COMPUTER SYSTEMS AND NETWORKS

Vol. 6, No. 1, 2024

DOI: https://doi.org/10.23939/csn2024.01.067

УДК 004.6

РОЗРОБЛЕННЯ ГІБРИДНОГО МЕТОДУ ПОБУДОВИ СХОВИЩА ДАНИХ

О. М. Коваль, О. І. Гарасимчук

Національний університет "Львівська політехніка",
кафедра захисту інформації
E-mail: oleh.koval.kb.2020@lpnu.ua, oleh.i.harasymchuk@lpnu.ua

Розглянуто підхід до побудови адаптивного та зручного сховища даних, зокрема не тільки для зберігання даних, але з метою їх обробки для різного роду звітів. Запропоновано гібридний метод побудови сховища даних, який поєднує переваги відомих методологій та надає найкращу взаємодію зі сховищем даних для всіх його користувачів. За допомогою запропонованого методу можна швидко розгорнути сховище даних, адаптивно його підтримувати та легко масштабувати в майбутньому.

Цей авторський гібридний метод складається з централізованого зберігання даних та розподілу деталізованих даних на маленькі довідникові таблиці для кращого та швидшого функціонування сховища. Також запропонований метод відзначається швидшим розгортанням системи завдяки обмеженій кількості зв'язків між таблицями.

На прикладі онлайн магазину було продемонстровано всі переваги запропонованого методу як під час записування даних та їх обробки, так і під час використання цих даних для майбутніх звітів. Підхід є перспективним та ефективним рішенням для компаній, які прагнуть знайти оптимальний компроміс між традиційними методологіями та сучасними вимогами бізнесу.

Ключові слова: сховище даних, Data Warehouse, метод Інмона, метод Кімбола, гібридний метод.

Вступ

Інформаційні ресурси будь-якого сучасного підприємства є надзвичайно цінним активом, і тому важливо організувати правильне їх зберігання та забезпечувати легкий доступ до них, коли це потрібно. Однак надмірна доступність великого обсягу даних може спричиняти труднощі у визначенні та отриманні важливої інформації, а іноді й унеможливлювати її повне та своєчасне отримання.

Появі поняття сховища даних сприяли постійне зростання обсягу електронних даних, що нагромаджувався впродовж останніх років, та потреба використання цих даних для досягнення стратегічних цілей, крім рутинних завдань щоденної обробки. Створення ефективного сховища даних стає вирішальним завданням для підприємств у забезпеченні безперервного доступу до важливої інформації. Це містить розробку систем управління базами даних та інструментів для забезпечення безпеки і конфіденційності даних. Крім того, важливо регулярно аналізувати й оптимізувати обсяги даних, щоб уникнути перенавантаження та забезпечити швидкий доступ до потрібної інформації. Розвиток ефективних стратегій резервного копіювання та відновлення даних дає можливість забезпечити надійність інформаційних систем, навіть у разі непередбачених

ситуацій або кібератак. Крім того, важливо регулярно навчати персонал у сфері кібербезпеки та свідомої роботи з інформаційними ресурсами для запобігання можливим загрозам інформаційної безпеки.

1. Огляд літературних джерел

Data Warehouse (DWH) [1, 2] – це система, яка зберігає дані з різних джерел для аналітики та складання звітів. Data Warehouse підходить для складних та комплексних обчислень краще, ніж база даних. Під час виконання складного запиту базу даних можна перевантажити. Тому є ризик втрати нової інформації – для її обробки не вистачить ресурсів.

Сховища відрізняються від баз даних низкою ознак:

– дані в DWH не обов'язково мають надходити в реальному часі (якщо інше не передбачено бізнес-завданням);

– дані можуть мати різну структуру (залежно від джерел);

– сховище не обов'язково має працювати швидко. Головне, щоб швидкості вистачало на розв'язання всіх аналітичних завдань.

Сучасне бізнес-середовище зазнає постійних змін, потребуючи від підприємств швидкої та виваженої реакції на нові виклики. У цьому контексті система DWH стає основним інструментом для ефективного управління та ухвалення стратегічних рішень. Data Warehouse є центральним сховищем даних, в якому інформація з різних джерел об'єднується та перетворюється на цільовий формат для подальшого аналізу. Воно не тільки допомагає зберігати великі обсяги даних, але і забезпечує їхню доступність та структурованість. DWH дає можливість підприємствам не лише здійснювати звітність та моніторинг, але й отримувати цінні інсайти з даних, що полегшує стратегічне планування та вдосконалює ухвалення управлінських рішень.

Цей інструмент є незамінним для аналітики та бізнес-інтелекту, даючи змогу виявляти тенденції, прогнозувати розвиток подій та визначати оптимальні стратегії розвитку [3–5]. В контексті ринкового середовища, що швидко змінюється, система DWH стає критичним інструментом для підтримки конкурентоспроможності та забезпечення стабільності бізнесу.

Загальна мета використання Data Warehouse полягає в тому, щоб надати підприємствам засоби для ефективного аналізу даних, що своєю чергою сприяє прийняттю обґрунтованих та стратегічних рішень. Застосування цієї технології в сучасних умовах дає можливість досягти значних висот у сфері управління та оптимізації бізнес-процесів.

У світі, де обсяги даних нестримно зростають, роль побудови DWH у сучасному бізнесі стає критично важливою. Відділи фінансів і маркетингу, а також експерти з Data Science залежать від надійних і ефективних систем управління даними для здійснення стратегічних рішень та досягнення бізнес-цілей. Наприклад, великі корпорації з численними філіями мають аналізувати, як кожна філія сприяє загальному розвитку бізнесу. Корпоративна база даних містить деталізовану інформацію про завдання, які виконують філії. Для задоволення потреб керівників можуть створюватися індивідуальні запити для отримання потрібних даних. Проте цей процес потребує від адміністраторів баз даних сформулювати бажаний запит, дослідивши каталоги баз даних, та обробити його, що може займати значний час через обсяг даних та складність запиту. Нарешті генерується звіт як електронна таблиця, що подає вищий керівник.

Фахівці з побудови баз даних та сховищ даних є основними гравцями в цьому контексті, маючи навички, потрібні для здійснення повного життєвого циклу даних. Вони вміло витягують інформацію з різноманітних джерел, розробляють структури баз даних та сховищ даних, точно визначають типи колонок та ключі, створюють ефективні зв'язки між таблицями та впроваджують передові методи ETL (Extract-Transform-Load) [6] та ELT (Extract-Load-Transform) [7] для оптимальної обробки та інтеграції даних.

Також однією з основних завдань є побудова моніторингу, спрямованого на виявлення неправильних даних та помилок у роботі процесів. Для підвищення ефективності фахівці також

впроваджують тригери для взаємодії з іншими системами, щоб підтримувати безперервність та автоматизацію бізнес-процесів.

Збільшення важливості сховищ даних у веденні бізнесу підкреслює потребу постійного вдосконалення і адаптації до нових технологій та викликів у галузі обробки інформації. Внаслідок збільшення обсягів даних і зростання їхньої складності перед фахівцями постають постійні виклики, але одночасно й відкриваються нові можливості для розробки ефективних стратегій управління даними. Навички роботи зі сховищами даних залишаються невід'ємною частиною, щоб забезпечити ефективне управління ресурсами та забезпечити стійкий успіх у динамічному світі сучасного бізнесу.

Перш ніж переходити до конкретних кроків створення сховища даних, доцільно розглянути методологічну основу цього питання, а також визначити відмінності, які є між базами даних та сховищами даних (див. табл. 1).

Таблиця 1

Основні відмінності між базами даних та сховищами даних

База даних	Сховище даних
Це сукупність взаємопов'язаних даних, яку можна спільно використовувати та керування якою здійснюється централізовано	Є інформаційною системою, що містить історичні та комутативні дані з одного або кількох джерел
Використовується для запису даних	Використовується для аналізу даних
Це орієнтований на використання набір даних	Це предметно орієнтовані зібрані дані
Використовує OLTP (англ. Online Transaction Processing) – онлайнову обробку транзакцій	Використовує OLAP (англ. Online Analytical Processing, аналітична обробка у реальному часі) – інтерактивну систему, що дає можливість переглядати різні підсумки з багатовимірних даних
Містить нормалізовані таблиці та з'єднання, які тому є складнішими	Містить денормалізовані таблиці та з'єднання, які тому є простішими
Для проєктування використовуються методи ER-моделювання	Для проєктування використовуються методи моделювання даних

На створення сховища даних фундаментально впливає методологія, якої дотримуються розробники. Загалом нині домінують два методи: Інмона і Кімбола [8–10]. Розглянемо основні процеси та особливості кожного з них.

Підхід Інмона: модель "зверху вниз"

Згідно з визначенням Білла Інмона, сховища даних – це "предметно-орієнтований, енергонезалежний, інтегрований, змінний у часі набір даних для підтримки управлінських рішень". Підхід Інмона [8, 10] наголошує на створенні централізованого сховища даних як на першому кроці, а пізніше на створенні вітрин даних. Ця низхідна модель спрямована на уніфіковане рішення даних для всього підприємства.

Етапи побудови та використання сховища даних у концепції моделі Інмона:

1. Збір неопрацьованих даних.

На першому етапі дані отримуються з джерел (OLTP Data Sources) неопрацьованими. Це можуть бути дані з операційних систем, онлайн транзакцій, сенсорів тощо. Опрацьовані дані ще не проходять через жодні етапи трансформації та зберігаються в сирому вигляді в сховищі даних (Data Warehouse). Це дає змогу забезпечити доступність первинних даних та запобігти втратам інформації.

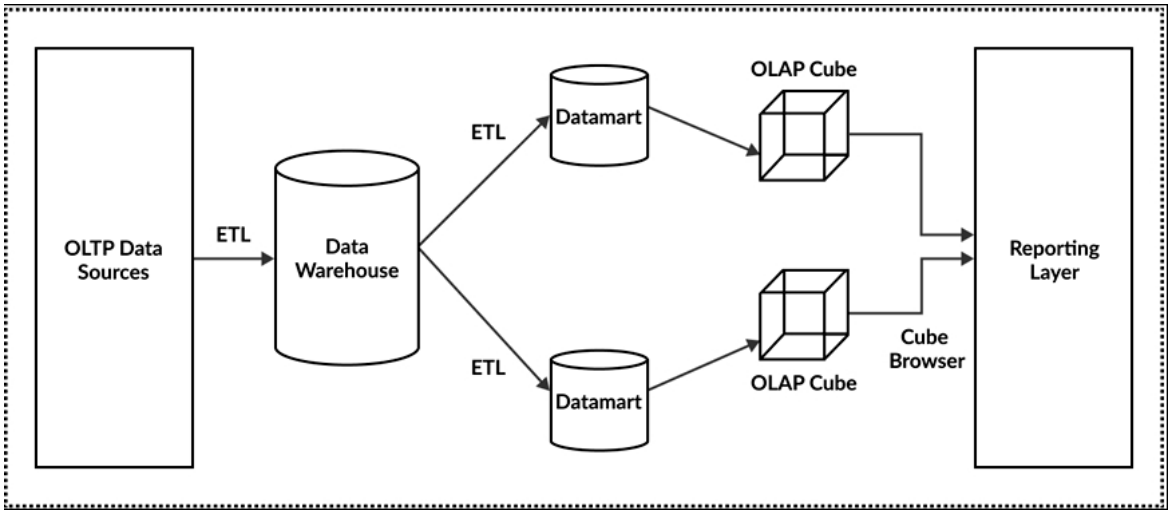

Рис. 1. Графік моделі Інмона

2. Завантаження-трансформація-виймання (ETL).

На другому етапі проводяться виймання, трансформація та завантаження даних (ETL). Цей процес дає змогу опрацьовувати та структурувати неопрацьовані дані так, щоб вони підходили для аналізу в різних вимірах. Важливим елементом є підготовка даних для швидкого та ефективного аналізу в різноманітних сценаріях.

3. Аналіз впорядкованих даних (OLAP Cube).

Третій етап передбачає створення так званого OLAP куба, що є багатовимірною моделлю даних. Цей куб дає можливість для швидкого різнопланового аналізу опрацьованих даних. Він дає змогу вибирати дані в різних напрямках та вимірах, що допомагає виявити закономірності та тренди, які можуть бути важливими для ухвалення управлінських рішень.

4. Відправлення даних для звітів та інших потреб (Reporting Layer).

На останньому етапі опрацьовані та оброблені дані використовуються для створення звітів, аналітичних документів та інших інструментів для ухвалення рішень. Ці дані передаються до різних інтерфейсів або платформ для забезпечення доступу користувачів до інформації, потрібної для їхньої роботи та управлінських ухвалень рішень.

Особливості побудови та використання сховища моделі Інмона:

1. Централізоване сховище даних.

Центральна точка – це централізоване сховище даних, яке слугує єдиним джерелом надійності для всієї організації. Усі організаційні дані зведено в цю єдину структуру. Але це своєю чергою може виявитися менш гнучким у реагуванні на швидкі зміни в потребах відділів чи бізнес-процесах.

2. Нормалізація даних.

Для підходу Інмона характерний, зокрема, акцент на нормалізації даних. Це означає, що дані в сховищі організовані так, щоб зменшити надмірність, таким способом забезпечуючи цілісність даних і полегшуючи ефективне надсилання запитів. Основною проблемою може виявитися складність написання запитів під час опитування даних, оскільки треба об’єднувати розподілену інформацію з різних таблиць.

3. Вітрини даних.

Після створення централізованого сховища даних модель Інмона рекомендує побудувати вітрини даних. Це менші та більш цілеспрямовані бази даних, створені для задоволення потреб окремих відділів, таких як маркетинг, фінанси чи кадри. Вітрини даних спрощують для цих відділів отримання конкретних відомостей, які стосуються їх діяльності. Мінусом цього є те, що створення та підтримка вітрин даних може потребувати значних витрат часу, робочої сили та фінансових ресурсів.

4. Високі початкові інвестиції.

Підхід Інмона потребує значних початкових інвестицій, але результатом є надійне уніфіковане сховище даних, здатне підтримувати складні запити та надавати глибоку аналітику в усій організації. На жаль, малі організації не завжди можуть собі дозволити витратити стільки ресурсів на початку.

Основні переваги методу Інмона такі:

– Сховище даних виступає як єдине джерело достовірної інформації для всього бізнесу, де всі дані інтегровані.

– Цей підхід має дуже низьку надмірність даних. Отже, знижується ймовірність порушень під час оновлення даних, що робить процес сховища даних на основі концепції ETL простішим і менш схильним до збоїв.

– Це полегшує бізнес-процеси, оскільки логічна модель представляє докладні бізнес-об'єкти.

– Цей підхід забезпечує більшу гнучкість, оскільки легше оновлювати сховище даних у разі будь-яких змін у бізнесових вимогах або вихідних даних.

– Він може задовольнити різноманітні вимоги до звітності у масштабі підприємства.

Основні хиби цього методу:

– Складність зростає у міру того, як згодом до моделі даних додається кілька таблиць.

– Попереднє налаштування та доставка займають багато часу.

– Потрібні ресурси, які дають змогу здійснювати моделювання даних сховища даних, знайти які може бути дорого і складно.

– Потрібна додаткова операція ETL, оскільки вітрини даних створюються після створення сховища даних.

– Цей підхід потребує від експертів ефективного управління сховищем даних.

Підхід Кімбола: модель "знизу вгору"

Підхід Кімбола підтримує методологію "знизу вгору" [8, 11–12], коли окремі вітрини даних спершу розробляються для конкретних бізнес-підрозділів, а пізніше інтегруються у повномасштабне сховище даних.

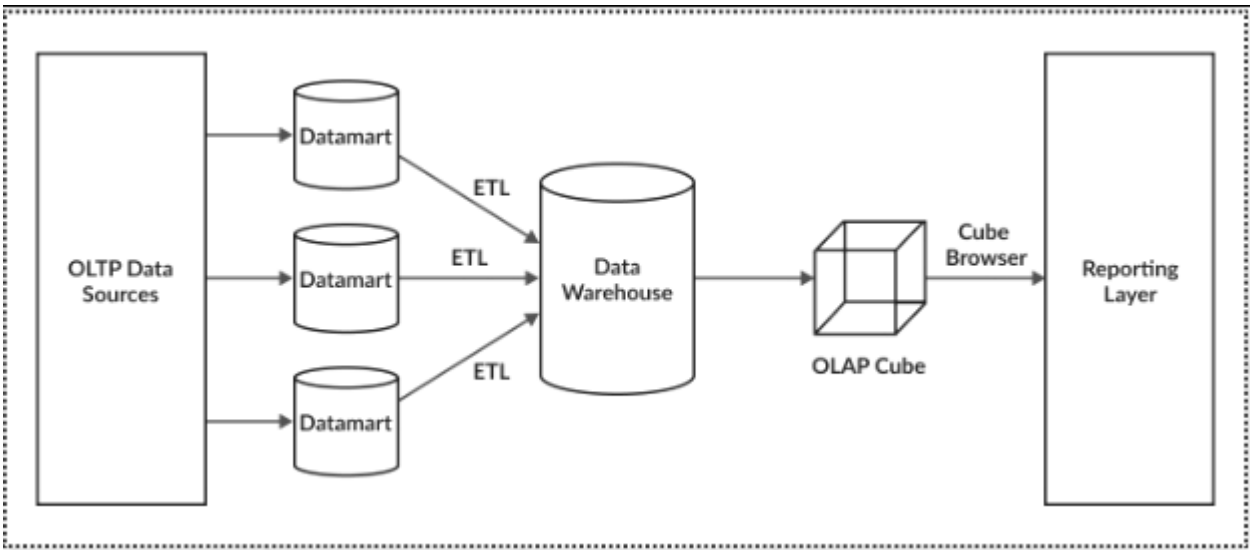

Рис. 2. Графік моделі Кімбола

Етапи побудови та використання сховища даних у моделі Кімбола:

1. Завантаження-трансформація-виймання даних під час завантаження (ETL).

У моделі Кімбола опрацювання даних починається уже під час завантаження їх з джерела (OLTP Data Sources). Це означає, що вони проходять через етапи завантаження, трансформації та

виймання, щоб забезпечити їх придатність для аналізу та звітності. Опрацювання на цьому етапі дає змогу виправити помилки, нормалізувати дані та структурувати їх для подальшого використання.

2. Опрацьовані дані в сховищі даних (Data Warehouse).

Важливою частиною підходу Кімбола є перехід від сирого вигляду даних до опрацьованого у сховищі даних (Data Warehouse). Це дає можливість зберігати дані в структурованій формі, що сприяє ефективному аналізу та роботі з ними. Сховище даних стає центром для подальших дій з аналітики та звітності.

3. Завантаження даних у структури для аналізу (OLAP Cube).

На етапі аналізу впорядкованих даних (OLAP Cube) додаткове опрацювання та чистка відбуваються для створення багатовимірної структури, яка сприяє швидкому та ефективному різноплановому аналізу. OLAP куб дає змогу взаємодіяти з даними в різних вимірах та використовувати їх для розуміння бізнес-тенденцій.

4. Відправка даних для звітів та інших потреб (Reporting Layer).

Опрацьовані дані використовуються для створення звітів та інших інструментів ухвалення рішень. Reporting Layer дає змогу розподіляти інформацію користувачам, щоб спростити їхній доступ до основної інформації та полегшити процес ухвалення рішень.

Особливості побудови та використання сховища моделі Кімбола:

1. Спершу вітрини даних.

Будівельними блоками є вітрини даних, розроблені для конкретних бізнес-функцій. Їх можна швидко розвивати, пропонуючи негайну цінність для бізнесу. Своєю чергою розробка окремих вітрин може призводити до фрагментації даних та їх надлишковості, ускладнюючи їх інтеграцію та зменшуючи загальну уніфікацію.

2. Зіркова схема.

Модель Кімбала використовує схему Зірка для організації даних. Ця структура є порівняно простою, але дуже ефективною для швидкого та гнучкого створення запитів, що робить її добре придатною для швидкої ітеративної аналітики. Проте використання зіркової схеми може бути обмеженим для складніших аналітичних опитувань та звітності.

3. Швидше розгортання.

Орієнтація на початкове створення вітрин даних дає змогу організаціям швидко розгортати функціональні елементи сховища даних, пропонуючи негайні переваги для бізнесу. Але можна виділити істотну проблему – брак загальної стратегії масштабування, що може призвести до розбіжностей та неоднорідності в розвитку окремих вітрин та їхній інтеграції.

4. Масштабованість.

Ці окремі вітрини даних можна масштабувати та плавно інтегрувати, щоб сформувати комплексне сховище даних, що робить цей підхід гнучким і адаптованим. Інтеграція великої кількості вітрин може бути складною та потребувати значних зусиль, особливо якщо їх розвиває велика організація з різними бізнес-процесами.

Основні переваги концепції сховища даних Кімбола такі:

– Багатовимірне моделювання Кімбола будується швидко, оскільки нормалізація не потрібна, що означає швидке виконання початкової фази процесу проєктування сховища даних.

– Перевага зіркоподібної схеми полягає в тому, що більшість операторів даних можуть легко її зрозуміти завдяки її денормалізованій структурі, яка спрощує запити та аналіз.

– Вплив системи сховища даних є тривіальним, оскільки воно зосереджено на окремих бізнес-областях і процесах, а не на всьому підприємстві загалом.

– Це дає змогу швидко витягувати дані зі сховища даних, оскільки дані поділені на таблиці фактів та вимірювання.

– Для управління сховищем даних потрібна досить невелика команда проєктувальників і планувальників, оскільки системи джерел даних стабільні, а сховище даних орієнтоване на процеси. Крім того, оптимізація запитів є простою, передбачуваною та контрольованою.

Основні хиби концепції сховища даних Кімбола такі:

– Дані не повністю інтегруються перед звітом; ідея “єдиного джерела істини втрачена”.

– Можуть виникнути певні помилки під час оновлення даних. Це пов’язано з тим, що з використанням методу денормалізації в таблиці бази даних додаються надлишкові дані.

– Можуть виникнути проблеми з продуктивністю через додавання стовпців до таблиці фактів, оскільки ці таблиці досить детальні. Додавання нових стовпців може розширити розміри таблиці фактів, що вплине на її продуктивність. Крім того, модель багатовимірного сховища даних ускладнює подальші зміни в ній.

– Оскільки модель Кімбола орієнтована на бізнес-процеси, а не на підприємство загалом, вона не може задовольнити всі вимоги до звітності.

– Процес долучення великих обсягів застарілих даних до сховища даних складний.

Обидва підходи, які розглянуті вище, порівнюючи один з одним, загалом мають свої переваги та хиби. Підхід Інмона забезпечує уніфіковане сховище даних, яке може підтримувати складні запити та надавати глибоку аналітику. Однак він потребує значних початкових інвестицій і може бути тривалим у реалізації. Підхід Кімбола дає змогу організаціям швидко розгортати функціональні елементи сховища даних і пропонувати негайні переваги для бізнесу. Однак він може призвести до відсутності уніфікації даних і нестабільності вітрин даних.

Оптимальний підхід залежить від конкретних потреб організації. Якщо організація потребує уніфікованого сховища даних для підтримки складних запитів, то підхід Інмона може бути найкращим вибором. Якщо ж організації потрібно швидше розгорнути функціональні елементи сховища даних і пропонувати негайні переваги для бізнесу, то підхід Кімбола може бути більш більш прийнятним.

Гібридний підхід побудови сховища даних

Є й інші підходи до побудови сховищ даних [13–15]. Наприклад, гібридний підхід побудови сховища даних [15] є інноваційним підходом, який узгоджує переваги визнаних методологій Інмона та Кімбола. Ця архітектура розробляє централізоване сховище даних, що відіграє роль головної точки істини для ефективної організації та використання як нормалізованих, так і денормалізованих даних залежно від конкретних потреб. Такий підхід надає гнучкість та адаптивність, особливо в контексті сховища даних для онлайн магазину, який взаємодіє з продажем та доставкою товарів клієнтам. Це революційне рішення дає змогу оптимально використовувати переваги обох методологій. Централізована точка істини забезпечує порядок та єдність в обробці даних, зокрема стабільну основу для високоефективного управління інформацією. Водночас гнучкість денормалізованих даних надає зручність в роботі з різноманітними бізнес-сценаріями та полегшує адаптацію до змінних вимог.

Гібридна модель [7] використовує схему зірки в побудові сховища даних. В її центрі розміщена центральна “фактична” таблиця, що представляє основні параметри, такі як продажі. Деталізовані дані, зокрема інформація про клієнтів чи товари, розташовані в окремих таблицях, які зв’язані з центральною таблицею через ключі. Це дає змогу уникнути дублювання даних, а також забезпечує зручність управління і аналізу. Всі дані про продажі зберігаються в денормалізованому вигляді, що полегшує підтримку запитів.

2. Постановка завдання

Основною метою цієї статті є аналіз наявних рішень побудови сховища даних, їх систематизація, виявлення переваг і недоліків та розроблення підходу до створення адаптивного та ефективного сховища даних, спрямованого не лише на їх зберігання, але й на обробку для подальшого аналізу і створення звітів. Розроблений метод побудови сховища даних має комбінувати переваги наявних методологій та забезпечувати оптимальну взаємодію з даними для всіх користувачів. Основні завдання дослідження – розгляд структури та особливостей

розробленого методу, аналіз його ефективності та гнучкості у роботі з даними, а також оцінка можливих економічних вигід та перспектив подальшого застосування в практиці.

3. Розроблення методу побудови сховища даних

Наведені вище підходи не можуть повною мірою забезпечити вимоги користувачів, які постійно зростають і які вони висувають до сховищ даних.

За результатами аналізу для покращення ситуації в цьому напрямку ми пропонуємо авторський гібридний метод, який відрізняється від інших відомих сьогодні. Цей підхід об'єднує обидві методології, що дає наступні переваги, використовуючи схему сніжинки для основних таблиць та денормалізацію для полегшення обробки даних, на відміну вже від наявних видів гібридного підходу, де використовують схему зірки. Наш підхід дає змогу зберігати загальну інформацію централізовано, а деталізовані дані розподіляються, забезпечуючи ефективність та гнучкість. Такий підхід оптимізує простір зберігання та полегшує управління даними у реальних умовах використання.

Результати порівняння за основними характеристиками наявних відомих підходів до створень сховищ даних із запропонованим нами наведено в табл. 2.

Таблиця 2

Порівняння підходів до створення сховищ даних

	Інмона	Кімбола	Гібридний	Авторський гібридний
1	2	3	4	5
Схема	*Сніжинка*	*Зірка*	*Зірка*	*Сніжинка*
Стан даних	Нормалізовані	Денормалізовані	Денормалізовані	Нормалізовані та денормалізовані
Управління зв'язками	Має багато зв'язків між таблицями	Мала кількість зв'язків між таблицями	Має багато зв'язків між таблицями	Для отримання головної інформації мала кількість зв'язків між таблицями, а для отримання детальнішої інформації велика кількість зв'язків
Зміна структури даних	Легше реагує на зміни у структурі даних	Може потребувати більше роботи під час зміни структури даних	Легше реагує на зміни у структурі даних	Легка підтримка змін структури, адже дані в нормалізованому вигляді
Підтримка таблиць та даних	Важко внести зміни в таблиці для отримання надійності та вірності даних	Легко внести зміни в таблиці для отримання надійності та вірності даних	Важко внести зміни в таблиці для отримання надійності та вірності даних	Легко внести зміни в таблиці для отримання надійності та вірності даних. Також легко перераховувати денормалізовані колонки, які мають згруповані дані
Час створення та моделювання	Потребує більше часу на проєктування та моделювання через нормалізовану структуру	Має простіший процес створення, оскільки використовує денормалізовану структуру	Потребує більше часу на проєктування та моделювання через нормалізовану структуру	На початку можна використовувати простішу схему, а після цього перейти на схему сніжинки
Підтримка звітності та аналітики	Важкість отримання даних у звітності через нормалізовану структуру та	Орієнтована на ефективність для аналітичних операцій та швидкий доступ	Орієнтована на ефективність для аналітичних операцій та швидкий доступ	Легкий доступ до даних, адже вони вже згруповані та лежать у фактичній таблиці і їх легко можна дістати

Продовження табл. 2

1	2	3	4	5
	багато зв'язків, але є невеликі витрати на групування та обчислення даних	до даних для аналітики	до даних для аналітики, але є невеликі витрати на групування та обчислення даних	
Розширюваність	Може потребувати більше зусиль для розширення та модифікації структури даних	Має більш просту структуру, але може бути менш гнучкою під час розширення або змін	Має простішу структуру, але може бути менш гнучкою під час розширення або змін	Може потребувати більше зусиль для розширення та модифікації структури даних, адже потрібно наперед продумати структуру, та стовпці, які будуть агреговані в основній таблиці

Для наглядного і реального порівняння запропонованого методу з відомими розробимо схеми для онлайн магазину у всіх методологіях: Кімбола, Інмона та авторського гібридного методу. Це дасть змогу краще зрозуміти переваги та недоліки у цих методологіях для реального застосування. Зв'язок таблиць для магазину за методологією Кімбола у схемі зірка наведено на рис. 3.

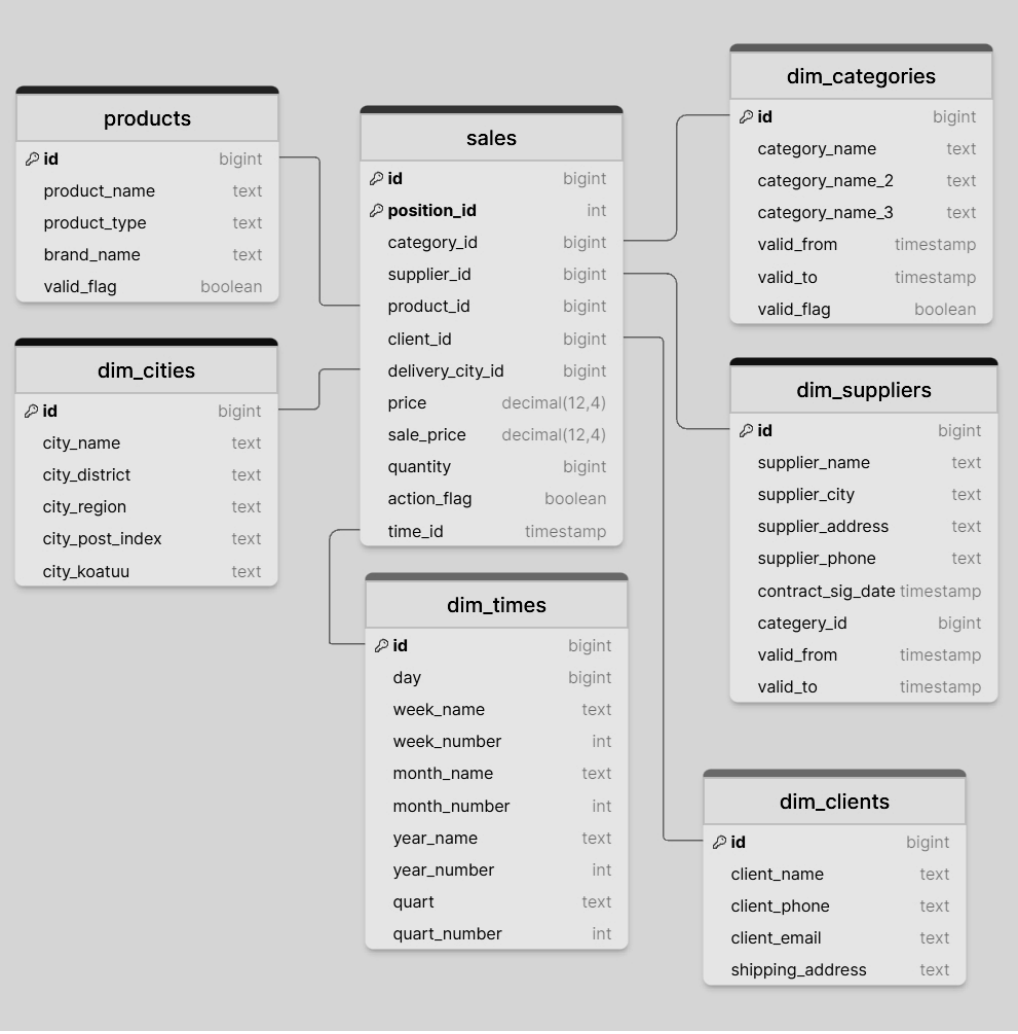

Рис. 3. Модель сховища даних за методологією Кімбола у схемі зірка (star schema)

Ця схема пропонує те, що кожен продаж буде записувати в таблицю “sales” та до кожного “id” буде “position_id”, яка відповідає за позицію товару в чеку. Всі інші дані заповнюються відповідно до схеми. Перед вставкою даних в основну таблицю “sales” всі додаткові дані (такі як товари, клієнти тощо) вставляються у відповідні довідники і пізніше з’єднуються з основною таблицею “sales”. Такий підхід дає можливість зберігати деталізовану інформацію та спрощує процес вставки даних в основну таблицю. Виникає додаткова складність управління багато-до-багатьох зв’язками між таблицями та потреба в оновленні кількох об’єктів під час внесення змін. Зв’язок таблиць для цього магазину за методологією Інмона у схемі сніжинка наведено на рис. 4.

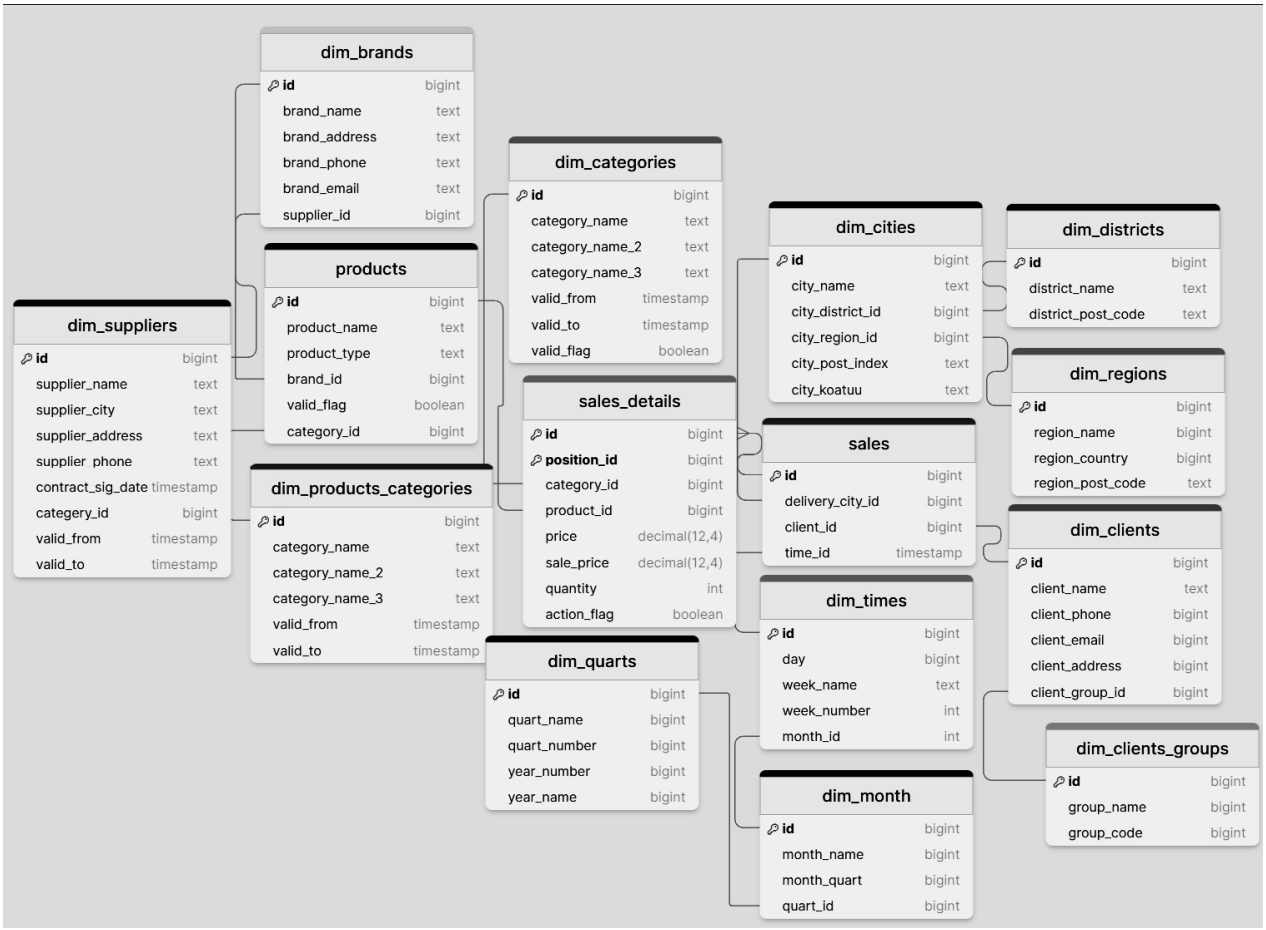

Рис. 4. Модель сховища даних за методологією Інмона у схемі сніжинка (snowflake schema)

На противагу схемі Кімбола, у таблиці “sales” зберігається лише загальна інформація про продаж, а всі деталі по кожній позиції в чеку йдуть в окрему таблицю “sales_details”. Таблиці використовуються у нормалізованому вигляді, що дає можливість ефективно управляти даними та забезпечує цілісність інформації. Нормалізація допомагає уникнути дублювання даних і забезпечує легкість у підтримці та зміні. Завдяки використанню нормалізованих таблиць виникає значна кількість зв’язків між різними таблицями. Кожен зв’язок відображає взаємозв’язок між даними, що може ускладнити дизайн бази даних та взаємодію між ними. Під час додавання нових продажів або оновлення наявних даних потрібно виконувати операції в таблицях, що може потребувати значних затрат ресурсів. А це своєю чергою може впливати на продуктивність та підтримку бази даних, особливо за великого обсягу транзакцій.

Розроблений авторський гібридний метод у сховищі даних є інтеграцією основних елементів підходів Інмона та Кімбола з використанням оптимальних аспектів кожного з них для досягнення

більшої ефективності та гнучкості в архітектурі. У конкретній реалізації запропонованого методу дані зберігаються за схемою сніжинки (snowflake schema), де у таблиці “sales” розміщується загальна інформація про продаж, така як ідентифікатор продажу, дата, клієнт тощо. Однак на відміну від схеми Кімбола, всі деталі щодо кожної позиції в чеку децентралізовано та зберігаються в окремій таблиці “sales_details”. Це дає можливість оптимізувати простір зберігання, забезпечуючи швидкий доступ до загальної інформації, а також полегшує управління та підтримку даних завдяки нормалізованій структурі. Такий підхід комбінує ефективність управління централізованим сховищем та гнучкість децентралізованих даних, що робить його потужним інструментом для різноманітних бізнес-сценаріїв. Зв’язок таблиць для цього магазину згідно з авторським гібридним методом наведено на рис. 5.

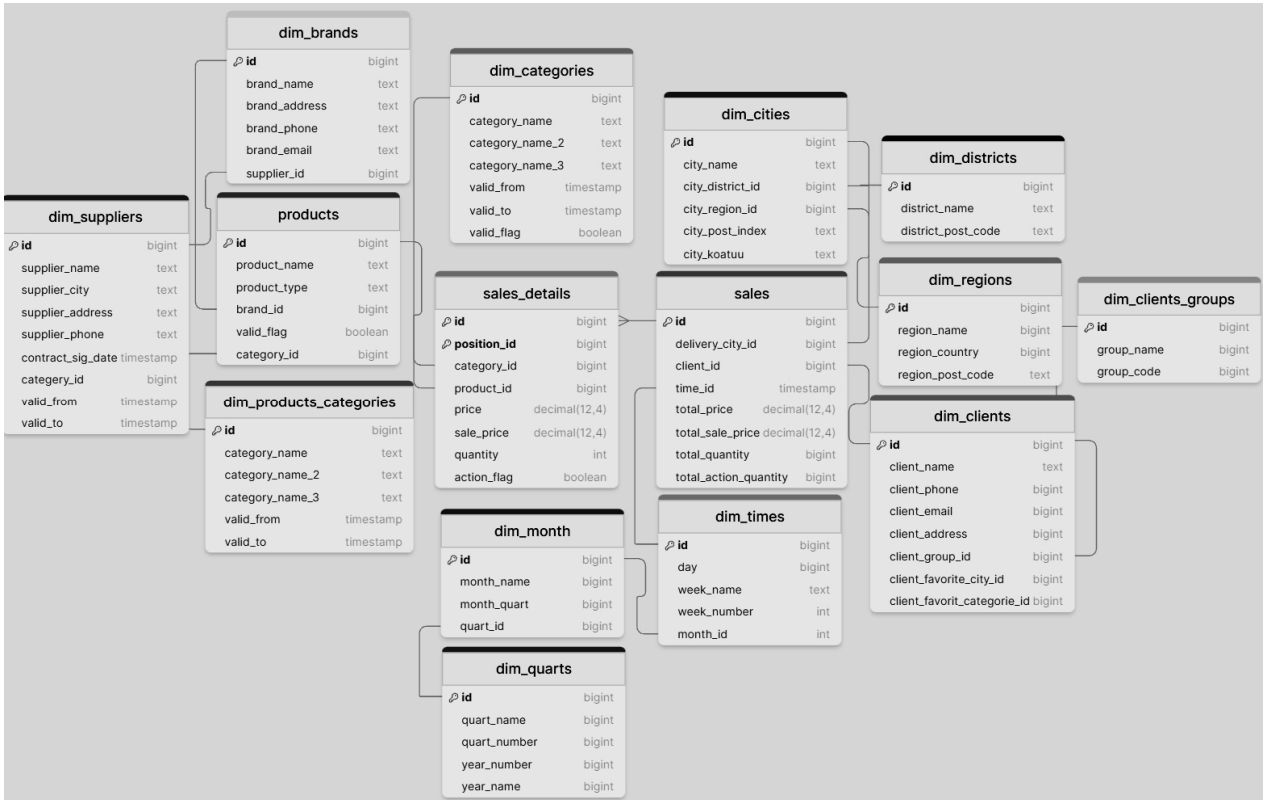

Рис. 5. Модель сховища даних за авторським гібридним методом

За результатами поданих прикладів можна виділити основні переваги та недоліки авторського гібридного методу порівняно з традиційними методами побудови та підтримки сховищ даних.

Основна перевага запропонованого методу полягає у більшій швидкодії розгортання системи. Це досягається завдяки обмеженій кількості зв’язків між таблицями, що значно полегшує процес проєктування таблиць і створення зв’язків та індексів. Такий підхід дає можливість оперативно налаштовувати базу даних для зберігання та обробки інформації. Важливо відзначити, що витягування даних з такого сховища потребує менше зусиль, оскільки не потрібно формувати складні запити до бази даних з великою кількістю зв’язків з іншими таблицями. Це сприяє ефективному використанню даних.

Однак разом із цією перевагою з’являється певний недолік – ускладнена підтримка такого сховища. Брак складних зв’язків допомагає прискорити процес розгортання, проте призводить до ускладнень під час внесення змін та розширення сховища даних. Наприклад, додавання нових колонок з інформацією може збільшити навантаження на базу даних, особливо якщо кількість записів до кожного id чеку зростає. Це може призвести до складнощів у підтримці та управлінні системою, особливо за великого обсягу транзакцій.

Ще однією важливою перевагою запропонованого авторського гібридного методу є зберігання агрегованих даних у загальних таблицях, таких як “sales” та “dim_clients”. Припустимо, що потрібно буде регулярно формувати звіти та створювати моделі для Data Science з метою передбачення різних метрик. Багато з цих звітів та моделей використовують загальні показники, такі як загальна сума продажу, загальна сума продажу зі знижкою або улюблене місто та категорія товарів покупця. У традиційних підходах щоразу під час створення звіту чи моделі доводилося б повторно розраховувати ці дані, що створює навантаження на базу даних та збільшує витрати ресурсів. Розроблений метод пропонує додати колонки “total_price” – загальна сума покупки клієнта всіх товарів, “total_sale_price” – загальна сума покупки клієнта акційних товарів у таблицю “sales”. Після занесення даних у таблицю “sales_details” оновлюються колонки “total_price”, “total_sale_price”, які розраховуються до кожного продажу. Аналогічно відбувається з колонками “client_favority_city_id” – найчастіше місце покупки клієнта за останній період та “client_favority_category_id” – найулюбленіша категорія товарів клієнта за останній час у таблиці “dim_clients”. Ці агреговані дані значно спрощують розробку вітрин, отримання даних для звітів та побудови Data Science моделей для навчання. Немає потреби повторного агрегування даних під час створення кожного звіту у розробленому методі, оскільки основні показники, які широко використовуються в звітах або моделях, розраховуються лише раз. Це істотно спрощує навантаження на базу даних, зменшуючи навантаження на систему та полегшуючи процес витягування потрібної інформації для аналізу.

Наступна важлива перевага запропонованого методу побудови сховищ даних – це оптимальний компроміс між традиційними методологіями Інмона та Кімбола, спрямований на оптимізацію продуктивності та ефективність запитів. Також однією з основних переваг є обмеження кількості зв’язків між таблицями, що спрощує структуру бази даних та полегшує виконання запитів. Зменшення зв’язків сприяє швидкому доступу до інформації та покращенню продуктивності системи. Порівнюючи запропонований метод з традиційними, варто відзначити, що ця модель дає можливість зберігати баланс між централізованим зберіганням даних та гнучкістю децентралізованої структури. Такий баланс особливо актуальний у великих сховищах даних, де швидкий доступ до інформації та оптимальна продуктивність важливі для успішної обробки великого обсягу інформації. Запропонований авторський гібридний метод надає гнучкість у роботі з даними, спрощуючи структуру, забезпечуючи ефективність в обробці запитів і дає можливість для легшого масштабування у майбутньому.

Остання перевага полягає в тому, що запропонований метод сприяє ефективному управлінню завданнями інтеграції та підтримки. Зберігання деталізованої інформації у нормалізованій формі дає змогу легко додавати нові колонки та розширювати сховище даних. Водночас наявність денормалізованих таблиць для основних показників забезпечує швидкий доступ до загальної інформації. Це робить розроблений метод потужним інструментом для адаптації до змін у вимогах бізнесу та легкої підтримки системи в часі.

Порівняно з методологією Кімбола, запропонований авторський гібридний метод вигідний завдяки меншій кількості зв’язків між таблицями, що полегшує дизайн бази даних та робить його менш складним. Оптимізація продуктивності та зменшення кількості зв’язків роблять запропонований підхід більш практичним у випадках, коли потрібно швидко налаштовувати та використовувати сховище даних.

Порівняно з методологією Інмона, авторський гібридний метод дає змогу зберігати деталізовану інформацію в нормалізованій формі, але водночас забезпечує швидкий доступ до основних показників через денормалізовані таблиці. Це робить його більш гнучким і ефективним у випадках, коли потрібно забезпечити швидкий доступ до основних показників без втрати можливості зберігання деталізованих даних у нормалізованому вигляді.

4. Результати дослідження

Результати нашого дослідження призвели до розробки нового та інноваційного підходу зі створення сховища даних, який ми називаємо "авторським гібридним методом". Цей метод вдало об'єднує переваги наявних методологій, таких як Інмона та Кімбола, забезпечуючи централізоване зберігання основної інформації в центральній таблиці або сховищі даних, яке містить основні дані для аналізу та звітності. Деталізовані дані про товари, клієнтів або інші сутності розподіляються в нормалізованих таблицях, що дає змогу ефективно керувати обсягом даних та забезпечує більшу гнучкість під час обробки та аналізу інформації. Його основною особливістю є зменшена кількість зв'язків між таблицями, що спрощує дизайн бази даних та робить його менш складним. Він не лише сприяє швидкому доступу до інформації та покращенню продуктивності системи, але й відзначається швидким розгортанням системи завдяки обмеженій кількості зв'язків між таблицями. Використання цього методу у побудові сховища даних дає змогу оперативно конфігурувати базу даних для ефективного зберігання та обробки інформації.

Для розробленого методу характерна висока гнучкість у роботі з даними, що дає можливість легко вносити зміни та адаптуватися до різних потреб бізнесу. Гнучкість полягає в можливості змінювати та розширювати сховище даних без значного впливу на його продуктивність чи структуру. Це дає змогу бізнес-користувачам швидко адаптувати систему до нових вимог та використовувати її в різноманітних сценаріях використання. Спрощена структура бази даних в авторському гібридному методі робить її більш зрозумілою та нею легше керувати. Зменшення кількості зв'язків між таблицями полегшує дизайн та підтримку бази даних. Це сприяє підвищенню продуктивності розробників та адміністраторів баз даних. Ефективність обробки запитів забезпечує швидкий доступ до інформації. Зменшення кількості зв'язків сприяє оптимізації виконання запитів та забезпеченню ефективнішого використання ресурсів системи.

Висновки

Запропонований авторський гібридний метод є перспективним та ефективним рішенням для компаній, що шукають оптимальний баланс між витратами на розгортання та управління інфраструктурою, часом розробки та впровадження, а також аналізом складності та гнучкості системи. Цей метод може забезпечити значні економічні вигоди, наприклад, вартість розгортання може бути на 30 % меншою порівняно з аналогічними рішеннями. Крім того, швидкість роботи системи може бути вдвічі вищою, а обсяг обробки даних може зростати в 1,5 раза порівняно з іншими рішеннями. Важливо відзначити, що цей метод відкриває можливості для подальшого масштабування, що дає компаніям можливість адаптуватися до потреб бізнесу, які зростають.

Список літератури

1. Minukhin S., Fedko V., & Gnusov Y. Enhancing the performance of distributed big data processing systems using Hadoop and PolyBase. Eastern-European Journal of Enterprise Technologies, 4(2–94), (2018), pp. 16–28. DOI:10.15587/1729-4061.2018.139630.

2. Praveen Kumar, Dr. Kavita The Study On Data Warehousing Different Concepts, Vol. 21, No. 16, (2019), pp. 3103–3109. Available at: http://gujaratresearchsociety.in/index.php/JGRS/article/view/3497 (Accessed: 10 March 2024).

3. Inmon W. H. Building the Data Warehouse,3rd Edition (3rd. ed.). John Wiley & Sons, Inc., USA. 2002. Avaible at: https://fit.hcmute.edu.vn/Resources/Docs/SubDomain/fit/ThayTuan/DataWH/Bulding%20the%20ata%20Warehouse%204%20Edition.pdf (Accessed: 10 March 2024)

4. Bhatia P. (2019). Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press, Cambridge. DOI:10.1017/9781108635592

5. Padmaja Potinen/ Oracle Database Data Warehousing Guide, 21c. Copyright © 2001, 2022, Oracle and/or its affiliates. Available at https://docs.oracle.com/en/database/oracle/oracle-database/21/dwhsg/preface.html#GUID-9CDC42C7-5BB2-4433-9F3E-ADE92929A0EA (Accessed: 10 March 2024).

6. Simitsis A., Skiadopoulos S., & Vassiliadis P. The History, Present, and Future of ETL Technology. Proceedings of the 25th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP)

co-located with the 26th International Conference on Extending Database Technology and the 26th International Conference on Database Theory (EDBT/ICDT 2023), Ioannina, Greece, March 28, 2023. Available at: https://ceur-ws.org/Vol-3369/invited1.pdf (Accessed: 10 March 2024).

7. Jaganathan Manonmani, Arun. Research Paper: The role of Software architecture for the design of scalable and secure Bigdata in Banking Sectors The role of Software architecture for the design of scalable and secure Bigdata in Banking Sectors. (2023). Available at: https://www.researchgate.net/publication/371491773_Survey_Paper_The_role_of_Software_architecture_for_the_design_of_scalable_and_secure_Bigdata_in_Banking_Sectors_The_role_of_Software_architecture_for_the_design_of_scalable_and_secure_Bigdata_in_Banki (Accessed: 10 March 2024).

8. Building a data warehouse: A step-by-step guide. Available at: https://www.n-ix.com/building-a-data-warehouse/ . (Accessed: 25 February 2024).

9. Aberer K., Hemm K. A Methodology for Building a Data Warehouse in a Scientific Environment, Cooperative Information Systems, 1996. Proceedings., First IFCIS International Conference, DOI: 10.1109/COOPIS.1996.555001

10. Manole V., Matei G. Building a Data Warehouse step by step DOAJ, 2007. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1028461 (Accessed: 10 March 2024).

11. Gardner D. R. Building the Data Warehouse Communications Of The Acm September 1998/Vol. 41, No. 9. Available at: https://web.archive.org/web/20060519201128id_/http://www.csun.edu:80/~chchen/ Appalachian/Database/Data%20Warehouse/Building%20a%20Data%20Warehouse.pdf (Accessed: 10 March 2024).

12. Bal B. Building Machine Learning Warehouse -A Myth or Reality 2023. Available at: https://www.researchgate.net/publication/370902095_Machine_Learning_Warehouse_-A_Myth_or_Reality (Accessed at 10 March 2024).

13. Vaisman Alejandro Ariel and Esteban Zimányi. "ata Warehouse Systems: Design and Implementation. Data Warehouse Systems (2022): n. pag. *DOI:10.1007/978-3-662-65167-4*

14. Томашевський В. М. Особливості проектування гібридних сховищ даних з врахуванням джерел даних / В. М. Томашевський, А. Ю. Яцишин // Вісник Національного університету "Львівська політехніка". 2011. № 715 : Інформаційні системи та мережі. С. 246–254. Available at: https://science.lpnu.ua/uk/sisn/vsi-vypusky/vypusk-715-2011/osoblyvosti-proektuvannya-gibrydnyh-shovyshch-danyh-z-vrahuvannyam (Accessed: 10 March 2024).

15. El Moukhi N., El Azami I., Hajbi S. Towards a new hybrid approach for building document-oriented data warehouses. International Journal of Electrical and Computer Engineering (IJECE) 12(6), 2022. DOI: 10.11591/ijece.v12i6.pp6423-6431

DEVELOPMENT OF A HYBRID METHOD FOR DATA WAREHOUSE CONSTRUCTION

O. Koval, O. Harasymchuk

Lviv Polytechnic National University,
Information Security Department
E-mail: oleh.koval.kb.2020@lpnu.ua, oleh.i.harasymchuk@lpnu.ua

The examined approach to building an adaptive and convenient data warehouse goes beyond simple data storage, focusing on processing data for various types of reports. A hybrid concept for constructing the data warehouse is proposed, combining the advantages of existing methodologies and providing optimal interaction with the data warehouse for all users. The proposed method allows for the quick deployment, adaptive maintenance, and easy scalability of the data warehouse in the future.

This proprietary hybrid concept includes centralized data storage and the distribution of detailed data into small lookup tables for improved and faster warehouse functionality. The method is also characterized by a faster system deployment due to a limited number of connections between tables.

Using an online store as an example, all the benefits of the proposed method were demonstrated, both in data recording and processing, as well as in utilizing the data for future reports. The approach proves to be a promising and effective solution for companies seeking an optimal compromise between traditional methodologies and modern business requirements.

Keywords: Data Warehouse, concept Inmon, concept Kimbal, hybrid concept.