УДК 004.624

Д.Д. Рахмани, старший преподаватель кафедры «Сетевых информационные технологии и сервисы», Московский технический университет связи и информатики, г. Москва

Ю.С. Дружинин, студент, 4 курс, факультет «Информационных технологии», Московский технический университет связи и информатики, г. Москва

К. А. Михайлов, студент, 4 курс, факультет «Информационных технологии», Московский технический университет связи и информатики, г. Москва

ТЕНДЕНЦИИ РАЗВИТИЯ И ВНЕДРЕНИЯ DWH В ИНФРАСТРУКТУРУ КОМПАНИЙ

Аннотация. *В статье рассматриваются современные тенденции развития и внедрения систем хранения данных (Data Warehouse, DWH) в инфраструктуру компаний. Актуальность исследования обусловлена возрастающей потребностью организаций в эффективной обработке, хранении и анализе больших объемов информации для принятия управленческих решений. В работе представлены архитектурные подходы к построению DWH, их виды и особенности применения. Особое внимание уделяется влиянию внедрения DWH на оптимизацию бизнес-процессов, аналитическую деятельность и визуализацию данных. В результате исследования определены ключевые направления развития DWH и сформулированы рекомендации по их интеграции в бизнес-процессы компаний различных масштабов и отраслей, что способствует повышению эффективности и конкурентоспособности организаций.*

Ключевые слова: *DWH, Data Warehouse, хранилище данных, бизнес процессы, ETL, бизнес-аналитика.*

Abstract. *The article examines current trends in the development and implementation of Data Warehouse (DWH) systems in company infrastructures. The relevance of the study is determined by the growing need of organizations for efficient processing, storage, and analysis of large volumes of information to support*

managerial decision-making. The paper presents architectural approaches to DWH design, their types, and specific features of application. Particular attention is given to the impact of DWH implementation on business process optimization, analytical activities, and data visualization. As a result of the study, key directions in the development of DWH are identified, and recommendations for their integration into the business processes of companies of various scales and industries are formulated, contributing to improved efficiency and competitiveness of organizations.

Keywords: *DWH, Data Warehouse, data storage, business processes, ETL, business analytics.*

1. Введение

В мире жёсткой конкуренции успех бизнеса во многом зависит от быстроты и эффективности принятия решений. Из-за развития технологий и постоянного увеличения объёма данных бизнес вынужден находить эффективные способы управления этими данными. DWH (Data Warehouse, хранилище данных) – это не просто место для хранения информации, но и фундамент для аналитики и принятия бизнес-решений. Цель статьи – комплексное рассмотрение теоретических основ, современных трендов и практических аспектов внедрения DWH в инфраструктуру компаний.

2. Что такое DWH

Система хранение данных (Data Warehouse, DWH) - это централизованная платформа для сбора, хранения и анализа больших объемов данных из различных источников. Основная цель DWH заключается в создании единого источника «истины» (single source of truth), который обеспечивает доступность консистентных и структурированных данных для аналитических задач и стратегического принятия решений в компании [1].

Современное хранилище данных, как правило, состоит из следующих основных логических слоёв:

- Слой источников данных (Data Sources Layer) - включает разнообразные внешние и внутренние источники: CRM, ERP, веб-приложения, базы данных, файлы Excel, API и т. д.
- Слой извлечения, трансформации и загрузки (ETL/ELT Layer) - здесь данные проходят обработку: извлекаются из источников, преобразуются к унифицированному виду и загружаются в хранилище. В классической ETL-парадигме обработка данных происходит до загрузки. В ELT-парадигме данные сначала загружаются, затем обрабатываются в самом DWH.
- Слой хранения данных (Data Storage Layer)
  Это центральный компонент — база данных, содержащая очищенные, агрегированные и исторические данные, готовые для анализа. Данные обычно хранятся в многомерной или звездной схеме.
- Слой представления и анализа (BI/Analytics Layer) - используется для построения отчетов, дашбордов и аналитики. Включает BI-инструменты: Power BI, Tableau, Qlik, Grafana и др [2].

Существует несколько типов хранилищ данных. Они отличаются в архитектурной организации:

- Классическое (on-premise) хранилище данных - DWH разворачивается на физических серверах внутри организации. Такой подход требует значительных инвестиций в оборудование и поддержку инфраструктуры. Это предоставляет полный контроль над системой и безопасностью данных. Также отсутствуют зависимости от сторонних провайдеров, что обеспечивает независимость системы от внешних источников и позволяет гибко настраивать сервис для интеграции с внутренними сервисами инфраструктуры.
- Облачное хранилище данных (Cloud DWH) - Решение размещается в облаке, что существенно снижает затраты на оборудование. При

использовании облачного сервиса можно автоматизировать резервное копирование, скорость разворачивания системы высокая [5].

- Гибридное хранилище данных (Hybrid DWH) - сочетание локальной и облачной инфраструктуры. Данные могут храниться и обрабатываться частично в локальных серверах, частично — в облаке. Это позволяет распределять нагрузку (load balancing), отделять сущности в логики процессов друг от друга [4].

DWH может включать в себя большое количество отдельных компонентов, но самая стандартная реализация - база данных, разделенная на слои. Каждый слой это логическая составляющая DWH, в зависимости от того, какого типа данные в этот слой собираются. К примеру:

- raw - это слой, в котором хранятся сырые данные, которые никак не изменялись
- ods/dds - это слой, в котором данные были каким-то образом обработаны и приведены к единому формату
- cdm - это слой витрины (слой представления), в который попадают данные, которые можно использовать в представлениях (к примеру, для визуализации на графиках).

3. Сравнительный анализ DWH, Data Lake, Database

В рамках архитектуры хранения и обработки данных различают несколько ключевых подходов, помимо DWH: Database и Data Lake. Каждый из них решает определенные задачи, связанные с хранением, обработкой и анализом данных.

Database (база данных) является одним из элементов DWH, предназначенным для хранения данных в упорядоченном виде. Базы данных бывают реляционными (SQL) и нереляционными (NoSQL), а их основная цель — обеспечение надежного и быстрого доступа к информации. В рамках ETL (Extract, Transform, Load) данные извлекаются из источников, трансформируются и загружаются в базы данных, обеспечивая их

структурированное хранение. Существуют также альтернативные подходы, такие как ELT (Extract, Load, Transform), при которых данные сначала загружаются в промежуточное хранилище, а затем подвергаются обработке.

Data Lake представляет собой архитектурное решение, в котором данные из различных источников загружаются в сырую, неизмененную форму. Это позволяет сохранять всю первичную информацию, обеспечивая гибкость для последующей обработки, нормализации, агрегации и компоновки. В отличие от DWH, Data Lake не требует жестко заданной структуры данных, что делает его более подходящим для работы с неструктурированной информацией, машинного обучения и анализа больших данных (Big Data). Однако при отсутствии необходимости хранения больших объемов исходных данных применение Data Lake может быть неоправданно с точки зрения затрат и сложности администрирования.

Data Lake и DWH предназначены для хранения данных из различных источников, но имеют разные цели и принципы организации.

- Data Lake аккумулирует структурированные и неструктурированные данные, обеспечивая возможность их обработки для задач машинного обучения и продвинутой аналитики.
- DWH ориентирован на обработку структурированных данных с четкой схемой, что делает его более удобным для бизнес-аналитики и отчетности. Data Lake предоставляет высокую гибкость, но требует значительных вычислительных мощностей, в то время как DWH обеспечивает эффективное и стандартизированное управление бизнес-данными.

Хотя и DWH, и традиционные базы данных предназначены для хранения структурированных данных, они различаются по масштабу и назначению.

- База данных обычно используется в рамках отдельных приложений или монолитных сервисов, обеспечивая оперативную обработку данных.
- DWH разрабатывается для работы с данными, поступающими из множества источников внутри организации, и поддерживает аналитические запросы на больших объемах данных.

На рисунке 1 представлен график сравнения DWH и Data Lake по критериям. Синяя линия – DWH, зелёная линия – Data Lake. Критерии:

1. Производительность (ETL/ELT задачи)
2. Гибкость хранения данных
3. Поддержка структурированных данных
4. Поддержка неструктурированных данных
5. Время извлечения данных для аналитики
6. Стоимость владения
7. Масштабируемость

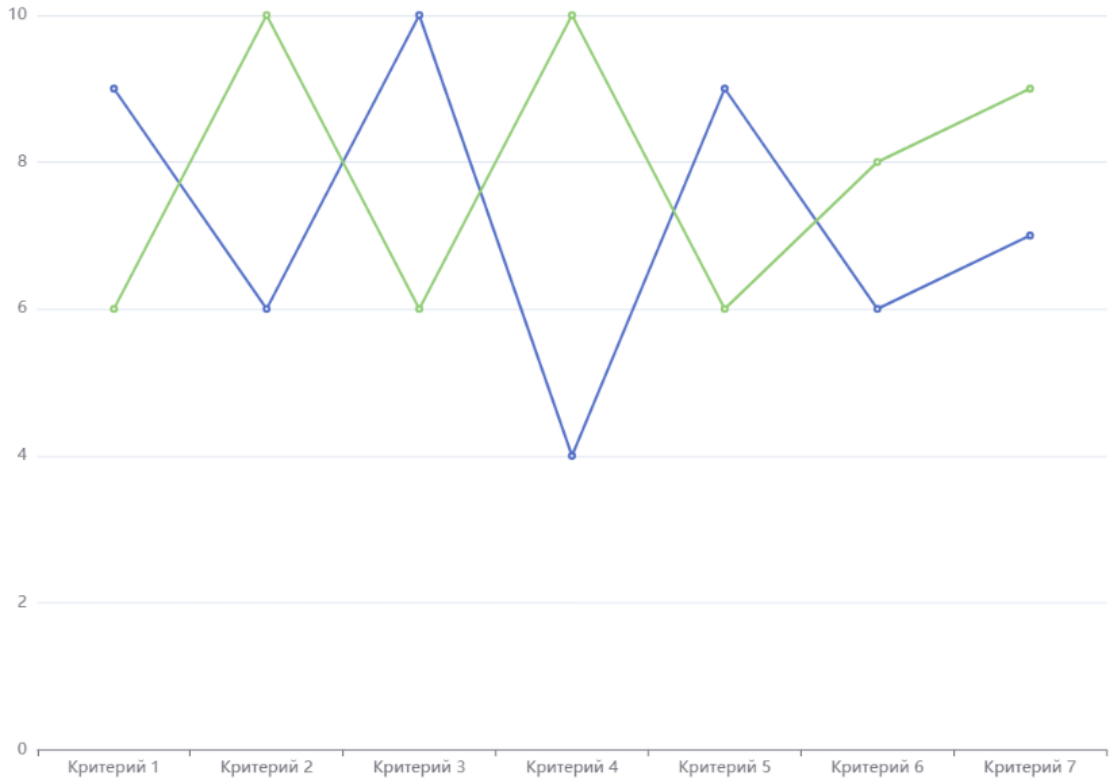

Рисунок 1 – График сравнения DWH и Data Lake

График, представленный на рисунке 1, иллюстрирует сравнительный анализ систем Data Warehouse (DWH) и Data Lake по семи ключевым критериям эффективности: производительности, гибкости хранения данных, поддержке структурированных и неструктурированных данных, времени извлечения информации для аналитики, стоимости владения и масштабируемости.

Data Warehouse демонстрирует уверенное превосходство в задачах, где важны высокая производительность ETL/ELT-процессов (оценка 9 из 10), качественная поддержка структурированных данных (10 из 10), а также минимальное время отклика при аналитических запросах (9 из 10). Это делает его оптимальным решением для зрелых бизнес-процессов и управленческой отчетности, особенно в отраслях с жесткими требованиями к целостности и доступности данных.

В то же время Data Lake показывает высокую эффективность в гибкости хранения (10 из 10), работе с неструктурированными и полуструктурированными данными (также 10 из 10), а также в масштабируемости и экономичности (8 и 9 баллов соответственно). Эти особенности делают его особенно привлекательным для сценариев, связанных с машинным обучением, Big Data, потоковой аналитикой и интеграцией IoT-источников.

Важно подчеркнуть, что хотя Data Warehouse выигрывает в стабильности и скорости аналитики, Data Lake предоставляет гораздо более широкие возможности в плане хранения разнообразных форматов данных и масштабируемости с минимальными затратами.

Таким образом, выбор между этими архитектурами должен опираться на бизнес-цели: для классической отчетности и BI-платформ предпочтителен DWH, а для гибких, исследовательских и экспериментальных сред — Data Lake. В гибридных архитектурах (Data Lakehouse) все чаще происходит слияние этих подходов, что позволяет достичь компромисса между скоростью, гибкостью и экономичностью.

4. Скорость обработки данных в DWH и DataLake

Если рассматривать DWH и Data Lake, можно оценить временные задержки (t) при обработке данных с помощью формул 1 и 2:

$$t_{DWH} = f_{index} + f_{aggregation} \quad (1)$$

$$t_{DL} = f_{scan} + f_{ETL} \quad (2)$$

где:

- f_{index} — время индексации данных.
- $f_{aggregation}$ — время агрегации в хранилище.
- f_{scan} — полное сканирование данных в Data Lake.
- f_{ETL} — время предварительной обработки.

Если $f_{index} + f_{aggregation} < f_{scan} + f_{ETL}$, то DWH будет быстрее.

Пример расчета:

Допустим:

- Индексация в DWH занимает 2 сек.
- Агрегация — 3 сек.
- Сканирование в Data Lake — 6 сек.
- ETL в Data Lake — 4 сек.

$t_{DWH} = 2 + 3 = 5$

$t_{DL} = 6 + 4 = 10$

На основе приведённых расчётов можно сделать вывод, что при заданных параметрах DWH обеспечивает более высокую скорость обработки данных по сравнению с Data Lake. Это связано с тем, что хранилище данных (DWH) изначально оптимизировано под аналитические запросы: оно использует индексацию и предварительно агрегированные структуры, что позволяет существенно сократить время отклика. В рассмотренном примере общее время обработки в DWH составляет 5 секунд, в то время как в Data Lake — 10 секунд. Разница объясняется архитектурными особенностями: DWH хранит структурированные данные с чёткими схемами и индексами, что делает возможным выполнение целевых аналитических операций быстрее. В то же время Data Lake чаще используется для хранения неструктурированных и полуструктурированных данных без предварительной подготовки, из-за чего любые аналитические задачи требуют сначала полного сканирования и ETL-подготовки, что увеличивает задержки. Таким образом, при сценариях, где важна скорость получения аналитического результата, DWH оказывается

предпочтительнее, особенно в случаях, когда структура и объём обрабатываемых данных заранее известны.

5. Применение DWH в инфраструктуре компаний

Современные исследования подтверждают, что внедрение Data Warehouse (DWH) трансформируется из технологического решения в стратегический фактор конкурентного преимущества. Эмпирические данные свидетельствуют о значительном улучшении ключевых показателей эффективности управления при корректной реализации DWH-решений: точность стратегических решений повышается на 23-41%, а скорость реагирования на рыночные изменения увеличивается в 1,7-2,3 раза (по данным McKinsey, 2023) [15].

Отраслевые кейсы применения:

1. Финансовый сектор:

В банковской сфере DWH-системы позволяют интегрировать разнородные данные из CRM, транзакционных систем и внешних источников. Строить предиктивные модели кредитных рисков (AUC 0,81-0,89). Оптимизировать кросс-селлинг за счет кластерного анализа клиентской базы.

2. Промышленные предприятия:

Анализ практики внедрения показывает снижение логистических затрат на 12-18%. Оптимизацию уровня запасов (снижение на 23-27 дней оборачиваемости). Повышение точности производственного планирования (MAPE < 8,5%).

3. Государственный сектор:

Реализация DWH в госучреждениях демонстрирует улучшение мониторинга бюджетных расходов (снижение нецелевого использования средств на 15-20%). Повышение прозрачности принятия решений. Возможность построения комплексных социально-экономических моделей.

Процесс реализации DWH требует последовательного выполнения этапов:

- Анализ требований и построение концептуальной модели (CRISP-DM)
- Разработка звездочной или снежинкообразной схемы данных
- Реализация ETL-процессов с учётом требований ACID
- Валидация качества данных (метрики DQ)
- Постреализационный мониторинг (KPI: data freshness, query performance, retention)

Критическим фактором успеха является формирование междисциплинарной команды, сочетающей экспертизу в предметной области и технических аспектах реализации.

Компания Qlever занимается внедрением DWH в различные компании и демонстрирует существенный прирост в оптимизации и производительности внутренний процессов компаний.

Например, после внедрений в компанию Orby корпоративного хранилища данных, которое включало в себе сырые и обработанные данные из маркетплейсов, CRM систем выросли следующие показатели:

- Экономия 3-х часов работы, ранее затрачиваемых на сбор таблиц и подсчет метрик в Excel;
- Своевременное выставление скидок и управление ассортиментов для повышения конверсии продаж;
- Устранение 80% ошибок, которые ранее возникали при планировании отгрузок товаров на маркетплейсы;
- Оптимизированная логистика между складами 2-х маркетплейсов для своевременной транспортировки излишек товаров между складами.

Рекомендации по выбору DWH-решений должны учитывать специфику отрасли, объем и структуру данных, а также бизнес-цели компании. Ниже приведены основные ориентиры, которые помогут компаниям различных секторов экономики подобрать подходящее хранилище данных:

1. Ритейл и e-commerce

   Рекомендовано использование облачных DWH (например, Google BigQuery, Snowflake), которые позволяют обрабатывать данные в режиме реального времени и масштабируются под сезонные пики. Ключевые задачи — консолидация данных из CRM, маркетплейсов, систем лояльности и аналитика продаж по SKU и каналам. Приоритетом должно быть наличие API-интеграций с внешними источниками и гибкие механизмы расчёта витрин.

2. Финансовый сектор

   Для банков, страховых компаний и финтеха важны безопасность, соответствие требованиям законодательства (например, 152-ФЗ, GDPR), возможность построения отчётности для ЦБ и автоматизированный расчёт финансовых показателей. Рекомендуется использовать решения с высокой степенью кастомизации и контролем доступа, такие как Microsoft SQL Server DWH или Oracle Exadata.

3. Производственные предприятия

   Оптимально применение решений с глубокой интеграцией в ERP и MES-системы (например, SAP BW/4HANA), способных обрабатывать телеметрию, логистические данные и сведения о поставках в реальном времени. DWH должен поддерживать агрегацию больших объёмов показателей по цехам, линиям и сменам для целей производственного контроля и прогнозирования.

4. Телеком и IT-компании

   Критична работа с потоковыми данными, логами и событиями. Эффективными будут DWH-платформы, поддерживающие распределённые вычисления и real-time ingestion (например, ClickHouse, Apache Druid). Такие решения позволяют проводить поведенческий анализ пользователей, мониторинг качества сервиса и оперативную реакцию на аномалии.

5. Государственные структуры и здравоохранение
   Требуются решения с повышенными требованиями к надёжности и защищённости данных, сертифицированные в соответствии с госстандартами (например, PostgreSQL с настройками отказоустойчивости или отечественные разработки, такие как Ланит БИ). DWH должны быть способны агрегировать данные из разрозненных ведомств, вести контрольные срезы и поддерживать историчность данных.

Независимо от отрасли, критически важно формировать междисциплинарную команду, сочетающую экспертизу в предметной области и технических аспектах реализации. Именно взаимодействие бизнес-аналитиков, инженеров данных и архитекторов позволяет настроить DWH так, чтобы он решал реальные задачи бизнеса и масштабировался по мере его роста.

6. Интеграция AI/ML в DWH

В 2024–2025 годах российские компании активно внедряют передовые технологии в области хранилищ данных (DWH), включая интеграцию искусственного интеллекта (AI/ML) и квантовых вычислений. Ниже представлены ключевые примеры и тенденции:

Сбербанк активно внедряет AI/ML в свои DWH-системы, особенно в блоке «Сеть продаж». Это позволяет автоматизировать процессы, улучшать клиентский опыт и повышать эффективность бизнес-решений.

Компания норильский никель создала единую ML-платформу, интегрированную с DWH, что позволило сократить время вывода ML-моделей в продакшн и повысить прозрачность процессов для специалистов по данным.

X5 Digital использует интеллектуальную систему Skillaz для автоматизации массового подбора персонала. Это позволило сократить время на найм и повысить конверсию из заявки в наём.

Компания видеоматрикс разрабатывает решения на базе AI/ML для промышленной видеоаналитики, что способствует повышению качества продукции и снижению производственных затрат.

Технологический прорыв и активное внедрение DWH в компаниях не обошли стороной также квантовые вычисления.

Корпорация «Росатом» заявила о планах создания первого отечественного квантового компьютера, что откроет новые возможности для обработки больших данных в DWH.

Компанией «Университет Иннополис» была разработана облачная платформа для квантовых вычислений, которая ускорит разработку и тестирование квантовых алгоритмов, потенциально применимых в DWH-системах.

АО «СМАРТС» и Университет ИТМО совместно реализуют проект по созданию системы управления географически распределёнными центрами обработки данных с использованием квантовых технологий для защиты линий связи.

Таким образом, отечественные компании активно внедряют AI/ML и квантовые технологии в свои DWH-системы, стремясь повысить эффективность, безопасность и конкурентоспособность в условиях цифровой трансформации.

7. Методы оптимизации и управления DWH

1. Ключевые аспекты оптимизации

Эффективность работы хранилищ данных определяется комплексом технологических решений, среди которых особое значение имеют:

- Индексирование и партиционирование данных для ускорения выполнения запросов;

- Сжатие информации с сохранением семантической целостности;

- Распределённые вычисления (на платформах Hadoop, Spark) для обработки больших массивов данных;

- In-memory технологии (например, SAP HANA) для оперативной аналитики.

Современные ETL-инструменты (Informatica PowerCenter, Talend, SSIS) обеспечивают:

- Автоматизацию загрузки данных с контролем качества (DQ);
- Минимизацию временных затрат на трансформацию данных;
- Снижение ошибок при миграции информации (до 0.01% от общего объема).

2. Интеграция AI/ML в DWH [3]

Применение методов машинного обучения расширяет функциональность DWH:

- Кластерный анализ (k-means, DBSCAN) для сегментации пользователей;
- Прогнозные модели (линейная регрессия, XGBoost) для анализа KPI;
- Нейросетевые алгоритмы для выявления аномалий в данных.

Системы мониторинга (Grafana, Prometheus) позволяют:

- Отслеживать производительность DWH в реальном времени
- Автоматически детектировать сбои (F1-score ≥ 0.95)
- Прогнозировать нагрузку на систему (RMSE < 5%)

3. Управление качеством и безопасностью данных

Критически важными являются:

- Контроль качества данных (DQ):
    - Стандартизация форматов (ISO 8000);
    - Очистка от дубликатов и аномалий;
    - Валидация по бизнес-правилам.
- Защита информации:
    - Шифрование (AES-256) при передаче и хранении;
    - Ролевая модель доступа (RBAC);
    - Аудит изменений (на основе блокчейн-технологий).

Математическое обоснование: допустим, у нас есть реляционная база данных с 1 млрд строк. Время выполнения запроса T зависит от:

- Способа индексирования (I)
- Наличия партиционирования (P)

- Использования кеширования (C)

Модель можно представить, как на формуле 3:

$$T = \frac{N}{I + P + C} \quad (3)$$

где:

- N — число строк в таблице.
- I — коэффициент ускорения индексации.
- P — коэффициент партиционирования.
- C — коэффициент кеширования.

Практический расчет: Допустим, без оптимизации запрос выполняется за 10 секунд. Введем индексы (I = 3), партиционирование (P = 2) и кеширование (C = 2):

$$T = \frac{1000000000}{3 + 2 + 2} = 142857142$$

Это дает прирост производительности почти в 7 раз.

Оптимизация DWH требует комплексного подхода, сочетающего технологические инновации, строгий контроль качества данных и современные методы защиты информации. Реализация этих мер позволяет повысить эффективность аналитических процессов на 30-40% (по данным Gartner, 2023) [13].

8. Сравнительный анализ гибридной и облачной архитектур

Современные хранилища данных развиваются в двух ключевых направлениях: полностью облачные архитектуры и гибридные решения, сочетающие локальные и облачные компоненты. Каждое из этих направлений имеет свои преимущества и ограничения, влияющие на эффективность эксплуатации в корпоративной среде.

Облачная DWH-архитектура предоставляет довольно высокий уровень гибкости и масштабируемости. Масштабируемость очень важна для организаций, которые быстро растут. Помимо этого, облачные решения имеют

предустановленные механизмы устойчивости к сбоям и оплату по факту использования. Данная архитектура имеет свои минусы. При постоянной нагрузке данный подход может привести к значительным расходам.

Когда бизнесу необходим полный контроль над данными, соответствие регуляторным требованиям и возможность использовать локальные ресурсы, на помощь приходит гибридная архитектура. Данный подход позволяет обрабатывать конфиденциальные и ресурсоёмкие данные на собственных серверах, а облачные мощности применять для менее чувствительных данных и задач. Помимо этого, упрощается интеграция с внутренними системами. Как облачная, так и гибридная архитектура имеет свои минусы. Например, гибридная архитектура требует большего объёма ресурсов, так как требует точной настройки взаимодействия между локальной и облачной средой.

На рисунке 2 представлен график сравнения архитектур по критериям. Красная линия – облачная архитектура, синяя линия – гибридная архитектура. Критерии:

1. Гибкость и масштабируемость;
2. Затраты;
3. Производительность;
4. Безопасность и контроль;
5. Интеграция с ИТ-ландшафтом;
6. Отказоустойчивость;
7. Скорость внедрения.

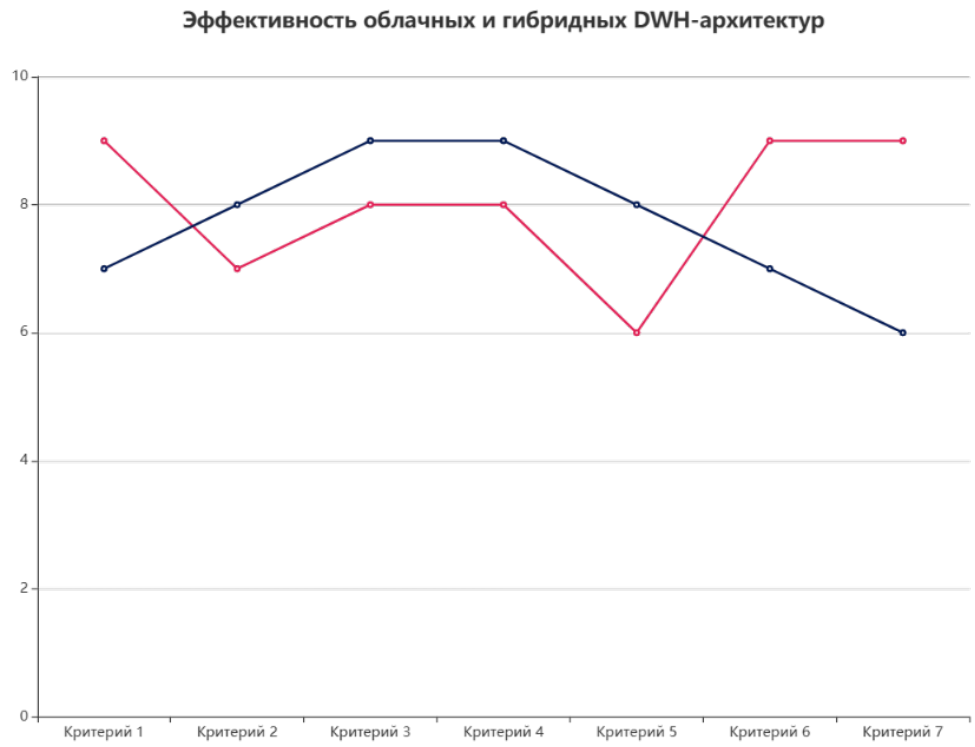

Рисунок 2 – График сравнения гибридной и облачных архитектур

Облачные архитектуры получают наивысшие оценки по таким параметрам, как скорость развертывания (9 из 10) и гибкость масштабирования (10 из 10), что делает их особенно привлекательными для быстрорастущих организаций и стартапов. Кроме того, они демонстрируют высокие показатели по стоимости владения и общему удобству эксплуатации благодаря автоматическому управлению инфраструктурой со стороны облачного провайдера.

В то же время гибридные архитектуры уверенно лидируют по показателям управляемости, безопасности и производительности при высоких нагрузках (оценки 9 из 10). Это делает их предпочтительным выбором для крупных предприятий с высокими требованиями к контролю над данными, соответствию регуляторным нормам и интеграции с локальными системами. Особенно важно, что по критерию интеграции с локальными решениями гибридные архитектуры значительно превосходят облачные (оценка 9 против 5).

Таким образом, облачные решения целесообразны в сценариях, где приоритетом является скорость внедрения и масштабируемость, а гибридные —

когда критичны локальный контроль, безопасность и интеграция с существующей ИТ-инфраструктурой. Выбор подхода напрямую зависит от зрелости бизнес-процессов, отраслевых требований и доступных ресурсов.

9. Будущее DWH: тенденции и прогнозирование нагрузки

Предсказание времени ответа

Можно применить регрессионный анализ для предсказания времени ответа запросов с помощью формулы 4:

$$T = \alpha \times loglog \log N + \beta \quad (4)$$

где:

- N — количество записей в БД.
- α , β — коэффициенты, зависящие от архитектуры.

Пример прогнозирования:

Таблица 1. Прогнозирование времени ответа запросов

N (млн строк)	T (сек)
10	1.2
50	2.8
100	4.1
500	9.3

По регрессии с помощью формулы 5:

$$T = 2.1 \times log \log N + 0.5 \quad (5)$$

Можно предсказать, что при 1 млрд строк:

$$T = 2.1 \cdot \log(1000) + 0.5 = 14.3$$

Это поможет оценить, насколько масштабируемо решение.

Развитие технологий в области обработки и анализа данных не стоит на месте. В ближайшие годы ожидается дальнейшая интеграция DWH с облачными платформами, что позволит компаниям более гибко масштабировать свои решения и использовать современные вычислительные мощности без

значительных капитальных вложений. Увеличение объёмов данных, поступающих из разнообразных источников, будет стимулировать развитие новых методов обработки и аналитики, включая технологии Big Data и Stream Processing.

Еще одной важной тенденцией является активное внедрение искусственного интеллекта в процессы анализа данных. Интеграция AI/ML-моделей в DWH позволит не только автоматизировать рутинные задачи, но и создавать прогнозные модели, способные выявлять скрытые зависимости и закономерности. Это, в свою очередь, откроет новые возможности для оптимизации бизнес-процессов и повышения конкурентоспособности компаний.

Также следует отметить, что развитие стандартов обмена данными и повышение уровня автоматизации ETL-процессов будут способствовать более быстрой интеграции разнородных источников данных. В результате компании смогут оперативно реагировать на изменения рыночной конъюнктуры, улучшая качество принимаемых решений и эффективность управления организацией.

10. Оптимизация ETL/ELT-процессов

Проблема избыточности данных

В процессе построения систем хранения и анализа данных одной из ключевых задач является борьба с избыточностью и дублированием данных. Особенно остро эта проблема встает при интеграции данных из различных источников. Когда данные поступают из нескольких независимых систем, есть высокий риск того, что одни и те же сущности (например, клиенты, сделки, продукты) будут загружены несколько раз под разными идентификаторами или с незначительными отличиями. Это приводит к искажению аналитики, увеличению объема хранилища и усложнению последующей обработки.

Чтобы количественно оценить вероятность возникновения дубликатов, можно воспользоваться следующей формулой 6:

$$P_{dup} = 1 - \prod_{i=1}^{n} (1 - p_i) \quad (6)$$

где p_i — вероятность дублирования записи в каждом источнике.

Если у нас есть 3 источника, в каждом из которых вероятность дубликата 5%:

$P_{dup} = 1 - (1 - 0.05)^3 = 0.14$

То есть, с вероятностью около 14% мы столкнемся с дублированием записей при интеграции этих источников. Это значение может казаться незначительным на первый взгляд, однако в масштабах хранилища, содержащего миллионы строк, даже такой процент приводит к серьезным последствиям: нагрузка на систему возрастает, снижается точность отчетности и аналитики, возникают сложности в синхронизации и актуализации данных.

Пути решения проблемы

1. Унификация и нормализация данных на этапе извлечения (ETL/ELT). Это может включать в себя стандартизацию форматов имен, адресов, телефонов и других ключевых полей.
2. Использование мастер-данных (MDM) — систем, которые помогают определить «золотую запись» (golden record) и устранить дубли.
3. Сопоставление и дедупликация записей с помощью алгоритмов fuzzy matching или ML-моделей, особенно если идентификаторы в разных источниках различаются.
4. Контроль качества данных (Data Quality) — регулярная проверка и мониторинг на наличие аномалий, дубликатов, пробелов и неконсистентности.

Концепция гибридной микробатч-потоковой архитектуры с динамической ML-регуляцией предлагает оригинальный метод оптимизации ETL-процессов. В её основе лежит объединение микробатчевой обработки, предназначенной для периодической загрузки крупных объемов данных, и потоковой обработки, которая обеспечивает реактивную обработку информации в режиме реального

времени. Такой подход позволяет достичь баланса между производительностью и актуальностью данных, особенно в системах со смешанными типами нагрузок.

Ключевая особенность метода заключается в применении ML-модели для автоматической регуляции параметров обработки данных: размера микробатчей, частоты обновлений, схемы загрузки. Регулятор, построенный, например, на базе CatBoost и оптимизируемый с помощью Optuna, обучается на исторических метриках загрузки, временных рядах производительности и событиях отказов. Это позволяет адаптировать поведение ETL-системы к текущему состоянию инфраструктуры и паттернам поступления данных.

Практическая реализация включает использование таких компонентов, как Kafka или отечественные аналоги для стриминга, ClickHouse или DuckDB для микробатчей, а также систем мониторинга и логирования для сбора телеметрии. Прогнозирование нагрузки может осуществляться через модели временных рядов (например, Prophet), что позволяет предсказывать пики и заранее подготавливать ресурсы.

Эффективность предложенного метода подтверждается рядом научных и прикладных публикаций. Так, в работе «Optimizing ETL Dataflow Using Shared Caching and Parallelization Methods» [16] описана схожая методика оптимизации, которая привела к увеличению производительности более чем в четыре раза. Другие исследования (например, «AI-Powered ETL Workflow Orchestration» [17]) также подтверждают потенциал использования машинного обучения для оркестрации ETL-операций.

Таким образом, интеграция гибридной архитектуры и алгоритмов машинного обучения в процесс ETL представляет собой перспективный вектор развития хранилищ данных, обеспечивая масштабируемость, устойчивость и адаптивность аналитических систем. И своевременная оптимизация ETL/ELT-процессов с акцентом на предотвращение дублирования позволяет значительно повысить надежность аналитики и снизить затраты на хранение и обработку информации.

11. Заключение

Анализ показывает, что DWH-системы играют ключевую роль в цифровизации бизнеса. Сегодня основные тренды — это массовый переход в облака (особенно гибридные и мультиоблачные решения), использование искусственного интеллекта и машинного обучения для прогнозной аналитики, а также внедрение новых подходов к хранению данных, таких как Data Vault 2.0 и Delta Lake. Кроме того, набирает популярность методология DataOps. Согласно исследованию Forrester (2024) [14], эти технологии помогают увеличить эффективность аналитики на 35%.

Чтобы успешно внедрить DWH-решение, важно учитывать отраслевые особенности, объединять данные из разных систем (ERP, CRM, IoT) и применять инструменты контроля качества данных. Сама архитектура хранилища должна быть гибкой и способной к масштабированию.

В будущем ожидается сближение технологий — интеграция DWH, Data Lake и потоковой обработки данных, а также автоматизация управления метаданными и переход к концепции Data Fabric. Уже сейчас такие решения сокращают время обработки данных на 40–50% и уменьшают совокупную стоимость владения (TCO).

По сути, DWH-системы превращаются в интеллектуальные аналитические платформы. Их использование ускоряет бизнес-процессы, улучшает точность прогнозов и усиливает конкурентные преимущества. Дальнейшие исследования стоит направить на оценку окупаемости (ROI) и создание типовых архитектур для разных отраслей.

12. Литература

1. Н. И. Иванова. Информационные технологии в бизнесе: современные тенденции и перспективы развития // монография – М.: Наука, 2020. – 250 с.

2. Vaisman A., Zimányi E. Data Warehouse Systems: Design and Implementation. 2-е изд. Berlin: Springer, 2023. — 625 с.

3. Bhatia P. Data Warehouse and Data Mining: Concepts, Techniques and Real Applications. 1-е изд. New Delhi: BPB Publications, 2023. — 350 с.

4. Serra J. Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh. 1-е изд. Sebastopol: O'Reilly Media, 2024. — 280 с.

5. Облачные технологии и большие данные: интеграция, анализ и управление // Сборник. – 2021. – № 1. – С. 15–25.

6. Inmon W. H. Building the Data Warehouse. 6-е изд. Indianapolis: Wiley, 2023. — 432 с.

7. Handler J. The Definitive Guide to OpenSearch: Discover Advanced Techniques and Best Practices for Efficient Search and Analytics with OpenSearch. 1-е изд. Birmingham: Packt Publishing, 2023. — 400 с.

8. Linstedt D., Olschimke M. Building a Scalable Data Warehouse with Data Vault 2.0. 1-е изд. Waltham: Morgan Kaufmann, 2023. — 684 с.

9. Astera Software. Cloud Data Warehousing Trends for 2025. Электронное издание. Torrance: Astera Software, 2024. — 45 с.

10. Corr L., Stagnitto J. Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema. 1-е изд. Leicester: DecisionOne Press, 2024. — 325 с

11. Ponniah P. Data Warehousing Fundamentals for IT Professionals. 2-е изд. Hoboken: Wiley, 2023. — 544 с.

12. Sen, A. Data Warehousing: Concepts, Techniques, Products and Applications: Springer, 2015. - 171 с.

13. Gartner. What's New in the 2023 Gartner Hype Cycle for Emerging Technologies [Электронный ресурс]: https://www.gartner.com/en/articles/what-s-new-in-the-2023-gartner-hype-cycle-for-emerging-technologies (дата обращения: 16.05.2025)

14. Forrester Research. The Data Management For Analytics Platforms Landscape, Q4 2024 [Электронный ресурс]: https://www.forrester.com/report/the-

data-management-for-analytics-platforms-landscape-q4-2024/RES181764 (дата обращения: 16.05.2025)

15. McKinsey & Company. New Year's Resolutions for Tech in 2023 [Электронный ресурс]: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/new-years-resolutions-for-tech-in-2023 (дата обращения: 26.05.2025)

16. Xiufeng Liu. Optimizing ETL Dataflow Using Shared Caching and Parallelization Methods, 2014 [Электронный ресурс]: https://arxiv.org/abs/1409.1639 (дата обращения: 21.05.2025)

17. Raghavender Maddali. AI-POWERED ETL WORKFLOW ORCHESTRATION WITH SELF-ADJUSTING DATA TRANSFORMATIONS, 2025 [Электронный ресурс]: https://www.academia.edu/128922260/AI_POWERED_ETL_WORKFLOW_ORCHESTRATION_WITH_SELFADJUSTING_DATA_TRANSFORMATIONS (дата обращения: 25.05.2025)

18. Kushal Shah. Real-Time Analytics in E-commerce: Strategies for Implementing Near Real-Time ETL Pipelines, 2025 [Электронный ресурс]: https://www.researchgate.net/publication/390326020_Real-Time_Analytics_in_E-commerce_Strategies_for_Implementing_Near_Real-Time_ETL_Pipelines (дата обращения: 22.05.2025)

19. Qlever. Как Qlever Solutions с помощью DWH и BI-аналитики помогли бренду Orby на 80% снизить количество ошибок при планировании отгрузок на маркетплейсы [Электронный ресурс]: https://www.qlever.ru/projects/dwh-i-analitika-marketplejsov-v-bi-dlya-orby (дата обращения: 24.05.2025)