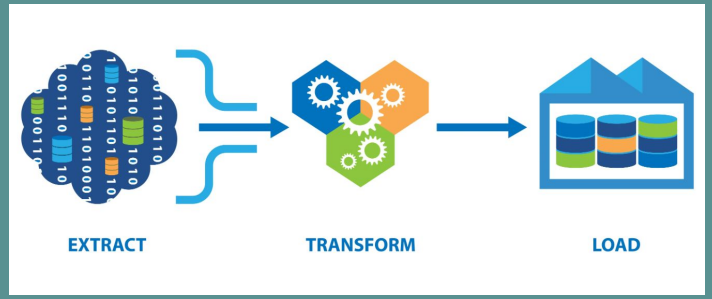


Co to jest proces ETL?

Ekstrakcja, Transformacja, Ładowanie danych
Witold Bazela

Wprowadzenie do ETL

ETL (**E**xtract **T**ransform **L**oad) to proces pobierania danych z różnych źródeł, ich transformacji oraz ładowania do systemu docelowego.

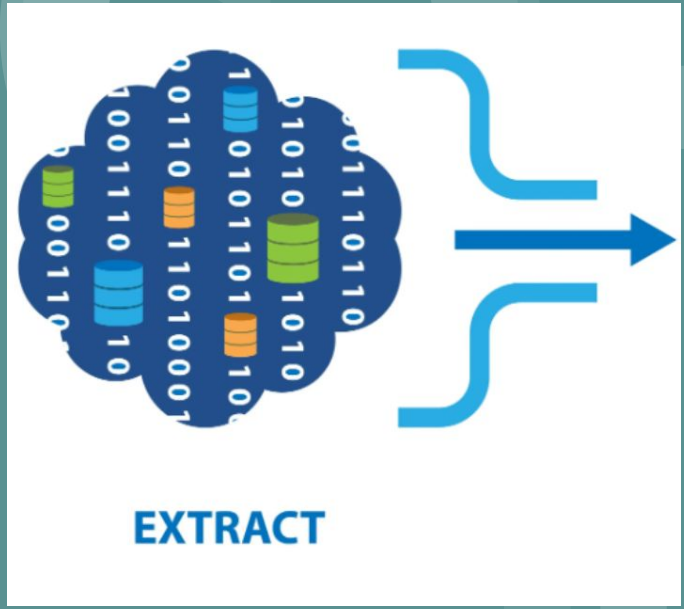
Ponieważ jest to proces jest on agnostyczny względem technologii.

Jest kluczowy w inżynierii danych i analityce biznesowej.

Etap 1: Ekstrakcja (Extract)

Ekstrakcja to pobieranie danych z różnych źródeł:

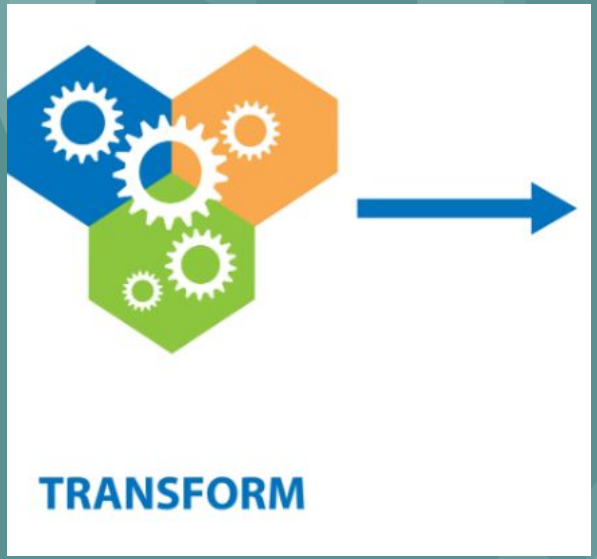
- Bazy danych
- Pliki CSV, XML, JSON
- API i usługi webowe
- Streamingi
- Systemy ERP i CRM



Etap 2: Transformacja (Transform)

Transformacja obejmuje:

- Czyszczenie danych (usuwanie duplikatów, brakujących wartości)
- Zmianę formatów i jednostek
- Mergowanie danych
- Agregację (sumowanie, grupowanie)
- **Szyfrowanie i anonimizacja danych**
- Obsługa błędów i wyjątków



Etap 3: Ładowanie (Load)

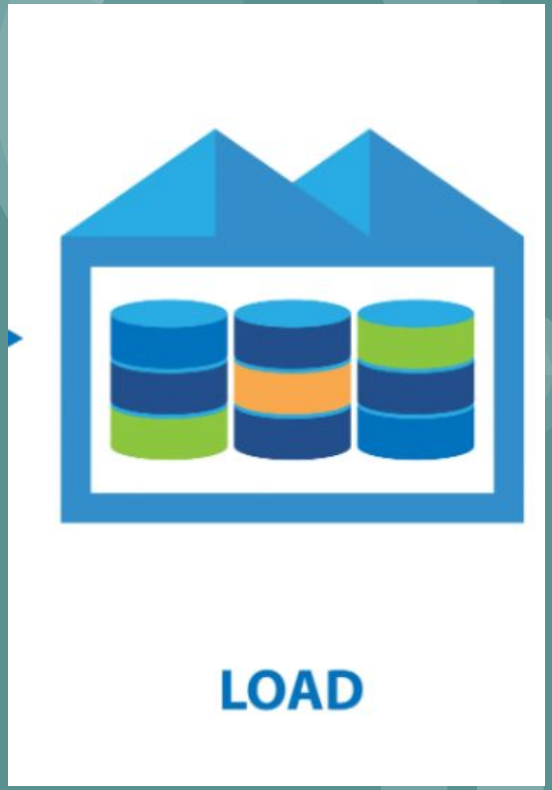
Ładowanie danych do systemu docelowego, np.:

- Hurtownie danych
- Bazy danych OLAP/OLTP

(My na początku będziemy rozładowywać dane do pliku .csv, to też jest ok)

Strategie ładowania:

- Pełne (wszystkie dane od nowa)
- Przyrostowe (tylko nowe lub zmienione dane)



Zastosowania ETL

ETL jest wykorzystywane w:

- Business Intelligence (BI)
- Analityce Big Data
- Integracji danych z różnych źródeł
- Przetwarzaniu danych w chmurze

Popularne narzędzia ETL

Przykładowe narzędzia ETL:

- AWS Glue
- Apache Spark
- Apache Beam
- DBT

Każde z nich pozwala na efektywne przetwarzanie i integrację danych.

ETL vs ELT

Oto kluczowe różnice między ETL (Extract, Transform, Load) a ELT (Extract, Load, Transform):

- Kolejność przetwarzania danych
  - ETL: Ekstrakcja → Transformacja → Ładowanie
  - ELT: Ekstrakcja → Ładowanie→ Transformacja
- Miejsce przetwarzania danych
  - ETL: Transformacja odbywa się przed załadowaniem do hurtowni danych (zazwyczaj na osobnym serwerze ETL).
  - ELT: Transformacja odbywa się w hurtowni danych lub w systemie Big Data.
- Zastosowanie
  - ETL: Tradycyjne hurtownie danych (np. Oracle, Teradata, SAP BW).
  - ELT: Chmurowe hurtownie danych i Big Data (np. Snowflake, Google BigQuery, AWS Redshift, Databricks).
- Wydajność
  - ETL: Może być wolniejszy, szczególnie przy dużych zbiorach danych.
  - ELT: Wykorzystuje moc obliczeniową chmurowych baz danych, co przyspiesza procesy na dużą skalę.
- Elastyczność
  - ETL: Lepsze do przetwarzania uporządkowanych danych i skomplikowanych transformacji przed załadunkiem.
  - ELT: Lepsze do analizy dużych, nieustrukturyzowanych zbiorów danych i pracy z Data Lakes.

ETL vs ELT: Understanding the Key Differences

Podsumowanie teorii

Podsumowanie:

- ETL to kluczowy proces w inżynierii danych
- Składa się z trzech etapów: Ekstrakcja, Transformacja, Ładowanie
- Wykorzystuje się go w analityce biznesowej i Big Data

Czym jest AWS Glue?

AWS Glue to w pełni zarządzana usługa ETL od Amazon Web Services.

Pozwala na ekstrakcję, transformację i ładowanie danych w chmurze bez potrzeby zarządzania infrastrukturą.

Kluczowe funkcje AWS Glue

- Automatyczne wykrywanie schematów danych (Glue Data Catalog)
- Generowanie kodu ETL w Pythonie/Scala
- Obsługa batch i streamingu
- Integracja z innymi usługami AWS (S3, RDS, Redshift)

Jak działa AWS Glue?

1. Ekstrakcja danych z różnych źródeł (S3, RDS, DynamoDB)

2. Transformacja i czyszczenie danych przy użyciu Spark

3. Ładowanie danych do hurtowni danych (Redshift, S3, Athena)

Zadanie

1. Pobierz ze strony https://www.investing.com/ dane zawierające codzienne ceny ETFu S&P 500 (SPX) w formacie csv.
2. Stwórz bucket na S3 oraz prześlij tam plik z cenami
3. Stwórz AWS Glue job który pobierze nasz plik csv(Extract), Wyciągnie tylko cenę(Price) przemnożoną przez dzisiejszą cenę złotówki oraz datę(Date) i zapisze w naszym bucketcie pilk z wynikiem.

S&P 500 Historical Data (SPX) - Investing.com

Zadanie (dla kursantów)

1. Pobierz ze strony https://www.investing.com/ dane zawierające codzienne ceny iShares MSCI World Momentum Factor UCITS (IWFM) w formacie csv.
2. Stwórz bucket na S3 oraz prześlij tam plik z cenami
3. Stwórz AWS Glue job który pobierze nasz plik csv(Extract), Wyciągnie tylko cenę(Price) przemnożoną przez dzisiejszą cenę złotówki oraz datę (Date) i zapisze w naszym bucketcie pilk z wynikiem.
4. *pobierz dane o codziennej cenie dolara Archiwum kursów średnich – tabela A (CSV, XLS) | Narodowy Bank Polski - Internetowy Serwis Informacyjny zmodyfikuj obecny job tak żeby cena każdego dnia była oparta na realnej cenie złotego danego dnia

IWFM ETF Stock Price History - Investing.com

Zadanie 2 (dla kursantów)

Stwórz nowy job który pobierze utworzony w poprzednim zadaniu plik S3, a następnie stworzy nowy plik z datą w której cena akcji była najwyższa.

Początkowy kod

```python
import boto3
import os

# Konfiguracja
SOURCE_BUCKET = "twoj-bucket-zrodlowy"
SOURCE_KEY = "sciezka/do/pliku.csv"
TARGET_BUCKET = "twoj-bucket-docelowy"
TARGET_KEY = "sciezka/do/nowej-nazwy-pliku.csv"

# Inicjalizacja klienta S3
s3 = boto3.client('s3')


def copy_csv():
    try:
        # Pobranie pliku z S3
        response = s3.get_object(Bucket=SOURCE_BUCKET, Key=SOURCE_KEY)
        data = response['Body'].read()

        # Zapisanie pliku z nową nazwą w S3
        s3.put_object(Bucket=TARGET_BUCKET, Key=TARGET_KEY, Body=data)
        print(f"Plik skopiowany z {SOURCE_BUCKET}/{SOURCE_KEY} do {TARGET_BUCKET}/{TARGET_KEY}")
    except Exception as e:
        print(f"Błąd podczas kopiowania pliku: {e}")


copy_csv()
```