Czy można zarządzać jeziorem danych?

Czy Big Data zmienia sposób pracy z danymi?

Plan na dziś

1. O co chodzi z tym jeziorem i jak to się ma do hurtowni danych?
2. Data Governance: administracja danymi?
3. Big Data Lifecycle
4. Chief Data Officer: kto to i po co?

Po co powstały hurtownie danych?

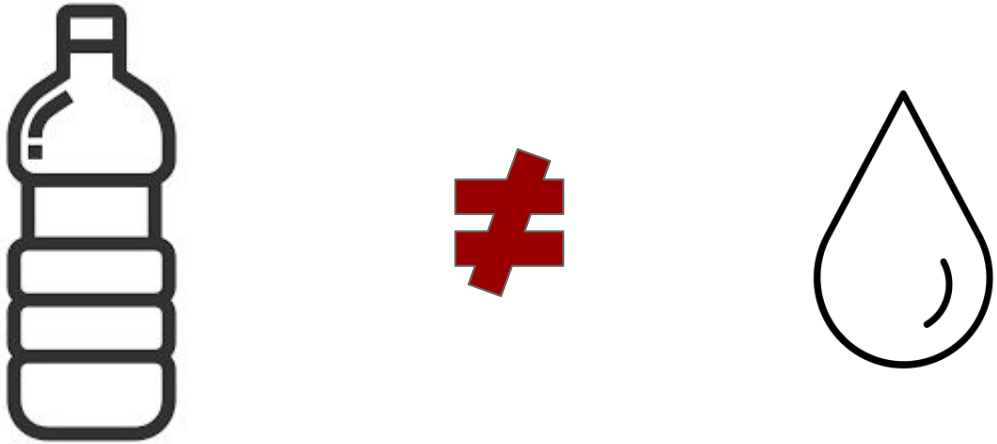
1. **Dostęp** - dostęp do danych sam w sobie jest sukcesem (dla wielu organizacji nadal).
2. **Szybkość** – dane gromadzone w centralnym miejscu zapewniają szybką odpowiedź na część pytań biznesowych.
3. **SVOT (single version of true)** - wielka obietnica hurtowni danych. Wartość obrotu za wybrany okres czasu jest jednoznaczna, profil wybranego klienta..... niekoniecznie.

To w czym problem?

1. SVOT nigdy nie został realnie osiągnięty (w wypadku niektórych informacji po prostu nie istnieje).
2. Nowe rodzaje informacji często stawiają architekturę hurtowni pod znakiem zapytania (więc lepiej uznać że ich nie ma lub są niepotrzebne).
3. Żaden projekt DWH nie doczekał oficjalnego zakończenia.
4. DWH zwykle okazuje się dostępna i szybka w odpytywaniu tylko dla wtajemniczonych.
5. Inkrementalne zmiany architektury DWH wymagają nieakceptowalnie dużego nakładu (z czasem ten problem eskaluje).

Stephen J. Smith: https://www.eckerson.com/articles/the-demise-of-the-data-warehouse

Data Lake



Data Mart: oczyszczony, spakowany, do bezpośredniej konsumpcji

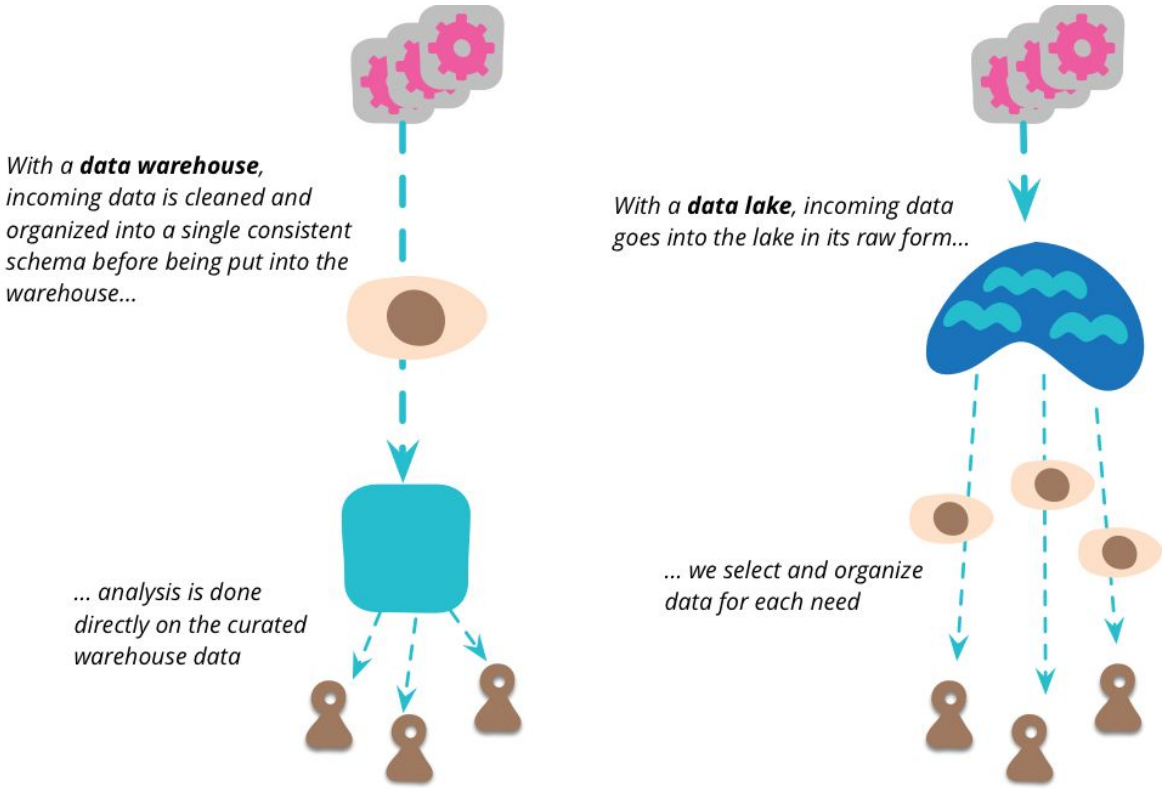
Data Lake: woda w stanie naturalnym = strumień danych: gotowy badania, próbkowania, zanurzenia.

Pentaho CTO James Dixon

Definicja na serio

Data lake to repozytorium, w którym jest przechowywana ogromna ilość nieprzetworzonych danych w oryginalnym formacie. Podczas gdy hierarchiczna hurtownia danych przechowuje informacje w plikach i folderach, data lake do przechowywania danych wykorzystuje płaską architekturę. Każdy element znajdujący się w repozytorium ma przypisany unikalny identyfikator i jest oznaczany zestawem znaczników metadanych.

http://www.it-professional.pl/archiwum/art,5912,repozytoria-data-lakes-zaawansowana-analiza-danych.html


With a **data warehouse**, incoming data is cleaned and organized into a single consistent schema before being put into the warehouse...
... analysis is done directly on the curated warehouse data
With a **data lake**, incoming data goes into the lake in its raw form...
... we select and organize data for each need

Czym to się różni?

DWH	co?	Data Lake
ustrukturyzowane, pre-procesowane	dane	ustrukturyzowane, częściowo ustrukturyzowane, nieustrukturyzowane, surowe
schema-on-write	przetwarzanie	schema-on-read
kosztowny dla dużych wolumenów danych	koszt zapisu	zoptymalizowany kosztowo pod duże wolumeny danych
niewielka, ustalony schemat	elastyczność schematu	wysoka, rekonfigurowalny ad-hoc
zaawansowane	bezpieczeństwo	coraz bardziej zaawansowane

Strategia 1: DL + MDM

data lake + master data management

MASTER DATA MANAGEMENT

Master data management (MDM) is a technology-enabled discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise's official shared master data assets. Master data is the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise including customers, prospects, citizens, suppliers, sites, hierarchies and chart of accounts.

https://www.gartner.com/it-glossary/master-data-management-mdm/

Data Lake = "leniwa" hurtownia danych?

Strategię DL+MDM (data lake plus master data management) można porównać do "lazy evaluation" w programowaniu (wartościowanie leniwe*). Wartość żadnego wyrażenia nie jest wyznaczana dopóki nie jest potrzebna. "Leniwy" program może okazać się znacznie bardziej wydajny, nie wykonując zbędnych obliczeń.

Analogiczna strategia może dotyczyć danych i pracy z hurtownią danych. W modelu DL+MDM dane są transferowane wyłącznie gdy zachodzi taka konieczność. Efektem jest "leniwa" hurtownia danych.

*https://pl.wikibooks.org/wiki/Haskell

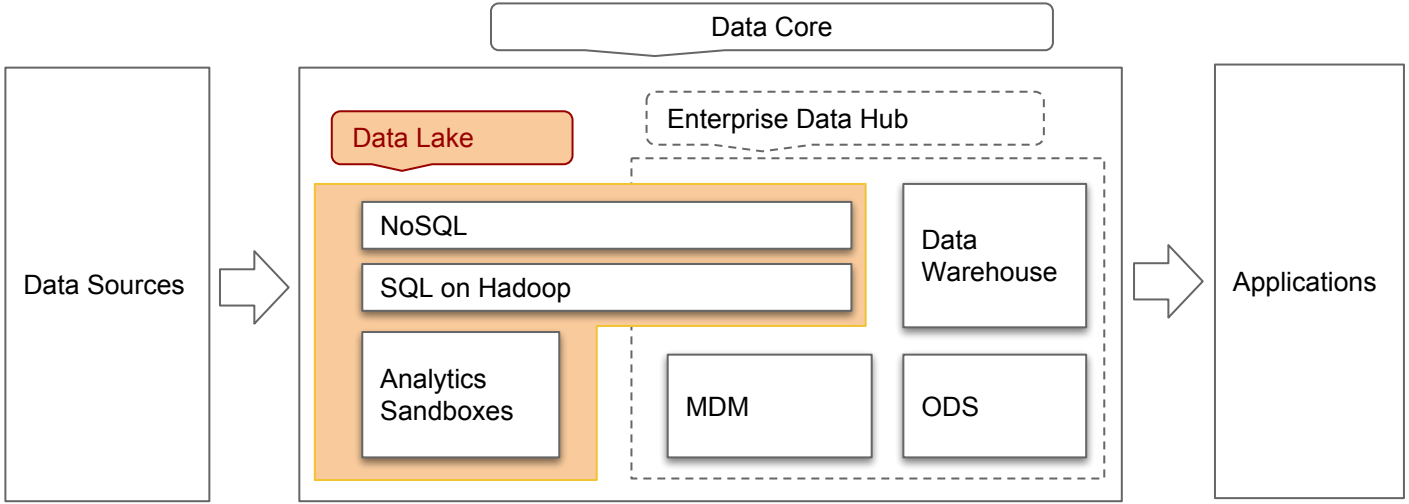
Antidotum na Ograniczenia DWH?

1. Weryfikacja hipotez jest procesem analizy zmiennych w czasie rzeczywistym, który przebiega inkrementalnie, w przemyślany sposób. Nie ma potrzeby rozstrzygania wszystkich problemów klasyfikacyjnych na wejściu.
2. Nowe dane (wygenerowane lub zebrane) są dostępne natychmiast, ponieważ nie wymagają dodatkowych transformacji.
3. Nie ma klauzuli “wykonany” dla projektu hurtowni danych. Nie jest on ograniczeniem funkcjonalnym: dane są dostępne od początku. Kolejne etapy to inkrementalna zmiana jakościowa.
4. Konfiguracja DL+MDM jest skrajnie elastyczna sprzętowo (nie ma potrzeby pre-konfiguracji). Zasoby mogą być dostarczane sukcesywnie w trakcie trwania projektu.
5. DL+MDM to strategia permanentnej zmiany. Zmienia się w czasie i zakłada zmiany.

Strategia 2: DL + MDM + DWH

data lake + master data management + data warehouse

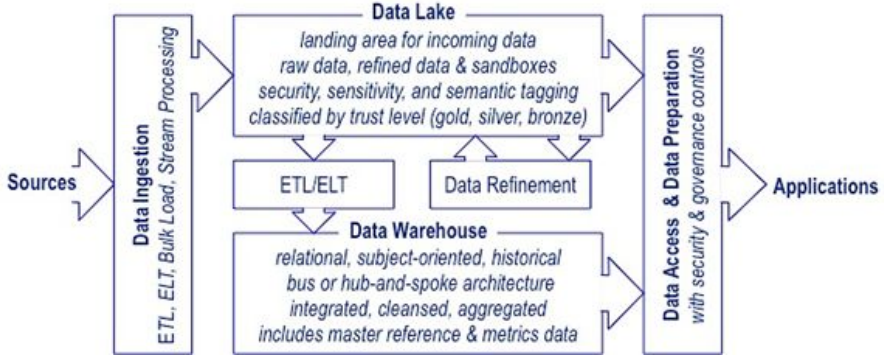
Enterprise Data Management



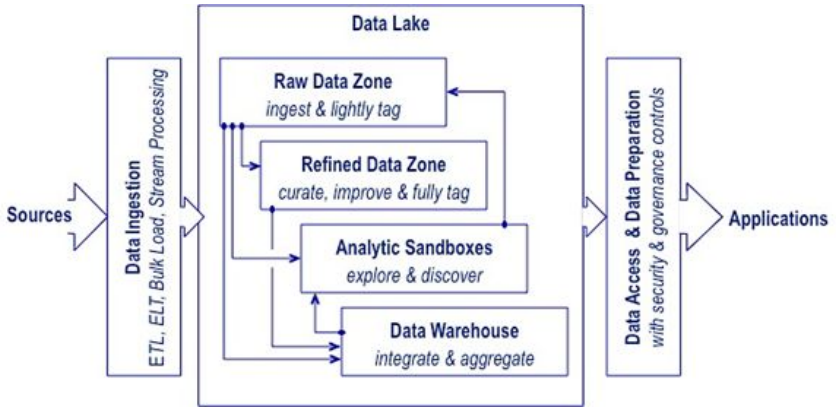
Dave Wells: https://www.eckerson.com/articles/the-future-of-the-data-warehouse

Razem czy osobno?

DWH poza Data Lake

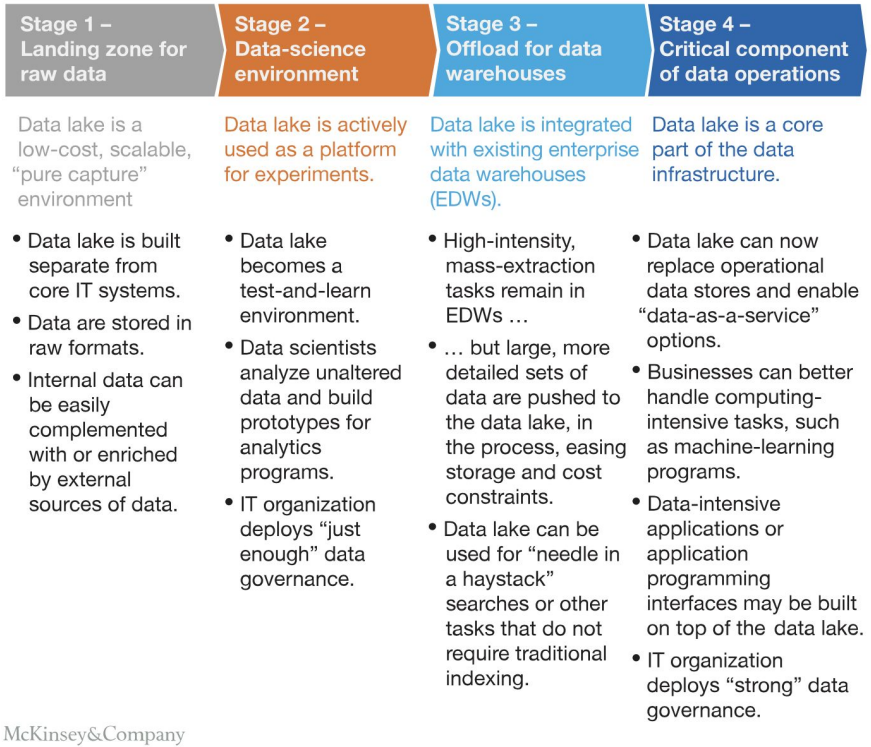


DWH wewnątrz Data Lake



https://www.eckerson.com/articles/the-future-of-the-data-warehouse

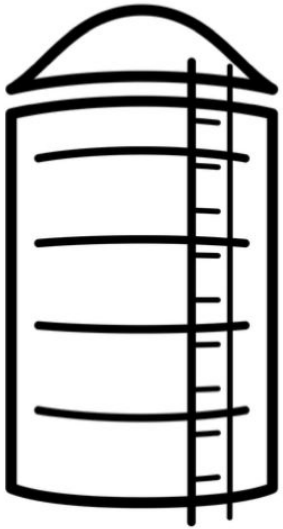
Proces implementacji Data Lake



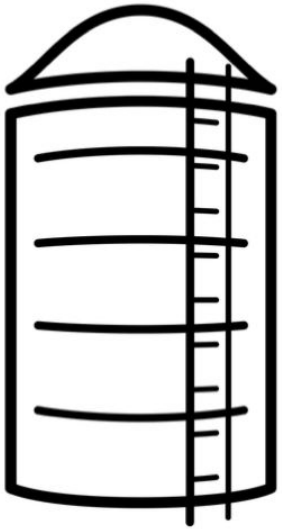
https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/a-smarter-way-to-jump-into-data-lakes

Realia są mniej zachęcające: dane są w silosach

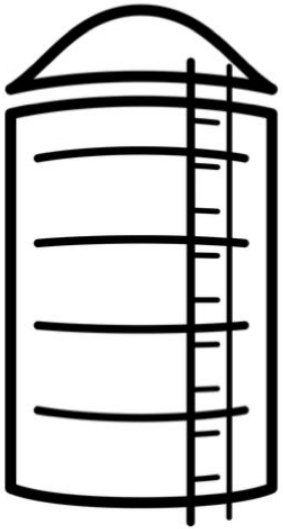
Data Folk



Analyst Folk



Manager Folk



Icon: Noun Project/David Waschbüsch

Dlaczego?

1. Firmy są zwykle podzielone na niezależne jednostki biznesowe. Te utrzymują własne data stores.
2. Firmy prowadzą akwizycje: kupują inne data stores.
3. Data legacy: nowe rozwiązania zastępują starsze, data stores zostają.
4. CTO ma zwykle dane arkuszach .xls (setkach arkuszy).
5. Część wartościowych danych to dane publiczne.

Nie ma standardowego modelu danych

1. Próby jego tworzenia są fiaskiem już na starcie.
2. Grzechy poprzedników kształtują rzeczywistość.
3. CEO nie wywodzi się z IT.
4. Dostęp do danych daje poczucie władzy.

Data Curation

1. zasilenie danymi
2. walidacja (kompletność)
3. transformacja (usd → pln)
4. potwierdzenie zgodności ze schematem
5. konsolidacja i deduplikacja (Mike = Michael)

Trzy generacje rozwiązań wspierających data curation

GEN-1: tradycyjny ETL (lata 90te)

GEN-2: ETL na sterydach (2000te)

GEN-3: skalowalna data curation

GEN1: TRADYCYJNY ETL

Jak to działa?

1. Człowiek definiuje docelowy schemat danych (z góry)
2. Każde źródło danych wymaga:
    a. zrozumienia logiki
    b. napisania skryptu integrującego schemat lokalny z globalnym
    c. napisania procedur oczyszczających dane
    d. sekwencyjnego uruchamiania procesu ETL-owego

Skaluje się do 25-50 źródeł danych.

Dlaczego?

1. Trudno utrzymać spójny schemat danych.
2. Zbyt dużo manualnej ingerencji w logikę skryptów
3. Brak automatyzacji

GEN2: ETL NA STERYDACH

1. Automatyczne systemy deduplikacji danych
2. Detekcja anomalii
3. Domenowy standard oczyszczania danych

Nadal skaluje się do 25-50 źródeł danych.

To nie starcza....

Duża firma farmaceutyczna:

1. Tradycyjna hurtownia danych
2. 10.000 biologów i chemików; 1.3 mln atrybutów
3. Wyniki zapisywane na laptopach w plikach
4. Brak współdzielonych słowników
5. Brak unifikacji na poziomie miar
6. Angielski nie jest domyślnym językiem komunikacji

Nie wystarcza ponieważ

1. Tradycyjny ETL ma ograniczone możliwości w skali big data.
2. Wymaga zbyt dużej ingerencji manualnej.
3. Inżynier danych nie ma kompetencji merytorycznych (czy ICE-50 ma związek z ICU-50?)

Subtelna zmiana miejsc

ETL ➡ ELT

Extract

Transform

Load

GEN 3: Skalowalna Data Curation

1. Machine Learning i modele statystyczne podstawą strategii klasyfikacyjnych
2. Model docelowy tworzony jest celowo w trybie bottom-up.
3. W procesie ewaluacji danych uczestniczą eksperci dziedzinowi (data stewards - nie inżynierowie danych!)

https://www.tamr.com/video/turing-award-winner-michael-stonebraker-on-scalable-data-curation/

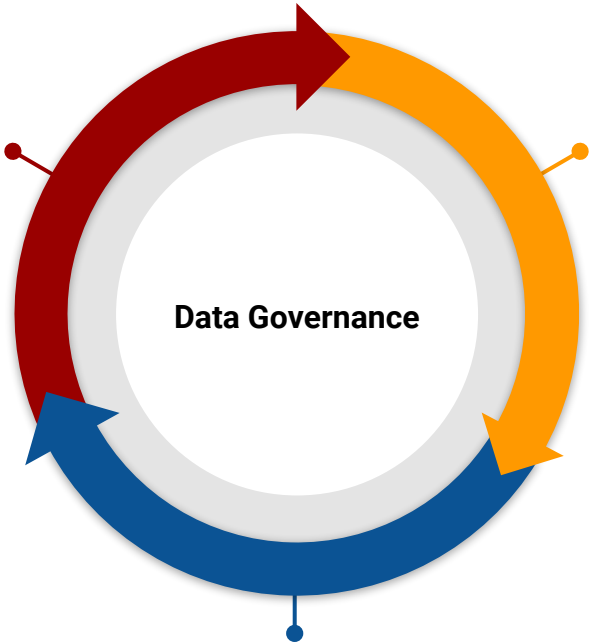
Data Curation to Proces

1. Nowe źródła danych pojawiają się nieustannie
2. Pojawiają się w nich nowe dane
3. Istniejące ulegają nieustannym modyfikacjom

Wnioski:

1. Globalny schemat budowany bottom-up
2. Aplikacja algorytmów odbywa się inkrementalnie
3. Kompetencje wewnętrzne

Big Data lifecycle



Analityka

Wykorzystanie danych wzbogaconych w poszukiwaniu odpowiedzi na pytania o fakty dotyczące organizacji.

Akwizycja

Pozyskiwanie danych z bogatego w nie i zmiennego otoczenia

Kontekstualizacja

Łączenie danych z różnych źródeł w spójną całość, w tym modelowanie zawartości, tworzenie kontekstu oraz zapewnienie możliwości ich wyszukiwania.

Akwizycja danych

HTML

XML

JSON

NoSQL

Hadoop

chmura publiczna

= Volume i Velocity nie są już fundamentalnym problemem

Kontekstualizacja danych

... jest nim Variety

Reużywalność danych to jedno z założeń Big Data.

Wymaga uzgodnienia semantyki natywnego modelu komunikatu z modelem docelowym.

To obszar o największym potencjale w procesie tworzenia wartości z danych.

{Linked Data, OWL, RDF, SPARQL, Rules Based Analysis, Natural Language Processing, Inferencing, Artificial Intelligence}

Kurt Cagle https://www.linkedin.com/pulse/four-keys-big-data-life-cycle-kurt-cagle/

Analityka danych

1. Potwierdza wewnętrzną spójność danych
2. Tworzy analogie ułatwiające innym rozumienie danych
3. Konstruuje modele predykcyjne na bazie danych historycznych

{Stochastics, Business Intelligence Tools, R, Visualization tools, Computational Mathematical Tools, Modeling Systems}

Data Governance

1. **Jakość**. Dane wykazują spójną strukturę, cechuje je integralność (również wstecz), są kompletne (w pełni lub w wystarczającym zakresie) i zapisane w formacie, który umożliwia efektywną pracę.
2. **Unikalne.** Mają istotną wartość informacyjną, nie są redundantne.
3. **Prawdziwe.** Źródło danych jest znane i ma reputację źródła informacji prawdziwych.
4. **Użyteczność.** Dane są bezużyteczne jeśli w żaden sposób nie nawiązują do hipotez i potrzeb w kontekście w jakim są gromadzone.

Chief Data Officer: kto to jest?