УДК 004

Сравнительное исследование подходов к моделированию хранилищ данных: Инмон, Кимбалл и Data Vault

Сухоненко София Андреевна
Российский экономический университет имени Г.В. Плеханова
Студент

Аннотация
Бизнес-аналитика - важная дисциплина для компаний, и задачи, с которыми она сталкивается, носят стратегический характер. Центральным понятием в BI является хранилище данных, которое представляет собой набор консолидированных данных из разнородных источников (обычно баз данных в 3NF). Для моделирования хранилища данных наиболее часто используются подходы Инмона и Кимбалла. Оба решения монополизируют рынок бизнес-аналитики. Однако третий подход к моделированию, названный «Data Vault» его создателя Линштедта, набирает обороты из года в год. Это позволяет построить хранилище сырых (необработанных) данных из разнородных источников. Цель данной статьи - представить сравнительное исследование трех прецедентных подходов. Сначала изучается каждый подход отдельно, а затем проводится сравнение между ними. Даны рекомендации по выбору наилучшего подхода.
Ключевые слова: моделирование хранилищ данных; подход Инмона; подход Кимбалла; подход Data Vault; сравнение.

A Comparative Study of Data Warehouse Modeling Approaches: Inmon, Kimball and Data Vault

Abstract
Business Intelligence is an important discipline for companies, and the challenges it faces are strategic. The central concept in BI is the data warehouse, which is a collection of consolidated data from disparate sources (usually 3NF databases). Inmon's and Kimball's approaches are most commonly used to model a data warehouse. Both solutions monopolize the business intelligence market. However, a third approach to modeling, called the "Data Vault" by its creator Linstedt, is gaining momentum from year to year. This allows you to build a repository of raw (raw) data from heterogeneous sources. The purpose of this article is to provide a comparative study of three case-based approaches. First, each approach is examined separately and then a comparison is made between them. Recommendations are given for choosing the best approach.
Keywords: data warehouse modeling; Inmon approach; Kimball approach; data vault approach; comparison.

Введение

Сегодня компании сталкиваются с множеством проблем, особенно в отношении использования и анализа операционных данных, которые они хранят в своих разнородных источниках. Конечная цель этих задач - получить полезную информацию для принятия решений. Бизнес-аналитика — это недостающее звено, которое может преобразовывать необработанные данные в полезную и актуальную информацию, на основании которой принимаются корпоративные решения лидеров. Центральным понятием в системе принятия решений является хранилище данных. Оно является основным компонентом информационной системы, которая предназначена для хранения операционных данных из нескольких источников и предоставляет их пользователям в определенных форматах для анализа. Создание хранилища данных требует подхода к моделированию, который учитывает все аспекты разработки, такие как моделирование данных, управление проектами, управление рисками, развертывание и многие другие важные аспекты. В течение нескольких лет в моделировании данных для хранилищ данных конкурируют два подхода: предметное моделирование Инмона [4] и подход пространственного моделирования Кимбалла [6]. Однако в последние годы появился третий подход, названный «Data Vault», который быстро распространяется, поскольку он улучшает гибкость, масштабируемость и производительность хранилищ данных [5]. В литературе редкие работы объединяют три подхода в сравнении. Именно поэтому в данной статье мною будет проведено сравнительное исследование подходов. Работа включает несколько критериев, обычно используемых в литературе для сравнения подходов Инмон и Кимбалла или обоих подходов с подходом Data Vault, например: методология разработки, архитектура хранилища данных, управление жизненным циклом и ETL и т. д. В рамках статья будет изучена тема хранилищ данных и их назначения. Также будут рассмотрены подходы к моделированию хранилищ данных. Поэтому для каждого подхода будут представлены его определение, философию и некоторые основные концепции. В рамках исследования будет проведен сравнительный анализ подходов Инмона, Кимбалла и Data Vault на основе критериев для предоставления рекомендаций для выбора наилучшего подхода.

Определение

Определение по Кимбаллу

Кимбалл определяет хранилище данных как «копию транзакционных данных, специально структурированных для запросов и анализа» [1]. Кроме того, по словам Кимбалла, цель хранилища данных - «предоставить информацию для поддержки принятия решений в компании». Таким образом, хранилище данных представляет собой специальную базу данных, используемую в контексте принятия решений и анализа.

Определение по Инмону

Со своей стороны Билл Инмон дает следующее определение: «Хранилище данных — это предметно-ориентированный, интегрированный, ограниченный во времени и долговременный инструмент для сбора данных

для поддержки процесса принятия решений руководством» [4]. Необходимо разобрать предложение на составляющее

- Предметно-ориентированный: данные связаны с бизнесом компании и организованы по функциям;
- Интегрированный: означает, что данные, полученные из нескольких операционных и внешних систем, должны быть удовлетворительными, что предполагает решение проблем, связанных с определением данных и различиями в содержании, такими как разные форматы и кодирование данных;
- Ограниченный во времени: данные идентифицируются за определенные периоды;
- Долговременный: данные используются для запросов и не могут быть изменены. Итак, операции обновления не разрешены, возможно только чтение.

Таким образом, хранилище данных — это централизованная база данных, которая хранит рабочие данные определенным образом и делает их доступными и пригодными для анализа.

Моделирование хранилища данных

Моделирование — это процесс настройки хранилища данных. Есть несколько подходов к реализации хранилища данных.

a. Подход Инмона

Подход был основан Биллом Инмоном в 90-х годах, чтобы соответствовать требованиям бизнеса и позволять им разрабатывать свои системы принятия решений. Он позволяет хранить все события компании и выделяет важные ресурсы для создания системы. Подход Инмона основан на диаграммах «сущность-взаимосвязь» операционных систем. Данные компании загружаются без предварительного знания требований пользователя. Архитектура хранилища данных Инмона (см. Рис. 1.) включает в себя все информационные системы компании с их базами данных вместо того, чтобы рассматривать только фрагменты информации. Инмон делит среду баз данных компании на четыре уровня, включая:

- операционный;
- атомарный (хранилище данных);
- ведомственный (витрины данных);
- индивидуальный уровень.

Последние три уровня относятся к хранилищу данных, в то время как первый уровень (операционный) поддерживает ежедневные операции и содержит транзакционные данные компании. Затем эти данные преобразуются и загружаются в хранилище атомарных данных с помощью ETL (извлечение, преобразование и загрузка). Согласно требованиям отделов компании, данные на уровне отделов являются предметно-тематически ориентированными и всегда непротиворечивыми, так как поступают из одного хранилища данных. Инмон считает, что хранилище данных и витрина данных физически разделены. Хранилище данных имеет собственное

физическое существование и ориентировано на хранение и масштабируемость в соответствии с новыми требованиями. Витрины имеют собственное физическое существование и предлагают структуру, ориентированную на производительность в ответ на требования пользователей.

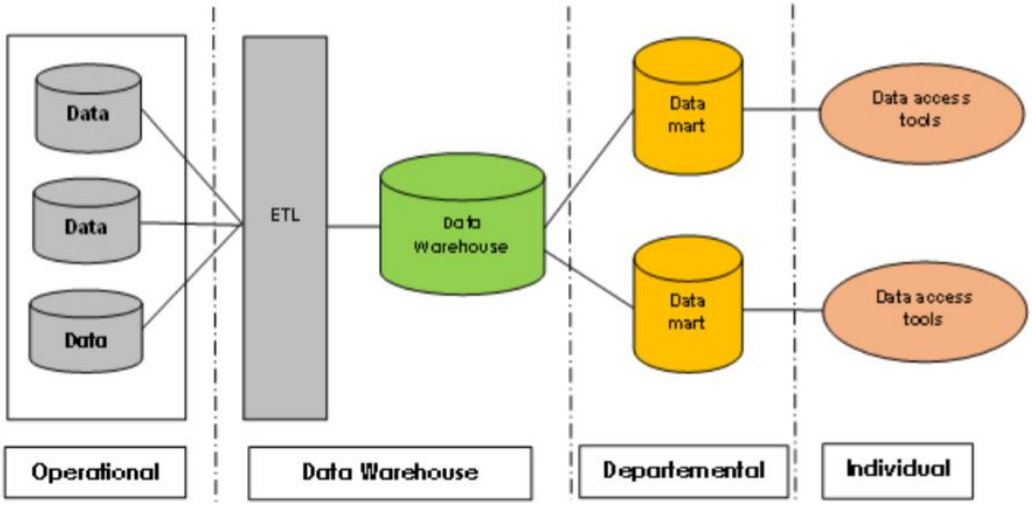

Рисунок 1. Подход к архитектуре хранилища данных Билла Инмона [4]

Пользователи создают последний уровень, когда они анализируют и используют загруженные данные в витрины данных (OLAP-анализ, отчеты, информационные панели и т. Д.). Для реализации хранилища данных Инмон предлагает спиральную методологию разработки, которая представляет стадии разработки хранилища данных. Модель данных компании является отправной точкой для внедрения хранилища данных. Для этого важно, чтобы оно было полным. Для создания этой модели Инмон предлагает три уровня моделирования данных, а именно [4]:

- Диаграммы сущностей-отношений (ERD): для моделирования высокоуровневой абстракции данных. Эта модель, как и при разработке операционных баз данных, состоит в том, чтобы, во-первых, идентифицировать бизнес-субъектов компании, а во-вторых, определить отношения между этими субъектами. Для каждого отдела компании, который хочет использовать хранилище данных, создается ERD, и все ERD будут составлять ERD компании.
- DIS (набор элементов данных): второй уровень моделирования - это место, где мы находим наибольшую информацию о модели данных компании. Для каждого бизнес-субъекта, указанного в ERD компании, создается DIS. Этот уровень моделирования содержит ключи, атрибуты, подтипы, группы атрибутов и соединители.
- Физическая модель: это последний уровень моделирования. Он создается на основе второго уровня модели данных. Для каждой части DIS это будет уникальная и отдельная физическая модель данных. Эта модель похожа на реляционные таблицы.

b. Подход Кимбалла

Кимбалл разработал свой подход в 90-х годах, предложив новую архитектуру, новое видение и инновационное моделирование хранилища данных. Этот подход основан на концепции размерного моделирования. Кимбалл выступает против принципа изоляции конечных пользователей, предложенного Инмоном. Действительно, его подход сильно вовлекает конечных пользователей на ранних этапах проекта, поэтому он называется подходом, основанным на требованиях пользователей. Кимбалл представляет другое видение хранилищ данных. Он считает, что хранилище данных можно рассматривать как набор согласованных витрин данных, основанных на общих согласованных измерениях (рисунок 2) [7].

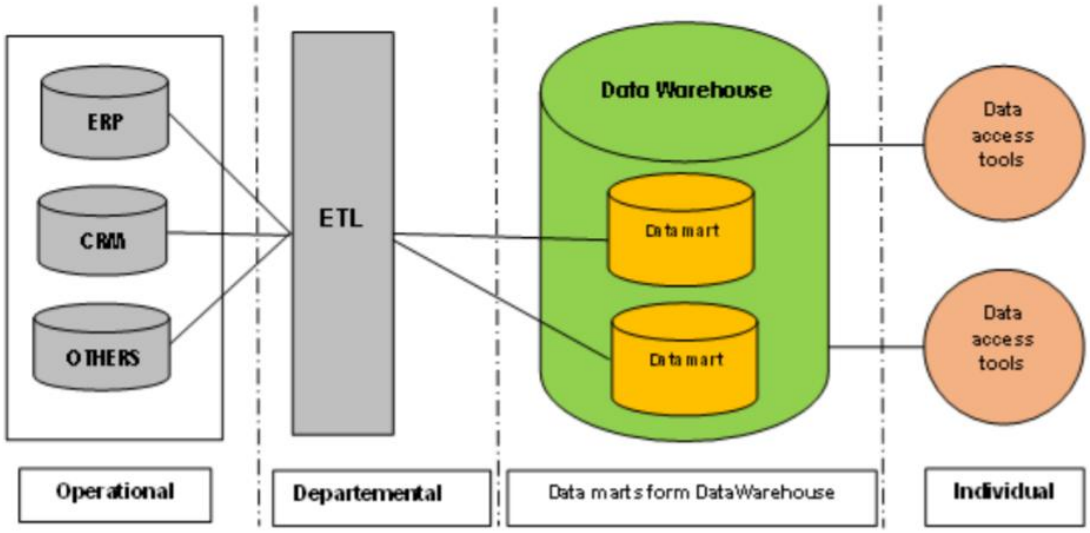

Рисунок 2. Подход к архитектуре хранилища данных Ральфа Кимбалла

Многомерный обзор данных подчеркивает анализируемый предмет и различные перспективы анализа [6]. Эта модель содержит несколько основных концепций: таблицу фактов, измерения, согласованные измерения и матрицу шины. Ниже представлены первые две основные концепции:

Таблица фактов: она включает наблюдаемые данные (факты) о предмете, который необходимо изучить, с использованием различных аналитических осей (измерений). Это может быть сумма продаж, количество проданных единиц продукта и т. Д. [[1]].

Измерение: содержит ось анализа (измерения), по которой мы хотим изучать факты. Подвергнутые многомерному анализу, эти данные предоставляют пользователям информацию, необходимую для принятия решений. Это могут быть клиенты или продукция компании и т. д. [7].

c. Подход Data Vault

В начале 2000-х Дэн Линстедт вступил в соревнование, предложив третий подход, названный «Моделирование Data Vault». Дэн Линстедт определяет Data Vault как «детально ориентированный, исторический отслеживающий и однозначно связанный набор нормализованных таблиц,

которые поддерживают одну или несколько функциональных областей бизнеса» [13]. Согласно Линстедту, 3NF Инмона и пространственное моделирование Кимбалла имеют слабые места, если объем данных увеличивается. Таким образом, «Data Vault» интересен скорее для изменений процессов и структур данных, чем для изменений бизнес-функций. Основными особенностями Data Vault являются [5]:

- Структурная информация отделена от описательной информации (атрибутов) по причинам гибкости и предотвращения реинжиниринга в случае изменения.
- Data Vault позволяет параллельную загрузку данных.
- Данные не обрабатываются и не фильтруются (отслеживание источника данных).
- Данные никогда не меняются (остаются неизменными).
- Структура Data Vault не допускает окончательного использования данных. Data Vault основан на трехуровневой архитектуре для разделения хранилища необработанных данных от конечных пользователей и различных уровней интеллектуального анализа данных.

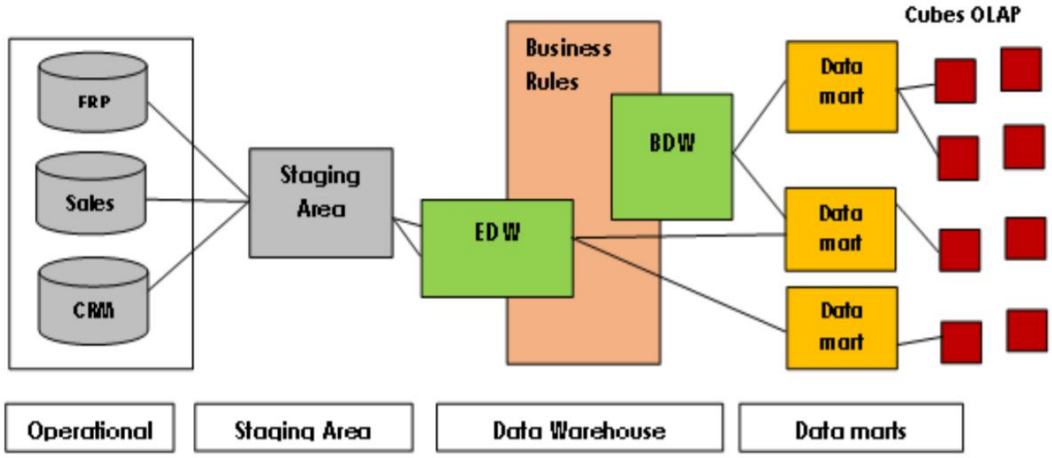

Рисунок 3. Подход к архитектуре хранилища данных «Data Vault»

Тремя уровнями «Data Vault» являются: область подготовки, хранилище данных Data Vault и витрины данных. На предыдущем рисунке показана архитектура Data Vault корпоративного хранилища данных (EDW для Enterprise Data Warehouse).

- Область подготовки: этот уровень поддерживает процесс загрузки данных из нескольких источников Data Vault.
- «Data Vault» и бизнес-ориентированное хранилище данных (BDW для Business Data Warehouse): «Data Vault» является основным уровнем; он содержит данные из разных источников без преобразований. Модель Data Vault состоит из набора хабов, каналов и сателлитов, которые являются основными элементами модели. В то время как BDW содержит данные, полученные в

результате применения различных бизнес-правил, как согласование данных с бизнес-ключами.

- Витрины данных: представляют уровень представления; они развернуты с использованием размерного моделирования. Витрины берут данные из бизнес-хранилища и Data Vault и предоставляют кубы OLAP для использования их в отчетах, интеллектуальном анализе данных и т. д.

В модели Data Vault есть три типа сущностей: хабы, линки и сателлиты.

- Хабы: «Хабы определяются уникальным списком бизнес-ключей, которые представляют основные бизнес-концепции, такие как клиент, поставщик, продажа или продукт. Эти бизнес-ключи жизненно важны для компаний, чтобы отслеживать, находить и идентифицировать свою информацию» [5]. Крайне важно, чтобы бизнес-ключи имели историческую и общую уникальность.
- Линк: «Линки являются связующим звеном между двумя или более бизнес-ключами, а иногда и другими линками»[10]. Гранулярность линка определяется хабами по отношению к линку. Линки как хабы не содержат описательных данных. Этот тип данных появляется в другом объекте, называемом сателлитом.
- Сателлит: «Сателлит - это временная таблица, содержащая подробную описательную информацию, которая обеспечивает контекст для бизнес-ключей Hub или Link»[10]. У сателлита может быть только одна родительская таблица. Он фиксирует описательные изменения данных хранилища данных, когда они происходят. Первичный ключ сателлита получается путем объединения ключа родительского (или хаб-канала) и даты загрузки.

Сравнительное исследование

Сравнение подходов Инмона, Кимбалла и Data Vault [2] показывает, что подходы Кимбалла и Инмона похожи, так как оба они используют ETL для подпитки хранилища данных. [8] обнаружил, что сходство между подходом Data Vault и подходом Кимбалла заключается в итеративной реализации решения и использовании области подготовки для восстановления и синхронизации. В то время как подход Data Vault и подход Инмона сходятся в том, что хранилище данных является крупнейшим репозиторием компании. Несмотря на эти сходства, различия между тремя подходами многочисленны и глубоки. Эти различия суммированы в таблицах.

1. Философия: каждый подход начинается с философии, которая может быть близкой или полностью отличаться от двух других подходов. Основные действующие лица, указанные в подходе, и цели каждого из трех подходов представлены в таблице 1.

Таблица 1. Сравнение между подходами со стороны философии

	Инмон	Кимбалл	Data Vault
Действующие лица	ИТ-специалисты	Конечные пользователи	Конечные пользователи
Цель	Предоставляет полное техническое решение, основанное на проверенных методах и технологиях	Предлагает решение, которое упрощает прямой запрос данных конечными пользователями.	Обеспечивает прочное и полное решение. В его основе проверенные методы.

2. Методология и архитектура. В таблице 2 показано сравнение трех подходов, основанных на архитектуре методологии разработки хранилища данных, общей сложности подхода, а также стоимости и времени развертывания хранилища данных.

Таблица 2. Сравнение между подходами со стороны философии

	Инмон	Кимбалл	Data Vault
Архитектура	Хранилище данных атомарного уровня питает витрины данных отдела	Набор витрин данных составляет хранилище данных	Data Vault питает витрины данных
Методология разработки	Вдохновленный спиральной методикой	На основе четырехэтапного процесса	На основе гибкой методологии
Затраты на развертывание	Первоначальные затраты выше, а затраты на последующие разработки ниже	Начальные затраты ниже. Каждый последующий шаг стоит столько же	Стоимость проекта ниже, чем при двух предыдущих подходах
Время разработки	Длительная продолжительность	Короткая продолжительность	Короткая продолжительность
Сложность	Довольно сложно	Простой	простой

3. Интеграция данных и ETL. Ключевым элементом при выборе подхода является возможность простой и эффективной интеграции нескольких источников данных. Этот элемент поясняется в Таблице.

Таблица 3. Сравнение между подходами по интеграции данных и ETL

	Инмон	Кимбалл	Data Vault
Интеграция нескольких источников	Правила преобразования должны быть реализованы в процессах ETL	Правила преобразования должны быть реализованы в процессах ETL	Разделение сателлитов и бизнес-ключей снижает сложность
Сложность процесса ETL	Правила преобразования просты, если модель данных аналогична моделям источников данных	Преобразования между OLTP-моделью и размерной моделью сложны	Правила просты для загрузки хабов, линков и сателлитов

4. Моделирование данных: Таблица 4 показывает сравнение того, как моделируются данные. Это сравнение основано на инструментах и участии конечного пользователя в процессе моделирования.

Таблица 4. Сравнение между подходами по моделированию данных

	Инмон	Кимбалл	Data Vault
Инструменты	Инструменты классического моделирования (ERD, DIS)	Размерное моделирование	Моделирование хабов, каналов и сателлитов.
Вовлечение конечного пользова-теля	Слабое	Сильное	Сильное

Как показывают предыдущие сравнительные таблицы, ни один из подходов не отвечает полностью всем требованиям. У каждого метода есть свои достоинства и недостатки.

Ниже на основе сравнительного исследования трех проведенных подходов предлагается несколько рекомендаций, которые отвечают на вопрос «Какой подход в какой ситуации?»

Подход Инмона рекомендуется, если требования к анализу не определены или целью витрин является предоставление информации о нескольких системах бизнес-аналитики. Также предпочтительно, чтобы структуры исходной системы были относительно стабильными.

Подход Кимбалла: этот подход настоятельно рекомендуется для витрин данных, поскольку многомерная модель обеспечивает высокую производительность запросов и понятна конечным пользователям. Кроме того, также целесообразно разработать хранилище данных, если требования известны и четко определены.

Подход Data Vault: это мощный подход к разработке хранилища данных при наличии нескольких источников данных с регулярными изменениями их структур. Это эффективно в среде гибких проектов. Если гибкость, производительность и масштабируемость хранилища данных являются основными заботами компании, то выбор Data Vault - лучший выбор.

Большинство существующих в литературе исследований в основном сосредоточено на сравнении моделей Инмона и Кимбалла. Однако в некоторых работах используется подход Data Vault.

Заключение

Компания, которая хочет выбрать систему принятия решений, должна глубоко задуматься над методологией разработки, потому что, как и любой проект, система принятия решений подвержена ошибкам. Выбор метода основан на знании компании и ее политики для достижения благоприятного подхода. В статье проведено сравнительное исследование трех подходов к моделированию хранилищ данных и сделан вывод, что ни один из них не является лучшим. Действительно, зная ситуацию в компании и ее направлениях, можно выбрать определенный подход. Подход Кимбалла предпочтительнее для моделирования витрин данных, поскольку он обеспечивает производительность запросов, особенно если требования стабильны и четко определены. Хотя подход Инмона рекомендуется, если требования не определены или очень масштабируемы. Оба подхода сталкиваются с множеством проблем, особенно если источники данных часто меняются, что подразумевает реорганизацию хранилища данных. Для решения этих проблем рекомендуется использовать Data Vault, поскольку он обеспечивает исключительную гибкость и масштабируемость.

Библиографический список

1. Ballard C., Herreman D., Schau D., Bell R., Kim E., Valencic A. Data modeling techniques for data warehousing. IBM,1998.
2. Breslin M. Data warehousing battle of the giants // Business Intelligence Journal. 2004. С. 3-14.
3. Inmon B. Building the Data Warehouse. Wiley Computer Publishing, 2016
4. Inmon B., William H. Building the data warehouse, John wiley & sons, 2005.
5. Mathiews D. Data Vault et BI URL: http://fr.slideshare.net/dlinstedt/prsentationdata-vault-et-biv20120508.
6. Kimball R., Ross M. The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling. Wiley Computer Publishing, 2002.
7. Kimball R.The Data Warehouse Toolkit: Practical Techniques for Building Dimensional DataWarehouses. Wiley Computer Publishing, 2016.
8. Schreuder M., Dimensional modelingandDataVault–a happy marriage? URL: http://blog.in2bi.com/business-intelligence/ dimensional-modeling-and-data-vault-ndash-a-happy-marriage/, 2011.

9. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. 336 с.
10. Дмитриева Е.О. Инструменты BI для поддержки принятия решений для руководителей высшего и среднего уровней с помощью решения SAS (опыт ВТБ24) // Вестник университета. 2017. №6.
11. Нефедов Ю.В., Староверова О.В., Уринцов А.И. Особенности организации систем поддержки формирования и использования решений // Матеріали III Мжнародної науково-практичної конференції. Киев: Университет экономики и права "КРОК", 2016. С. 144-146.
12. Уринцов А.И., Дик В.В. Системы формирования и принятия решений в условиях информатизации общества. М., 2008.
13. Уринцов А.И, Староверова О.В., Афанасьев М.А. Компьютерный инструментарий управления эффективностью бизнеса. М., 2016.
14. Data Vault Series 1 // TDAN.com URL: http://tdan.com/data-vault-series-1-data-vault-overview/5054 (дата обращения: 17.12.2020).