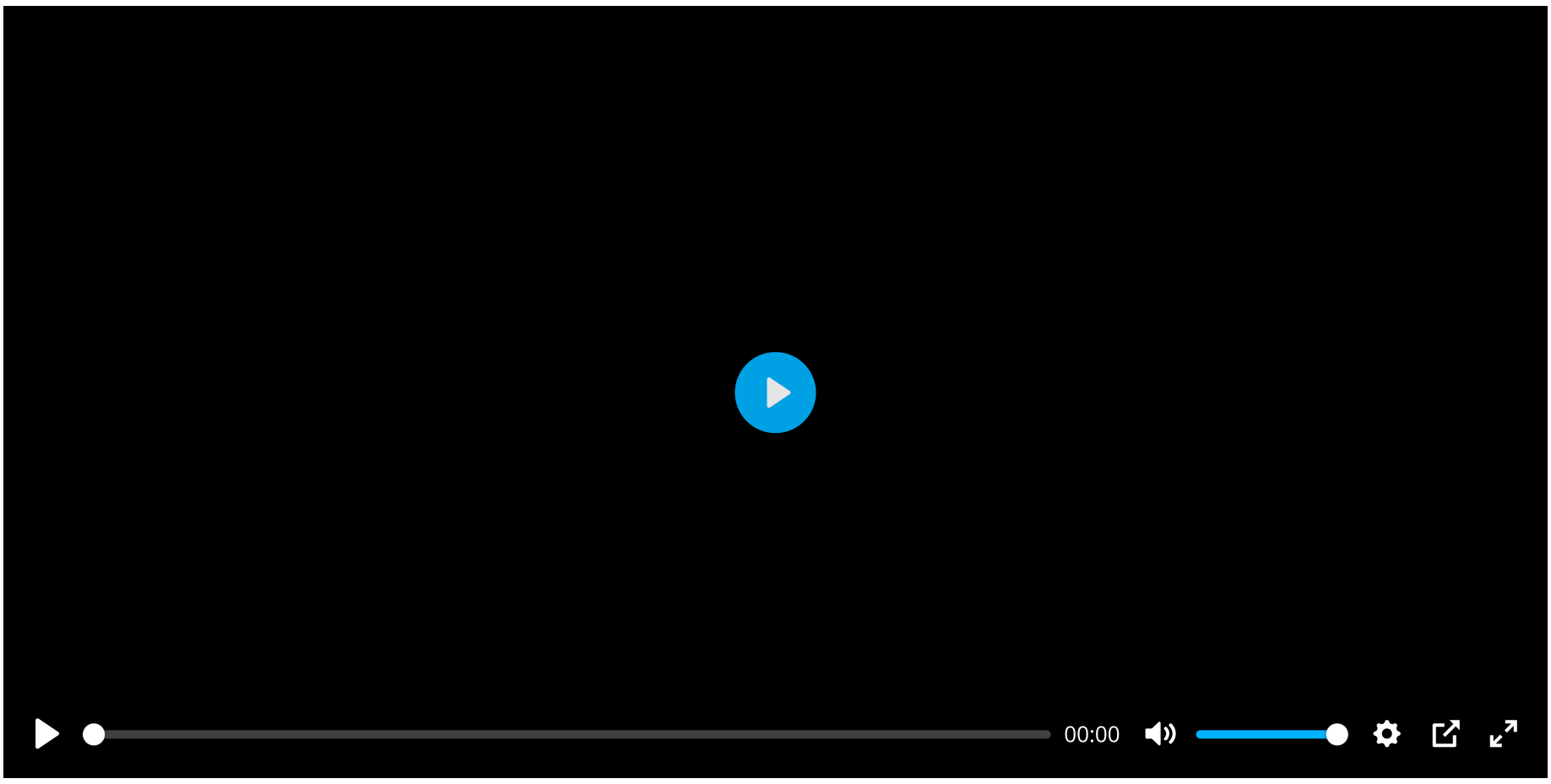
7.2 Процессы ETL и ELT



Текстовая расшифровка видео:

ЗАЧЕМ НУЖЕН ETL

План:

- Перекладывание данных;
- Что мы хотим по итогу;
- ETL;
- Зачем мы все это делаем;
- Reverse ETL.

Перекладывание данных

Как выглядит работа с пайплайнами:

- У организации есть **источники данных**, например, логи поведения пользователей на сайте.
- Эти данные нужно **выгрузить и положить в хранилище** неструктурированных (**Data Lake**) или структурированных (**Data Warehouse**) данных.
- Также может потребоваться переложить данные **из неструктурированного в структурированное**.

Весь этот процесс по перекладыванию данных будет состоять из трех стадий:

- **Извлекаем** (Extract);
- **Преобразуем** (Transform);
- **Загружаем** (Load).

Что мы хотим по итогу

Мы преследуем цели:

- **Достать** данные из источника в виде **пакетов/файлов** или **потока**.

- **Отфильтровать** данные по необходимым признакам согласно тому, что требуется для анализа.
- **Стандартизировать** и привести к нужному виду для заливки дальше, например, в формате **CSV**, **Parquet** или **Avro**.
- **Залить** данные в хранилище.

ETL

Аббревиатура «ETL» используется практически для любого процесса перекладывания данных. Однако чаще всего имеется в виду именно **заливка данных из источников в хранилище**.

Зачем мы все это делаем

Очевидно, что мы перегружаем все в аналитическое хранилище. Однако есть и другие причины:

- Бизнесу необходимо **быстро** получать **релевантные инсайты** на основе своих данных в процессе **Master Data Management**.

Ранее мы говорили о книге «DATA-DMBOK» (Data Management Body of Knowledge), где описываются разные бизнес-термины, связанные с перекладыванием данных.

- Для этого должен существовать **доверенный** источник данных, на основе которого можно делать аналитику – **SSOT (Single Source of Truth)**.

SSOT – это набор данных, который является основным для того, чтобы на его основе делать дальнейшие выводы.

- В некоторых случаях дополнительно используется термин «**Golden Record**» – это эталонный набор данных, например, о клиентах.

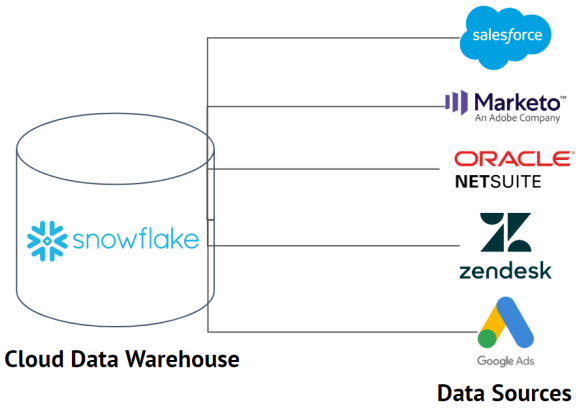
Проще говоря, ETL забирает данные и превращает их в нужный для бизнес-аналитики вид.

Reverse ETL

Когда компании начали выгружать свои данные в разные хранилища, они поняли, что у них есть свои платформы с продуктовыми данными (Salesforce, Zendesk, разные CRM и т.д.). Однако то, что оно есть в том же Salesforce, не очень помогает делать кросс-доменную аналитику; так или иначе данные нужно извлечь в хранилище и уже там что-либо считать.

Когда мы уже что-то посчитали, нам вероятнее всего придется эти посчитанные данные вернуть в систему, которую использует определенный отдел. Соответственно, получается **обратный процесс**.

На нашей практике подобное встречается нечасто.



Как вам урок?

Изучил, далее