DATA ENGINEERING

Решение задач с помощью озер данных

АРТЁМ БАДАНОВ

Data Engineer

ЧТО ТАКОЕ DATA LAKE?

На русский язык **data lake** переводится как «озеро данных».

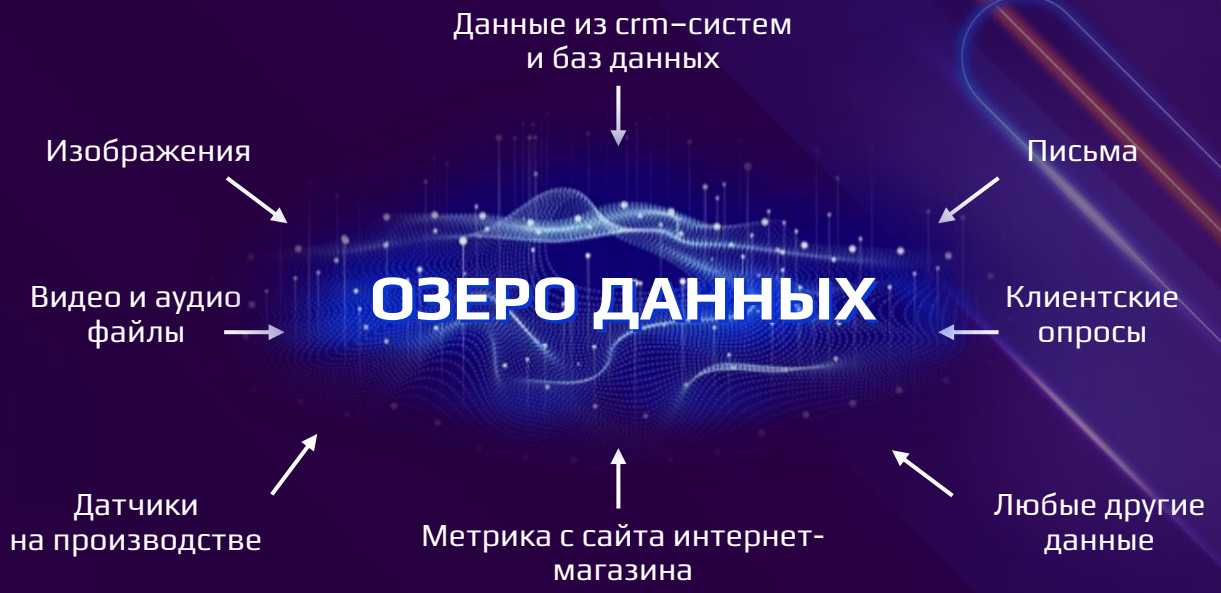
Оно представляет собой огромное хранилище, в котором разные данные хранятся в «сыром», то есть неупорядоченном и необработанном виде.

Данные в data lake как рыба в озере, которая попала туда из реки, — вы точно не знаете, какая именно там рыба и где она находится. А чтобы «приготовить» рыбу, то есть обработать данные, ее нужно еще поймать.

ЧТО ТАКОЕ DATA LAKE?

Мы в своей жизни чаще всего сталкиваемся именно с неструктурированными данными.

Видеоролики, книги, журналы, документы Word и PDF, аудиозаписи и фотографии — все это неструктурированные данные, и все они могут хранится в **Data Lake**.



КАК РАБОТАЕТ ОЗЕРО ДАННЫХ?

Data lake — это огромное хранилище, которое принимает любые файлы всех форматов.

Источник данных тоже не имеет никакого значения. Озеро данных может принимать данные из CRM- или ERP-систем, продуктовых каталогов, банковских программ, датчиков или умных устройств — любых систем, которые использует бизнес.

Уже потом, когда данные сохранены, с ними можно работать — извлекать по определенному шаблону в классические базы данных или анализировать и обрабатывать прямо внутри **data lake.**

Для этого можно использовать Hadoop — программное обеспечение, позволяющее обрабатывать большие объемы данных различных типов и структур. С его помощью собранные данные можно распределить и структурировать, настроить аналитику для построения моделей и проверки предположений, использовать машинное обучение.

КОМУ И ЗАЧЕМ НУЖНЫ ОЗЕРА ДАННЫХ?

Озёра данных предназначены для того, чтобы собирать, хранить и обрабатывать большое количество информации, поступающей практически непрерывным потоком. Такую информацию называют **Big Data**, или большими данными.

Data Lake полезны всем компаниям, которые планируют анализировать большие данные любой области, например ретейла, IT, промышленности или логистики.

Само по себе озеро данных бесполезно, потому что это просто хранилище. Чтобы с ним работать, нужны инструменты для очистки, структурирования, извлечения и анализа данных, и специалисты для работы с этими инструментами.

Попробовать себя в роли такого специалиста можно **на курсе «Аналитик данных».**

КОМУ И ЗАЧЕМ НУЖНЫ ОЗЕРА ДАННЫХ?

Без Data Lake можно обойтись, если компания:

- Вообще **не работает с Big Data** и не планирует делать это в ближайшем будущем. Обычно это характерно для небольшого бизнеса с минимальной IT-инфраструктурой и небольшим объёмом поступающих данных.
- В основном **собирает структурированные данные,** например, из баз или систем сбора метрик. В таком случае их сразу можно помещать в хранилища и использовать
для аналитики.

КАК УСТРОЕНО ОЗЕРО ДАННЫХ?

Озеро представляет собой файловое хранилище на нескольких серверах, в котором лежат данные. Как правило данные распределены между серверами, чтобы хранилище можно было быстро масштабировать — подключить новые серверы для расширения места.

К серверам настраивают подключение разных источников данных, доступных компании. Каналы поставки данных называют **пайплайнами**, а всю схему подключения — **ETL-процессом**. Обычно всё настроено так, чтобы данные загружались автоматически.

КАК УСТРОЕНО ОЗЕРО ДАННЫХ?

Хотя Data Lake и неструктурированное, порядок в нём всё-таки должен быть, иначе спустя время накопится огромное количество данных, в которых невозможно будет разобраться. Поэтому перед добавлением в озеро данные размечают и запоминают, откуда и в каком формате они поступили.

В итоге внутри озера данных хранятся не только сами объекты, но и метаданные, то есть информация об объектах. Это облегчает поиск, извлечение и анализ данных в будущем.

В архитектуре озера данных должны быть предусмотрены инструменты резервного копирования, чтобы информация не терялась.

КАК УСТРОЕНО ОЗЕРО ДАННЫХ?

КАК УСТРОЕНО ОЗЕРО ДАННЫХ?

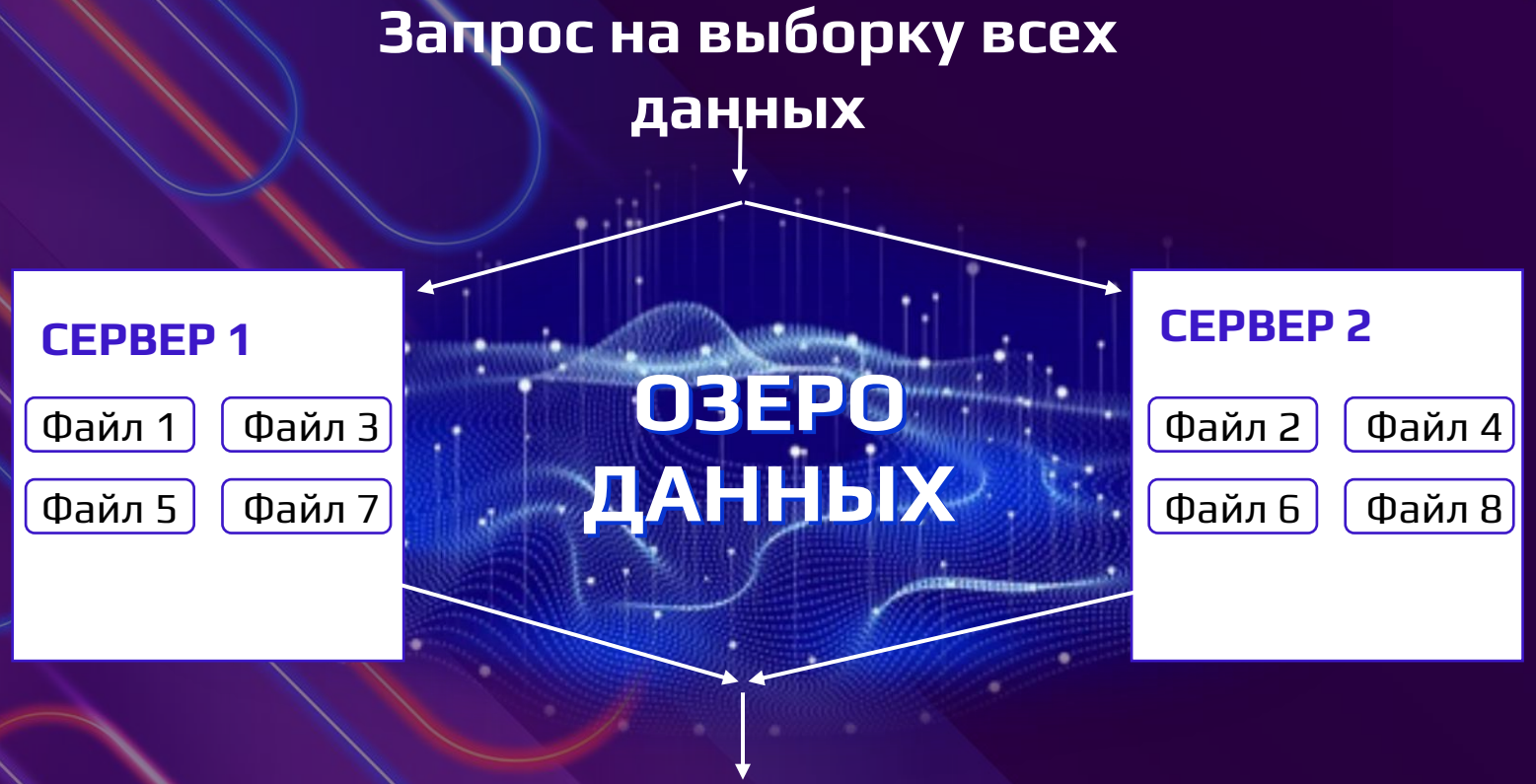
Озеро данных не существует само по себе.
К нему примыкают другие инструменты:

- **Источники,** в которых данные генерируются и собираются. Это могут быть базы данных, CRM, ERP, IoT и другие системы и сервисы.
- **Аналитические сервисы,** которые отбирают, сортируют и анализируют информацию. Например, это BI-инструменты для построения дашбордов. Или сервисы машинного обучения для создания ML-моделей и нейросетей.
- **Хранилища,** в которых лежат уже структурированные и очищенные данные из озера.

Аналитики данных взаимодействуют именно с инструментами, а не с Data Lake напрямую. Современные озёра данных, как правило, строят с помощью инструмента Hadoop.

Он позволяет хранить поступающую информацию на разных подсерверах и обрабатывает её параллельно, что значительно ускоряет работу.

КАК УСТРОЕНО ОЗЕРО ДАННЫХ?



Озера данных и хранилища данных

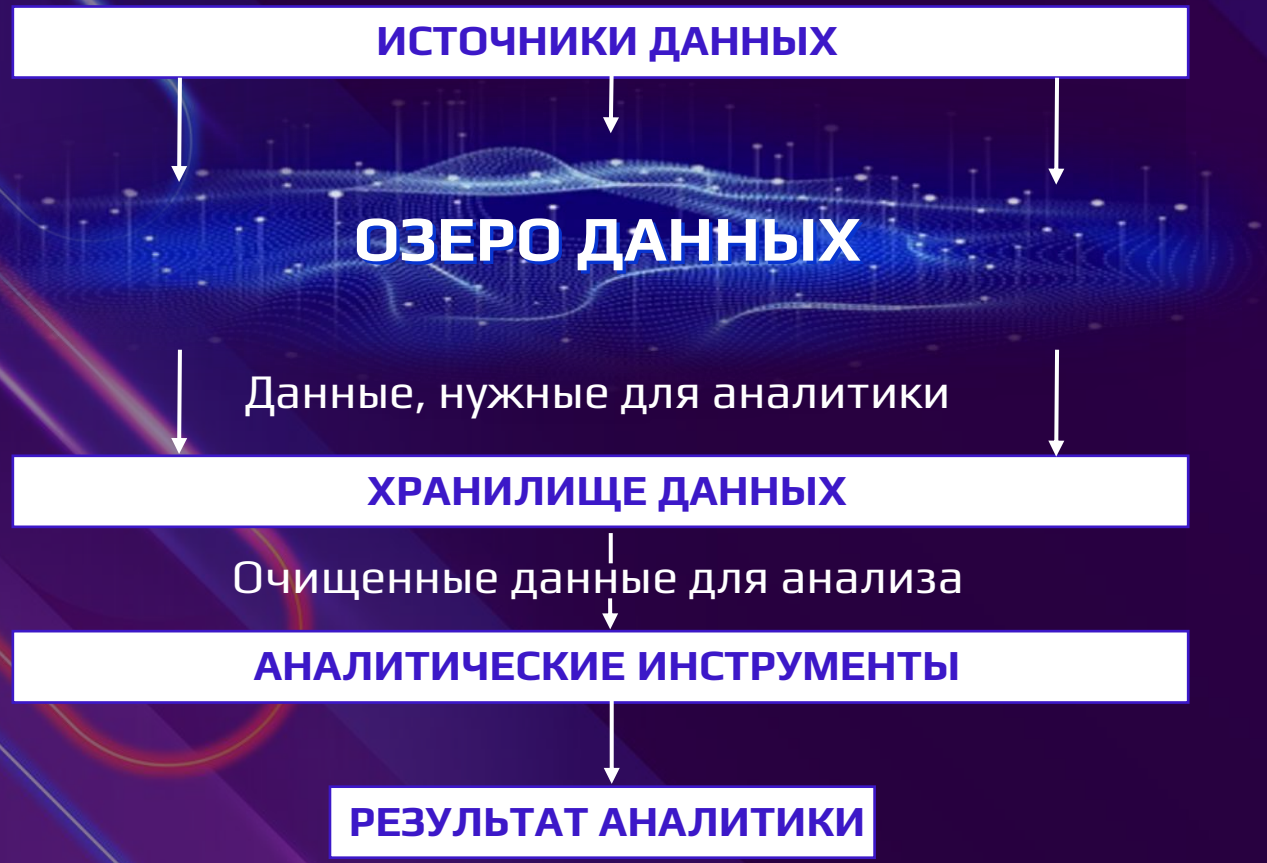
Озеро данных, Data Lake.

Предназначены для хранения данных любых типов. Перед аналитикой их нужно обязательно найти, очистить и структурировать, то есть поместить в озеро просто, а извлекать — сложнее. Из-за отсутствия структуры и простого обслуживания озеро данных обходится дешевле, чем хранилище.

Хранилище данных, Data Warehouse.

Построено на основе распределённых баз данных — как классических, так и специальных, типа ClickHouse. Оно содержит уже отсортированную, преобразованную и структурированную информацию. Данные из хранилища можно сразу использовать в анализе. Помещать информацию в хранилище занимает больше времени, потому что её нужно предварительно структурировать. Из-за структуры данные в хранилище занимают больше места и требуют более сложного обслуживания, поэтому само хранилище обходится дороже, чем озеро данных.

ОЗЕРА ДАННЫХ И ХРАНИЛИЩА ДАННЫХ



РАЗНИЦА ОЗЕР И ХРАНИЛИЩ ДАННЫХ

Ключевое отличие озер данных от обычных баз данных — структура. В базах данных хранятся только четко структурированные данные, а в озерах — неструктурированные, никак не систематизированные и неупорядоченные.

Пример: представим, что есть вольное художественное описание вашей целевой аудитории: «Девушки возрастом 20–30 лет, незамужние, обычно без детей, работающие на низких руководящих должностях. И мужчины 18–25 лет, женатые, без детей, без четкого места работы». Такое описание — неструктурированные данные, которые можно загрузить в data lake.

Чтобы эти данные о целевой аудитории стали структурированными, их нужно обработать и преобразовать **в таблицу:**

	Пол	Возраст	Семейный статус	Дети	Работа
Портрет 1	женский	20-30	в браке	нет	низкая руководящая должность
Портрет 2	мужской	18-25	в браке	нет	любая

РАЗНИЦА ОЗЕР И ХРАНИЛИЩ ДАННЫХ

В классической базе данных вы должны определить тип данных, проанализировать их, структурировать — и только потом записать в четко определенное место базы данных. Мы можем создать алгоритм, который работает с конкретными ячейками, потому что четко знаем, что хранится в этих ячейках.

В случае с озером данных информацию структурируют на выходе, когда вам понадобится извлечь данные или проанализировать их. При этом процесс анализа не влияет на сами данные в озере — они так и остаются неструктурированными, чтобы их было также удобно хранить и использовать для других целей.

	Пол	Возраст	Семейный статус	Дети	Работа
Портрет 1	женский	20-30	в браке	нет	низкая руководящая должность
Портрет 2	мужской	18-25	в браке	нет	любая

РАЗНИЦА ОЗЕР И ХРАНИЛИЩ ДАННЫХ

Есть и другие различия между базами данных и озерами данных:

Полезность данных

В базах данных все данные полезны и актуальны для компании прямо сейчас. Данные, которые пока кажутся бесполезными, отсеиваются и теряются навсегда.

В озерах хранятся в том числе и бесполезные данные, которые могут пригодиться в будущем или не понадобиться никогда.

Типы данных

В базах данных хранятся таблицы с конкретными цифрами и текстом, распределенными по четкой структуре.

В озерах лежат любые данные: картинки, видео, звук, файлы, документы, разнородные таблицы.

Гибкость

У базы данных гибкость низкая — еще на старте нужно определить актуальные для нее типы данных и структуру. Если появятся данные новых форматов — базу придется перестраивать.

У озер гибкость максимальная, потому что ничего не нужно определять заранее. Если вы вдруг решите записывать новые данные, например, видео с камер для распознавания лиц, озеро не придется перестраивать.

РАЗНИЦА ОЗЕР И ХРАНИЛИЩ ДАННЫХ

Стоимость

Базы данных стоят дороже, особенно если требуется хранить много данных. Нужно организовывать сложную инфраструктуру и фильтрацию, все это требует денег.

Озеро данных стоит намного дешевле — вы платите исключительно за занятые гигабайты.

Понятность и доступность данных

Данные в базе легко смогут прочитать и понять любые сотрудники компании, с ними могут работать бизнес-аналитики.

Чтобы структурировать **данные в озере** требуются технические специалисты, например Data Scientist.

Сценарии использования

Базы данных идеальны для хранения важной информации, которая всегда должна быть под рукой, либо для основной аналитики.

В озерах данных хорошо хранить архивы неочищенной информации, которая может пригодиться в будущем. Еще там хорошо создавать большую базу для масштабной аналитики.

НЕДОСТАТКИ И ОПАСНОСТЬ ОЗЕР ДАННЫХ

У озер данных есть одна серьезная проблема. Любые данные, попадающие в data lake, попадают туда практически бесконтрольно. Это значит, что определить их качество невозможно.

Если у компании нет четкой модели данных, то есть понимания типов структур данных и методов их обработки, плохо организовано управление озером, в нем быстро накапливаются огромные объемы неконтролируемых данных, чаще всего бесполезных.

Уже непонятно, откуда и когда они пришли, насколько релевантны, можно ли их использовать для аналитики.

В итоге наше озеро превращается в болото данных — бесполезное, пожирающее ресурсы компании и не приносящее пользы. Чтобы озеро не стало болотом, нужно наладить в компании процесс управления данными — data governance. Главная составляющая этого процесса — определение достоверности и качества данных еще до загрузки в data lake.

Есть несколько способов это сделать:

- *отсекать* источников с заведомо недостоверными данными;
- *ограничить* доступ на загрузку для сотрудников, у которых нет на это прав;
- *проверять* некоторые параметры файлов, например не пропускать в озеро картинки, которые весят десятки гигабайт.

Настроить такую фильтрацию проще, чем каждый раз структурировать данные для загрузки в базу данных

НЕДОСТАТКИ И ОПАСНОСТЬ ОЗЕР ДАННЫХ

Потеря качества данных

Озеро имеет склонность становиться «болотом» — накапливать данные, которые плохо размечены и никому не нужны. Это может привести к тому, что Data Lake просто больше нельзя будет использовать для аналитики — его придется полностью стирать и наполнять заново, уже более аккуратно.

Техническая сложность

Создание озера данных — непростая задача. Нужна инфраструктура: мощные серверы, надёжные каналы связи, большие объёмы дискового пространства, а ещё опытные инженеры, которые будут это поддерживать. Технология Data Lake для России относительно новая, и специалистов на рынке не очень много, поэтому их придётся долго искать и много им платить.

НЕДОСТАТКИ И ОПАСНОСТЬ ОЗЕР ДАННЫХ

Дополнительные затраты на извлечение данных

Помещать данные в озеро можно почти мгновенно, а для извлечения часто нужны сложные инструменты поиска и очистки, которые придётся настраивать отдельно. В этом плане озеро уступает хранилищу данных, в котором всё хранится по заранее проработанной структуре.

Хранение лишнего

Данные в озеро часто поступают бесконтрольно. Из-за этого в нём может быть много дублей и файлов, которые вообще не нужны ни для какой аналитики. Из-за этого озеро может разрастись и потреблять слишком много ресурсов бизнеса.

ГЛАВНОЕ ОБ ОЗЕРАХ ДАННЫХ

Data lake — это озеро данных, хранилище, в котором собрана неструктурированная информация любых форматов из разных источников.

Озера данных дешевле обычных баз данных, они более гибкие и легче масштабируются.

Озера данных можно использовать для любых целей: анализов, прогнозов, оптимизации бизнес-процессов.

Данные можно извлекать из озера по определенным признакам или анализировать прямо внутри озера, используя системы аналитики.

Если собирать слишком много данных «просто так» и никак с ними не работать, озеро может стать бесполезным болотом. Поэтому важно заранее определить, для чего именно вы собираете данные, и не накапливать их просто так.

ЗАДАНИЕ

Задача — изобразить схему работы фреймворка, с учетом расширения функционала с точки зрения дополнительных источников данных.

Подробное описание смотрите под QR

Решения — отправляйте на почту **stagi@1t.ru**

СПАСИБО ЗА ВНИМАНИЕ

Баданов Артем
Data Engineer

Telegram : @artem5240
+7 (977) 699-82-41