Обзор методологии и архитектуры Data Vault
Злобина А. В.

Злобина Александра Владимировна / Zlobina Alexandra Vladimirovna - младший научный сотрудник,
кафедра радиоэлектроники информационных систем,
Уральский Федеральный университет имени первого Президента России Б. Н. Ельцина,
г. Екатеринбург

Аннотация: *в статье выполнен обзор методологии Data Vault, которая используется для моделирования бизнес-процессов. Демонстрируются отличительные качества рассматриваемой методологии и основные компоненты: Hub, Link, Satellite, а также их атрибуты.*
Ключевые слова: *Data Vault, hub, link, satellite, бизнес-процесс, data warehouse, big data.*

С каждым днем, несомненно, растет объем информации, а применение таких приложений, как ERP, CRM, SCM, EAM, ECM приводит к увеличению объема данных в хранилищах, что влечет сложность с масштабируемостью, гибкостью представления и степенью детализации данных.

Одной из методик моделирования данных для корпоративных хранилищ данных является Data Vault, которая была спроектирована Даном Линстедтом (Dan Linstedt).

Data Vault - набор уникально связанных нормализованных таблиц, содержащих детальные данные, отслеживающих историю изменений и предназначенных для поддержки одной или нескольких функциональных областей бизнеса.

Это - гибридный подход, обобщающий лучшие свойства третьей нормальной формы (3NF) и схемы Звезда (Star schema). Дизайн Data Vault – гибок, масштабируем, последователен и приспосабливаем к потребностям предприятия. Архитектура Data Vault предназначена для удовлетворения потребностей хранилищ данных (data warehouse) [1].

В методологии Data Vault применяются три компонента, это: Hub, Link и Satellite, что в свою очередь позволяет сохранить структуру хранилища данных простым и в тоже время изящным.

Сущность Hub позволяет отобразить функциональную область предметной области. Таблицы типа Hub содержат определенный набор бизнес-ключей. Бизнес-ключ – это уникальный идентификатор, который бизнес использует в своих повседневных операциях [2]. Примером может служить номер сотрудника, номер клиента, номер договора и т.д. При утере данного ключа теряется вся информация об объекте. Атрибутами Hub являются:

– Суррогатный ключ (Surrogate Key) – является необязательным компонентом (смарт-ключ или порядковый номер).

– Временная отметка даты загрузки (Load Date Time Stamp) – регистрация первоначальной загрузки ключа в хранилище.

– Источник данных (Record Source) – регистрация исходной системы применяется для обратной трассировки (отслеживания) данных.

Сущность Link представляет связь (транзакцию) типа «многие ко многим» третьей нормальной формы (3NF). Данный компонент содержит такие атрибуты, как:

– Суррогатный ключ (Surrogate Key) – является необязательным компонентом (смарт-ключ или порядковый номер). Используется при условии, что существует более двух Hub-таблиц, связанных сущностью Link, или если составной первичный ключ может привести к проблемам производительности.

– Ключи Hub-таблиц (от 1-го до N-го) – переносятся в сущность Link, образуя составной ключ, и отображают взаимодействия и связи между Hub-таблицами.

– Временная отметка даты загрузки (Load Date Time Stamp) – регистрация первоначальной загрузки ключа в хранилище.

– Источник данных (Record Source) – регистрация исходной системы, применяется для обратной трассировки (отслеживания) данных.

При наличии нескольких Hub- и Link-таблиц модель уже может описывать бизнес-процессы.

Необходимо заметить тот факт, что данная техника моделирования разработана для хранилищ данных, а не для OLTP систем.

Сущность Satellite включает в себя описательную информацию ключа Hub. В течение времени информация будет изменяться, поэтому структура данной сущности должна позволять хранить как измененную так и новую информацию. В таблице Satellite имеются следующие обязательные атрибуты:

– Первичный ключ Satellite-таблицы, который представляет собой первичный ключ Hub- или Link-таблицы (переносится в Satellite из Hub или Link);

– Временная отметка даты загрузки (Load Date Time Stamp) – регистрация первоначальной загрузки ключа в хранилище.

– Последовательный (sequence) суррогатный номер – используется для Satellite, которые имеют несколько значений (например, адрес выставление счетов и домашний адрес). Является необязательным полем.

– Источник данных (Record Source) – регистрация исходной системы, применяется для обратной трассировки (отслеживания) данных.

Таблица Satellite наиболее близка медленноменяющимся измерениям формата SCD II в определении Ральфа Кимбалла. Она хранит изменения на детальном уровне, а ее функция заключается в описании контекста экземпляров Hub и Link.

Проектирование Satellite-таблиц должно основываться на математических принципах сокращения избыточности данных и на скорости изменения данных. Таким образом, Satellite-таблицам отводится роль описания бизнес-ключа на наиболее доступном детальном уровне. Это обеспечивает основу для развития контекста, описывающего бизнес. На основе данных простых компонентов можно построить как простое хранилище, состоящее из одной пары Hub- и Satellite-таблиц, так и огромное корпоративное хранилище данных, содержащее сотни Hub [2].

Наличие огромного объема данных приводит к проблемам с запросами, это относится к схеме «Звезда», но не относится к 3NF. При больших объемах информации нарушается продуктивность выполнения запросов в согласованных измерениях и таблицах фактов. Для решения данной проблемы зачастую необходимо делать секционирование (разделение), либо менять структуру хранилища данных, для предоставления дополнительной детализации пользователям. Перезагрузка постоянно меняющихся Звезд является трудной задачей (не говоря уже о попытках выполнить это с большим объемом данных, например, свыше 1 Тб) [2].

Модель Data Vault уходит корнями в математические основы нормализованной модели данных и потому лишена этих недостатков. Сокращение избыточности данных и учет темпов изменения наборов данных способствует повышению производительности и простоте в управлении. Архитектура Data Vault не ограничивается применением какой-либо одной платформы.

Возможности масштабирования модели Data Vault можно продемонстрировать на следующем примере. Предположим, компания, торгующая компьютерами, имеет хранилище данных, состоящее из Hub-таблицы «Компьютеры», Hub-таблицы «Счета» и Link-таблицы связей между ними. Затем компания решает продавать автомобили. Модель данных Data Vault позволяет ввести новый Hub «Автомобили» и создать новый Link «Автомобили-Счета». В результате никакие данные не потеряны, вся информация, накопленная в течение долгого времени, сохранена и отражает изменения бизнеса. Это лишь одна из многочисленных возможностей для обработки подобной ситуации.

Реализация данного примера в нотации Data Vault приведена на рисунке 1.

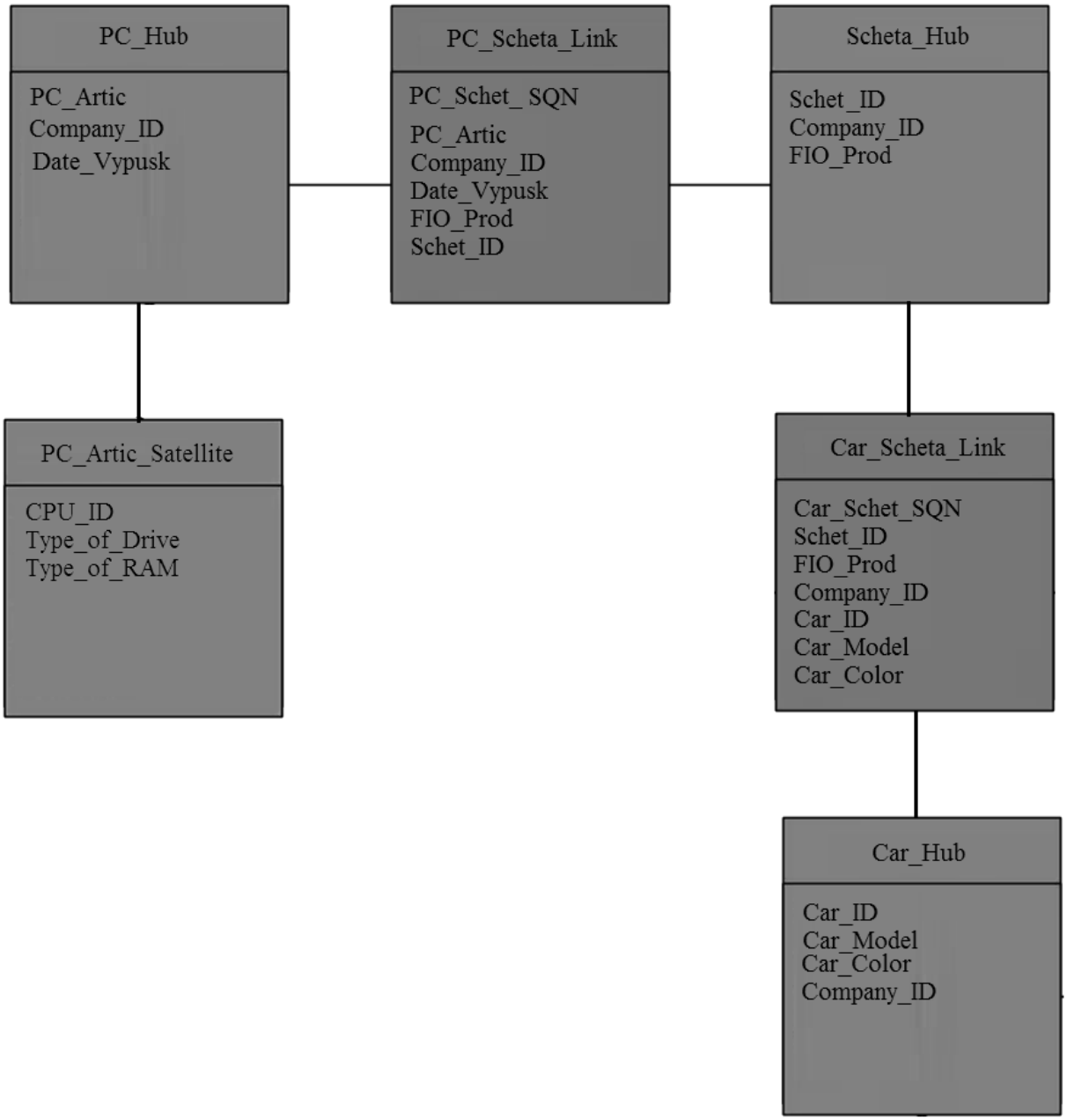

Рис. 1. Реализация примера в нотации Data Vault

Одним из главных преимуществ метода Data Vault является динамическое представление взаимосвязей предметной области хранилища данных. Взаимосвязи определяются через бизнес-ключи Hub и фиксируются в сущностях Link. Они существуют во времени, и их история сохраняется в Satellite. Это позволяет отображать динамику развития взаимосвязи [3].

Метод Data Vault целесообразно использовать в следующих случаях:

– для создания динамических хранилищ данных (Dynamic Data Warehousing), когда возникает необходимость учитывать динамику изменения как обработки данных, так и структур данных;

– для создания Data mining/Exploration Warehousing, когда пользователям нужно менять структуру данных без потери информации;

– при встраивании процедур DM в хранилище данных [3].

Следует отметить, что объем и структура хранилища, построенного по технологии Data Vault, всегда может быть изменена или дополнена с минимальными трудозатратами.

Гибкость архитектуры, заложенная в модель Data Vault, позволяет создавать хранилище данных итерационно, без существенных изменений созданной структуры. Практическое применение модели Data Vault в различных бизнес-областях успешно это доказывает.

Литература

1. [Электронный ресурс]: Статьи о хранилищах данных. URL: http://www.dwh-club.com/ru/dwh-bi-articles/data-warehouse-business-intelligence-podborka-statei.html/ (дата обращения: 18.08.2016).
2. [Электронный ресурс]: Современные подходы к архитектуре хранилищ данных. Модель Data Vault. URL: http://www.remmag.ru/admin/upload_data/remmag/10-3/ /Lanit.pdf/ (дата обращения: 21.08.2016).
3. [Электронный ресурс]: ИНТУИТ. Национальный открытый университет. Лекция 13: Метод моделирования.
4. «Свод данных». [Электронный ресурс]. URL: http://www.intuit.ru/studies/courses/599/455/lecture/10179/ (дата обращения: 25.08.2016).