

ПРИМЕНЕНИЕ КОМПОНЕНТНОГО ПОДХОДА В ЗАДАЧЕ ПРОЕКТИРОВАНИЯ ХРАНИЛИЩА ДАННЫХ

Солянов К.С.,
Финансовый университет при Правительстве Российской Федерации

Аннотация: в статье рассмотрена роль хранилищ данных в информационном ландшафте организаций. Предметом исследования являются хранилища данных как ключевой элемент информационно-аналитической системы организации. Целью является выработка нового подхода к проектированию хранилищ данных, позволяющего повысить эффективность внедрений систем такого рода и качества итогового продукта. Повышение качества данных в хранилище позволяет руководителям организации принимать более взвешенные и обоснованные управленческие решения, что в конечном счете трансформируется в одно из конкурентных преимуществ. В статье рассмотрены существующие подходы к построению хранилищ данных, проведен анализ их недостатков и дана оценка. Вопрос проектирования хранилищ данных рассмотрен как с точки зрения направления проектирования (от бизнеса к данным или наоборот), так и с точки зрения организации данного процесса в масштабах организации. Также в исследовании предложена авторская концепция компонентного подхода в задаче проектирования хранилищ данных. Данная концепция предусматривает последовательный переход от высокоуровневого описания предметной области и бизнес-процессов к физической модели данных и последующей реализации хранилища. Применение предложенного подхода позволяет сократить время, затрачиваемое на внедрение хранилища данных, а также повысить качество конечного продукта и, как следствие, окупаемость проектов построения хранилищ данных. Результаты исследования могут быть применены в проектах по проектированию и внедрению корпоративных хранилищ данных в организациях различных отраслей. Представленные в статье результаты могут применяться как организациями-владельцами будущих хранилищ данных, так и компаниями-интеграторами, реализующими хранилища для своих заказчиков.

Ключевые слова: хранилище данных; проектирование; компонентный подход; анализ данных; информационно-аналитические системы; моделирование данных

Введение

Проблемам управления информационными потоками в организациях и вопросам проектирования хранилищ данных (ХД) в частности посвящено большое количество научных трудов. [1, 4, 5, 7, 8] Идея создания референтной модели хранилищ данных для разных отраслей разрабатывается многими крупными консалтинговыми компаниями и вендорами на протяжении многих лет. [3, 6] Например, в статье «Построение хранилищ данных IBM Data Warehouse для различных индустрий» [3] представлены модели данных, процессов, интеграции и другое для телекоммуникационных компаний, банков и финансовых организаций, страховых компаний, торговых организаций. Однако, западные модели не всегда успешно могут быть адаптированы под современные особенности и потребности российского бизнеса. Среди научных трудов отечественных авторов следует обратить внимание на работу Михалькевича И.С. [9] В своей работе автор недостаточно проработал вопрос разработки архитектуры корпоративного ХД: рассмотрение ХД с точки зрения решения одной узкоспециализированной задачи; поверхностный анализ предметной области; опущен вопрос о способе проектирования ХД. С другой стороны, к вопросу проектирования ХД подошел Ладин Е.Г.

[10] Недостатком предложенного автором подхода является отсутствие адаптивности ХД – автор не рассматривает проблему необходимости изменений и расширения ХД во времени.

Целью данного исследования является разработка универсальной методики проектирования ХД, основанной на бизнес-потребностях организации. Главной задачей в данном случае является повышение эффективности проектов внедрения ХД и повышение качества самих ХД. Основная гипотеза, выделенная автором, заключается в том, что любое ХД должно разрабатываться исходя из бизнес-требований, т.к. такой подход с большей вероятностью удовлетворит ожидания заинтересованных сторон и в конечном счете минимизирует их затраты.

1. Роль хранилища данных в информационно-аналитическом ландшафте организации

Хранилище данных (ХД) – это особый класс систем, вокруг которых строится информационно-аналитический ландшафт организации. Базовой задачей ХД является консолидация всей имеющейся в компании информации, ее унификация и верификация, а также интеграция имеющихся информационных систем – источников данных. Помимо этого, хранилище призвано накапливать исторические данные, позволяя производить ретро-

спективный анализ. На сегодняшний день наличие достоверного источника информации в масштабах всей организации является жизненной потребностью для любой крупной организации. Компания любой отрасли может добиться успеха только в случае принятия качественных управленческих решений, которые невозможны без обеспечения лиц, принимающих решения, (ЛПР) достоверной и полной информацией. ХД выполняют роль инструмента, позволяющего решить данную задачу.

ХД с одной стороны является приемником разнородной информации из множества источников, которая преобразуется и складывается в хранилище в единую модель. С другой стороны, ХД – это источник «вычищенной» и достоверной информации, которая может быть использована для решения комплексных бизнес-задач. Как правило, для этого в рамках ХД разрабатываются узконаправленные представления данных – *Витрины данных*, в которых информация приводится к требуемому для решения конкретной задачи формату.

Технически ХД представляет собой несколько слоев данных, которые могут быть организованы по-разному в зависимости от архитектуры конкретного решения. В общем случае ХД обязательно содержит следующие слои (или уровни) данных.

- **Операционный слой данных (ODS).** Задача данного слоя заключается в сборе всех необходимых данных из всех источников. При этом на этом уровне преобразования данных минимальны: информация хранится в моделях систем-источников, данные дублированы, справочная информация не унифицирована.

- **Детальный слой данных (DDS).** При загрузке данных в DDS выполняется основная часть преобразований данных. Данные ODS приводятся к единой модели, унифицируются и дедуплицируются. На этом уровне информация хранится в наиболее детальном виде, чтобы при необходимости ее можно было агрегировать определенным образом или же использовать без преобразований. Архитектурно DDS может быть реализован в виде нескольких слоев данных, на каждом из которых производится часть преобразований.

- **Слой витрин данных (DM).** Для решения большинства конечных бизнес-задач, связанных с информацией, используются витрины данных. Здесь данные могут быть преобразованы различным образом с целью удовлетворения требований потребителя соответствующей информации. В данном случае потребителем могут являться некоторые сотрудники, но в большинстве случаев бизнес-пользователи не работают с ХД напрямую, а используют различные аналитические приложе-

ния. Соответственно, с этой точки зрения потребителем информации в витрине данных является аналитическое приложение.

Важным аспектом при проектировании ХД является подход, по которому будет выстраиваться вся архитектура хранилища. Глобально проектирование ХД может быть организовано двумя диаметрально-разными способами.

2. Существующие подходы к проектированию ХД

2.1. Дата-ориентированный подход

Проектирование, ориентированное на данные – за основу берется структура, наполнение систем-источников и их особенности. Далее они сводятся воедино, формируя единый источник информации – ХД. Такой подход эффективен с точки зрения трудоемкости построения ХД, так как модель и архитектура изначально разрабатываются таким образом, чтобы реализация ETL была оптимальной. Однако, к сожалению, ХД, спроектированные таким образом, зачастую не приносят своим владельцам ожидаемых результатов. Возникают сложности в получении итогового business-value для компании, снижается степень доверия к данным в хранилище и в итоге такие ХД не используются и подвергаются серьезному рефакторингу или вовсе замене [1].

Одной из проблем здесь может оказаться тот факт, что зачастую в проектах построения ХД недостаточное внимание уделяется сбору и анализу требований. Т.е., разумеется, формируется список объектов и их характеристик, которые необходимы будущим пользователям ХД для решения своих бизнес-задач. Однако, далеко не всегда за этим следует высокоуровневый анализ этих требований, их консолидация и формирование целостной «картины мира». На это существует ряд причин:

- проекты построения хд очень долгосрочные и на начальных этапах сложно собрать и предугадать все требования к столь сложной системе;
- требования к данным довольно гибкие и часто изменяются во времени.

2.2. Бизнес-ориентированный подход

Проектирование, ориентированное на цели – за основу берутся бизнес-требования, т.е. те задачи, которые должны будут решаться с помощью данного ХД. Далее производится их анализ и поиск необходимых систем-источников. При этом модель данных в первую очередь ориентирована на конечные бизнес-задачи. Такой подход является более трудоемким, и требует большей квалификации от проектировщиков ХД, но взамен владелец ХД получает именно те данные, и именно в том виде, которые он ожидает, и которые позволяют решать его бизнес-задачи.

Рассмотрим второй подход более подробно. В основу сформулированного подхода в том числе лег метод GRAnD. GRAnD – goal-oriented approach to requirement analysis in data warehouses (подход к анализу требований к ХД, ориентированный на достижение целей бизнеса). [2] При таком подходе важной характеристикой ХД как информационной системы становится ее бизнес-ориентированность, т.е. чтобы конкретная реализация ХД была успешной, оно должно быть спроектировано исходя из особенностей и потребностей конкретной организации. Очевидно, что эти особенности и потребности в существенной мере одинаковы, или по крайней мере, значительно пересекаются в разных компаниях одной отрасли. В результате, при рассмотрении структур и моделей ХД организаций одной отрасли всегда можно выявить общие черты, или фрагменты. Таким образом, можно заключить, что при проектировании конкретного ХД есть некоторые общие компоненты, а есть частные, которые возникают от потребностей и предпочтений конкретной организации и лиц принимающих решения (ЛПР). Как было сказано, изучив и проанализировав модели ХД различных организаций одной отрасли можно выделить и формализовать общие компоненты, которые будут актуальны при проектировании ХД для любой организации данной отрасли.

3. Способы организации процесса построения ХД

Организационно процесс построения ХД в компании также может осуществляться по-разному. Вне зависимости от того, на что первоначально ориентируются проектировщики ХД (на бизнес-цели или данные), в компании может быть выбран один из трех способов решения задач по обеспечению своего бизнеса консолидированными данными.

3.1. Децентрализованный подход

Одной из распространенных проблем создания ХД в российских организациях является децентрализованный подход к решению этой комплексной задачи. Почему развитие ХД идет экстенсивным путем? Во-первых, ответственность за качество отчетности, полученной из ХД, является персональной. Чтобы доверять цифрам, руководители стремятся держать в своих руках данные для расчетов вместе с их источником — хранилищем данных. Во-вторых, у подразделения, инициировавшего создание ХД, мотивация делиться внутри банка своими наработками отсутствует во многом потому, что новые пользователи не разделяют с владельцем системы понесенные на проект расходы.

Все это приводит к лоскутной автоматизации и выращиванию в компании карманных ХД. В пер-

вую очередь такой подход противоречит одному из основополагающих постулатов ХД – «ХД должен быть единственным источником непротиворечивой информации». Наличие нескольких ХД в одной организации приводит к противоречивости данных и отчетов, что вводит лиц, принимающих решение в заблуждение или недоумение.

В дополнение, следует отметить, что наличие нескольких небольших ХД приводит не только к несогласованности и к дополнительным расходам, связанным с дублированием функционала и увеличению совокупных операционных затрат на сопровождение.

3.2. Единовременное внедрение полномасштабного корпоративного ХД

Кардинально другим подходом можно назвать централизованное внедрение корпоративного ХД в масштабах всей организации. Как правило, такие ХД проектируются и реализуются как авторская разработка, предназначенная исключительно для конкретной организации. В целом подход более удачный, т.к. лишен обозначенных выше недостатков, но с высокой вероятностью влечет за собой существенный перерасход ресурсов. Он обусловлен рядом причин:

- При столь масштабном проекте как внедрение корпоративного ХД практически никогда не удается собрать и правильным образом интерпретировать все требования к ХД;
- В попытке решить задачу сразу целиком зачастую допускаются архитектурные ошибки, которые вскрываются на конечных этапах и чрезвычайно сложны в устранении;
- В процессе реализации проекта требования неоднократно меняются и дополняются, а стоимость реализации таких изменений бывает весьма высокой в силу недостаточной гибкости данного подхода.

4. Компонентный подход

При бизнес-ориентированном подходе к проектированию ХД можно применять процесс перехода от бизнес-требований в терминах объектов предметной области и объектам бизнес-процессов к техническим требованиям и реализации, постепенно детализируя и уточняя требования. На основе научных трудов [1, 2, 6, 7], посвященных проектированию информационных систем и ХД в частности, а также на опираясь на результаты ряда проектов по построению корпоративных ХД в российских компаниях, была разработана методика проектирования ХД на основе конфигурируемой многокомпонентной семантической модели, которая позволяет повысить экономическую эффективность проектов разработки ХД и увеличить business value от их внедрения. Концептуально

данная методика должна включать следующие шаги (см. рис. 1):

1. Формирование вербальной модели в виде словесного описания предметной области;
2. Формирование набора компонент будущего ХД, которые должны соответствовать объектам предметной области и объектам бизнес-процессов;
3. Создание семантической модели, т.е. обогащение связей между компонентами (ключевыми бизнес-объектами) смысловой нагрузкой;

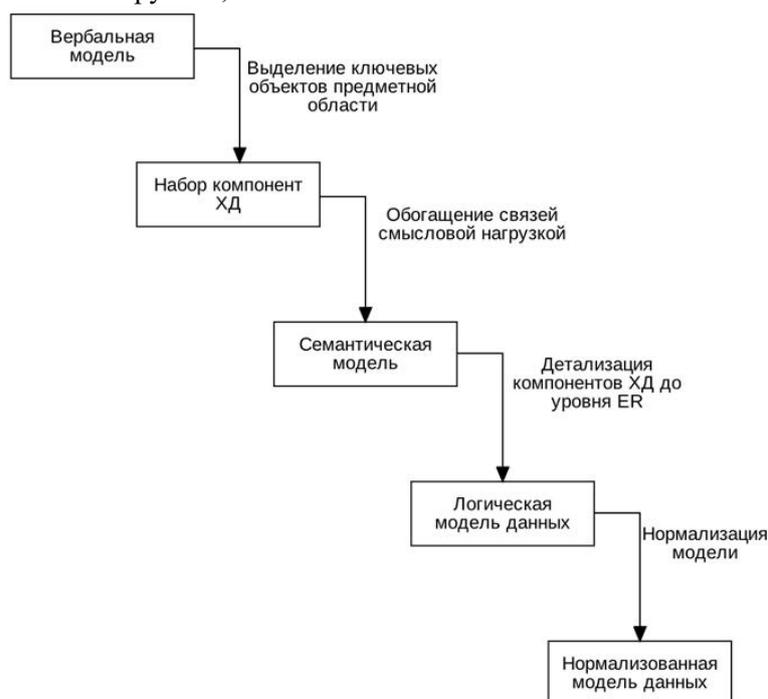


Рис. 1. Процесс проектирования ХД (составлено автором по материалам исследования)

Заключение

Предлагаемый подход предполагает использование разработанной универсальной референтной модели ХД, которая состоит из набора компонент. В данном случае минимизируются те недостатки и риски, которые присущи единовременному внедрению корпоративного ХД. При использовании компонентного подхода каждая организация получает возможность с минимальными усилиями сконфигурировать ХД, базирующееся на лучших практиках, под свои потребности. Это позволяет существенно снизить расходы на внедрение и получить MVP (Minimal Viable Product (продукт, обладающий минимальными, но достаточными для удовлетворения первых потребителей функциями)). Кроме того, ХД, построенное на основе компонентной модели легко масштабируемо, что позволяет наращивать скоуп задач, решаемых с его помощью, при этом распределяя затраты на его доработку во времени. Данное преимущество имеет большое значение с точки зрения бюджетирования и снижению барьеров для старта столь необходимой инициативы как внедрение ХД.

4. Формализация «концептуальной» ER-модели, т.е. логическая модель данных. На этом уровне каждый бизнес-объект (компонент ХД) представляет собой набор сущностей и связей между ними, а каждая связь с уровня семантики трансформируется в отношения между сущностями или таблицы связей (bridge entity);

5. Создание многомерной нормализованной модели данных.

Использование компонентной модели при проектировании ХД позволяет опираться на лучшие практики внедрения, адаптируя их под решение бизнес-задач конкретной организации. Адаптивность достигается за счет возможности скомбинировать компоненты необходимым образом, так каждый из них является целостным и с легкостью интегрируемо с другим. Иными словами, компоненты представляют собой своего рода модули ХД. Таким образом, предложенная концепция позволяет достигнуть максимального бизнес-эффекта при минимальных затратах на внедрение. Минимизация затрат обеспечивается:

- Использованием лучших практик и опыта внедрения ХД, которые легли в основу универсальной компонентной модели;
- Универсальными компонентами ядра ХД, которые требуют минимальной настройки;
- Возможностью простой конфигурации необходимых компонент и отказа от невостребованных в данной организации;
- Простой интеграции компонентов друг с другом.

Литература

1. Ладин Е.Г. Семантическая модель хранилища данных по технологическим характеристикам процессов переработки и кондиционирования радиоактивных отходов низкой и средней активности: дис. ... канд. техн. наук / Московская Государственная Академия тонкой химической технологии им. М.В. Ломоносова. М., 2006. 92 с.
2. Михалькевич И.С. Инструменты и методы анализа слабоструктурированных данных в оптимизации маркетинговых коммуникаций: дис. ... канд. экон. наук / Финуниверситет. М., 2016. 141 с. Автореферат. М., 2018. 25 с. Свободный доступ из сети Интернет (чтение, печать, копирование). URL: http://elib.fa.ru/avtoreferat/mihalkevich_diss.pdf
3. Точилкина Т.Е., Громова А.А. Хранилища данных и средства бизнес-аналитики: учебное пособие. М.: Финансовый университет. 2017. 161 с.
4. Alhyasat E.B., Al-Dalahmeh M. Data Warehouse Success and Strategic Oriented Business Intelligence: A Theoretical Framework // *Journal of Management Research*. 2013. Vol. 5.
5. *Data Modeling Techniques for Data Warehousing* / C. Ballard, D. Herreman, D. Schau, R. Bell at al. Valencic, 1998.
6. *Evolving the Data Warehouse: The Next Generation for Financial Services Institutions* [Электронный ресурс] / An Oracle White Paper. 2011. Режим доступа: <http://www.oracle.com/us/industries/financial-services/evolving-datawarehouse-wp-400257.pdf> (дата обращения: 25.07.2019)
7. Giorgini P., Rizzi S., Garzetti M. GRAN D: A goal-oriented approach to requirement analysis in data warehouses // *Decision Support Systems – Bologna, Italy*, 2008. №45. P. 4 – 21.
8. *IBM Industry Models for Banking. Banking Data Warehouse. General Information Manual* [Электронный ресурс]. IBM. 2012. Режим доступа: ftp://public.dhe.ibm.com/software/data/sw-library/industry-models/brochures/IBM_banking_data_warehouse_GIMv85.pdf (дата обращения: 25.07.2019)
9. Inmon W.H. *Building the Data Warehouse* / Inmon W.H. // Third Edition, John Wiley & Sons Inc. New York, 2002.
10. Kimball R., Ross M. *The Data Warehouse Toolkit. Second Edition. The Complete Guide to Dimensional Modeling* // Wiley Computer Publishing. 2002.

References

1. Ladin E.G. *Semanticheskaya model' hranilishcha dannyh po tekhnologicheskim harakteristikam processov pererabotki i kondicionirovaniya radioaktivnyh othodov nizkoj i srednej aktivnosti: dis. ... kand. tekhn. nauk / Moskovskaya Gosudarstvennaya Akademiya tonkoj himicheskoy tekhnologii im. M.V. Lomonosova. M., 2006. 92 s.*
2. Mihal'kevich I.S. *Instrumenty i metody analiza slabostrukturirovannyh dannyh v optimizacii marketingovyh kommunikacij: dis. ... kand. ekon. nauk / Finuniversitet. M., 2016. 141 s. Avtoreferat. M., 2018. 25 s. Svobodnyj dostup iz seti Internet (chtenie, pechat', kopirovanie). URL: http://elib.fa.ru/avtoreferat/mihalkevich_diss.pdf*
3. Tochilkina T.E., Gromova A.A. *Hranilishcha dannyh i sredstva biznes-analitiki: uchebnoe posobie. M.: Finansovyj universitet. 2017. 161 s.*
4. Alhyasat E.B., Al-Dalahmeh M. *Data Warehouse Success and Strategic Oriented Business Intelligence: A Theoretical Framework // Journal of Management Research. 2013. Vol. 5.*
5. *Data Modeling Techniques for Data Warehousing* / C. Ballard, D. Herreman, D. Schau, R. Bell at al. Valencic, 1998.
6. *Evolving the Data Warehouse: The Next Generation for Financial Services Institutions [Elektronnyj resurs] / An Oracle White Paper. 2011. Rezhim dostupa: <http://www.oracle.com/us/industries/financial-services/evolving-datawarehouse-wp-400257.pdf> (data obrashcheniya: 25.07.2019)*
7. Giorgini P., Rizzi S., Garzetti M. GRAN D: A goal-oriented approach to requirement analysis in data warehouses // *Decision Support Systems – Bologna, Italy*, 2008. №45. P. 4 – 21.
8. *IBM Industry Models for Banking. Banking Data Warehouse. General Information Manual [Elektronnyj resurs]. IBM. 2012. Rezhim dostupa: ftp://public.dhe.ibm.com/software/data/sw-library/industry-models/brochures/IBM_banking_data_warehouse_GIMv85.pdf (data obrashcheniya: 25.07.2019)*
9. Inmon W.H. *Building the Data Warehouse* / Inmon W.H. // Third Edition, John Wiley & Sons Inc. New York, 2002.
10. Kimball R., Ross M. *The Data Warehouse Toolkit. Second Edition. The Complete Guide to Dimensional Modeling* // Wiley Computer Publishing. 2002.

COMPONENT APPROACH APPLICATION IN THE PROBLEM OF DATA WAREHOUSE DESIGN

Solyanov K.S.,

Financial University under the Government of the Russian Federation

Abstract: the role of data warehouses in the information landscape of organizations is discussed in the paper. The subject of the research are data warehouses as a key element of the organization's information-analytical system. The goal is development of the new approach to the data warehouse design, which allows to increase the efficiency of the such systems implementation and the final product quality. Improving the quality of data in the warehouse allows chiefs to make more informed and reasonable management decisions, what finally transforms into one of the competitive advantages. The article considers the existing approaches to building data warehouses, analyzes their weaknesses and provides an assessment. The problem of data warehouse design is considered both from the point of view of the direction of design (from business to data or vice versa), and from the point of view of organizing this process across the organization. The study also proposed the author's concept of a component approach in the problem of data warehouse design. This concept provides a coherent transition from a high-level description of a subject area and business processes to a physical data model and subsequent warehouse implementation. The proposed approach application allows to reduce the time spent on the data warehouse implementation, as well as to improve the quality of the final product and, as a result, the payback of the data warehouse building projects. The research results can be applied in projects for the design and implementation of corporate data warehouses in organizations of various industries. The results presented in the article can be used both by organizations-owners of future data warehouses, and by integrator companies that implement warehouses for their customers.

Keywords: data warehouse; design; component approach; data analysis; information analytical systems; data modelling