УДК 004.623 – 004.652

# КОМПОНЕНТЫ ТИПОВОЙ АРХИТЕКТУРЫ ДЛЯ ХРАНЕНИЯ И ОБРАБОТКИ БОЛЬШИХ МАССИВОВ ДАННЫХ

*Павлович Н.В.*

*Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь*

*Научный руководитель: Тонкович И.Н. – канд.хим.наук, доцент, доцент кафедры ПИКС*

**Аннотация.** В статье рассматриваются основные понятия проектирования и разработки хранилища данных, описываются лучшие практики проектирования архитектур хранения и обработки больших массивов данных.

**Ключевые слова:** *DWH*, слой загрузки данных, *ETL*.

***Введение***. Большие массивы данных требуют специального подхода к их хранению и обработке. Для этого используются различные компоненты типовой архитектуры, которые позволяют эффективно управлять большими объемами данных.

В статье рассматриваются свод требований и рекомендаций по проектированию модели данных и *ETL*-процессов, включая связанные аспекты разработки.

***Основная часть***. При загрузке данных из систем-источников в хранилище данные подвергаются ряду последовательных преобразований. Любую цепочку преобразований можно разбить на несколько этапов, результатом выполнения каждого из которых является обновление постоянной таблицы хранилища данных. Каждому этапу соответствует слой хранилища данных, объединяющий в себя все множество таблиц, обновляемых на соответствующем этапе преобразований.

Экосистема хранилища данных включает в себя следующие слои:

1 Слой загрузки данных. Физическое перемещение данных из систем-источников в экосистему *DWH*. В целях оптимизации возможно выполнение технических преобразований.

2 Слой детальных данных. Интеграция данных из нескольких систем-источников, приведение в соответствие с концептуальной моделью хранилища, построение версионности.

3 Слой *enterprise*-витрин. Вычисление производных аналитических показателей, востребованных широким кругом пользователей хранилища данных.

4 Слой *custom*-витрин. Преобразование данных в соответствие с требованиями отдельного процесса, в том числе: бизнес-анализ, построение отчетности, экспорт данных.

5 Слой справочников и правил трансформаций. Сохранение соответствий исходных ключей и ключей хранилища, исходных значений бизнес-атрибутов и внутрихранилищных, а также правила получения одних атрибутов на основе других.

Каждый слой включает в себя набор компонентов. Каждый компонент в логической модели представлен логической схемой или, в некоторых случаях, набором логических схем. Объект логической схемы, в свою очередь, может быть физически реализован в одной или нескольких различных системах хранения.

Слой детальных данных, слой *enterprise*-витрин, слой справочников и слой *custom*-витрин содержат презентационные компоненты. Они представляют собой совокупность сущностей, их свойств, характеристик, классификаций и связей между ними, однозначно описывающих бизнес-процессы.

Слой загрузки данных является одним из основных слоев стандарта проектирования хранилища данных. Он предназначен для обеспечения эффективной интеграции данных из различных источников в хранилище данных. В этом слое могут быть определены следующие компоненты [1]:

1 *Operational Data Storage* (*ODS*). Копии таблиц-источников, загружаемые с помощью систем репликации. Поставка данных через *ODS* простой и удобный способ для систем-

источников, имеет ряд ограничений: размер таблицы, СУБД системы-источника, определенные ограничения на типы и методы заполнения полей на стороне источника.

2 *Staging* (*STG*). Копии таблиц-источников, не подлежащих репликации. Компонент *STG* содержит копии таблицы-источника, обновляемых с помощью стандартного *ETL*-процесса. Таблица создается в *STG* в тех случаях, когда по каким-либо причинам невозможна репликация таблицы-источника в *ODS*. Таким образом, *STG* является второй по приоритету входной точкой данных в хранилище данных. Структура таблицы *STG* должна совпадать со структурой таблицы-источника. Данные загружаются один в один без преобразований. Но, бывают исключения - в случае, когда точно известно, какие поля будут использоваться в *ETL* и *BI* процессах, в целях оптимизации допустимо сохранять не всю информацию, а только нужную.

3 *NiFi*. Накопление данных, получаемых из систем-источников с помощью *streaming-ETL* средства *NiFi*.

Слой детальных данных – это один из ключевых компонентов стандартного подхода к проектированию хранилища данных, который содержит детализированные данные, полученные из различных источников данных.

Этот слой содержит наиболее подробную информацию о бизнес-процессах и операциях, которые происходят в организации. Слой детальных данных обеспечивает сохранение данных в исходном виде без каких-либо преобразований и агрегации. Это позволяет аналитикам и бизнес-пользователям получить максимально точную и полную информацию из хранилища данных.

Слой детальных данных является центральным элементом в хранилище данных, так как он предоставляет основу для анализа бизнес-процессов и принятия управленческих решений. Другие слои хранилища данных, такие как слой *enterprise*-витрин, используют данные из слоя детальных данных для создания сводных таблиц и представлений данных на разных уровнях агрегации.

Слой состоит из двух компонентов:

1 *Data Vault* (*DV*). Декомпозиция объектов *DDS* схемы для оптимизации загрузки.

*Data Vault* состоит из трех основных компонентов – *Hub*, *Link*, *Satellite*.

Таблица типа *Hub* является составляющей частью декомпозиции сущности *DDS* типа измерение или факт. Общее предназначение *Hub* – сохранение множества ключей декомпозируемой сущности *DDS* и некоторых особых ее неизменяемых атрибутов.

Таблица типа *Link* представляет из себя аналог *Hub* для сущности типа связь. Она также является центральной составляющей структуры, создаваемой в *DV* для связи.

Таблицы типа *Satellite* предназначены для сохранения атрибутного состава декомпозируемых сущностей. Каждый *Satellite* может быть привязан к одному и только одному *Hub*. У каждого *Hub* должно быть не меньше одного привязанного *Satellite*. Каждый *Satellite* содержит один или несколько бизнес-атрибутов декомпозируемой сущности. Контекст из разных систем-источников принято размещать в отдельные сателлиты [2].

2 *Detail Data Store* (*DDS*). Базовый компонент *DWH*, предназначенный для хранения детальной информации о бизнес-процессах.

Компонент *DDS* содержит сущности 3 видов: измерение – объект, содержащий историю изменений характеристик бизнес-сущности.

В таблицах типа измерение хранится атрибутный состав бизнес-сущности. Любая бизнес-сущность, участвующая в процессах организации, должна быть отражена в таблице-измерении *DDS*. Отличительной особенностью измерения является то, что любой её экземпляр с момента создания существует определенное время, в течение которого значения его атрибутов могут меняться.

1 Факт – объект, содержащий характеристики определенного события.

Таблица типа факт содержит информацию о событиях, происходящих одномоментно или в течение короткого интервала времени. Конечное состояние факта представляет инте-

рес для потребителей данных. Это является отличительной особенность сущности этого типа.

2 Связь – объект, характеризующий связь двух или более измерений, или фактов в рамках бизнес-процессов.

Таблица типа связь содержит информацию о нахождении двух или более сущностей в прямой взаимосвязи друг с другом в рамках какого-либо бизнес-процесса, а также набор бизнес-атрибутов, описывающих контекст этой взаимосвязи. Каждая таблица-связь определяется парой связываемых сущностей, количество сохраняемых типов связей при этом не ограничено. Типы связываемых сущностей также не ограничены – ими могут быть и факты, и измерения.

Слой *enterprise*-витрин предназначен для сборки денормализованных аналитических витрин, востребованных большим количеством пользователей хранилища данных из разных подразделений.

Основная цель *enterprise*-витрины – предоставить бизнес-пользователям единую точку доступа к централизованным данным. Это позволяет пользователям быстро и легко получать доступ к необходимой информации, а также обеспечивает надежную, однородную и консистентную информацию.

В проектировании хранилища данных, слой *enterprise*-витрины является критически важным, так как он предоставляет централизованный и однородный доступ к данным для всей организации. Он помогает снизить риск ошибок и улучшить качество данных, что может привести к улучшению производительности и принятию более обоснованных решений.

Слой *custom*-витрин в проектировании хранилища данных – это слой, который содержит предопределенные представления данных для определенных пользовательских потребностей. Витрины данного слоя предназначены для решения локальных аналитических или интеграционных задач. Такие витрины создаются на основе существующих данных в хранилище данных, и обычно настраиваются в соответствии с требованиями конкретных пользователей или групп пользователей [3].

Эти пользовательские витрины могут включать в себя:

1 Представления для конкретных бизнес-пользователей, которые требуют определенных аналитических данных для своих ежедневных задач.

2 Визуализации данных для менеджеров и руководителей, которые могут требовать быстрого и простого доступа к ключевым метрикам и показателям.

3 Отчеты для финансовых аналитиков, которые должны получать финансовые данные в формате, удобном для анализа и представления.

Пользовательские витрины обычно создаются с использованием существующих данных, но могут также включать и вычисляемые данные, которые не хранятся в хранилище данных, но могут быть рассчитаны на основе существующих данных.

Создание пользовательских витрин помогает упростить процесс анализа данных для конечных пользователей, увеличивает их удовлетворенность и повышает производительность бизнеса. Кроме того, пользовательские витрины могут уменьшить количество времени, затрачиваемого на создание отчетов и анализ данных, что увеличивает эффективность работы сотрудников.

Слой справочников – это один из важных слоев при проектировании хранилищ данных. Этот слой содержит справочную информацию, которая используется для описания данных в других слоях хранилища данных.

Справочники содержат статическую информацию, которая остается постоянной на протяжении времени, в отличие от оперативных данных, которые могут изменяться с течением времени. Справочники содержат информацию о сущностях, таких как клиенты, продукты, поставщики и другие объекты, с которыми взаимодействует бизнес-система.

Справочники могут использоваться для описания и хранения метаданных, которые описывают структуру данных в хранилище. Эти метаданные могут включать в себя описание таблиц, столбцов, типов данных и связей между таблицами.

Один из примеров использования справочников – это нормализация данных, что означает разделение данных на отдельные таблицы и связывание их между собой. Это уменьшает дублирование данных и повышает качество данных в хранилище.

Слой справочников может содержать справочники-словари, которые могут использоваться для стандартизации данных. Например, справочник-словарь может содержать список стран и их кодов, чтобы обеспечить единообразное заполнение этого поля в таблицах, где используется этот справочник.

Слой правил трансформации – это слой в хранилище данных, который используется для преобразования данных из исходных систем в формат, необходимый для анализа и принятия решений. Этот слой содержит правила трансформации, которые определяют, как данные будут преобразовываться, чтобы соответствовать требованиям аналитики.

Правила трансформации включают в себя: фильтрация данных для исключения ненужной информации, создание производных данных, таких как вычисляемые поля и агрегированные данные, стандартизация данных, чтобы обеспечить единообразие в формате и содержании данных, обработка ошибок и исключений, которые могут возникнуть при трансформации данных.

Слой правил трансформации важен для обеспечения качества данных в хранилище. Этот слой позволяет стандартизировать данные, устранить дублирование и ошибки, а также обеспечить соответствие данных требованиям аналитики. Благодаря этому аналитики могут получать точные и надежные данные для анализа и принятия решений.

Слой правил трансформации является важным компонентом в проектировании хранилища данных, так как он обеспечивает качество данных, необходимых для анализа и принятия решений, а также упрощает обработку данных и обновление хранилища.

***Заключение***. Выполнен анализ требований, предъявляемых к стандарту проектирования хранилища данных, определены основные понятия, которыми необходимо оперировать при формировании архитектуры обработки и хранения данных. Выявлены основные методы проектирования и разработки хранилищ данных.

Практическое применение рассмотренных теоретических аспектов позволит в значительной степени повысить степень удовлетворенности потребителей данных. Что, в свою очередь, сделает хранилище данных гибким и легко расширяемым.

### *Список литературы*

1. *DWH [Электронный ресурс]. – Режим доступа: https://www.xelent.ru/blog/chto-takoe-dwh/ – Дата доступа: 10.02.2023.*

2. *Павлович, Н.В. Data vault: преимущества и недостатки / Павлович Н.В. // Новые информационные технологии в научных исследованиях: материалы XXVI Всероссийской научно-технической конференции студентов, молодых ученых и специалистов, Рязань, 2021 г. / Рязанский государственный радиотехнический университет имени В.Ф. Уткина. – Рязань, 2021. – С. 35–36.*

3. *Data Warehouse [Электронный ресурс]. – Режим доступа: https://www.itweek.ru/infrastructure/article/detail.php?ID=48156 – Дата доступа: 11.02.2023.*

UDC 004.623 – 004.652

## COMPONENTS OF A TYPICAL ARCHITECTURE FOR STORING AND PROCESSING LARGE ARRAYS OF DATA

*Paulovich N.V.*

*Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus*

*Tonkovich I.N. – PhD, associate professor, associate professor of the Department of ICSD*

**Annotation**. The article discusses the basic concepts of designing and developing a data warehouse, describes the best practices for designing architectures for storing and processing large data arrays.

**Keywords**: *DWH, data warehouse, ELT.*