DATA VAULT КАК СПОСОБ ОРГАНИЗАЦИИ ХРАНИЛИЩ ДАННЫХ

А.В. Свирновский
Научный руководитель – Алексеев В.Ф.
канд. техн. наук, доцент
Белорусский государственный университет информатики и радиоэлектроники

Для большинства компаний важным является накопление различных данных которые получены в процессе работы. Часто данные приходят из различных источников – структурированные 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, готовить отчеты, вовремя замечать какие-то аномалии [1, 2]. Для этих целей проектируется хранилище данных.

Существует несколько подходов к построению такого универсального хранилища, которые помогают архитектору избежать распространенных проблем, а самое главное обеспечить должный уровень гибкости и расширяемости. Автором предлагается один из таких подходов.

Data Vault – гибридный подход, объединивший достоинства схемы «звезды» и третьей нормальной формы.

Data Vault состоит из трех основных компонентов – Хаб, Ссылка и Сателлит.

Хаб – основное представление сущности (Клиент, Продукт, Заказ) с позиции бизнеса. Таблица-Хаб содержит одно или несколько полей, отражающих сущность в понятиях бизнеса. В совокупности эти поля называются «бизнес ключ». Идеальный кандидат на звание бизнес-ключа это ИНН организации или VIN номер автомобиля, а сгенерированный системой ID будет наихудшим вариантом. Бизнес ключ всегда должен быть уникальным и неизменным.

Хаб так же содержит мета-поля load timestamp и record source, в которых хранятся время первоначальной загрузки сущности в хранилище и ее источник (название системы, базы или файла, откуда данные были загружены). В качестве первичного ключа Хаба рекомендуется использовать MD5 или SHA-1 хеш от бизнес ключа. Пример таблиц-Хабов представлен на рисунке 1.

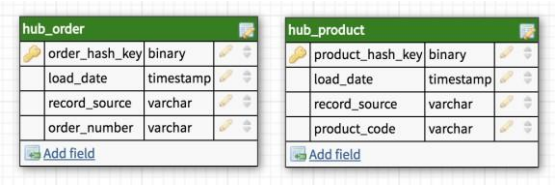

Рисунок 1 – Таблицы-Хабы

Таблицы-Ссылки связывают несколько хабов связью многие-ко-многим. Она содержит те же метаданные, что и Хаб. Ссылка может быть связана с другой Ссылкой, но такой подход создает проблемы при загрузке, так что лучше выделить одну из Ссылок в отдельный Хаб. Пример таблицы-Ссылки представлен на рисунке 2.

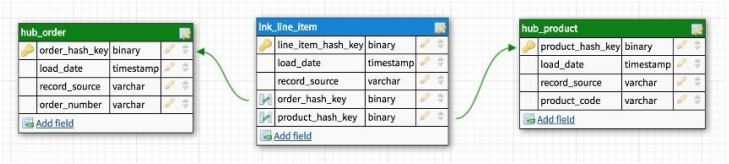

Рисунок 2 – Таблица-Ссылка

Все описательные атрибуты Хаба или Ссылки (контекст) помещаются в таблицы-Сателлиты. Помимо контекста Сателлит содержит стандартный набор метаданных (load timestamp и record source) и один и только один ключ «родителя». В Сателлитах можно без проблем хранить историю изменения контекста, каждый раз добавляя новую запись при обновлении контекста в системе-источнике. Для упрощения процесса обновления большого сателлита в таблицу можно добавить поле hash diff: MD5 или SHA-1 хеш от всех его описательных атрибутов. Пример таблиц-сателлит представлен на рисунке 3.

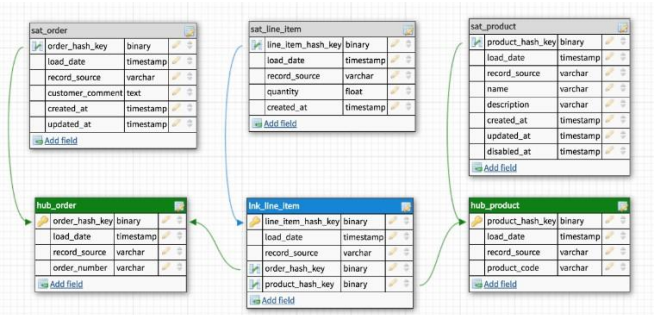

Рисунок 3 – Таблицы-Сателлиты

Для Хаба или Ссылки может быть сколь угодно Сателлитов, обычно контекст разбивается по частоте обновления. Контекст из разных систем-источников принято класть в отдельные Сателлиты.

Библиографический список

1. Bluesoft [Электронный ресурс]: база данных. – Режим доступа: https://bluesoft.com/en/data-vault-architecture-2/. – Дата доступа: 10.10.2019.

2. Inmon W.H., Linstedt D. Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault – Elsevier, 2015. — 342 p.