Озеро данных (Data Lake)

> Разделы: Источники данных

Data Lake (Озеро данных) — метод хранения структурированной, полуструктурированной и неструктурированной информации, а также организации больших объемов данных, поступающих из различных источников, таких как логи, события, медиафайлы и т.д.

Озеро данных ориентировано на консолидацию «сырых» данных, которые в дальнейшем могут быть преобразованы и использованы для аналитики, машинного обучения и других целей. Оно обычно используется в формате «храни все», то есть информация, поступающая в систему, складируется без изменений. Данный метод обеспечивает только базовую согласованность данных.

Озеро, как и хранилище данных, решает задачу консолидации, но отличается от него в фундаментальных подходах к работе с информацией.

Сравнительная характеристика методов хранения информации представлена в таблице:

Характеристики	Озеро данных	Хранилище данных
Хранение данных	Содержит все данные организации в независимости от их структуры и источника, а также может хранить информацию неограниченный период времени	Содержит только обработанные структурированные данные, подготовленные для конкретных бизнес задач
Пользователи	Аналитики и инженеры данных используют для изучения информации в сыром виде, для выявления тенденций и формирования новых уникальных бизнес-идей	Менеджеры и конечные бизнес-пользователи используют для получения ответов на поставленные вопросы

Характеристики	Озеро данных	Хранилище данных
Анализ	Предсказательная аналитика, машинное обучение, BI и аналитика big date	Визуализация данных, BI, аналитика данных
Схема хранения	Определяется после сохранения информации	Задается до сохранения информации
Обработка	Использует процесс ELT	Использует процесс ETL

Озеро данных имеет ряд преимуществ, выделяющих его на фоне других способов хранения информации:

- дешевизна реализации;
- быстрая адаптивность к изменениям;
- централизация различных источников данных;
- гибкий доступ к данным из любого места.

Но, несмотря на ряд весомых преимуществ, существуют определенные риски. В частности, нельзя быть уверенным в достоверности результатов анализа, так как часто нет информации о том, откуда были взяты исходные сведения. К недостаткам также можно отнести появление сомнительных данных, которые трудно проверить. Никто не ведет контроль при их заливке, что позволяет удешевить сбор и хранение данных, но ввиду этого существует риск превратить озеро в «болото».

Организация озера данных — сложный процесс, требующий компетентного подхода, но универсальность и высокая польза для бизнеса делает Data lake одним из популярных методов хранения информации.