DOI 10.37882/2223-2966.2025.08.09

ИНТЕГРАЦИЯ ДАННЫХ В ХРАНИЛИЩЕ ИЗ РАЗНОФОРМАТНЫХ СИСТЕМ-ИСТОЧНИКОВ: МЕТОДЫ И ПРАКТИКИ

INTEGRATION OF DATA INTO STORAGE FROM MULTI-FORMAT SOURCE SYSTEMS: METHODS AND PRACTICES

E. Dementieva

Summary. The purpose of our work is to describe the testing of some architectural solutions and methods that ensure reliable, scalable and high-quality data integration into a centralized repository. The relevance of the research is due to the need for effective integration of data from diverse and heterogeneous sources in the context of rapid growth in the volume, speed of receipt and diversity of data, which is a key challenge for corporate and information and analytical systems. The article discusses approaches to combining information from relational databases, APIs, and file storages, followed by uploading it to a ClickHouse-based analytical system. Technologies and tools (Apache NiFi, Apache Airflow, Apache Kafka in conjunction with Debezium, dbt, Great Expectations) used to implement ETL/ELT processes and data quality control are presented. The paper presents the results of modeling various integration scenarios, including batch and streaming downloads. The results of the study confirm the effectiveness of ELT and CDC architectures for building modern data integration platforms capable of providing high performance, fault tolerance and reliability of analytical information.

Keywords: data integration, data warehouse, streaming, ETL/EL, data architecture.

Дементьева Елена Максимовна
Старший разработчик, Московский технический университет связи и информатики
lenuka98@mail.ru

Аннотация. Целью нашей работы является описание апробации некоторых архитектурных решений и методов, обеспечивающих надёжную, масштабируемую и высококачественную интеграцию данных в централизованное хранилище. Актуальность исследования обусловлена необходимостью эффективной интеграции данных из разнотипных и разнородных источников в условиях стремительного роста объёмов, скорости поступления и разнообразия данных, что является ключевым вызовом для корпоративных и информационно-аналитических систем. В статье рассматриваются подходы к объединению информации из реляционных СУБД, API-интерфейсов и файловых хранилищ с последующей загрузкой в аналитическую систему на базе ClickHouse. Представлены технологии и инструменты (Apache NiFi, Apache Airflow, Apache Kafka в связке с Debezium, dbt, Great Expectations), применяемые для реализации процессов ETL/ELT и контроля качества данных. В работе приведены результаты моделирования различных сценариев интеграции, включая пакетную и потоковую загрузку. Результаты исследования подтверждают эффективность ELT— и CDC-архитектур для построения современных платформ интеграции данных, способных обеспечить высокую производительность, устойчивость к сбоям и достоверность аналитической информации.

Ключевые слова: интеграция данных, хранилище данных, потоковая обработка, ETL/EL, архитектура данных.

Введение

Современные информационные системы функционируют в условиях стремительного увеличения объемов обрабатываемых данных, возрастания скорости их поступления, а также роста разнообразия форматов и источников информации. Эти характеристики объединяются в так называемую концепцию «3V» (Volume, Velocity, Varienty), которая описывает ключевые аспекты больших данных. В контексте цифровизации и масштабируемых вычислений данная модель приобретает особую актуальность для организаций различных сфер деятельности — от государственного управления и здравоохранения до промышленных предприятий и коммерческих структур [1]. Одной из наиболее острых проблем, возникающих при работе с данными в таких условиях, является высокая степень фрагментации информационных источников. Это выражается в использовании разнородных программных решений, различающихся по типу баз данных, формата хранения, протоколам передачи и уровням структурированности информации [2,3].

Указанная гетерогенность порождает необходимость создания унифицированных механизмов для централизации, консолидации и стандартизации информации, поступающей из распределённых и технически неоднородных систем. Это, в свою очередь, требует построения архитектур, обеспечивающих устойчивую, масштабируемую и согласованную интеграцию данных, что является неотъемлемым условием для проведения комплексного аналитического и прогностического анализа в режиме, приближенном к реальному времени [4,5].

Одним из наиболее эффективных решений в данной области выступают корпоративные хранилища данных (Data Warehouse, DWH), обеспечивающие централизованное накопление, структурирование и долговременное хранение информации, поступающей из множества внешних и внутренних источников [6,7]. Концепция DWH предполагает не просто агрегацию данных, но и их логическую нормализацию, очистку, унификацию и обогащение в соответствии с бизнес-требованиями и аналитическими задачами. Ключевым компонентом такого подхода является использование специализированных

инструментов и методологий для интеграции данных, включая технологии извлечения, трансформации и загрузки (ETL/ELT), механизмы отслеживания изменений в источниках (Change Data Capture, CDC), а также платформы потоковой обработки данных с минимальной задержкой (real-time stream processing) [8–10].

Несмотря на обилие практических инструментов и решений, задача интеграции данных сохраняет свою сложность как в методологическом, так и в технологическом измерении. Среди наиболее значимых вызовов можно выделить обеспечение высокого качества и целостности данных, согласованность семантики между источниками, оптимизацию производительности обработки, устойчивость и масштабируемость интеграционных конвейеров, а также способность адаптироваться к изменениям в структуре и логике исходных систем.

Краткий обзор некоторых подходов

В последние десятилетия задачи интеграции данных из разнообразных по структуре и формату источников в централизованные хранилища приобрели стратегическое значение в процессе построения и развития информационно-аналитических систем различного масштаба и назначения. Обострение проблемы гетерогенности источников обусловлено широким распространением различных программных платформ, стандартов представления информации, форматов хранения и передачи данных. Эта неоднородность усложняет реализацию эффективности процессов агрегации, нормализации и унификации информации, что, в свою очередь, стимулирует развитие научных и прикладных исследований в области интеграционных технологий и архитектур [11].

Традиционным и широко применяемым подходом к решению данной задачи остается использование процессов ETL (Extract, Transform, Load), позволяющих последовательно извлекать, преобразовывать и загружать данные в целевые хранилища. Методология ETL на протяжении долгого времени служила основой для построения корпоративных хранилищ данных [12–14]. Однако, с распространением облачных вычислений, масштабируемых аналитических платформ и подходов к обработке данных в реальном времени наметился переход к альтернативной парадигме ELT (Extract, Load, Transform), в рамках которой операции преобразования осуществляются непосредственно в среде хранилища. Такой подход позволяет снизить издержки на промежуточную обработку, уменьшить задержки и повысить общее быстродействие аналитических систем.

Современные источники информации уже не ограничиваются традиционными реляционными базами данных. Всё чаще в архитектуре ИТ-систем используются REST— и GraphQL-API, событийные шины микросервисов, лог-файлы приложений, телеметрические потоки с IoT-устройств и данные из полу структурированных и нереляционных хранилищ, включая NoSQL-системы [15]. Такая расширяющаяся мультиформатность требует от интеграционных решений высокой адаптивности, расширяемости и способности к динамической маршрутизации данных. Важным направлением является разработка и применение универсальных средств парсинга и трансформации полу структурированных данных, представленных в формате JSON, XML, Avro и пр., с возможностью дальнейшего семантического обогащения [16–18].

Особую актуальность в условиях роста требований к своевременности данных приобретает использование методов захвата изменений — Change Data Capture (CDC). Эта технология позволяет в режиме, близком к реальному времени, регистрировать и передавать изменения, происходящие в исходных транзакционных системах, в аналитическое хранилище. Согласно данным различных исследований, CDC-подход обеспечивает низкие задержки синхронизации и минимальную нагрузку на источники, что делает его особенно востребованным в высоконагруженных системах, ориентированных на оперативную аналитику [19].

Отдельным направлением, тесно связанным с эффективной интеграцией, является обеспечение качества данных. Согласно ряду эмпирических исследований, свыше 80 % проблем, возникающих на этапе анализа данных, связаны с ошибками, допущенными в процессе их извлечения, преобразования и загрузки [20]. Среди наиболее распространенных нарушений — некорректные типы данных, пропуски в обязательных полях, дублирование и несоответствие форматов. Для их предотвращения целесообразно применять средства автоматизированной валидации и тестирования, такие как Great Expectations, а также использовать концепции DataOps, предполагающие автоматизацию, стандартизацию и мониторинг всех этапов работы с данными.

Таким образом, совокупный анализ современных теоретических работ и практических кейсов демонстрирует устойчивый вектор развития в сторону гибких, потоковых и событийно-ориентированных архитектур интеграции. Приоритетными направлениями становятся внедрение ELE— и CDC-стратегий, автоматизация контроля качества, снижение уровня ручного программирования за счет оркестрации процессов с использованием специализированных инструментов и платформ, а также развитие само адаптирующихся решений, способных динамически адаптироваться к изменяющимся структурам и объемам входных данных.

Материалы и методы

Наше исследование было направлено на анализ и моделирование процессов интеграции гетерогенных

источников данных в унифицированное аналитическое хранилище, с целью обеспечения масштабируемости, отказоустойчивости и высокого качества данных. В рамках исследования мы предлагаем некоторую архитектуру, включающую в себя механизмы объединения данных из разнородных источников с последующей их трансформацией и загрузкой в централизованную аналитическую систему.

При проведении аналитического исследования и практического моделирования процессов интеграции данных, были классифицированы три основных типа источников:

1. Реляционные системы управления базами данных (СУБД), такие как PostgreSQL и Oracle, использовались для получения строго структурированных данных, хранящихся в таблицах с четко определенными схемами.
2. Полуструктурированные API-интерфейсы, включая REST и GraphQL, обеспечивали доступ к данным с гибкой структурой, характерной для современных веб-приложений и микросервисной архитектуры.
3. Файловые хранилища и NoSQL-системы, в том числе MonqoDB, а также форматы данных JSON, CSV, XLSX представляли собой источники слабоструктурированной или нестабильной по формату информации.

В качестве целевого хранилища была выбрана высокопроизводительная аналитическая колоночная СУБД ClickHouse, оптимизированная под выполнение сложных аналитических запросов на больших объемах данных.

Для реализации интеграционных процессов применялся современный стек технологий, включающий следующие инструменты и платформы:

1. Apache NiFi — инструмент потоковой обработки и маршрутизации данных, использовавшийся для базовой трансформации и управления потоками на уровне источников;
2. Apache Airflow — система оркестрации задач, применяемая для координации процессов извлечения, трансформации и загрузки (ETL/ELT);
3. Apache Kafka в связке с Debezium — использовалась как платформа потоковой передачи данных и реализация CDC (Change Data Capture) для захвата изменений в источниках в реальном времени;
4. dbt (Data Build Tool) — применялся для организации ELT-логики внутри хранилища, в частности, для построения моделей трансформации на уровне SQL;
5. Great Expectations — система автоматизированной валидации данных, обеспечивающая контроль качества и соответствие загружаемых данных заданным правилам.

Для комплексной оценки эффективности предложенной архитектуры были определены следующие критерии:

1. Средняя длительность полного интеграционного цикла (время от извлечения до финальной загрузки);
2. Процент успешно загруженных записей относительно общего объема;
3. Количество ошибок на этапах трансформации и загрузки;
4. Степень соответствия данных сформулированным правилам контроля качества;
5. Устойчивость системы к отказам со стороны источников данных из сетевой инфраструктуры (тестирование на сбои и восстановление).

С целью проверки универсальности и надежности решений были смоделированы и протестированы различные сценарии интеграции, включая:

1. Пакетную загрузку больших объемов данных (свыше 10 миллионов записей), характерную для миграционных процессов и исторической репликации;
2. Потоковую синхронизацию в режиме реального времени с применением CDC;
3. Комбинированный подход, предусматривающий предварительную буферизацию данных с последующей асинхронной загрузкой в хранилище.

Эмпирические результаты аналитического исследования и практического моделирования

В результате проведенного аналитического исследования и практического моделирования процессов интеграции данных были получены следующие эмпирические результаты, отражающие особенности функционирования и ограничения используемого технологического стека:

1. Применение Apache NiFi продемонстрировало высокую гибкость при подключении к разнородным источникам данных, а также удобство в реализации базовых операций по извлечению и первичной трансформации информации. Однако, в ходе моделирования стало очевидно, что при увеличении объема данных свыше 10 миллионов записей производительность системы начинает снижаться. Особенно заметное падение наблюдалось при попытках масштабирования за счет повышения степени параллелизма обработки, что указывает на ограниченную масштабируемость данного инструмента в условиях интенсивной загрузки.
2. Интеграция dbt с ClickHouse показала высокую эффективность в реализации ELT-подхода, при котором вычислительная нагрузка смещается в сторону аналитического хранилища. Перенос

логики преобразования непосредственно в вычислительное ядро СУБД позволил существенно сократить время обработки и обеспечить более стабильную масштабируемость при росте объёмов данных. Вместе с тем, данный подход предъявлял повышенные требования к унификации и строгому контролю схем данных на этапе загрузки, поскольку даже незначительные отклонения в структуре могли привезти к сбоям в выполнении моделей трансформации.

3. Система потоковой передачи данных на базе Apache Kafka и Debezium показала себя как надежное решение для реализации Change Data Capture (CDC) при интеграции с реляционными СУБД, в частности PostgreSQL. Задержка доставки изменений в систему потребления данных составила в среднем менее 5 секунд, что удовлетворяет требованиям к near-real-time обработке. Архитектура продемонстрировала устойчивость к временным сбоям источников и сетевой инфраструктуры, сохраняя целостность потока сообщений благодаря встроенным механизмам ретрансляции и буферизации.
4. Механизмы контроля качества данных, реализованные с использованием библиотеки Great Expectations, выявили, что от 10 до 15 % загружаемых записей содержат те или иные нарушения — как структурного характера (например, несоответствие типов, отсутствие обязательных атрибутов), так и семантического (включая дублирование в логические несоответствия). Внедрение автоматизированных проверок позволило своевременно идентифицировать и отсеивать некорректные данные до их включения в аналитический контур, что, в свою очередь, существенно повысило надёжность принимаемых на основе этих данных решений и снизило частоту отказов аналитических сценариев.

Обсуждение

Сравнительный анализ применяемых архитектур и инструментов интеграции выявил, что выбор конкретного технического решения должен быть обусловлен совокупностью факторов, включая особенности предметной области, структуру и тип источников данных, а также требования к временной актуальности и объему обрабатываемой информации.

В частности, ETL-подходы (извлечение, трансформация, загрузка) оказываются наиболее целесообразными в условиях регламентированной и регулярной пакетной обработки данных, при этом эффективность достигается при стабильной и слабо изменяемой структуре исходных источников. Такие сценарии характерны для систем, в которых допустим определенный лаг времени между обновлением данных и их аналитической обработкой.

ETL-архитектуры, предполагающие выполнение трансформаций уже после загрузки данных в хранилище, демонстрируют наилучшую масштабируемость и гибкость в условиях изменчивых и больших объемов данных. Перенос вычислительной нагрузки в сторону высокопроизводительных аналитических СУБД позволяет оптимизировать ресурсоёмкие операции и ускорить время отклика.

Потоковые интеграционные решения, основанные на технологиях Change Data Capture (CDC), проявили высокую эффективность при необходимости обеспечения почти непрерывной синхронизации данных, особенно в контексте транзакционных информационных систем. Данный подход минимизирует нагрузку на исходные базы данных и обеспечивает минимальную задержку передачи изменений. Вместе с тем он требует значительных усилий по конфигурации, постоянному мониторингу и обеспечению отказоустойчивости.

Использование универсальных инструментов оркестрации процессов, таких как Apache Airflow, позволяет стандартизировать управление сложными интеграционными пайплайнами и обеспечить их прозрачность, воспроизводимость и управляемость. Включение в архитектуру средств автоматической валидации данных, таких как Great Expectations, повышает доверие к загружаемой информации и способствует снижению вероятности возникновения критических ошибок на аналитическом уровне. Применение стандартизированных фреймворков трансформации, таких как dbt, способствует унификации логики обработки данных и облегчает сопровождение ELT-процессов в рамках быстро развивающихся информационных систем.

Таким образом, гибкий комбинированный подход, сочетающий возможности пакетной и потоковой обработки, а также применение открытых и расширяемых инструментов с высокой степенью автоматизации, представляет собой наиболее перспективную стратегию для построения современных платформ интеграции данных, удовлетворяющих требованиям как к производительности, так и к качеству.

Заключение

Интеграция данных из разнотипных и разноформатных источников в централизованное хранилище является важнейшим этапом построения корпоративных информационно-аналитических систем. В ходе настоящего исследования была проведена классификация источников данных, апробированы современные инструменты и методы интеграции, а также проведена их сравнительная оценка.

Полученные результаты демонстрируют высокую эффективность ELT— и CDC-архитектур в условиях боль-

шого объема данных и необходимости актуализации в реальном времени. Потоковые технологии обеспечивают устойчивость и гибкость, особенно при синхронизации с неструктурированными источниками.

Кроме того, интеграция механизмов автоматической валидации и контроля качества позволяет существенно снизить риски на аналитическом и отчетном уровнях. Практическая значимость работы заключается в возможности использования полученных результатов при проектировании и оптимизации процессов интеграции в реальных корпоративных системах.

В перспективе целесообразным направлением дальнейших исследований является разработка универсальных адаптивных интеграционных платформ с использованием искусственного интеллекта для автоматической идентификации структур данных и само настраиваемых трансформационных пайплайнов.

ЛИТЕРАТУРА

1. Трухонин А.А., Первых Е.А., Осипова. Методика интеграции корпоративного хранилища данных с ИИ-системами // Информатизация и виртуализация экономической и социальной жизни: материалы XII Международной студенческой научно-практической конференции, Иркутск, 31 марта 2025 года. — Иркутск: Иркутский национальный исследовательский технический университет, 2025. — С. 170–178.
2. Wnęk K., Boryło P.A Data Processing and Distribution System Based on Apache NiFi // Photonics. — 2023. — Vol. 10, No. 2. — P. 210. — DOI: 10.3390/photonics10020210.
3. Boyko N.I., Chernenko A.V. Modern approaches to data storage: comparison of relational and cloud data warehouses using ETL and ELT methods // Reporter of the Priazovskyi State Technical University. Section: Technical sciences. — 2024. — No. 48. — P. 7–19. — DOI: 10.31498/2225-6733.48.2024.310669.
4. Prabhu A. Leveraging Event-Based Architecture, AWS Step Functions, AWS Batch, and DynamoDB to Run ETL or ELT Jobs Concurrently While Allowing Granular Replay Capabilities // International Journal of Science and Research. — 2024. — Vol. 13, No. 9. — P. 25–28. — DOI: 10.21275/sr24829051214. — EDN KBYJGD.
5. Баданов А.А., Тугой И.А. Миграция больших данных: необходимость и особенности // Социально-гуманитарные знания. — 2025. — № 4. — С. 66–71.
6. Ганеев А.Р., Тугой И.А., Баданов А.А. Миграция больших данных между хранилищем данных HDFS и базой данных Clickhouse с использованием операций преобразования // Наука и бизнес: пути развития. — 2024. — № 6. — С. 156.
7. Баданина О.В., Гиндин С.И. Оценка оперативности передачи больших данных на примере базы данных PostgreSQL, платформы Hadoop и системы Sqoop // Интеллектуальные технологии на транспорте. — 2020. — № 2. — С. 18–26.
8. Белов В.А., Ильин Д.Ю., Никульчев Е.В. Оценка эффективности обработки больших объемов данных в реляционных и колоночных форматах // Вычислительные технологии. — 2022. — № 3(27). — С. 46–65.
9. Лабинский А.Ю. Программные средства обработки больших объемов данных // Природные и техногенные риски (физико-математические и прикладные аспекты). — 2023. — № 4(48). — С. 45–52. — DOI: 10.61260/2307–7476-2024-2023-4-45-52.
10. Брюхова Е.М., Данилов А.С. Технологии хранения и обработки больших данных для обучения скоринговых моделей // Международный журнал гуманитарных и естественных наук. — 2024. — № 12–3(99). — С. 55–59. — DOI: 10.24412/2500–1000-2024-12-3-55–59.
11. Батура М.П., Шнейдеров Е.Н. Система мониторинга показателей образовательного процесса в области информационных технологий в телекоммуникациях // СВЧ-техника и телекоммуникационные технологии. — 2022. — № 4. — С. 62–63.
12. Карташев В.И., Фахми Ш.С., Антонова А.А. Подход к построению системы обработки открытых данных дистанционного зондирования Земли для мониторинга наводнений с использованием технологий больших данных // Космическая техника и технологии. — 2025. — № 1(48). — С. 116–142.
13. Tellman B., Sullivan J.A., Doyle C.S. Global flood observation with multiple satellites: applications in Rio Salado (Argentina) and the Eastern Nile Basin // Global drought and flood: observation, modeling, and prediction. — 2021. — P. 99–121.
14. Петрова Л.А., Бадеева Е.А., Малахова Ю.В. Конвергенция ключевых цифровых технологий в бизнес-практике // Цифровая экономика. — 2024. — № 2. — DOI: 10.24412/2071–6435-2024-2-31–52.
15. Бегишев И.Р. Семантический анализ термина «цифровая безопасность» // Юрислингвистика. — 2021. — № 20. — С. 24–38.
16. Петрова Л.А., Кузнецова Т.Е. Цифровизация банковской системы: цифровая трансформация среды и бизнес-процессов // Финансовый журнал. — 2020. — № 3(12). — С. 91–101.
17. Ахметов Р.Р. Возникающие тенденции и возможности в интернет-технологиях // Актуальные исследования. — 2023. — № 23(153). — С. 56–61.
18. Долганова О.И., Козырев Д.А. Зарубежный опыт цифровизации превентивного государственного финансового контроля (на примере США, Китая, Канады, Индии и Австралии) // Государственное управление. Электронный вестник. — 2024. — № 104. — DOI: 10.55959/MSU2070-1381-104-2024-147-161.
19. Баланова М., Гусарова Л. Новые направления развития автоматизированного инструментария в системе внутреннего государственного финансового контроля // Экономика и управление: проблемы, решения. — 2023. — № 5(4). — С. 77–83. — DOI: 10.36871/ek.up.p.r.2023.05.04.010.
20. Исаев Э.А. Актуальные вопросы к цифровизации контроля в финансово-бюджетной сфере // Вестник университета. — 2022. — № 8. — С. 139–144. — DOI: 10.26425/1816–4277-2022-8-139-144.
21. Уласов Д.О. Цифровой аудит // Акционерное общество. — 2020. — № 02. — С. 30–33.