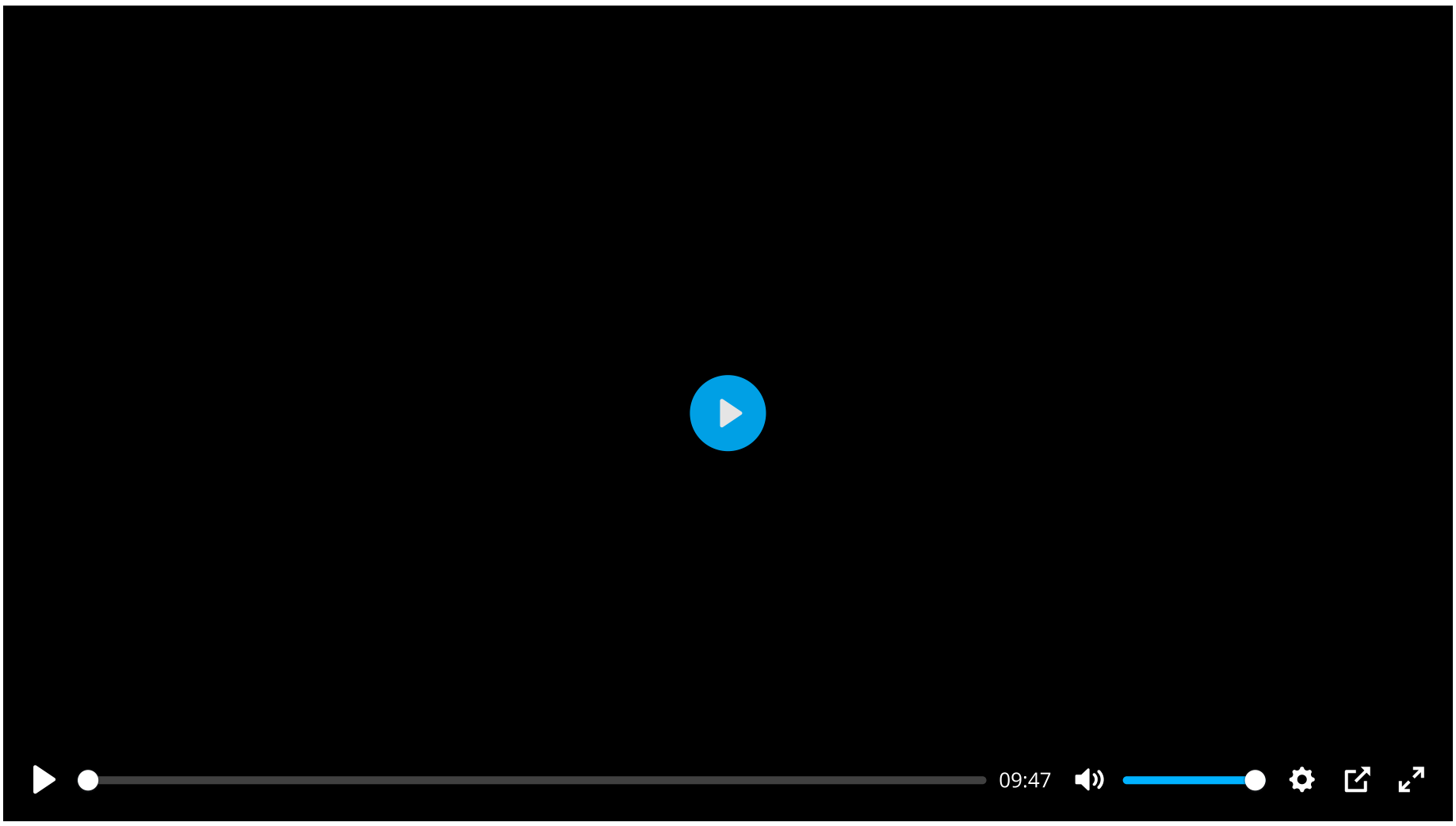
7.2 Процессы ETL и ELT



Текстовая расшифровка видео:

СУТЬ ЭТАПОВ «EXTRACT», «TRANSFORM» И «LOAD»

План:

- E – Extract;
- T – Transform;
- L – Load;
- Резюмируем;

E – Extract

Extract – это этап извлечения данных из первичных источников.

Данные из источников могут извлекаться разными способами, однако базовым способом является **пакетная обработка (batching)**, при которой:

- исходные данные извлекаются большими порциями из источника данных в целевую систему по расписанию (через запланированные промежутки времени).

Альтернативный способ извлечения данных – **потоковая обработка (streaming)**, при которой:

- данные как непрерывно извлекаются из источника данных в режиме реального времени, так и поступают в источник данных.

Один из вариантов стриминга – **потоковая обработка маленьких порций данных (microbatching)**:

- способ предназначен для того, чтобы в режиме реального времени в процессе передачи данных перед загрузкой в хранилище осуществлять над ними преобразования.

T – Transform

Transform – это этап преобразования извлеченных данных.

Это очистка и валидация данных:

- Исправление любых ошибок и заполнение отсутствующих значений;
- Удаление дубликатов;
- Логирование невалидных записей, полученных из источника.

Именно здесь подготавливаются данные для дальнейшего использования кем-либо.

Data Janitor – профессия, направленная на преобразование данных, на приведение их в должный вид.

Этапы подготовки данных:

- нормализация;
- обогащение;
- кодирование, анонимизация и обезличивание;
- объединение данных из разных источников;
- структурирование, преобразование одного формата в другой;
- генерация бизнес-идентификаторов.

Базовые операции (оптимизация данных для потребителей):

- сортировка;
- фильтрация;
- агрегирование.

L – Load

Load – это этап загрузки данных в хранилище.

Здесь может быть как **первичная загрузка** данных в хранилище, так и **инкрементальная загрузка**.

Инкрементальная загрузка – это периодическое (по расписанию или событию) добавление новых данных или модификация существующих в соответствии с изменениями в первичных источниках данных.

Иногда возможен вариант, при котором данные представляют из себя внешний справочник, и мы не сильно отслеживаем в нем изменения. Если он небольшой, то можем перезаливать целиком, сносить данные из таблички, заливать новую порцию и т.д.

Полное обновление данных в хранилище – это удаление содержимого одной или нескольких таблиц и загрузка свежих данных.

На этапе «Transform» или «Load» (чаще «Load») могут затесаться дополнительные проверки – Data Quality.

Data Quality – это набор подходов для оценки качества данных.

Резюмируем

- С появлением инструментов для потоковой обработки данных все чаще используется для обработки данных о событиях в реальном времени.
- ETL чаще всего как аббревиатура используется для описания процесса по выгрузке данных из источника хранилища.
- Исторически использовался для реализации пакетной обработки данных в больших масштабах.