BUSINESS INTELLIGENCE TRENDS

Nowe wyzwania w obszarze hurtowni danych

Warszawa, 26 czerwca 2018

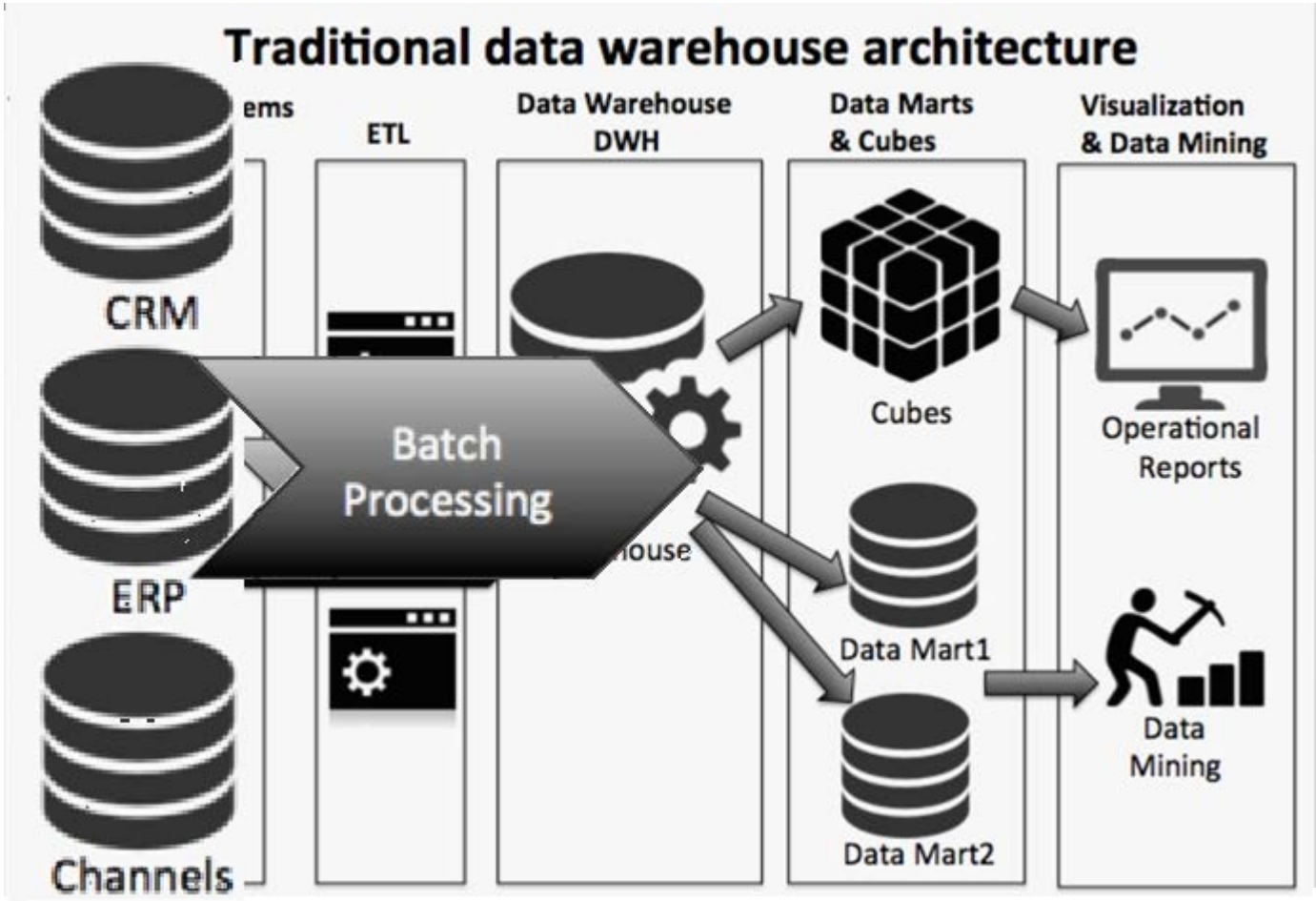
Zdzisław Dec

Co się zmieniło w ostatnich latach ?

- nowe źródła danych: portale społecznościowe, portale zakupowe, geolokalizacja
- nowe kanały sprzedażowe: Internet, mobile, bankomaty, hipermarkety
- nowe formy płatności: PayPass, Blik, NFC, SMS, Facebook
- szeroka gama produktów bankowych (nowe typy kart, ubezpieczenia, fundusze itp.)
- zmiana profilu lojalnościowego klientów (wyszukiwarki, rankingi, e-banking)
- wymogi GDPR/RODO oraz inwestorów strategicznych dot. precyzyjnego śledzenia jakie dane i do czego są wykorzystywane (data lineage, data governance)

Klasyczny model przetwarzań hurtowni

.. oparty o ETL i ładowanie batchowe jest coraz mniej niewydajny



Nowa rzeczywistość w IT banków

I. Zwiększenie liczby źródeł i wielkości danych w systemach bankowych

II. Wydłużenie procesów zamknięcia dnia

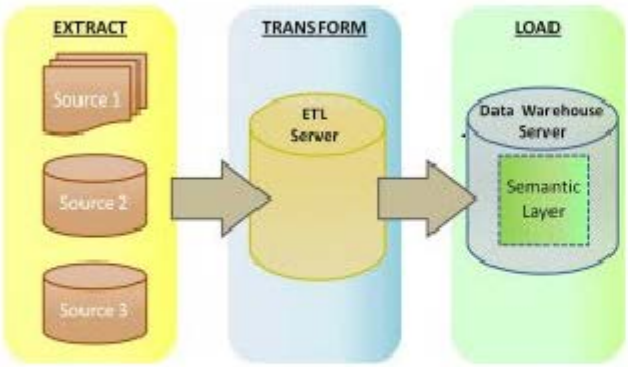
III. Zwiększenie zależności pomiędzy systemami
(jeden czeka, aż drugi skończy przetwarzanie)

W naszym Banku na przestrzeni ostatnich 10 lat:

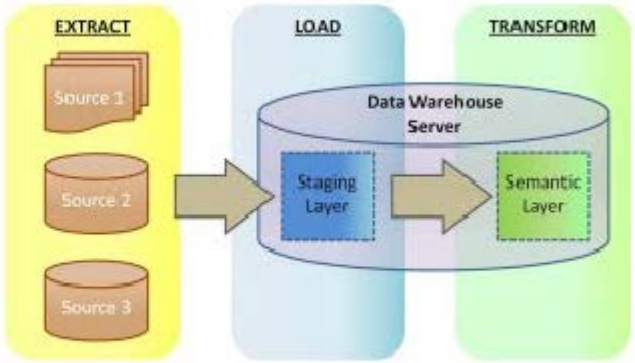
- Liczba aktywnych klientów: x7
- Liczba przetwarzanych rekordów danych w ciągu miesiąca: x400

OD CZEGO ZACZĄĆ ?

Przejście z modelu ETL na ELT

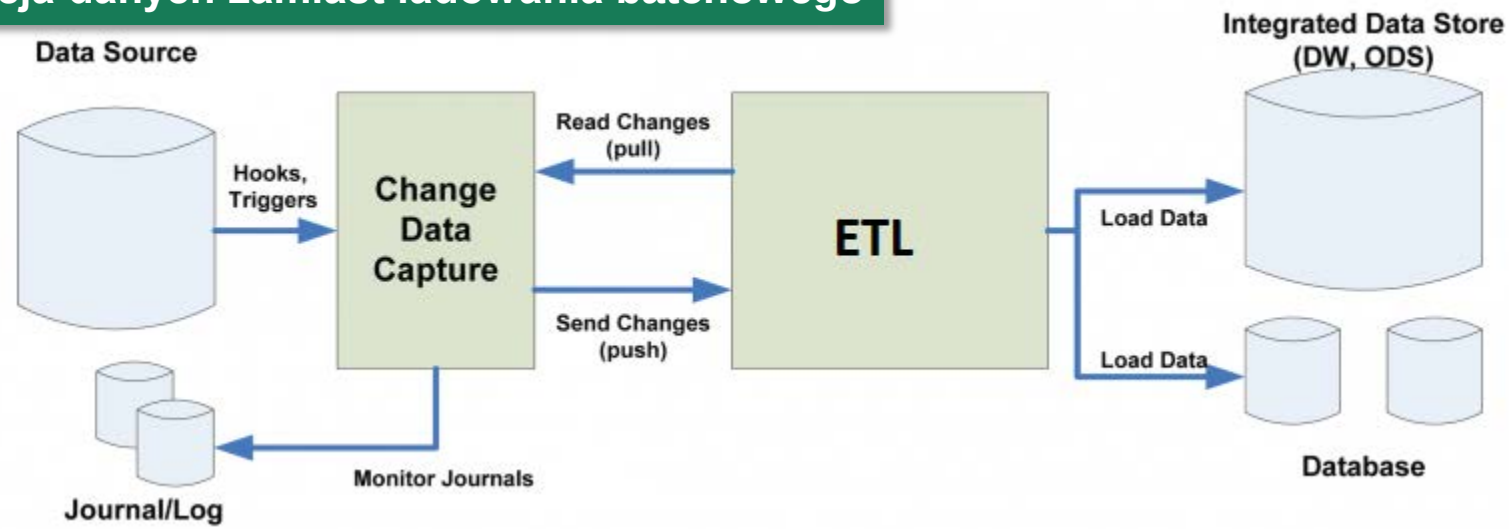


ETL Diagram



ELT Diagram

Replikacja danych zamiast ładowania batchowego



Mamy zebrane dane źródłowe co dalej ?

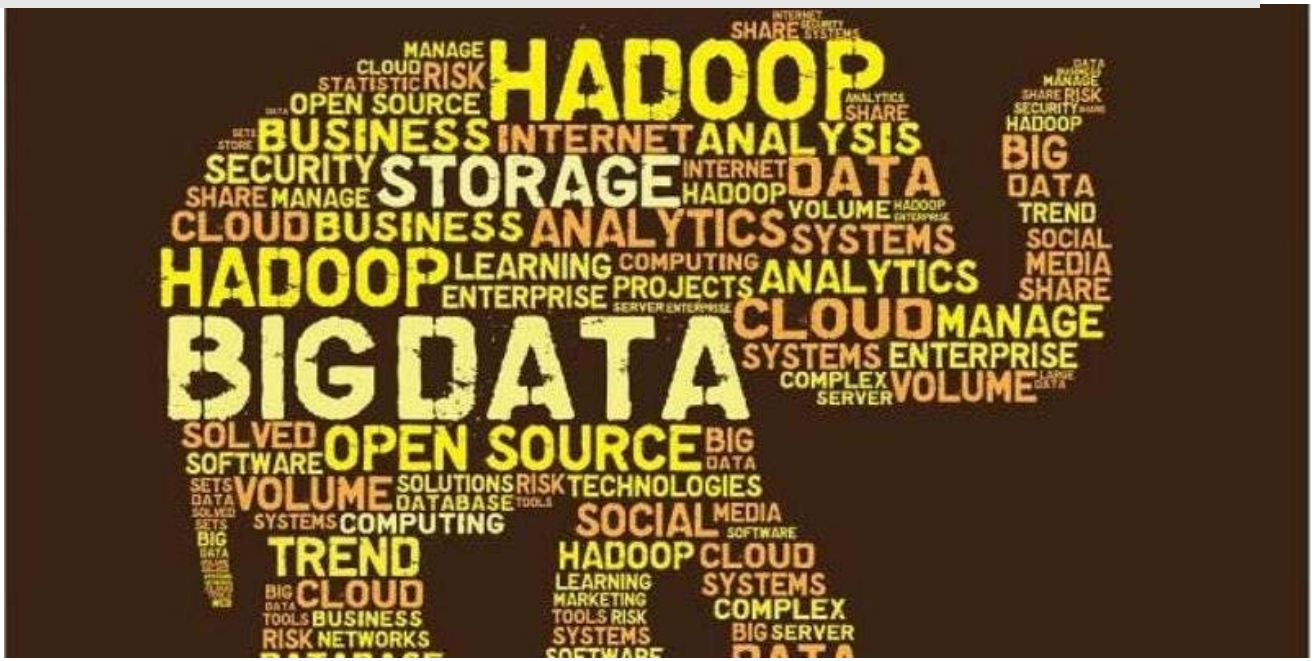
MINY i ZASADZKI ELT:

- skomplikowane przetwarzania to wiele transformacji i wiele wyników pośrednich
- silniki bazodanowe średnio nadają się do szybkich, masowych i równoległych zapisów
- „data warehouse appliance” to droga zabawka
- „pushdown” to nie ELT

Inwestujemy w klaster BIG DATA

MINY i ZASADZKI HADOOP:

- zoptymalizowane pod zapis i odczyt dużych plików
- optymalne zapytania tylko na pojedynczym zbiorze
- dużo narzędzi open source i brak ekspertów
- problemy z „data lineage”
- konieczność „uczłowieczania” (Hive LLAP, Kudu, Sqoop)



Nowoczesny model hurtowni danych:

- Zmiana w podejściu do modelowania danych (Data Vault ?)
- Drogie narzędzia maintstreamowe czy open source ?
- Jak pogodzić dane batchowe i w trybie real-time ?
- Jak pogodzić dane tabelaryczne z niestrukturalnymi ?
- I jak w tym wszystkich zachować data lineage ?