УДК 004.01, 004.622, 004.623, 004.627

ОБЗОР МЕТОДИК ОБРАБОТКИ И ХРАНЕНИЯ БОЛЬШИХ ДАННЫХ

Самсонов Алексей Дмитриевич

Аспирант;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: aleksey689@mail.ru

В данной статье был произведён обзор организации работ больших данных. Как показывает деятельность человека, большие данные используются для проведения аналитических расчётов. Полученные расчёты нужны для принятия важных управленческих или иных решений. Поэтому очень важно, чтобы эти большие данные были организованы эффективным образом. От грамотно поставленных задач организации больших данных, их обработки и хранения, зависит эффективность не только самой информационной системы, но и бизнес-процессов в целом.В статье показаны этапы развития идеи организации больших данных. Рассмотрены этапы развития идеи их организации, причины их возникновения, проблемы методик, а также сделаны предложения по развитию методик их обработки и хранения.

Ключевые слова: большие данные, хранилище данных, нормальные формы, витрины данных, интеллектуальный анализ данных.

Для цитирования:

Самсонов А. Д. Обзор методик обработки и хранения больших данных // Системный анализ в науке и образовании: сетевое научное издание. 2022. №4. С. 15-21. URL: https://sanse.uni-dubna.ru/index.php/sanse/article/view/557.

OVERVIEW OF METHODS FOR HANDLING AND STORING BIG DATA

Samsonov Alexey D.

PhD student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: aleksey689@mail.ru.

In this article, an overview of the organization of big data work was made. As human activity shows, big data is used to carry out analytical calculations. The resulting calculations are needed to make important management or other decisions. Therefore, it is very important that this big data is organized in an efficient way. The efficiency of not only the information system itself, but also business processes in general, depends on well-set tasks for organizing big data, handling and storing them. The article shows the stages of development of the idea of development of the idea of organizing big data. The stages of development of the idea of their organization, the reasons for their occurrence, the problems of methods are considered, as well as proposals are made for the development of methods for their handling and storage.

Keywords: big data, data warehouse, normal forms, data marts, data mining.

For citation:

Samsonov A. D. Overview of methods for handling and storing Big Data. *System analysis in science and education*, 2022;(4):15-21(in Russ). Available from: https://sanse.ru/index.php/sanse/article/view/557.

Введение

Работа с большими данными – один из важнейших аспектов в развитии предприятий и общества в целом. Без них не были б возможны выявления закономерностей и принципов, теорий и их доказательства. Работы с большими данными помогают науке, людям в их деятельности: от математиков до инженеров, от социологов до метеорологов и так далее. Поэтому важно иметь ресурсы, благодаря

которым можно обработать эти данные и в дальнейшем их выводить в определённом формате, нужный для эксперта по его задачам.

В данной работе будут рассмотрены существующие способы обработки больших данных, а также анализ того, почему тот или иной способ появился и почему его активно используют в бизнесе.

Большие данные

Термин «большие данные» используется в различных сферах по-разному. В зависимости от аспектов, предметных областей и задач понимается так или иначе данный термин. В Оксфордском словаре *OED* 2013 г. было зафиксировано такое определение при переводе на русский язык: «Данные очень большого размера, как правило, в том смысле, что представляют серьезные трудности в материально-техническом обеспечении по манипуляциям и управлению ими; (также) направление вычислений с использованием такого типа данных». Впоследствии от технического обеспечения был произведён переход в сторону прикладного понимания: выяснение закономерностей, тенденции и взаимосвязей [1].

Однако для выявления таких закономерностей нужно организовать большие данные приемлемым образом. Поэтому в статье затронем вопрос построения таких данных и последующий процесс их обработки. Для них зададим первоначальные условия:

- Большие данные организованы в виде таблиц;
- Использование реляционной модели данных;
- Непрекращающийся темп пополнения больших данных;
- Условно неограниченный объём хранения и обработки больших данных.

Методики анализа и хранения больших данных

В данной статье рассмотрим 4 методики работы с рассматриваемыми данными:

- Методика Инмона;
- Методика Кимбалла;
- *Data Vault*;
- Якорное моделирование.

Далее в статье в описаниях способов термины «большие данные» и «данные» по смыслу будут равнозначны.

Методика Инмона

Базы данных и другие системы обработки данных создавались, в первую очередь, для поддержания работоспособности вычислительных систем. Однако было замечено, что добавлялись всё больше и больше тех данных, от которых есть какое-то прикладное применение, но мешали при поддержании приложений. Поэтому были предложены подходы по обработке таких данных. Эти данные служили для отчётов, анализа и других задач понимания развития организации, для которой система реализовывалась. Иначе говоря, интеллектуальное хранение этих данных и их дальнейшее использование уже стали неотъемлемой частью успешно развивающегося бизнеса [2].

Одним из первых предложил данный способ обработки данных Билл Инмон. Он предоставил свою методику обработки данных (рис. 1):

1. Копии данных, поддерживающие работоспособность приложения (*OLTP Data Sources*), направить (*ETL*) на общее хранилище данных (*Data Warehouse*). Данное хранилище реализовать по правилам третьей нормальной формы (*3NF*);
2. Копии данных с хранилища направить (*ETL*) в таблицы, называемые «витринами данных» (*Data Marts*). Данные таблицы также создаются по правилам третьей нормальной формы (*3NF*) и предназначены как фундамент для аналитики;

3. Данные с витрин можно интерпретировать для анализа соответствующим образом. В этом нет ограничений. Но на момент создания методики было предложено преобразовать данные с витрин в кубы *OLAP* (*OLAP Cube*);
4. Полученные данные собрать в единый отчёт (*Reporting layer*) для последующего принятия решений.

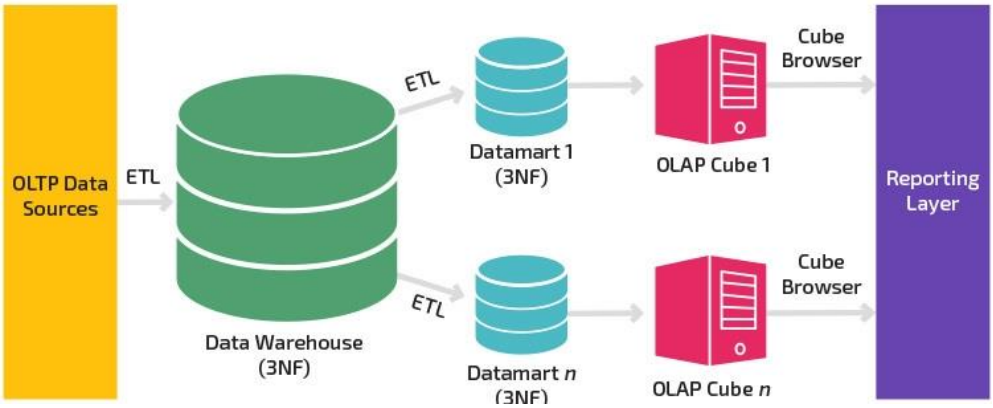

Рис. 1. Схема методики Инмона

Методика Кимбалла

Инмон не давал чёткого ответа, а как именно хранить таблицы в хранилище данных. Только известно про хранение в определённой нормальной форме. И ещё не было ответа, как связать данные с разных витрин для демонстрации и описания взаимодействия сущностей. Для решение этого вопроса была предложена методика Ральфа Кимбалла. Кимбалл разработал свой подход в 90-х годах, предложив новую архитектуру, новое видение и инновационное моделирование хранилища данных [3]. Он предложил следующее (рис. 2):

1. Копии данных, поддерживающие работоспособность приложения (*OLTP Data Sources*), направить на различные витрины данных. Эти витрины данных организовать по третьей нормальной форме (*3NF*).
2. Копии данных с витрин направить (*ETL*) на одно хранилище данных (*Data Warehouse*). Кимбалл предложил хранить таблицы по «Схеме Звезда» (*Star Schema*), где все таблицы витрин связаны с помощью одной таблицы. Эта таблица по методике Кимбалла называется «таблица фактов». В «Схеме Звезда» допускается использование второй или третьей нормальной формы.
3. Данные с хранилища можно интерпретировать для анализа соответствующим образом. В этом нет ограничений. Но на момент создания методики было предложено преобразовать данные с витрин в кубы *OLAP* (*OLAP Cube*);
4. Полученные данные собрать в единый отчёт (Reporting layer) для последующего принятия решений.

Второй шаг по созданию хранилища Ральф Кимбалл расписал очень подробно, чтобы бизнес понимал, как для него создать такую систему. Вопросов зачастую не возникало у лиц, проектирующих хранилища данных. Долгое время бизнес опирался на его модель.

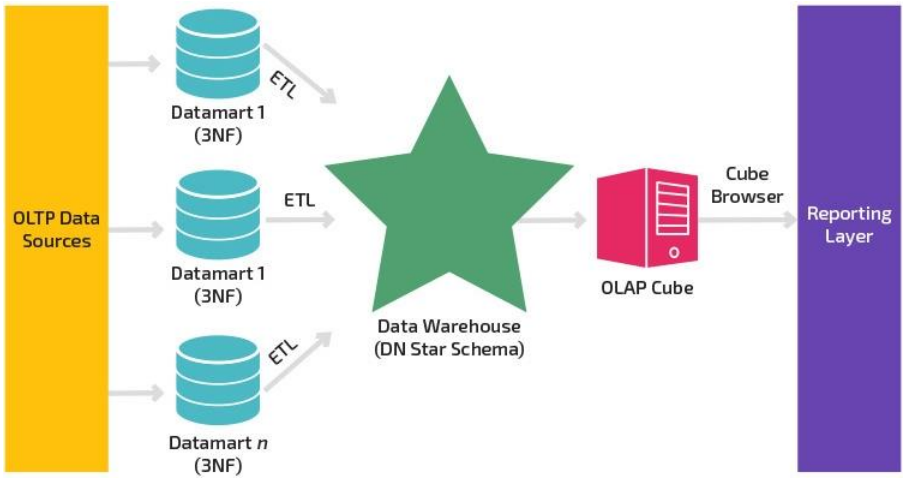

Рис. 2. Схема методики Кимбалла

Data Vault

В методиках Инмона и Кимбалла есть свои преимущества, так и свои недостатки. Они были пригодны в тех бизнесах, где цель неизменна на протяжении работы системы. То есть разработанные по каскадной модели жизненного цикла системы.

Однако возникали трудности при внедрении спиральной модели жизненного цикла системы:

- Как дальше сопровождать продукт, когда требуются новые таблицы для последующей обработки данных;
- Как учитывать преимущества обоих методик;
- Как без вреда для бизнеса отказаться от поддержания таблиц с хранилища.

На эти вопросы даёт ответы методика Дэна Линстеда: *Data Vault*. *Data Vault* – набор уникально связанных нормализованных таблиц, содержащих детальные данные, отслеживающих историю изменений и предназначенных для поддержки одной или нескольких функциональных областей бизнеса [4].

Предлагается введение 3 классов таблиц: хаб (*hub*), линк (*link*) и сателлит (*satellite* / *sat*). Названия таблиц указывать с первой буквы класса *h, l, s* соответственно. Их предназначение:

- Хаб – основная таблица;
- Линк – связь между таблицами хаба;
- Сателлит – связь с хабом для описания дополнительной информации.

Для данной методики нужно учесть правила проектирования таких хранилищ:

- Хаб не связывается с другим хабом. Он связывается только с линком или с сателлитом. Возможна привязка хаба к нескольким линкам или сателлитам;
- Линк связывает хабы или другие линки. Допускается связка линка и сателлита;
- Сателлит привязан только к одному хабу или линку. Между собой сателлиты или один сателлит к 2 или более хабам или линкам не допускается.

Недостатки методологии *Data Vault*:

- Множество соединений. За счет большого количества операций *join* теряется производительность запросов, в сравнении с традиционными хранилищами данных, где таблицы более денормализованы.
- Избыточность [5].

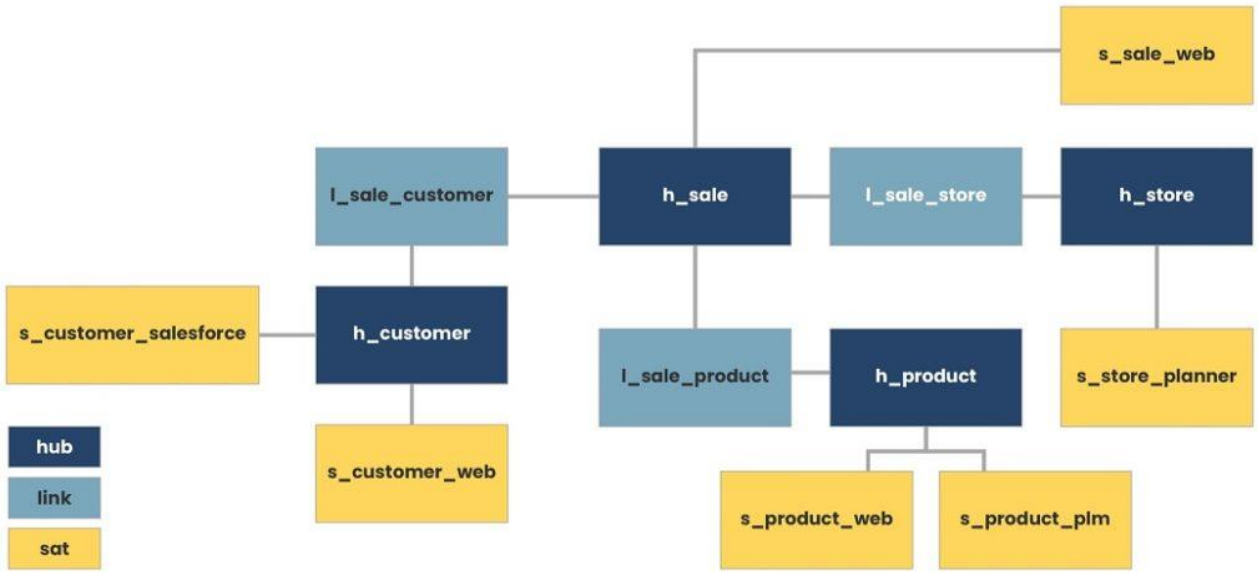

Рис. 3. Пример организации хранилища данных по Data Vault

Якорное моделирование

После того, как бизнес внедрил *Data Vault*, возникли вопросы о максимизации преимущества нормализации. Основная идея нормализации в том, чтобы минимизировать необходимое использование памяти. А также с активным внедрением микросервисной архитектуры для поддержания приложений возникали вопросы того, как для одной таблицы без потерь загрузить данные. То есть столбцы одной таблицы могли быть заполнены с разных источников. И был риск потери при заполнении данных.

Во избежание данных трудностей было предложено якорное моделирование хранилища данных от Ларса Рённбэка. Методика напоминает *Data Vault*, ибо правила идентичные. Цель методики заключается в достижении сильно нормализованной модели, которая эффективно обрабатывает изменения и увеличивает хранилища данных без отмены предыдущих действий [6]. Он предложил такие таблиц и правила их взаимодействия:

- Вместо хаба – якорь (*anchor*);
- Вместо линка – связи (*ties*);
- Вместо сателлита – атрибут (*attribute*);
- И дополнительно – узел (*knot*). Он зачастую никогда не меняется. Прикрепляется к атрибуту.

А также дополнительные опции для каждой из таблиц, отображённые на схеме (рис. 4).

Зачастую в проектах используются в плане организаций данных один из 2 используемых методов хранения данных: *Data Vault* и якорное моделирование. Они берут лучшие практики от различных систем, совершенствуются и возникают версии этих методик. Есть попытки гибрида этих 2 методов, но пока профессиональную пригодность на массовом рынке ещё не показали. Поэтому используется то, что описано ранее.

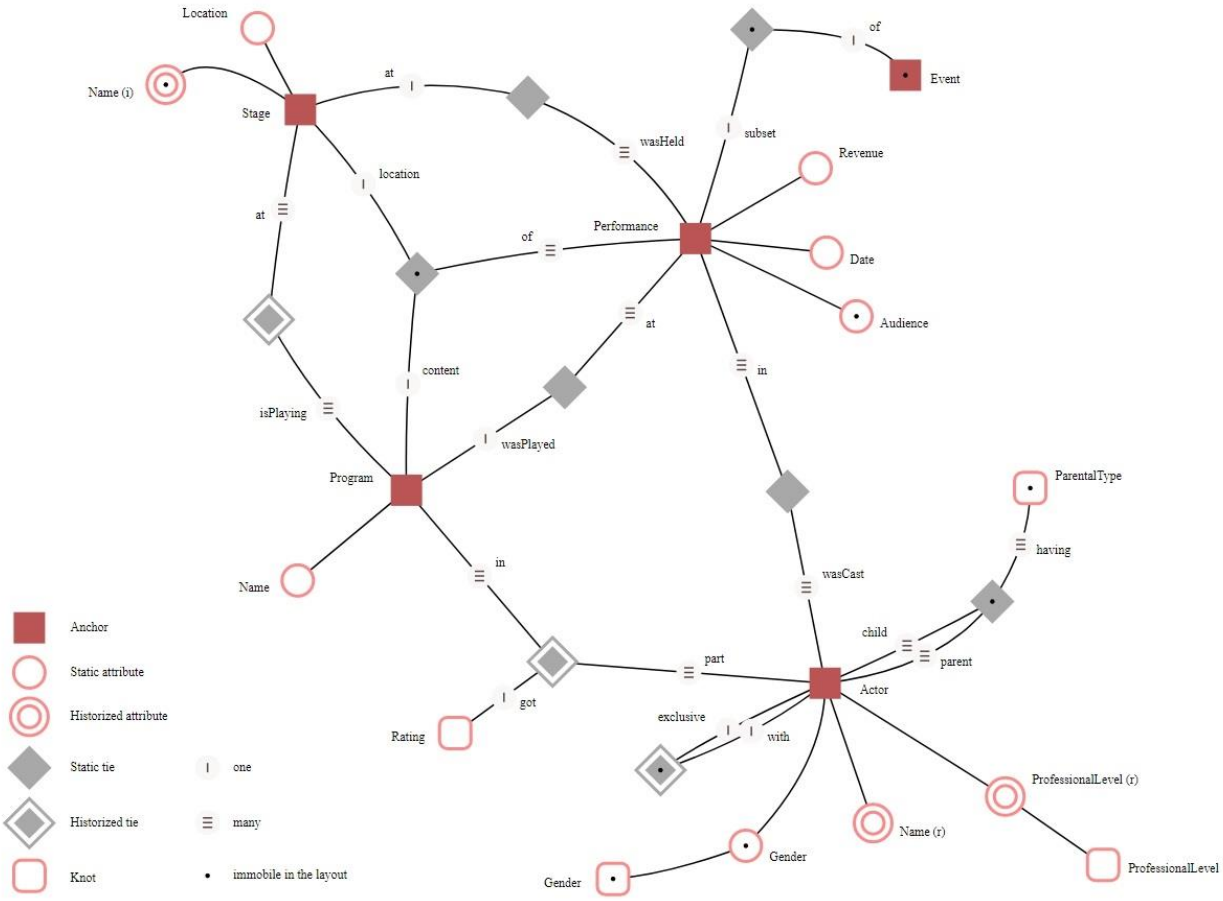

Рис. 4. Пример схемы якорного моделирования

Таким образом, исследования предшественников, несмотря на активное их использование в проектах, обладают следующими недочётами:

- Низкая методологическая гибкость обработки и хранения данных;
- Ограниченный выбор структуры данных для их обработки. Он ограничен только таблицей;
- Дорогая технологическая инфраструктура, как аппаратная, так и программная [7].

В данной работе предлагается:

- Повысить методологическую гибкость обработки и хранения данных;
- Обдумывание использования методик обработки и хранения данных с учётом тенденции рынка.

Заключение

В реальном мире идеи хранения данных с последующей их обработкой специалистами в области аналитики показали высокую эффективность развития бизнеса. Данные, которые не поддерживали работоспособность приложения, способны передать сведения для построения будущих систем, для будущих задач организаторов этих систем.

В итоге было проведены исследовательские работы по следующим пунктам:

- Разбор больших данных, а также отбор определения и условий для исследования рассматриваемого вопроса;
- Анализ способов обработки больших данных: принцип работы, компоненты и связи между ними;
- Анализ перехода от одного способа к другому;
- Выявление вопросов, необходимые для последующих исследований.

В данной работе хотелось бы пожелать научному сообществу раскрыть вопросы в исследовании способов обработки и анализа больших данных, а также предложить свои методики, которыми будут пользоваться большинство компаний, занимающейся работой с большими данными.

Список источников

1. Корнев, М. С. История понятия «большие данные» (Big Data): словари, научная и деловая периодики // Вестник РГГУ. Серия: Литературоведение, Языкознание. Культурология. – 2018. – № 1. – С. 81.-85. – DOI: https://doi.org/10.28995/2073-6355-2018-1-81-85.
2. Кошлич А. Д., Гулаков К. В Современные подходы к построению хранилищ данных // Sciences of Europe. – 2019. – Т. 1. – № 45-1. – С. 32-37.
3. Суховенко, С. А. Сравнительное исследование подходов к моделированию хранилищ данных: Инмон, Кимбалл и Data Vault // Постулат – 2021. – № 2.
4. Павлович, Н. В. Data vault: преимущества и недостатки // 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.
5. Джабраилов, Ш. В. Сравнительный анализ методологий проектирования хранилищ данных / Ш. В. Джабраилов, Ю. А. Орлова, Ш. Ш. Камбарова // Молодой ученый. – 2020. – № 20 (310). – С. 27-30. – URL: https://moluch.ru/archive/310/70003/ (дата обращения: 17.11.2022).
6. Кислицын, Е. В. Особенности использования методологии Anchor modeling для решения задач больших данных / Е. В. Кислицын, Р. С. Жернаков // Интернет-журнал «e-FORUM». – 2018. – № 2 (3).
7. Data Engineering: ETL, ELT, Data Pipeline, Data Warehouse, Data Lakes, Data Marts. – Текст: электронный // ivan-shamaev.ru: [сайт]. – URL: https://ivan-shamaev.ru/data-engineering-etl-pipeline-data-warehouse-datalake/ (дата обращения: 17.11.2022).