

Zebras have stripes as distinctive as fingerprints.

They don't see how their individual patterns set them apart.
But you can. With proven customer intelligence software and services from SAS.

www.sas.com/zebras

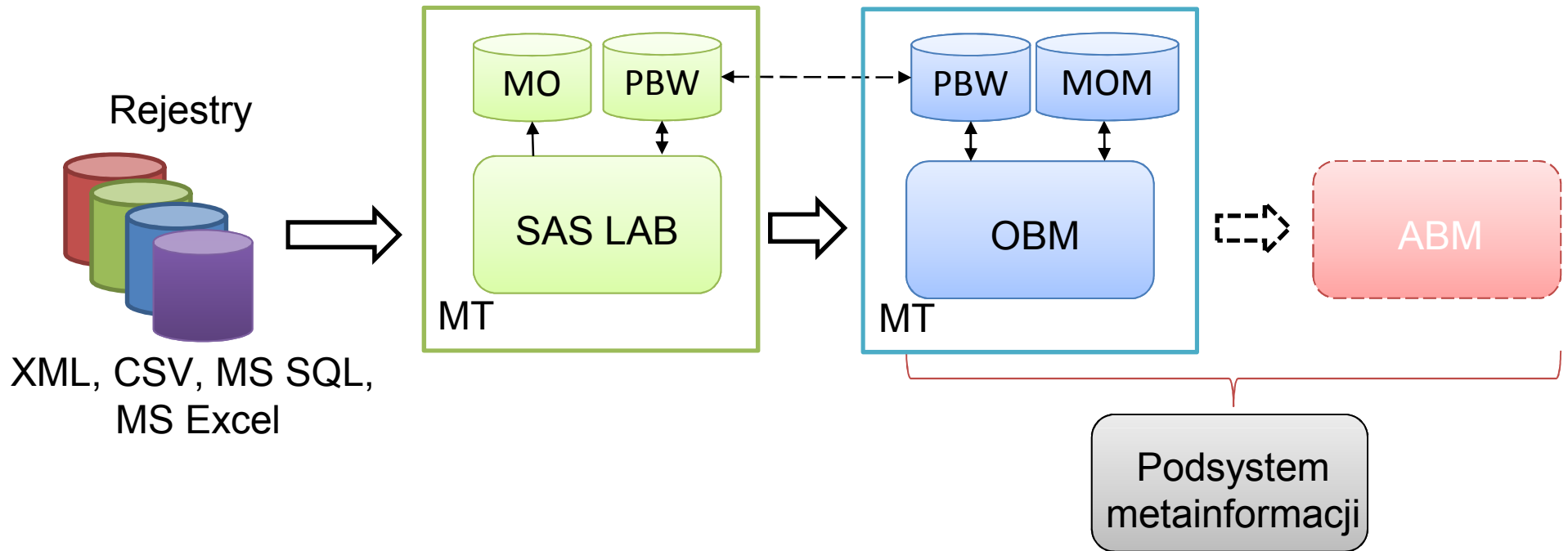


**PRZEKSZTAŁCENIE ZBIORÓW
PUBLICZNYCH W REJESTRY
STATYSTYCZNE PRZY
POMOCY NARZĘDZI ETL**

sas

THE
POWER
TO KNOW.

Architektura SAS na potrzeby spisów



Wykorzystywane produkty:

- SAS Enterprise DI Server
 - SAS Data Quality Solution
 - SAS/STAT, SAS/IML

Polska Baza Wiedzy (PBW)

Repozytorium przechowujący algorytmy, służące do przeprowadzania wybranych transformacji na danych

Możliwości:

- Parsowanie
- Standaryzacja
- Integracja
- Identyfikacja

Składowe definicji:

- Wyrażenia regularne
- Tablice dzielące (chop table)
- Słowniki
- Schematy standaryzacyjne
- Reguły gramatyczne
- Reguły fonetyczne

Metadane

Metadane operacyjne (MO):

- rejestrowanie przebiegów procesów ETL
- statystyki z przebiegów procesów i danych wynikowych

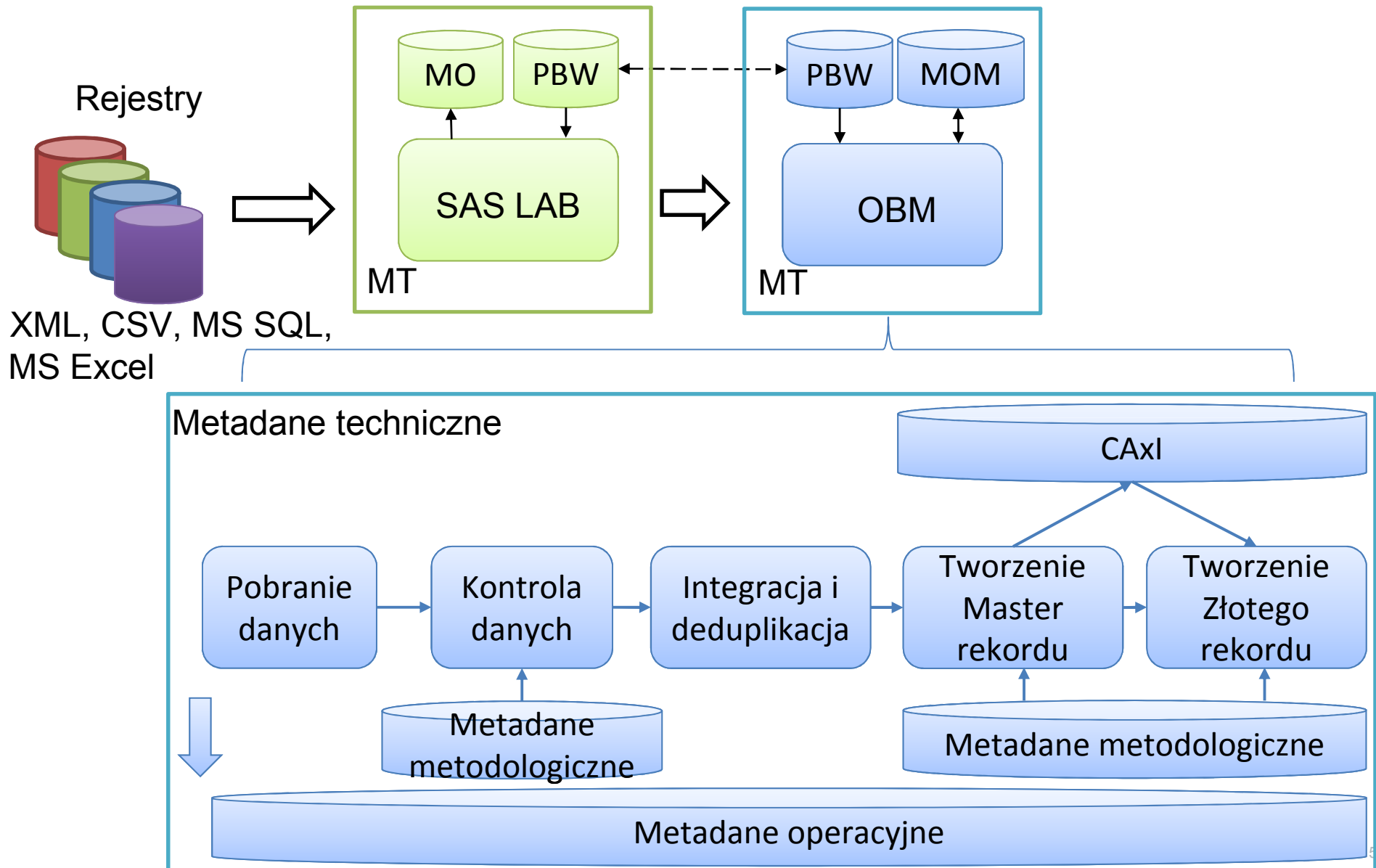
Metadane metodologiczne (MM)

- sterowanie procesami ETL poprzez zewnętrzne reguły bez potrzeby ingerencji w samą strukturę zadań
- struktura danych wynikowych (np. Master rekord, Złoty rekord) zdefiniowana poza procesem ETL

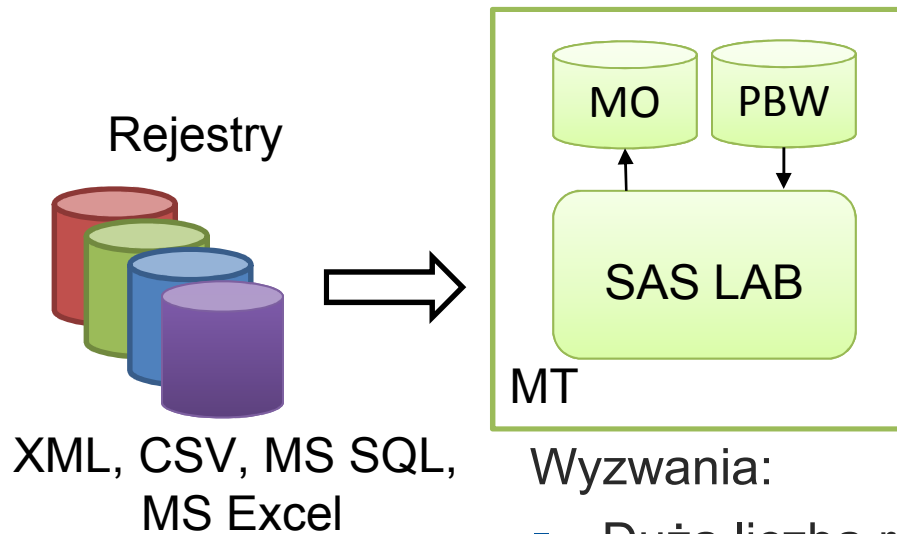
Metadane techniczne (MT)

- Bezpieczeństwo - warstwa uprawnień
- Informacje o środowisku (zbiory, serwery,...)

Środowisko OBM



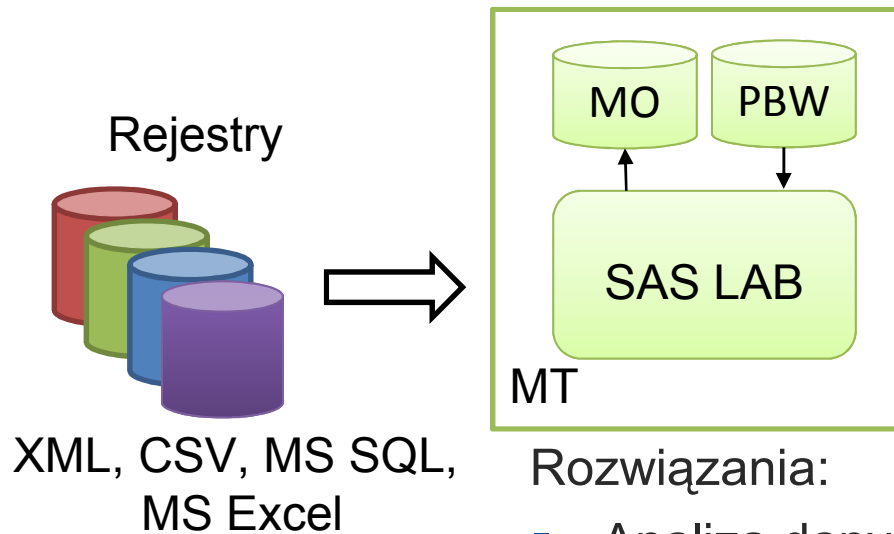
Środowisko SAS LAB



Wyzwania:

- Duża liczba rejestrów
- Zmienność rejestrów w czasie
- Rozbudowane modele danych
- Dane w różnych formatach
- Jakość danych
- Różne typy zadań czyszczenia

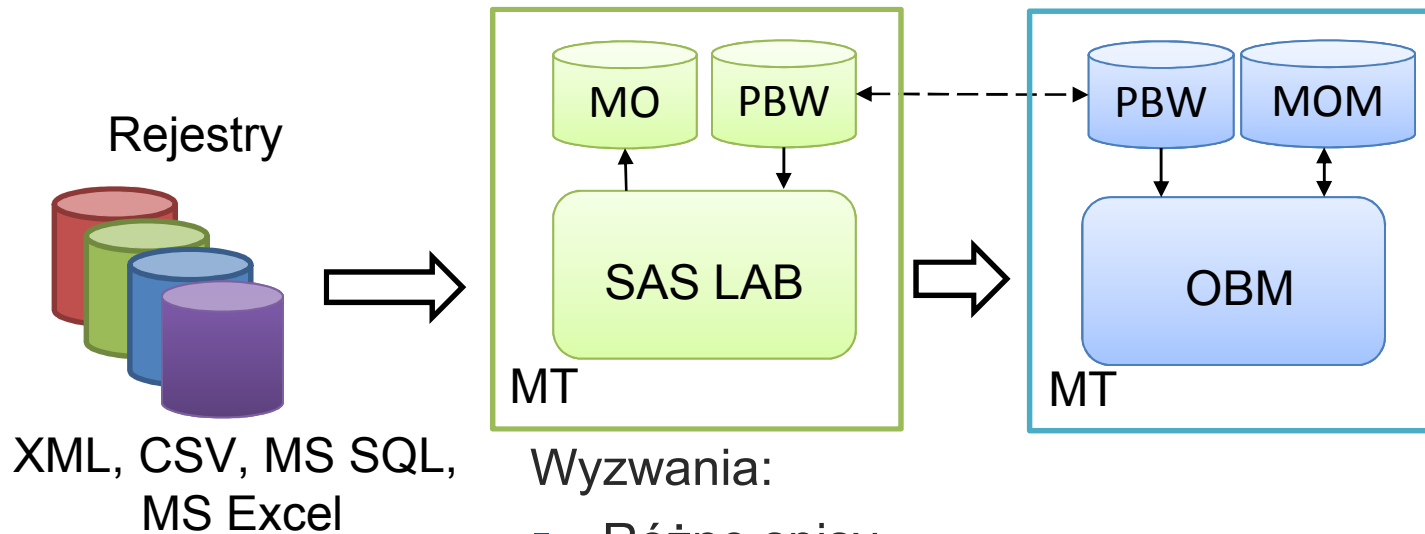
Środowisko SAS LAB



Rozwiązania:

- Analiza danych - profilowanie danych
- Wykorzystanie polskiej baza wiedzy (PBW)
 - Tworzenie schematów standaryzacyjnych
 - Modyfikacja istniejących reguł na potrzeby danych spisowych
- Dodatkowe algorytmy podobieństwa ciągów

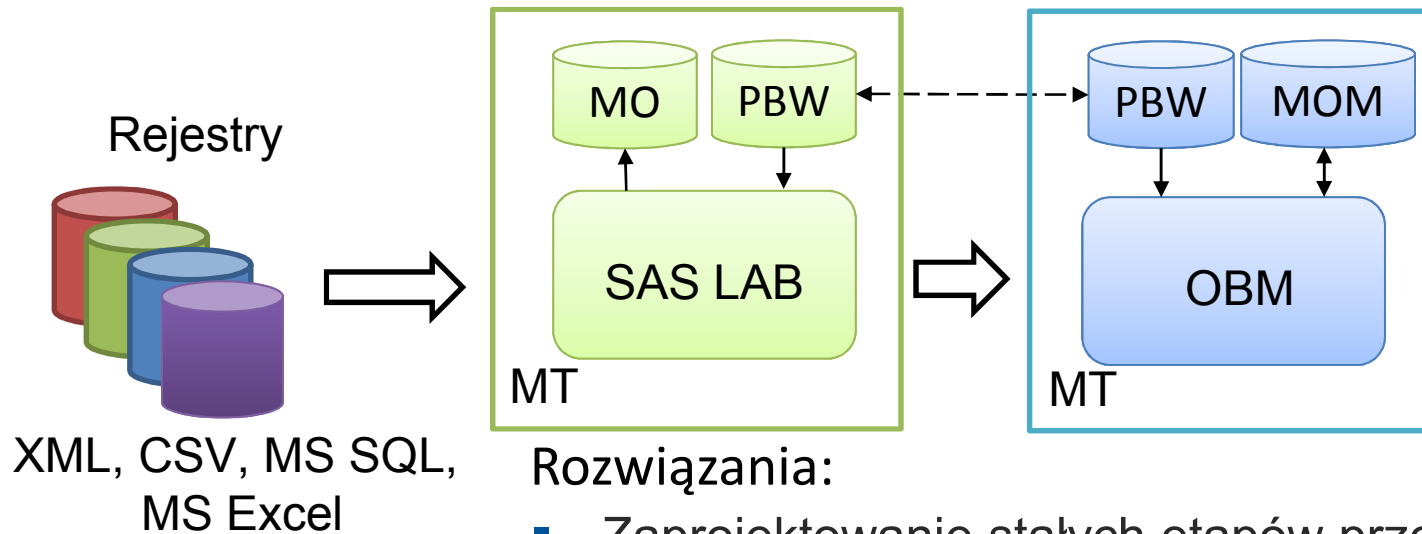
Środowisko OBM



Wyzwania:

- Różne spisy
- Zmienna struktura master rekordu i złotego rekordu
- Duża liczba rejestrów – duża liczba zadań
- Krótkie terminy realizacji
- Duży wolumen danych
- Integracja pomiędzy różnymi rejestrami – brak wspólnego identyfikatora

Środowisko OBM



Rozwiązania:

- Zaprojektowanie stałych etapów przetwarzania
- Metadane metodologiczne:
 - Zawarcie struktur master rekordu i złotego rekordu
 - Opracowanie autorskich transformacji danych sterowanych regułami zewnętrznymi
- Wykorzystanie Podsystemu Metainformacji
- Wykorzystanie algorytmów równolegających
- Wydajna infrastruktura sprzętowa

