DATA VAULT: ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

Павлович Н.В., магистрант, e-mail: mpaulovich00@gmail.com

2021

Белорусский государственный университет информатики и радиоэлектроники

Ключевые слова: Data Vault, Data Warehouse, Detail Data Store, Hub, Link, Satellite, Link-Satellite.

Аннотация: В статье рассматривается компонент слоя детальных данных Data Vault, назначение, правила проектирования, преимущества и недостатки данного подхода. Описываются составляющие части компонента, такие как: Hub, Link, Satellite, Link-Satellite, а также внутренние потоки в слое детальных данных.

Сегодня большинство компаний накапливают различные данные, полученные в процессе работы. Часто данные приходят из различных источников – структурированных и не очень, иногда в режиме реального времени, а иногда они доступны в строго определенные периоды. Все это разнообразие нужно структурированно хранить, чтобы потом успешно анализировать, создавать требуемые отчеты и вовремя замечать аномалии. Для этих целей проектируется Data Warehouse.

Существует несколько подходов к построению такого универсального хранилища, которые помогают архитектору избежать распространенных проблем, а самое главное обеспечить должный уровень гибкости и расширяемости DWH. Одним из таких подходов является Data Vault [1].

Data Vault – гибридный подход для оптимизации загрузки объектов Detail Data Store.

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 (атрибуты типизации или интеграционные ключи). В качестве бизнес-ключа Hub всегда наследует RK декомпозируемого измерения или факта, включая название. Hub также содержит мета-поля processed_dttm и source_system_cd, в которых хранятся время загрузки сущности в DV и ее источник (название системы, базы или файла, откуда данные были загружены). Таблица типа Hub по определению не исторична.

Таблица типа Link представляет из себя аналог Hub для сущности типа связь. Она также является центральной составляющей структуры, создаваемой в DV для связи. Требование о взаимно-однозначном соответствии таблиц типа Link и сущностей типа связь также актуально: Link создается всегда при декомпозиции связи и при том в единственном числе. В Link сохраняется множество наборов ключей связываемых сущностей. Аналогично Hub, Link наследует набор RK сущностей, входящих в состав декомпозируемой связи. Иногда в состав бизнес-ключа Link дополнительно входит атрибут тип связи. В

таблице типа Link допускается ведение истории по необходимости. Например, для ведения истории изменения флага удаления 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. Satellite наследует RK декомпозируемого измерения или факта аналогично Hub. Таблица типа Satellite может быть исторична, в случае, если необходимо отслеживать изменение атрибутов, либо не исторична, в случае отсутствия такой необходимости.

Таблицы типа Link-Satellite предназначены для сохранения атрибутного состава сущностей типа связь. Каждый Link-Satellite может быть привязан к одному и только одному Link. Link-Satellite создается в случае невозможности отслеживания нужных атрибутов в Link. Бизнес-ключ наследуется аналогично Link [2].

Схема внутренних потоков слоя детальных данных представлена на рисунке 1.

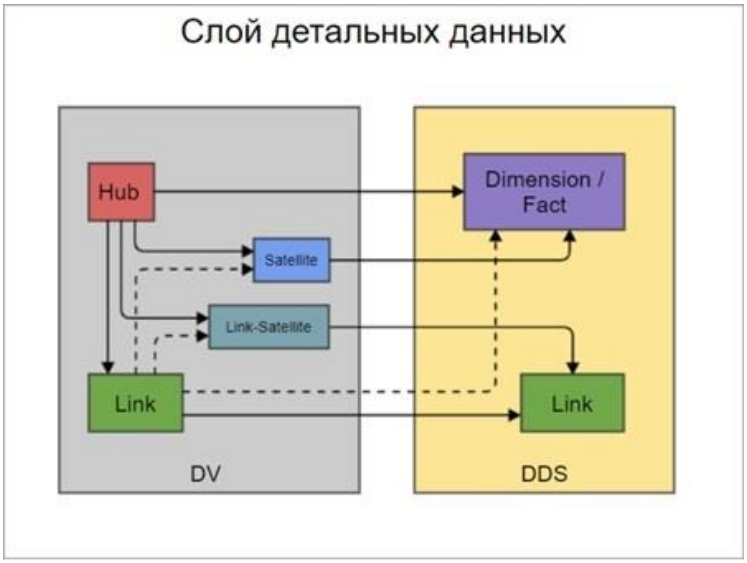

Рисунок 1 – Схема внутренних потоков слоя детальных данных

Внутренние потоки в слое детальных данных возникают лишь в тех случаях, когда в целях оптимизации загрузки сущности создается структура Data Vault. При этом основными являются потоки от объектов DV к объектам DDS соответствующего типа. Совокупность этих потоков является по сути сборкой сущности DDS из элементов соответствующей структуры DV.

В результате оптимизации процесса загрузки сущности в слое детальных данных возникают внутренние потоки. При этом допустимы как потоки от объектов DV к объектам DDS (непосредственно сборка сущности DDS из элементов ее декомпозиции), так и некоторые внутренние потоки компонента DV.

В заключении стоит отметить, что подход Data Vault сам по себе достаточно сложный. Из-за большого количества таблиц в последующем возникает большое количество операций join. Запросы могут быть медленнее, чем в традиционных DWH, где таблицы денормализованы.

Список использованных источников

1. Linstedt, D. Building a Scalable Data Warehouse with Data Vault 2.0 / M. Kaufmann. – Elsevier, 2015. – 684 p.

2. Data Vault 2.0 Modeling Basics [Электронный ресурс]. – Режим доступа: https://www.vertabelo.com/blog/data-vault-series-data-vault-2-0-modeling-basics/. – Дата доступа: 15.10.2021.