А.В. Свирновский,
магистрант 1 курса,
e-mail: ***vladimirovich1998@gmail.com,***
науч. рук.: ***В.Ф. Алексеев,***
к.т.н., доц.,
БГУИР,
г. Минск, Беларусь

ПРОЦЕСС ЗАГРУЗКИ В ХРАНИЛИЩЕ ДАННЫХ, ОРГАНИЗОВАННОЕ ПО ТИПУ DATA VAULT

Аннотация: в данной статье рассматривается организация процесса загрузки в хранилище данных, которое организовано по типу Data Vault.

Ключевые слова: хранилище данных, data vault, процесс загрузки

Сначала данные из операционных систем поступают в staging area. Staging area используется как промежуточное звено в процессе загрузки данных. Одна из основных функций Staging зоны – это уменьшение нагрузки на операционные базы при выполнении запросов. Таблицы здесь полностью повторяют исходную структуру, но любые ограничения на вставку данных, вроде not null или проверки целостности внешних ключей, должны быть выключены с целью оставить возможность вставить даже поврежденные или неполные данные (особенно это актуально для excel-таблиц и прочих файлов). Дополнительно в stage таблицах содержатся хеши бизнес ключей и информация о времени загрузки и источнике данных. Полный процесс поступления данных представлен на рисунке 1.

Business Vault – опциональная вспомогательная надстройка над Raw Data Vault. Строится по тем же принципам, но содержит переработанные данные: агрегированные результаты, сконвертированные валюты и прочее. Разделение чисто логическое, физически Business Vault находится в одной базе с Raw Data Vault и предназначен в основном для упрощения формирования витрин. Пример организации Бизнес-Сателлита представлен на рисунке 2.

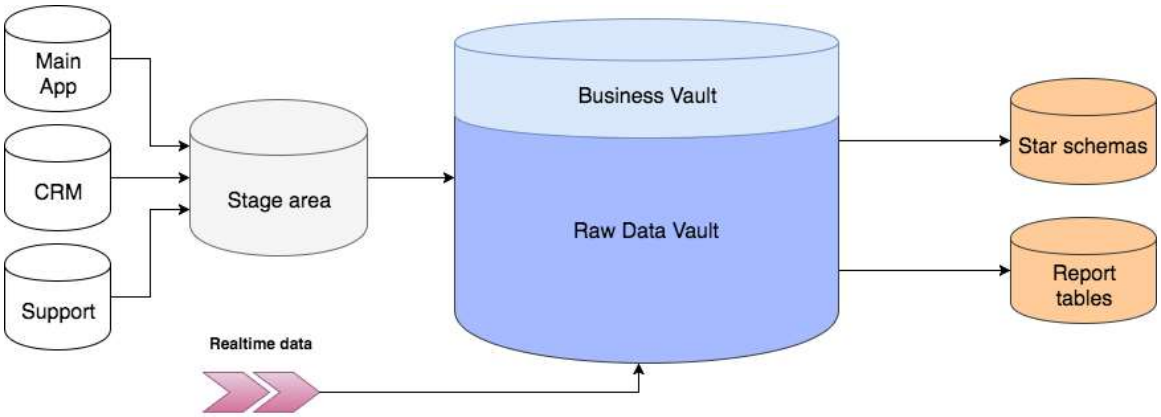

Рисунок 1 – Процесс поступления данных

Когда нужные таблицы созданы и заполнены, наступает очередь витрин данных. Каждая витрина – это отдельная база данных или схема, предназначенная для решения задач различных пользователей или отделов. В ней может быть специально собранная «звезда» или коллекция денормализованных таблиц. Если возможно, таблицы внутри витрин лучше делать виртуальными, то есть вычисляемыми «на лету». Для этого обычно используются SQL представления.

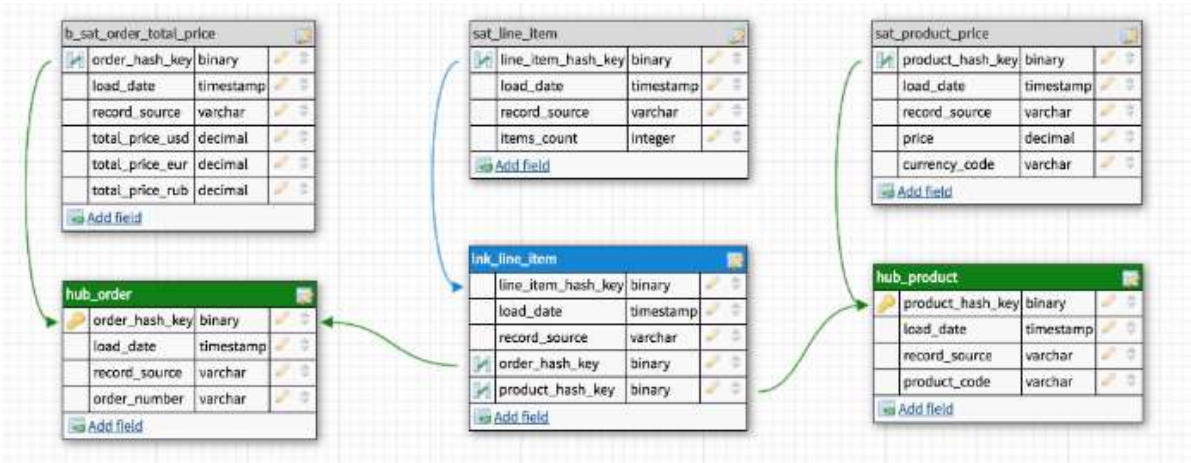

Рисунок 2 – Бизнес-Сателлит

Процесс заполнения данных следующий: сначала загружаются Хабы, потом Ссылки и затем Сателлиты. Хабы можно загружать параллельно, так же, как и Сателлиты и Ссылки, если, конечно, не используется связь link-to-link.

Преимущества Data Vault:

– гибкость и расширяемость: с Data Vault перестает быть проблемой как расширение структуры хранилища, так и добавление, и сопоставление данных из новых источников. Максимально полное хранилище «сырых» данных и удобная структура их хранения позволяют нам сформировать витрину под любые требования бизнеса, а существующие решения на рынке СУБД хорошо справляются с огромными объемами информации и быстро выполняют даже очень сложные запросы, что дает возможность виртуализировать большинство витрин;

– agile-подход из коробки: моделировать хранилище по методологии Data Vault довольно просто. Новые данные просто «подключаются» к существующей модели, не ломая и не модифицируя существующую структуру. При этом мы будем решать поставленную задачу максимально изолированно, загружая только необходимый минимум, и, вероятно, наша временная оценка для такой задачи станет точнее.

Недостатки Data Vault:

– обилие присоединений: за счет большого количества операций присоединений запросы могут быть медленнее, чем в традиционных хранилищах данных, где таблицы денормализованы;

– сложность: малое количество информации в интернете и почти полное отсутствие материалов на русском языке. Как следствие, при внедрении Data Vault возникают проблемы с обучением команды, появляется много вопросов относительно нюансов конкретного бизнеса. К счастью, существуют ресурсы, на которых можно задать эти вопросы. Большой недостаток сложности – это обязательное требование к наличию витрин данных, так как сам по себе Data Vault плохо подходит для прямых запросов;

– избыточность: в 3 раза больше ETL-процессов.

Архитектура Data Vault построена для решения конкретных задач и не является единственным возможным решением всех проблем.

Литература и примечания:

[1] Linstedt, D. Building a Scalable Data Warehouse with Data

Vault 2.0 / D. Linstedt. – USA: Morgan Kaufmann, 2015. – 684 p.

[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.