = 1 \dots L$. Значение k-го факта $f_{kd_{1j_1}d_{2j_2}\dots d_{nj_n}}$ соответствует комбинации значений $\left(d_{1j_1}, d_{2j_2}, \dots, d_{nj_n}\right)$ измерений $(\mathrm{D}_1, \mathrm{D}_2, \dots, \mathrm{D}_n)$.

Концепция аналитических витрин данных

Построение логических моделей данных в централизованном хранилище, зачастую, выполняется с учетом двух основных требований: исключить избыточность и максимально повысить надежность данных, которые вытекают из подходов к коллективному использованию витрин данных хранилища группой пользователей.

Концепция аналитических витрин данных заключается в выделении профильных данных по определенному направлению деятельности. И, поскольку, выбор сущностей, атрибутов и фиксация взаимосвязей между сущностями зависит от семантики предметной области и выполняется системным аналитиком субъективно в соответствии с его личным пониманием специфики прикладной задачи, то множества атрибутов, описывающих информационные объекты, могут пересекаться, частично пересекаться или полностью не пересекаться при построении аналитических витрин.

На уровне бизнес-приложений содержатся показатели, приведенные в понятия конкретной бизнес-логики, а также проецируются в сгруппированные по бизнес-задачам отчеты конечных пользователей, на базе подготовленных и рассчитанных на уровне бизнес-логики объектов. Также выполняется вычисление узкоспециализированных показателей, строятся иерархии, и для каждой бизнес-области связываются в схемы таблицы фактов и таблицы измерений.

При построении витрин данных для заданного множества бизнес-процессов $\{M_1, M_2, \dots, M_n\}$ должны быть определены классы иерархии атрибутов сущностей в соответствии с потребностями аналитиков. Уровни детализации данных должны быть доступны в соответствии с самым низким уровнем детализации данных.

Так для куба H_n, представляющего собой решение пользовательской задачи для бизнес-процесса M_i, кортеж аналитических витрин данных определяется как:

$$\left(M_i, H_{n_i}\right) = \langle F_{d_i}, F_{a_i}, F_{m_i} \rangle \quad (9)$$

где F_{d_i} – витрина детальных (оперативных) данных, переносимых непосредственно из источников данных или оперативного хранилища. Они соответствуют элементарным транзакционным событиям;

F_{a_i} – агрегированные витрины, представляющие собой обобщенные значения атрибутов информационных объектов;

F_{m_i}– витрина метаданных, выступающая как индекс содержимого детальных данных для связи данных между витринами.

Общее описание функциональных блоков движения и трансформации данных в системе бизнес аналитики QlikSense концептуально может быть представлено в виде следующих основных уровней (Рисунок 3).

1) К источникам данных могут быть отнесены любые объекты, содержащие как структурированные, так и не структурированные данные, которые могут оказаться полезными для решения аналитических задач. Аналитическая платформа должна иметь доступ к данным из источника напрямую, либо после их преобразования в другой формат.
2) ETL слой выполняет извлечение данных из различных источников, преобразование к согласованному виду. Данные преобразуются из одного вида информации в другой с помощью методов обработки данных. В методах обработки данных применяются операции сбора данных, формализации данных и их фильтрации, сортировки данных, группировки и архивации, транспортировки и преобразования данных.

   Внутреннее хранилище аналитических витрин строится на базе концепции плоских индексированных таблиц QVD, обеспечивающее сжатие данных и высокую скорость чтения данных до 100 раз в сравнении с другими источниками данных.
3) Ядро системы выполняет построение ассоциативной модели для взаимосвязи данных, обеспечивая агрегацию и индексирование данных «на лету», перестраивая модель интерактивно под запросы пользователей без необходимости перегрузки исходных данных из источников.

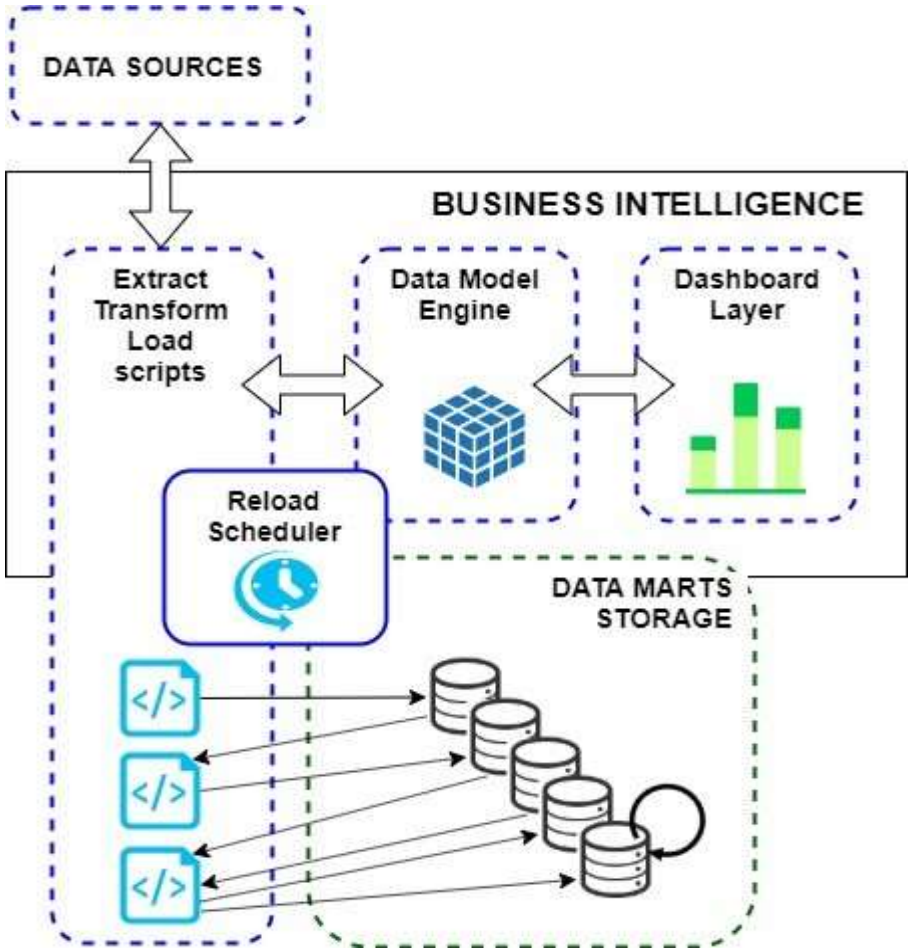

Рисунок 3. Информационная структура для BI системы QlikSense

Результаты

В качестве хранилища для оперативных витрин данных используется система на массивно параллельной архитектуре MPP GreenPlum. Заложенная в программно-аппаратном комплексе GreenPlum Database архитектура основана на разбиении полного массива данных на отдельные сегменты, работа с которыми может выполняться одновременно.

Архитектура GreenPlum изначально разрабатывалась для бизнес-аналитики и аналитической обработки данных на стандартном оборудовании. Сегменты данных автоматически распределяются между несколькими серверами сегментов, каждый из которых владеет и управляет отдельной частью общего массива данных.

Конфигурация «segment node»: [CPU] XEON 12 ядер 2.66GHz [RAM] 64GB [HDD] Hitachi scsi 3x146GB 10000rpm RAID-5. Сконфигурировано 36 сегментов (по 12 на сервер).

Конфигурация «master node»: [CPU] XEON 6 ядер 2.66GHz [RAM] 64GB [HDD] Hitachi scsi 146GB 10000rpm.

Результаты выполнения тестовых операций выгрузки срезовых массивов тематической информации из источника Hadoop Apache Hive в витрину GreenPlum представлены в Таблице 1.

Таблица 1. Время выполнения тестовых операций на кластере GreenPlum

Операция	Среднее время выполнения, мс
Создание external table для hive data warehouse для тестового набора данных 585 000 000 строк, 50 столбцов (int, float, text, datetime)	160,262
Создание физической таблицы distributed randomly из external table в 1Gbit сети	1733700,608
Выполнение операций join тестовых наборов 505 000 000 и 7 757 000 000 записей физически распределенных на сегментах по ключу randomly	884237,118
Сортировка сгенерированного массива 7 700 000 000 строк по одному полю	761334,002

Необходимо отметить, что в отличие от традиционных СУБД, хранящих данные строго построчно, Greenplum может хранить обрабатываемые данные, как в строках, так и в колонках. За счёт этого радикально снижается нагрузка на дисковую подсистему в процессе статистического анализа данных. Кроме того, одно из главных преимуществ Greenplum – это использование фреймворка PXF, позволяющего каждому сегменту параллельно обмениваться данными с источниками.

BI платформа Qliksense развернута на сервере следующей конфигурации: [CPU] XEON 24 ядра 2.33GHz [RAM] 1TB [HDD] Hitachi scsi 278GB 10000rpm.

В качестве тестовых наборов данных были сгенерированы два набора данных #1 = 9 000 000 строк и #2 = 11 000 000 строк содержащие 25 и 95 столбцов соответственно с типами данных int, float, varchar (<= 100 char), datetime. «Источник1» – локальное дисковое хранилище сервера QlikSense. «Источник2» – кластер GreenPlum.

Результаты выполнения тестовых операций для указанных источников и наборов данных представлены в Таблице 2.

Таблица 2. Время выполнения тестовых операций BI QlikSense

Операция	Источник1, мин		Источник2, мин	
	#1	#2	#1	#2
Загрузка данных без индексирования полей	0,23	0,27	6,8	9,88
Загрузка данных с индексированием полей	0,5	0,65	7,16	10,25
Загрузка данных с синхронизацией визуальной модели данных QlikSense без индексирования полей	1,77	1,83	8,1	11,13
Загрузка данных с синхронизацией визуальной модели данных QlikSense с индексированием полей	1,8	1,85	8,2	11,1

Для работы с различными источниками в Qliksense используются соответствующие ODBC коннекторы. При работе с источниками данных через визуальный редактор любые изменения в загрузочных скриптах требуют обновления данных и синхронизации с источником. Для анализа данных и построения моделей в визуальной среде целесообразно сохранение исходных данных во внутреннем формате QVD на локальное дисковое хранилище, что ускоряет операции чтения данных до 100 раз.

Заключение

Применение BI платформ как среды для визуального контроля за процессами трансформации и перехода данных от источников до виджетов является необходимым элементом для анализа данных и визуального контроля над потоками данных.

Невозможность оперативной загрузки больших объемов данных из источников для анализа нивелируется использованием внутреннего хранилища витрин данных, которое может быть инициализировано посредством полной выгрузки срезовых массивов данных, так и посредством частичной (инкрементальной) догрузки.

Кластер Greenplum в качестве оперативного хранилища витрин данных позволяет эффективно работать с большими данными выполняя параллельную обработку данных из источников на сегментах кластера. Greenplum содержит встроенные библиотеки аналитических алгоритмов с открытым исходным кодом, реализующие вычисления с параллельной обработкой в математических, статистических методах и методах машинного обучения для структурированных и неструктурированных данных.

Предложенная концептуальная модель оперативно-аналитических витрин данных апробированная на базе программно-аппаратного комплекса MPP GreenPlum и BI QlikSense позволяет разным группам специалистов за короткий промежуток времени (близкий к realtime) перестраивать витрины на основе больших данных, получать оперативный доступ к данным и, как следствие, максимально быстро перестраивать модель данных под решаемые задачи. Кроме того, это позволяет не только лучше понимать данные, получать о них более глубокое представление, но добиваться лучшей визуализации для конечных пользователей.

ЛИТЕРАТУРА

1. Эмиров Н.Д., Батталова С.С. Информационные услуги в современном информационном обществе: роль библиотек и их корпораций. *Экономика и предпринимательство*. 2017;11(88): 894-897.
2. Головина Т.А., Романчин В.И., Закиров А.И. Развитие технологий бизнес - аналитики на основе концепции Business Intelligence. *Известия Тульского государственного университета. Экономические и юридические науки*. 2014;5-1: 416-424.
3. Лубенец Н.А., Улитина Т.И., Акимова А.О. Современные ИТ-инструменты инновационного развития компании. *Экономические аспекты технологического развития современной промышленности*. 2017;:105-111.
4. Асадуллаев С. Архитектуры хранилищ данных-1. 2009. Доступно по адресу: https://www.ibm.com/developerworks/ru/library/sabir/axd_1/index.html (дата обращения 10.10.2019 г.).
5. Орешков В.И., Паклин Н.Б. Бизнес-аналитика: от данных к знаниям. ИД «Питер». 2013.
6. Ахрем А.А., Рахманкулов В.З., Южанин К.В. О сложности редукции моделей многомерных данных. *Искусственный интеллект и принятие решений*. 2016;(4): 79-85.
7. Homan J.V. et al. A comparison of the relational database model and the associative database model. *Issues in Information Systems*. 2009;10(1): 208-213.
8. Moving Towards Real-Time Analytics: All About In-Memory Computing and Self-Service BI. *Financial and credit activity: problems of theory and practice*. 2019;1(28): 272-278.
9. Зайко Т.А., Олейник А.А., Субботин С.А. Ассоциативные правила в интеллектуальном анализе данных. *Вестник Национального технического университета Харьковский политехнический институт. Серия: Информатика и моделирование*. 2013;39(1012).

REFERENCES

1. Emirov N.D., Battalova S.S. Informatsionnye uslugi v sovremennom informatsionnom obshchestve: rol' bibliotek i ikh korporatsiy. *Ekonomika i predprinimatel'stvo*. 2017;11(88): 894-897.
2. Golovina T.A., Romanchin V.I., Zakirov A.I. Razvitie tekhnologiy biznes - analitiki na osnove kontseptsii Business Intelligence. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Ekonomicheskie i yuridicheskie nauki*. 2014;5-1: 416-424.
3. Lubenets N. A., Ulitina T. I., Akimova A. O. Sovremennye IT-instrumenty innovatsionnogo razvitiya kompanii. *Ekonomicheskie aspekty tekhnologicheskogo razvitiya sovremennoy promyshlennosti*. 2017;:105-111.
4. Asadullaev S. Arkhitektury khranilishch dannykh-1. 2009. Dostupno po adresu: https://www.ibm.com/developerworks/ru/library/sabir/axd_1/index.html (data obrashcheniya 10.10.2019 g.).
5. Oreshkov V. I., Paklin N. B. Biznes-analitika: ot dannykh k znaniyam. ID «Piter». 2013.
6. Akhrem A.A., Rakhmankulov V.Z., Yuzhanin K.V. O slozhnosti reduktsii modeley mnogomernykh dannykh. *Iskusstvennyy intellekt i prinyatie resheniy*. 2016;(4): 79-85.
7. Homan J.V. et al. A comparison of the relational database model and the associative database model. Issues in Information Systems. 2009;10(1): 208-213.
8. Moving Towards Real-Time Analytics: All About In-Memory Computing and Self-Service BI. *Financial and credit activity: problems of theory and practice*. 2019;1(28): 272-278.

9. Zayko T.A., Oleynik A.A., Subbotin S.A. Assotsiativnye pravila v intellektual'nom analize dannykh. *Vestnik Natsional'nogo tekhnicheskogo universiteta Khar'kovskiy politekhnicheskiy institut. Seriya: Informatika i modelirovanie*. 2013;39(1012).

ИНОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Раевич Алексей Павлович, аспирант, кафедра системы искусственного интеллекта, ФГАОУ ВО "Сибирский Федеральный университет" Институт космических и информационных технологий, Красноярск, Российская Федерация.
ORCID: 0000-0002-9358-0651

Aleksey P. Raevich, PhD Student, Systems of Artificial Intelligence Department, Federal State Autonomous Educational Institution of Higher Education "Siberian Federal University", Krasnoyarsk, Russian Federation

Добронец Борис Станиславович, д-р. физ-мат. наук, профессор, кафедра системы искусственного интеллекта, ФГАОУ ВО "Сибирский Федеральный университет" Институт космических и информационных технологий, Красноярск, Российская Федерация.
ORCID: 0000-0002-0167-1637

Boris S. Dobronets, Dr. Sci. (Phys.–Math.), Professor, Systems of Artificial Intelligence Department, Federal State Autonomous Educational Institution of Higher Education "Siberian Federal University", Krasnoyarsk, Russian Federation