
Архитектура данных на основе Lakehouse: современные подходы к построению модели управления образовательных организаций на основе анализа данных

С.Г. Ермаков, О.С. Мандрика, Д.С. Патеев

Петербургский государственный университет путей сообщения Императора Александра I, г. Санкт-Петербург

Аннотация: В настоящее время образовательные организации сталкиваются с необходимостью эффективного управления растущими объемами разнородных данных – от академической успеваемости и цифровых образовательных ресурсов до административных процессов. Статья посвящена исследованию современных подходов к архитектуре данных с использованием LakeHouse, совмещающим элементы корпоративного хранилища данных (КХД) и озеро данных. В статье рассматривается интеграция традиционных методов хранения структурированных данных с гибкостью и масштабируемостью озера данных, что позволяет эффективно работать с большими объемами разнородных данных. Приводится описание архитектуры КХД, адаптированное для образовательных учреждений. Рассматривается первый уровень архитектуры КХД. Приводится описание платформы Apache Airflow.

Ключевые слова: Data Lake, корпоративное хранилище данных, Apache Airflow, Greenplum.

Введение

В условиях цифровой трансформации образовательные организации сталкиваются с необходимостью эффективного управления большими объемами данных, поступающих из разнородных источников. Традиционные системы хранения и обработки информации не справляются с задачами анализа неструктурированных и полуструктурированных данных, таких как логи образовательных платформ, данные с устройств Интернета вещей (Internet of Things - IoT), мультимедийные материалы и информация из социальных сетей. Это создает потребность в новых подходах к организации корпоративных хранилищ данных, которые могли бы обеспечить гибкость, масштабируемость и поддержку современных аналитических методов.

Технология озеро данных (Data Lake - DL), позволяющая хранить данные в их исходном формате и обрабатывать их по мере необходимости, становится ключевым инструментом для решения этих задач. Интеграция

озера данных с традиционными корпоративными хранилищами данных открывает новые возможности для образовательных организаций, включая прогнозирование успеваемости, оптимизацию учебных процессов, персонализацию обучения и улучшение управления ресурсами. Эти черты в себе совмещает архитектура LakeHouse. Комбинированная технология (озеро данных и корпоративное хранилище данных) – это LakeHouse.

Цель данной статьи — рассмотреть современные подходы к построению модели управления образовательными организациями на основе анализа данных с использованием корпоративного хранилища данных и технологии озера данных.

Технология озеро данных

Data Lake – это озеро данных, которое позволяет хранить данные в их первоначальном виде без предварительной обработки. Как правило, данные могут быть структурированными, неструктурированными или полуструктурированными. Озера данных, особенно облачные, отличаются доступной стоимостью, а также простотой масштабирования. В большинстве случаев, они применяются в сочетании с прикладной аналитикой на основе машинного обучения [1].

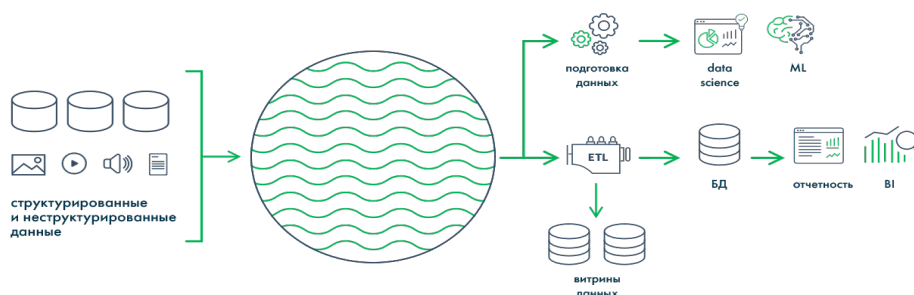


Рисунок 1 – Технология озеро данных

Под неструктурированными данными понимаются данные, которые не имеют определенной схемы, формата или организации.

Неструктурированные данные не упорядочены в строгом виде и могут быть представлены в различных форматах. Примеры таких данных включают: документы, электронные письма, изображения, аудитория, видео, потоки данных с устройств Интернета вещей.

Технология озеро данных особенно полезна для организаций, которые работают с большими данными и стремятся извлекать из них ценную информацию для улучшения бизнес-процессов и принятия решений.

Для того, чтобы хранить большой объем информации, поддерживать аналитику, машинное обучение и искусственный интеллект (ИИ), интегрировать данные из разных источников, а также оптимизировать затраты на хранение и обработку данных – используется озеро данных. Если данные не управляются должным образом, тогда озеро данных может превратиться в «болото» - хранилище с бесполезными или некачественными данными [2].

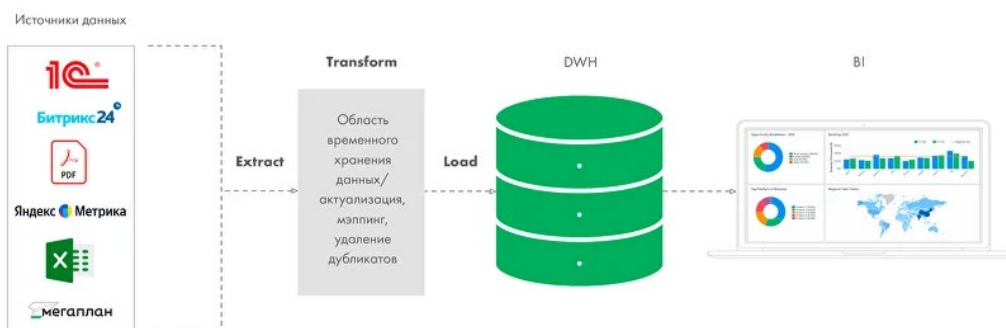


Рисунок 2 – Data Warehouse

Data Warehouse

Под понятием Data Warehouse (далее – DWH) понимается централизованная система для хранения, организации и анализа структурированных данных, предназначенная для поддержки процессов бизнес-аналитики и принятия решений [3].

Корпоративное хранилище данных DWH обновляет, стандартизирует, дополняет и консолидирует информацию из различных систем, форм и внешних источников в единую структуру. По итогу, не требуется дополнительная подготовка данных. Это значительно облегчает доступ к ним аналитиков и бизнес-пользователей.

В отличие от баз данных, которые оптимизированы для операционной обработки, DWH ориентирован на аналитическую обработку и предоставляет удобный доступ к историческим и агрегированным данным.

Рост объемов данных и необходимость обработки разнообразной информации – как упорядоченной, так и хаотичной – привели к изменению подходов к обработке данных. В последующем, это стимулировало развитие методологии ELT, которая отличается от традиционного ETL (Extract, Transform, Load) -подхода. В отличие от традиционного подхода ETL, который предполагает последовательное извлечение, преобразование и загрузку данных, метод ELT начинается с извлечения данных в их исходном, «сыром» состоянии. Эти необработанные данные затем загружаются в хранилище данных, известное как озеро данных. Преобразование данных в нужный формат осуществляется уже на финальном этапе, внутри озера данных. Таким образом, ELT откладывает трудоёмкую операцию преобразования до момента, когда данные уже находятся в целевом хранилище, что позволяет работать с большими объёмами информации эффективнее и быстрее, чем при использовании ETL.

Сравнение озера данных и Data Warehouse

В отличие от озера данных, хранилище данных предоставляет инструменты для эффективного управления информацией и включает в себя предварительно обработанные и отфильтрованные данные, которые уже адаптированы для решения определённых бизнес-задач или сценариев использования.

Хранилища и озера данных нередко используются совместно, дополняя друг друга. Когда нужно решить конкретную бизнес-проблему, необработанные данные из озера данных извлекаются, подвергаются очистке и преобразованию, а затем загружаются в хранилище данных для последующего анализа. Это лишь один из примеров, демонстрирующий разницу между этими двумя подходами к хранению и обработке данных. Помимо различий в типах данных и способах их использования, озеро данных и хранилище данных имеют ряд ключевых отличительных характеристик (Таблица 1).

Таблица № 1

Сравнение Data Lake и Data Warehouse

Характеристика	Data Warehouse	Data Lake
Тип данных	Структурированные	Структурированные, полуструктурированные, неструктурированные
Схема	Схема определяется перед загрузкой	Схема определяется при чтении
Оптимизация	Для аналитики	Для хранения и обработки больших данных
Обработка	ETL включает в себя процесс извлечения данных из источников, их очистку и структурирование, что в итоге делает данные готовыми для анализа.	ELT подразумевает извлечение данных из исходного источника, их сохранение в хранилище данных, а преобразование осуществляется лишь при возникновении необходимости.
Стоимость	Высокая	Экономичная
Использование	Бизнес-аналитика	Машинное обучение, исследование данных

Преимущества интеграции КХД и озера данных:

- *Универсальность хранения:* озеро данных предоставляет возможность сохранять данные в различных форматах и типах, что идеально дополняет традиционное КХД, которое сосредоточено на работе со структурированными данными.

- *Масштабируемость:* озеро данных обеспечивает простое масштабирование, позволяя эффективно обрабатывать огромные массивы данных.

- *Экономическая эффективность:* Хранение информации в озере данных обычно обходится дешевле, чем в традиционных системах хранения.

- *Поддержка аналитики:* озеро данных предоставляет возможности для глубокого анализа, включая машинное обучение и работу с неструктурированными данными.

Архитектура КХД с технологией озера данных

Всего в КХД существует 4 уровня. Первый уровень – уровень данных (озеро данных). На первом уровне происходит подключение к различным источникам данных, настройка процессов импорта, сбор и хранение данных из всех информационных систем Университета.

Основные задачи:

- Создание единого репозитория для всех исходных данных.
- Формирование структуры, обеспечивающей легкий доступ к данным в их исходном виде.
- Сохранение данных без изменений для возможности их дальнейшей обработки и анализа.

Этот уровень представляет собой своеобразное «зеркало» всех данных Университета.

Применяемые инструменты/модули: модуль DPM, Apache Airflow, Greenplum.

Apache Airflow представляет собой платформу с открытым исходным кодом, разработанную для создания, планирования и отслеживания сложных рабочих процессов. Этот гибкий фреймворк на Python предоставляет возможность разрабатывать рабочие процессы, которые могут взаимодействовать с практически любыми технологиями. Удобный веб-интерфейс облегчает управление состоянием этих процессов. Airflow можно развернуть различными способами: от запуска одного процесса на ноутбуке до настройки распределенной системы, способной поддерживать даже самые масштабные рабочие процессы.

Платформа имеет следующие преимущества:

- Процессы позволяют динамически генерировать потоки данных.
 - Airflow содержит разные типы операторов, позволяющие объединять разные технологии.
 - Параметры процессов определяются при помощи встроенного шаблонизатора Jinja.
 - Благодаря тому, что Airflow имеет открытый исходный код, есть возможность работать с компонентами, разработанными, протестированными и используемыми многими другими разработчиками по всему миру.
 - Рабочие процессы могут храниться в системе контроля версий.
 - Компоненты являются расширяемыми, и есть возможность построить собственную большую коллекцию компонентов.
 - Мониторинг и оповещение о событиях. Обеспечивается интеграция с инструментами Statsd и FluentD для сбора и передачи метрик и логов. Кроме того, доступен Airflow-exporter для взаимодействия с Prometheus.
 - Возможность настройки ролевого доступа. По умолчанию в Airflow предусмотрено 5 ролей, каждая из которых имеет свой уровень доступа: Admin, Public, Viewer, Op, User. Кроме того, поддерживается создание
-

создание пользовательских ролей с доступом к ограниченному набору DAG. Также доступна интеграция с Active Directory и гибкая настройка прав доступа с использованием Role-Based Access Control (Role-Based Access Control - RBAC).

- Поддержка тестирования. Можно включить базовые Unit-тесты, которые будут проверять как целостность пайплайнов, так и отдельные задачи внутри них.

- Наличие веб-интерфейса.

Веб-интерфейс призван максимально упростить управление рабочими процессами, а инфраструктура Airflow постоянно совершенствуется, чтобы сделать работу разработчиков максимально удобной. Однако философия Airflow заключается в том, чтобы определять рабочие процессы как код, поэтому написание кода будет требоваться всегда [4].

Процессы ETL

Airflow используется как планировщик ETL/ELT-процессов.

Процессы ETL занимают центральное место в современных системах бизнес-аналитики, обеспечивая интеграцию различных корпоративных информационных систем для унификации и анализа хранящихся в них данных. ETL является важным компонентом корпоративной инфраструктуры, основанной на технологиях обработки больших данных, где исходные «сырые» данные преобразуются в структурированную информацию, готовую к использованию в аналитике бизнеса [5].

ETL включает в себя несколько ключевых этапов:

- Извлечение (Extract) – на этом этапе данные извлекаются из различных источников, включая пользовательские и системные логи, реляционные базы данных, а также внешние наборы данных, такие как информация из социальных сетей и других веб-ресурсов;

- Преобразование (Transform) – здесь к данным применяются различные операции, основанные на бизнес-логике, включая фильтрацию, группировку и агрегирование. Этот процесс позволяет преобразовать «сырые» данные в структурированный набор данных, который готов к аналитике;
- Загрузка (Load) – на данном этапе обработанные данные помещаются в конечное хранилище, такое как озеро данных, реляционная база данных, витрина данных или облачное приложение и т.д.

На рисунке 3 представлена архитектура Apache Airflow.

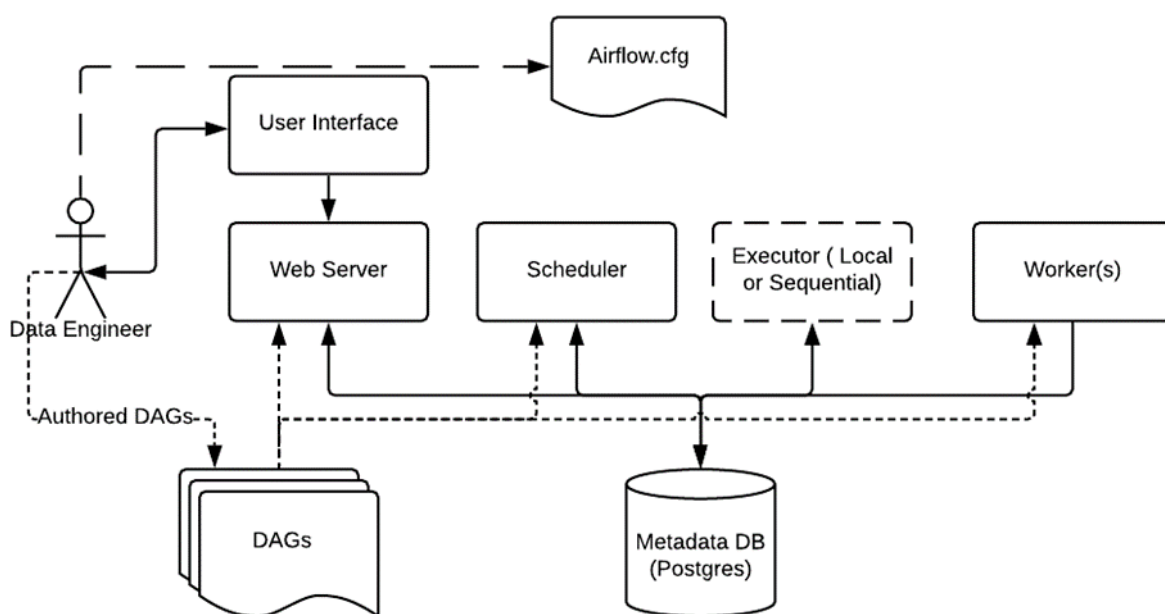


Рисунок 3 – Архитектура Airflow

- Metadata DB - Airflow использует SQL-базу данных для хранения метаданных о выполняемых пайпланах. На рисунке она обозначена как Postgres, которая широко распространена в Airflow. В качестве альтернативы Airflow также поддерживает другие базы данных, такие как MySQL [6].
- Scheduler (планировщик) контролирует все группы DAG и связанные с ними задачи. Когда выполняются все зависимости для задачи, планировщик запускает ее. Он регулярно проверяет задачи, готовые к выполнению [7].
- Executor - исполнитель, отвечающий за обработку запущенных задач. Он определяет, каким образом будет выполняться работа. Существуют

различные типы исполнителей, которые могут быть применены в зависимости от конкретных сценариев. Хотя исполнитель представлен как отдельный блок, на самом деле он не функционирует как независимый процесс, а работает внутри планировщика.

- Web server — это интерфейс пользователя в Airflow. Он отображает статус задач и предоставляет возможность взаимодействовать с базами данных, а также просматривать логи из удаленных хранилищ, таких как S3, Google Cloud Storage, большие двоичные объекты Microsoft Azure и т.д.

- Worker(s) представляют собой автономные процессы, которые взаимодействуют с другими компонентами архитектуры Airflow, а также с хранилищем метаданных [8].

- airflow.cfg – это файл конфигурации Airflow, к которому обращаются веб-сервер, планировщик и рабочие процессы.

- DAG Directory - Папка с файлами DAG, которые читаются планировщиком и исполнителем [9].

Заключение

В эпоху цифровой трансформации образовательные учреждения сталкиваются с необходимостью обработки, анализа и извлечения практической пользы из больших объемов разнородных данных. Технология озера данных, выступая в качестве дополнения к традиционным корпоративным хранилищам данных, обеспечивает эффективное управление как структурированной, так и неструктурированной информацией. Она также открывает возможности для применения современных аналитических инструментов, включая искусственный интеллект и машинное обучение.

Интеграция озера данных и Data Warehouse открывает новые возможности в сфере управления образованием: прогнозирование успеваемости и улучшение академических результатов, персонализация учебного процесса, предсказание пиковых нагрузок на ключевые системы,

оптимизация ресурсов и повышение качества принятия решений. Использование инструментов управления данными, таких как Apache Airflow, упрощает процессы ETL/ELT и обеспечивает гибкость в работе с большими данными.

Таким образом, внедрение современных подходов к организации корпоративных хранилищ данных в образовательных учреждениях становится не только стратегически важной задачей, но и необходимым условием повышения эффективности образовательных процессов в цифровую эпоху.

Литература

1. Горелик А. Корпоративное озеро больших данных: новый подход к использованию Big Data и Data Science в бизнесе. Москва: Эксмо, 2023. 272 с. — (Data Science. Лучшие книги о науке о данных). URL: cdn.img-gorod.ru/nomenclature/29/857/2985757.pdf
2. Ермаков С. Г., Бухарова К. А., Мандрика О. С., Лабузов С. В. Анализ и обработка больших данных на платформе QuAnt. Санкт-Петербург: ФГБОУ ВО ПГУПС, 2024. С.115
3. Вичугова А. Как устроено Lakehouse: архитектура и принципы работы. Москва: Школа больших данных, 2022. С. 4
4. Сабуров А. РСХБ: шаг к архитектуре Lakehouse. Москва: Открытые системы, 2025. С. 7
5. Ведерникова О. Как построить архитектуру аналитики данных и превратить обычную компанию в data-driven. Москва: Эпсилон Метрикс, 2024. С.7
6. Искалиева О. Б. Современные подходы к управлению образовательной организацией. Журнал «Актуальные исследования», 2023. № 23 (153). С. 12-20.

7. Хасанов И. Т. Особенности современных методов управления данными. Труды конференции, 2022. С. 111-114.
8. Johnson M., Smith T. Data Lakes in the Age of Big Data: A Comprehensive Guide. New York: Tech Press, 2022. 400 p.
9. Kaczmarek M., Nowak J. Data Lake vs Data Warehouse: A Comparative Study for Big Data Applications. Journal of Computer Science and Technology, 2023. Vol. 38. No. 4. Pp. 789-805.
10. Brown A., Davis R. The Impact of Big Data on Business Strategies: Trends and Insights. Journal of Business Research, 2022. Vol. 105. Pp. 123-135.

References

1. Gorelik A. Korporativnoe ozero bol'shix danny'x: novy'j podxod k ispol'zovaniyu Big Data i Data Science v biznese. Moskva: E'ksmo, 2023. 272 p. (Data Science. Luchshie knigi o nauke o danny'x). [Corporate Data Lake: A New Approach to Using Big Data and Data Science in Business]. URL: cdn.img-gorod.ru/nomenclature/29/857/2985757.pdf
2. Ermakov S. G., Bukharova K. A., Mandrika O. S., Labuzov S. V. Analiz i obrabotka bol'shix danny'x na platforme QuAnt.[Analysis and Processing of Big Data on the QuAnt Platform] Sankt-Peterburg: FGBOU VO PGUPS, 2024. p.115. .
3. Vichugova A. Moscow: Shkola bol'shix danny'x, 2022. p. 4.
4. Saburov A. Moscow: Otkry'ty'e sistemy', 2025. p. 7.
5. Vedernikova O. Moscow: Epsilon Metrics, 2024. p.7.
6. Iskalieva O. B. Zhurnal «Aktual'ny'e issledovaniya», 2023. № 23 (153). pp. 12-20.
7. Xasanov I. T. Trudy' konferencii, 2022. p. 111-114.
8. Johnson M., Smith T. New York: Tech Press, 2022, 400 p.
9. Kaczmarek M., Nowak J. Journal of Computer Science and Technology, 2023, Vol. 38, No. 4, pp. 789-805.



10. Brown A., Davis R. Journal of Business Research, 2022, Vol. 105, pp. 123-135.

Дата поступления: 11.02.2025

Дата публикации: 27.05.2025