Fundusze Europejskie
Wiedza Edukacja Rozwój

Politechnika Wrocławska

Unia Europejska
Europejski Fundusz Społeczny

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

Hurtownie danych

Podstawy procesu ETL

dr inż. Bernadetta Maleszka

Hurtownia danych - definicja

Hurtownia danych to:

- **tematycznie zorientowana**
- **zintegrowana**
- **chronologiczna**
- **trwała**

kolekcja danych do wspomagania procesów podejmowania decyzji

ETL

- Extract
- Transform
- Load

- Zastosowanie reguł biznesowych do istniejących danych w celu uzyskania użytecznych informacji
- Czyszczenie i standaryzacja danych
- Integracja różnych danych (wewnętrznych i zewnętrznych)
- Agregacja danych
- Nawet 70% - 80% wysiłku budowy hurtowni danych

ETL

- Pobierz dane ze źródła i załaduj do hurtowni
  - kopiowanie danych pomiędzy bazami
- Dane są wyciągane z bazy OLTP, przekształcane tak, aby pasowały do schematu hurtowni i ładowane do hurtowni
- Źródłowe dane mogą nie być przechowywane w tej samej bazie
- Myśl o procesie ETL, a nie o fizycznej implementacji tego procesu!

ETL

- Złożona kombinacja procesu i technologii wymagająca nakładów sił i energii:
  - analityków biznesowych
  - projektantów baz danych
  - developerów aplikacji

- Nie mylić procesu ETL z jednorazowym czy nawet okresowym dodawaniem danych do bazy!
- Proces:
  - zautomatyzowany
  - udokumentowany
  - łatwo modyfikowalny

Extraction

- integracja wszystkich systemów przedsiębiorstwa
- heterogeniczne źródła danych
- każde źródło danych ma swoją charakterystykę:
  - DBMS
  - system operacyjny
  - hardware
  - protokoły komunikacji

- Logiczna mapa danych
  - określa relacje pomiędzy skrajnymi etapami procesu ETL

Ekstrakcja – mapa logiczna

Cel			Źródło			Przekształcenie
Tabela	Kolumna	Typ danych	Tabela	Kolumna	Typ danych	

- dokładnie wiadomo, co dzieje się z danymi
- przekształcenie – zazwyczaj SQL

Mapa logiczna - przykład

Target					Source				Transformation
Table Name	Column Name	Data Type	Table Type	SCD Type	Database Name	Table Name	Column Name	Data Type	
EMPLOYEE_DIM	EMPLOYEE_KEY	NUMBER	Dimension	1				NUMBER	Surrogate key.
EMPLOYEE_DIM	EMPLOYEE_ID	NUMBER	Dimension	1	HR_SYS	EMPLOYEES	EMPLOYEE_ID	NUMBER	Natural Key for employee in HR system
EMPLOYEE_DIM	BIRTH_COUNTRY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	COUNTRIES	NAME	VARCHAR2(75)	select c.name from employees e, states s, countries c where e.state_id = s.state_id and s.country_id = c.country
EMPLOYEE_DIM	BIRTH_STATE	VARCHAR2(75)	Dimension	1	HR_SYS	STATES	DESCRIPTION	VARCHAR2(255)	select s.description from employees e, states s where e.state_id = s.state_id
EMPLOYEE_DIM	DISPLAY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(75)	select initcap(salutation) || ' ' || initcap(first_name) || ' ' || initcap(last_name) from employee
EMPLOYEE_DIM	BIRTH_DATE	DATE	Dimension	1	HR_SYS	EMPLOYEES	DOB	DATE	trunc(DOB)
EMPLOYEE_DIM	SALUTATION	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SALUTATION	VARCHAR2(12)	initcap(salutation)
EMPLOYEE_DIM	FIRST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(30)	initcap(first_name)
EMPLOYEE_DIM	LAST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	LAST_NAME	VARCHAR2(30)	initcap(last_name)
EMPLOYEE_DIM	MARITAL_STATUS	VARCHAR2(12)	Dimension	2	HR_SYS	MARITAL_STATUS	DESCRIPTION	VARCHAR2(12)	select nvl(m.name,'Unknown') from employee e marital_status m where e.marital_status_id = m.marital_status_id
EMPLOYEE_DIM	DIVERSITY_CATEGORY	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	EEO_CLASS	VARCHAR2(30)	decode(eeo_class,null, 'Not Stated', decode(eeo_class,'N', 'Not Stated',eeo_class))
EMPLOYEE_DIM	GENDER	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SEX	VARCHAR2(12)	nvl(sex, 'Unknown')
EMPLOYEE_DIM	EMPLOYEE_STATUS	VARCHAR2(24)	Dimension	1	HR_SYS	EMPLOYEES	STATUS	VARCHAR2(24)	select es.name from employee e employee_statuses where e.employee_status_id = m.employee_status_id
EMPLOYEE_DIM	POSITION_CODE	VARCHAR2(12)	Dimension	2	HR_SYS	POSITIONS	POSITION_CODE	VARCHAR2(12)	select p.code from employees e, positions p where p.position_id = e.position_id
EMPLOYEE_DIM	POSITION_CATEGORY	VARCHAR2(30)	Dimension	2	HR_SYS	POSITIONS	POSITION_CATEGORY	VARCHAR2(30)	select p.category from employees e, positions p where p.position_id = e.position_id
EMPLOYEE_DIM	HIRE_DATE	DATE	Dimension	1	HR_SYS	EMPLOYEES	DATE_HIRED	DATE	trunc(date_hired)

Fazy ekstrakcji

1. Wykrywanie danych:
    - czystość danych
    - spójność danych
    - identyfikacja i sprawdzenie źródła pod kątem założonego celu

    - dokumentacja systemu źródłowego
    - śledzenie zmian w systemie
    - określenie miejsca pochodzenia danych
    - świadomość redundancji danych (dane kopiowane, przekształcane, czyszczone, itp.)

Fazy ekstrakcji

2. Detekcja anomalii:
    - NULL (operacje złączenia tabel)
    - wartości kluczowe
    - daty
    - audit columns – używane przez DB, warunkowo uaktualniane

3. Eliminacja anomalii:
    - tworzenie dwóch tabel (dane z poprzedniego i bieżącego ładowania)
    - obliczanie różnicy pomiędzy tabelami w celu wykrycia zmian

Transformation

- udokumentowany etap modyfikacji danych do pożądanej postaci

- paradygmaty jakości danych:
  - poprawność
  - jednoznaczność
  - spójność
  - kompletność

- dwukrotne sprawdzenie:
  - po ekstrakcji
  - po czyszczeniu i potwierdzeniu dodatkowych warunków

Transformation - Czyszczenie danych

- detekcja anomalii:
  - próbkowanie danych
  - zliczanie rekordów

- sprawdzenie własności kolumn:
  - wartości NULL w miejscu kluczy
  - wartości numeryczne poza oczekiwanym zakresem
  - zbyt długie/krótkie długości danych
  - dane poza zakresem zbioru
  - dane odstające od wzorca

Transformation - zatwierdzenie

- Sprawdzenie struktury
  - klucze główne i obce
  - integralność referencyjna kluczy

- Sprawdzenie danych i reguł
  - prostych reguł biznesowych
  - na poziomie logicznym

Loading
Ładowanie danych do wymiarów

- minimalizacja zbioru komponentów
- prosty klucz główny
- denormalizacja tabel

- slowly changing dimentions
  - zapis wymiaru jako fizycznej tabeli na dysku

- przypisanie kluczy zastępczych

Loading
Ładowanie danych do tabeli faktów

- w tabeli faktów przechowywane są miary
- uproszczone relacje pomiędzy tabelą faktów a wymiarami

- tworzenie klucza tabeli faktów
  - tworzenie klucza zastępczego

ETL – zasilanie hurtowni danych danymi

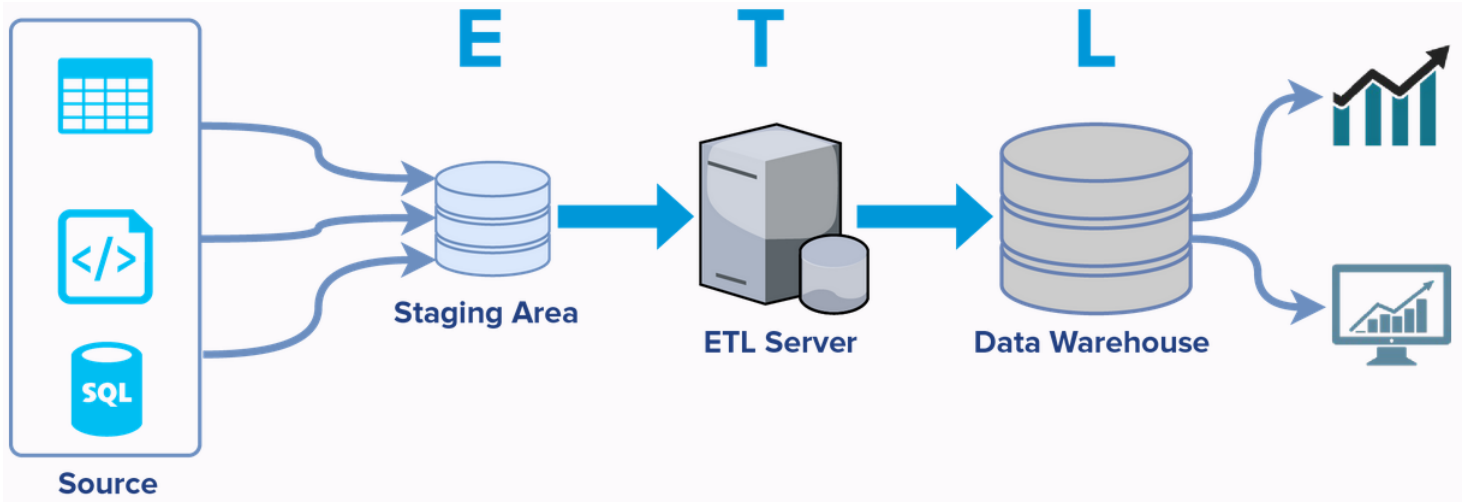
- **ekstrakcja danych** z systemów źródłowych (SAP, ERP, inne systemy transakcyjne), dane z rożnych systemów są konwertowane do wspólnego, jednolitego formatu danych hurtowni danych
- **transformacja danych**:
  - zastosowanie logiki biznesowej,
  - czyszczenie danych,
  - filtrowanie,
  - rozdzielenie jednej kolumny na kilka i odwrotnie,
  - łączenie danych z kilku źródeł (lookup, merge),
  - transpozycje kolumn i wierszy,
  - odrzucanie danych niespełniających zdefiniowanych wymagań/założeń
- **załadowanie danych** do hurtowni danych lub repozytoriów danych innych aplikacji raportujących

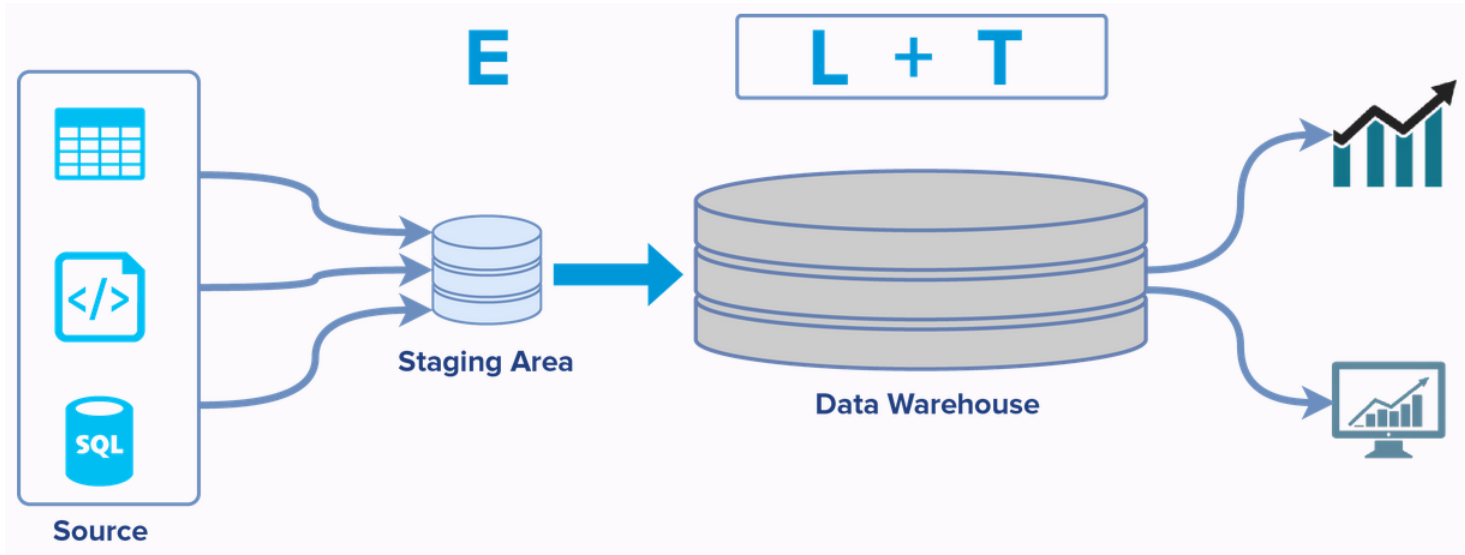
ELT

- Dane ekstraktowane z systemów źródłowych bezpośrednio ładowane w oryginalnym formacie do bazy danych hurtowni danych
- Przy pomocy wygenerowanych poleceń i procedur SQL serwer bazy danych (DBMS) wykonuje transformacje danych
- Zasila tabele docelowe hurtownie

- Wymagania:
  - bardzo wydajny
  - wysoce skalowalny
  - i dobrze dostrojony serwer DBMS
- Stosowany przy bardzo dużych wolumenach danych

ETL vs ELT





https://blog.gft.com/pl/2017/02/24/etl-vs-elt-czyli-rozne-podejscia-do-zasilenia-hurtowni-i-repozytoriow-danych/

ETL vs ELT

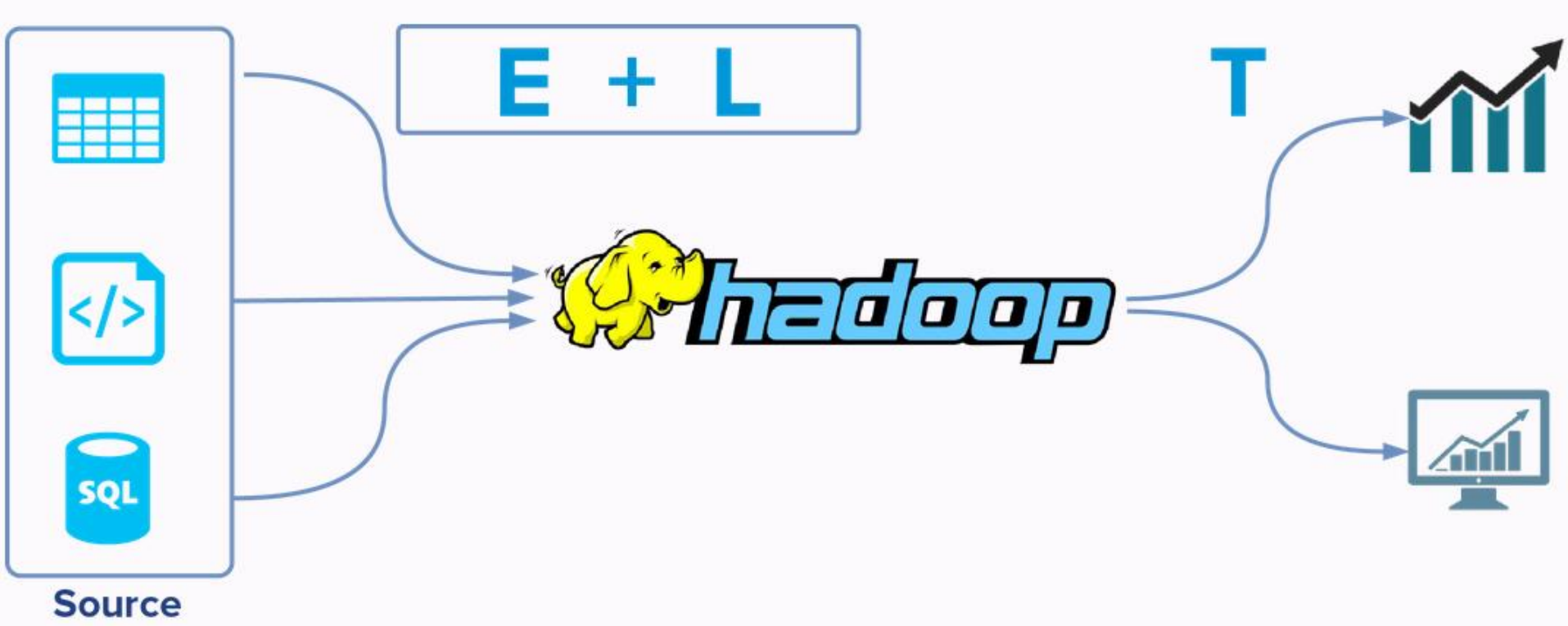
ETL

- Extract – wyładowanie danych i załadowanie ich do przestrzeni tymczasowej (ang. staging)
- Wada: niezbędny serwer na potrzeby narzędzia SQL
- Transform – przygotowanie modelu i przekształcenie danych do pożądanej postaci (ang. schema-on-write)
- Load

ELT

- Exctract – przygotowanie danych, ale bez definiowania, jak mają wyglądać dane wyjściowe (ang. schema-on-read)
- Load – załadowanie surowych danych do centralnego repozytorium danych (ang. Data Lake)
- Transform - wykorzystanie technologii pozwalającej przetwarzać dane nierelacyjne, w różnych formatach i strukturach

Przykład ELT – Big Data



Zalety i wady

Kryterium	ETL	ELT
Schemat	Podczas tworzenia hurtowni.	ELT nie wyklucza podejścia Schema-on-Write. Decyzja o formie danych podczas ich odczytu z repozytorium danych.
Zmiany w modelu hurtowni	Często musimy zmienić przepływ ETL oraz model hurtowni.	Zmiana może ograniczyć się do warstwy hurtowni danych i kroku transformacji.
Infrastruktura	Potrzebne dodatkowe maszyny.	Całość procesu realizowana na docelowym wystarczająco wydajnym serwerze.
Kompetencje	Wymagane dodatkowe kompetencje związane z procesami i narzędziami ETL.	L+T -> znajomość baz danych. W pozostałych przypadkach wymagana jest znajomość technologii, wykorzystywana do przechowywania i procesowania danych.

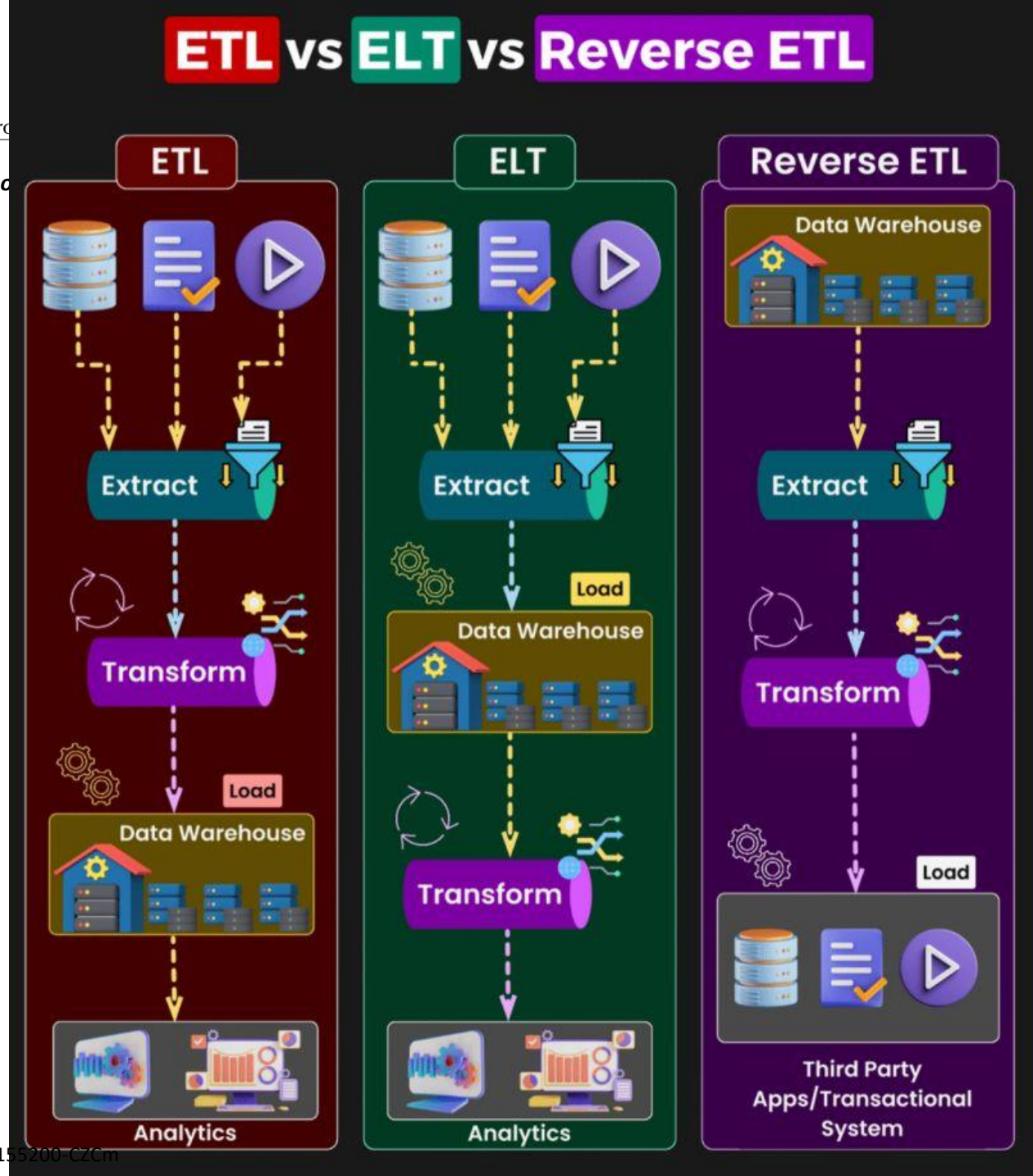
Zalety i wady

Kryterium	ETL	ELT
Czas dostępu do danych	Zazwyczaj dane dostępne po ukończeniu całego procesu.	Dane szybciej dostępne na docelowej maszynie. Możemy mieć dostęp do danych surowych przed transformacją.
Zastosowanie	Rozwiązanie popularne i optymalne przy dużych wolumenach danych oraz skomplikowanych transformacjach. Może nie być optymalne kosztowo dla małych rozwiązań.	Zysk widoczny przy przetwarzaniu potężnych zbiorów danych opartych o rozwiązania nastawione na skalowalność oraz dane nieustrukturyzowane.

Reverse ETL

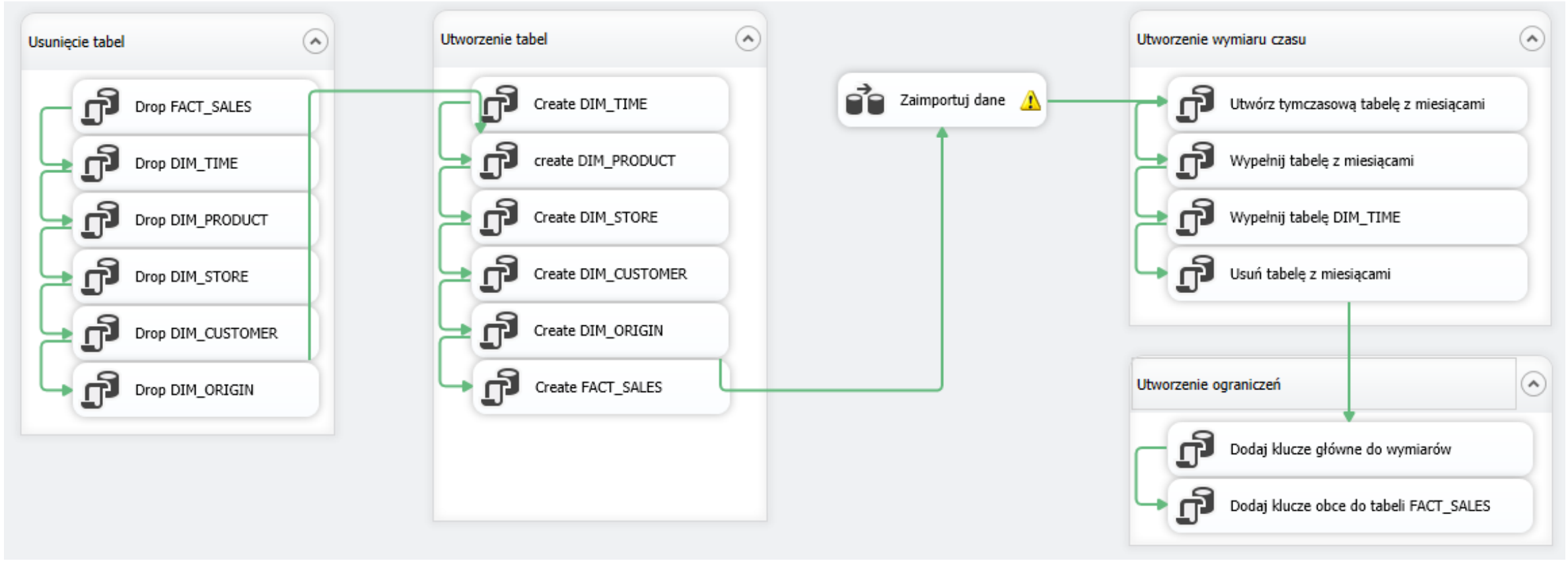
- Źródło danych: nowoczesna hurtownia danych
- Cel: analiza operacyjna w wybranych zakresach

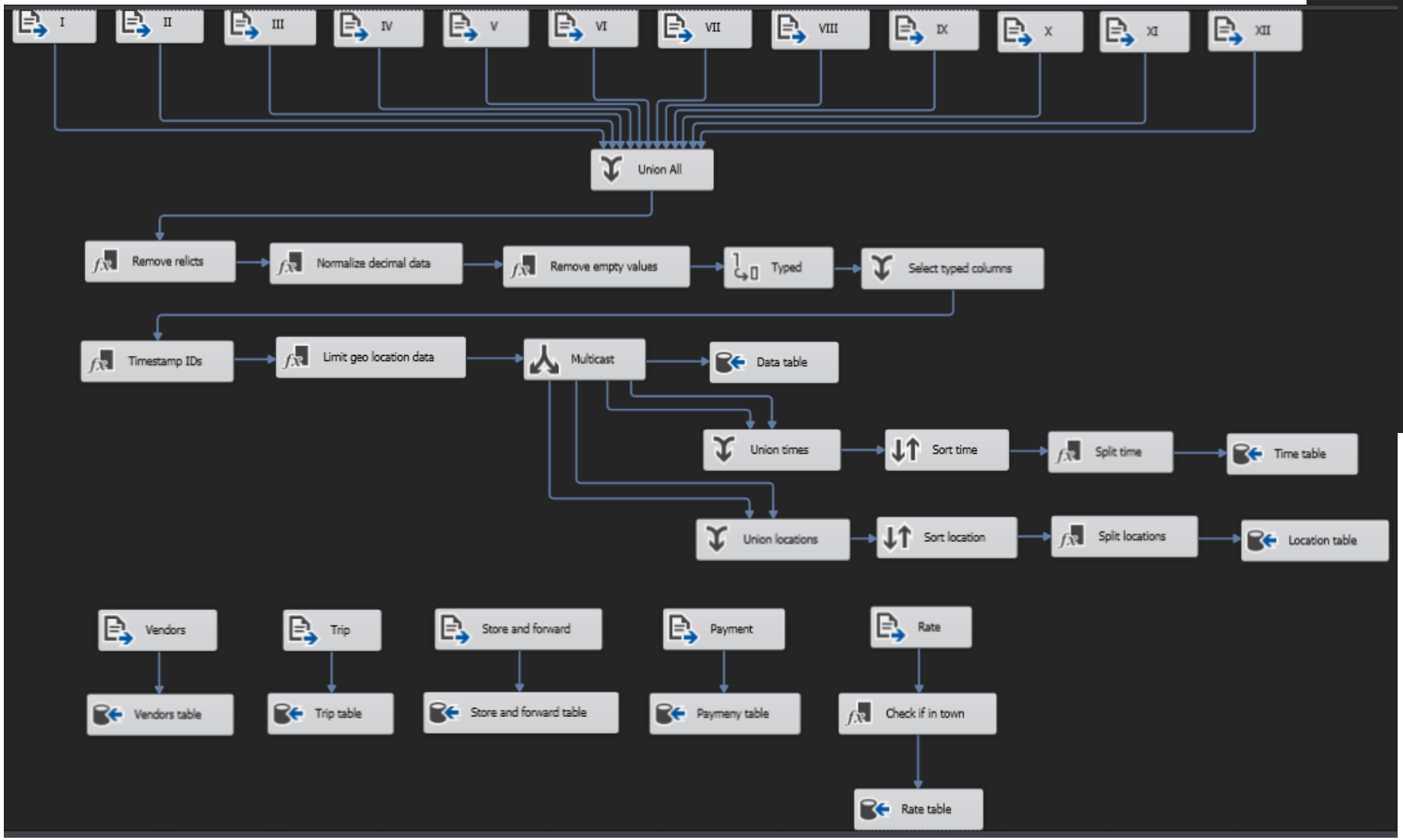
- Software-as-a-service (SaaS)
- Wykorzystanie danych z hurtowni (KPI) przez zewnętrzne systemy

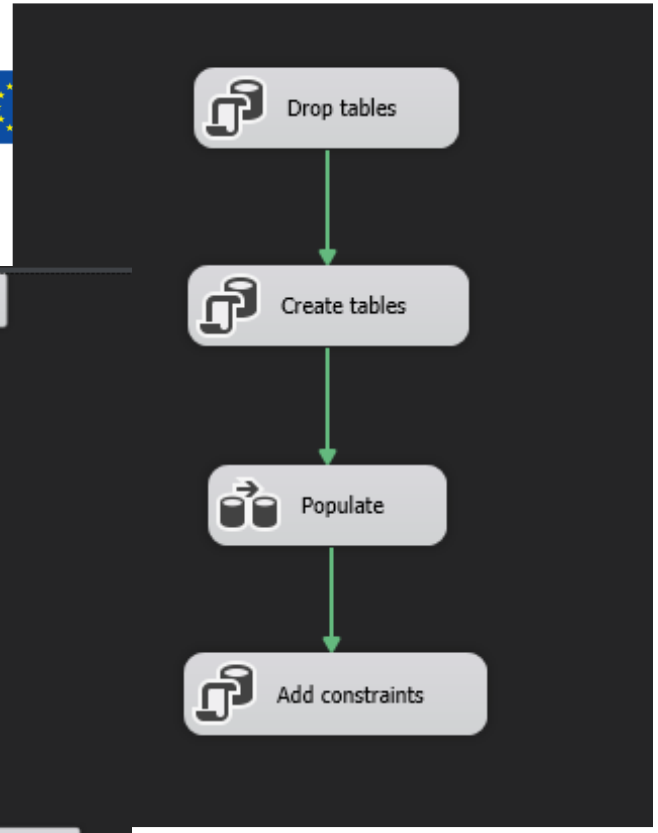


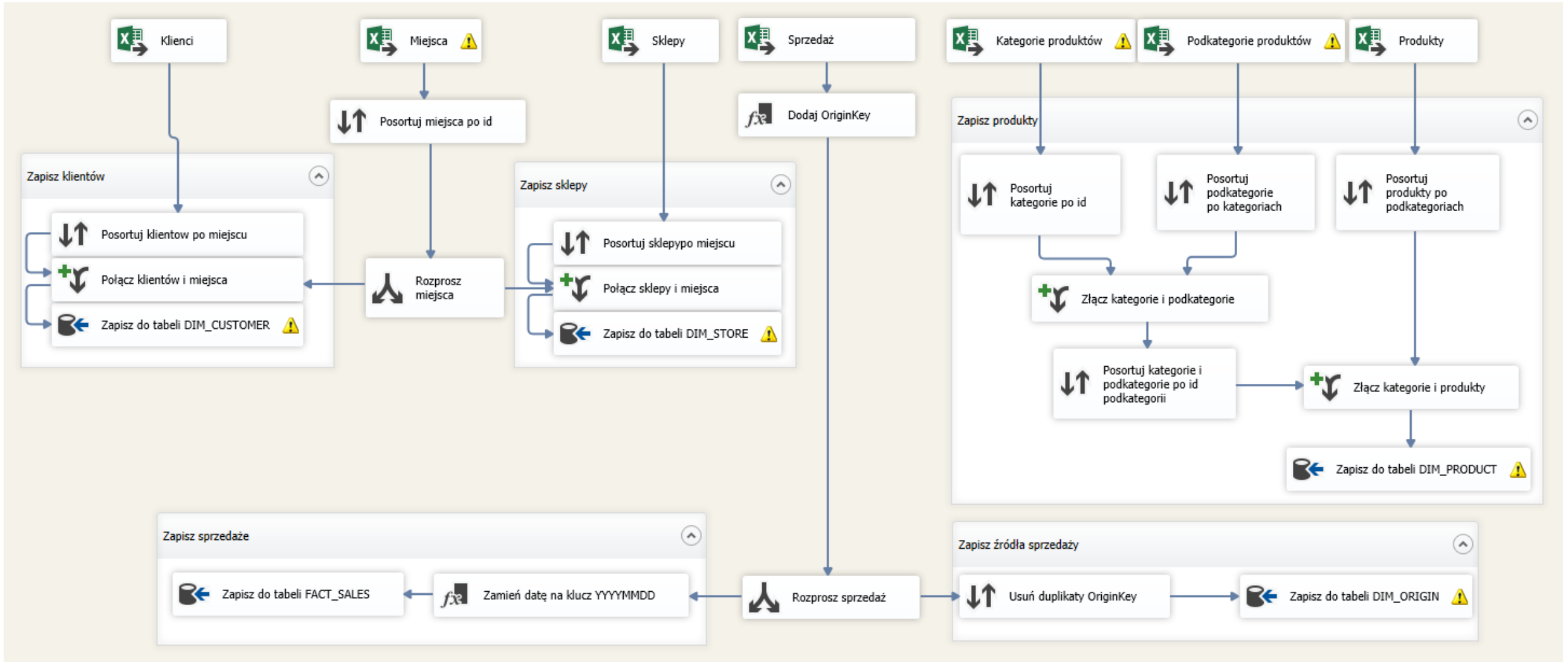
https://www.linkedin.com/posts/brijpandeyji_bigdata-dataengineering-softwareengineering-activity-7109498805263155200-CZCm

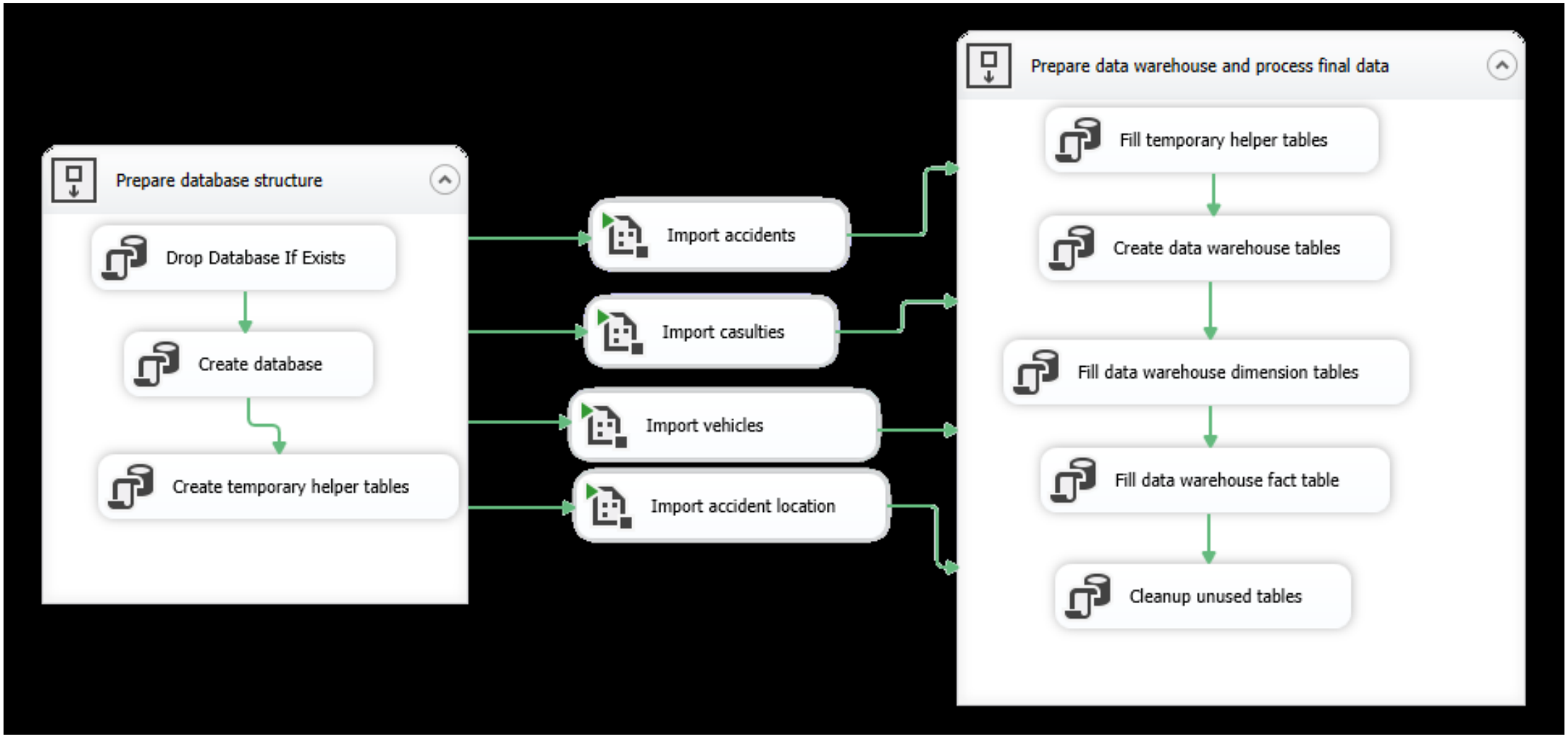
Przykłady procesów ETL



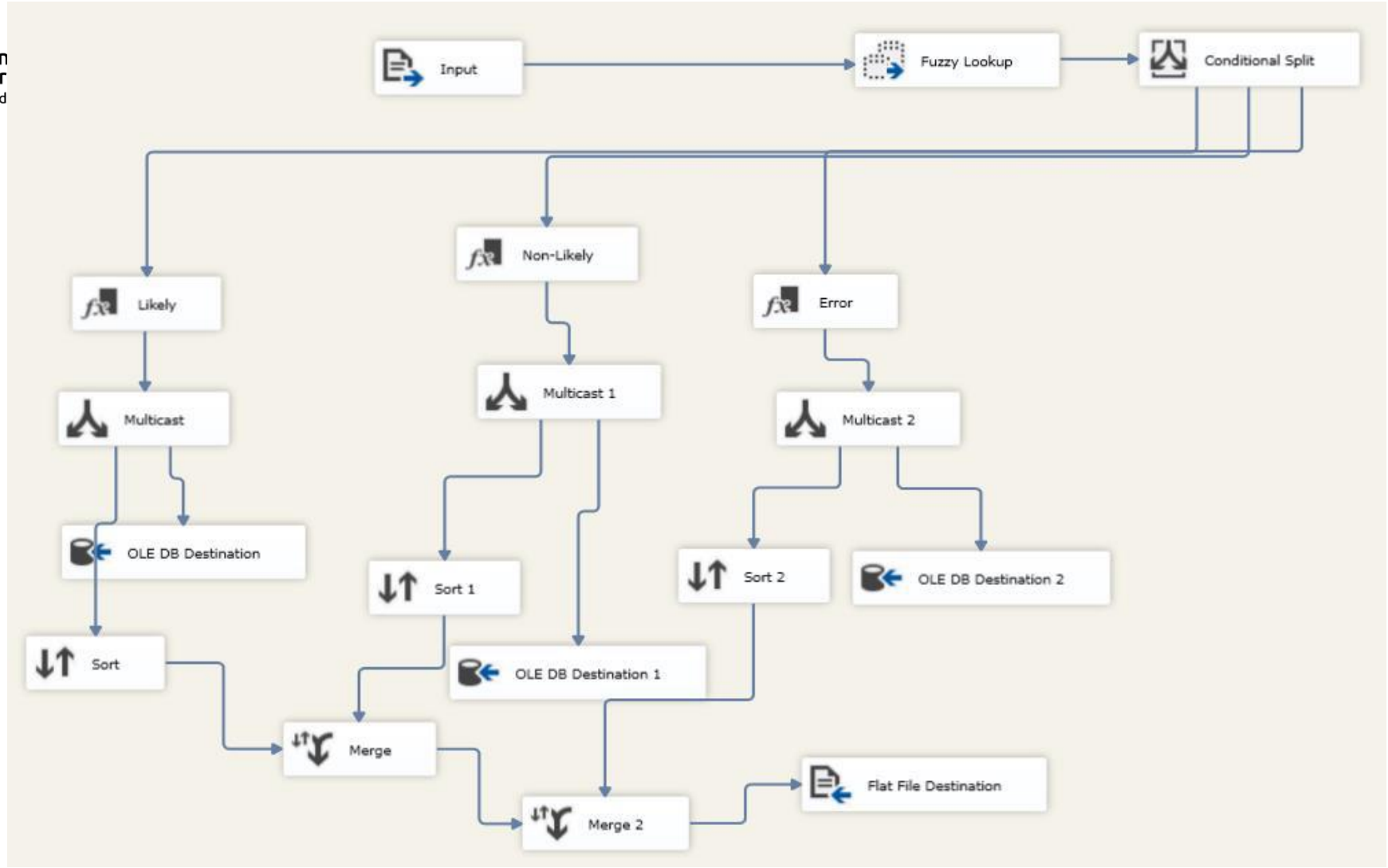

I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
Union All
Remove relicts
Normalize decimal data
Remove empty values
Typed
Select typed columns
Timestamp IDs
Limit geo location data
Multicast
Data table
Union times
Sort time
Split time
Time table
Union locations
Sort location
Split locations
Location table
Vendors
Trip
Store and forward
Payment
Rate
Vendors table
Trip table
Store and forward table
Paymeny table
Check if in town
Rate table


Drop tables
Create tables
Populate
Add constraints


Klienci
Miejsca
Sklepy
Sprzedaż
Kategorie produktów
Podkategorie produktów
Produkty
Posortuj miejsca po id
Dodaj OriginKey
Zapisz produkty
Zapisz klientów
Posortuj klientow po miejscu
Połącz klientów i miejsca
Zapisz do tabeli DIM_CUSTOMER
Rozprosz miejsca
Zapisz sklepy
Posortuj sklepypo miejscu
Połącz sklepy i miejsca
Zapisz do tabeli DIM_STORE
Posortuj kategorie po id
Posortuj podkategorie po kategoriach
Posortuj produkty po podkategoriach
Złącz kategorie i podkategorie
Posortuj kategorie i podkategorie po id podkategorii
Złącz kategorie i produkty
Zapisz do tabeli DIM_PRODUCT
Zapisz sprzedaże
Zapisz do tabeli FACT_SALES
Zamień datę na klucz YYYYMMDD
Rozprosz sprzedaż
Zapisz źródła sprzedaży
Usuń duplikaty OriginKey
Zapisz do tabeli DIM_ORIGIN


Prepare database structure
Drop Database If Exists
Create database
Create temporary helper tables
Import accidents
Import casulties
Import vehicles
Import accident location
Prepare data warehouse and process final data
Fill temporary helper tables
Create data warehouse tables
Fill data warehouse dimension tables
Fill data warehouse fact table
Cleanup unused tables

Fuzzy Lookup



Error	Package	String	ERROR
likely	Package	String	LIKELY
non_match	Package	String	NON-MATCH

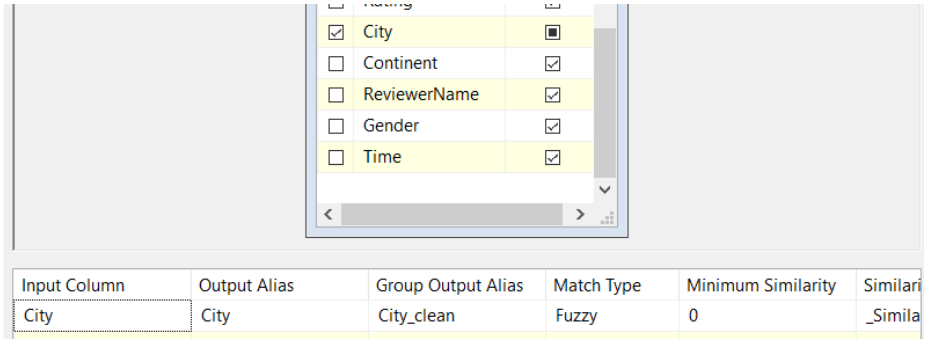
94	"Hex Nut 20	Hex Nut 20	0.9875	0.9007731	0.9875	LIKELY
95	"HL Touring Seat/Saddle	HL Touring Seat/Sad...	0.9875	0.6114928	0.9875	LIKELY
96	"Lock Washer 2	Lock Washer 2	0.9875	0.9578876	0.9875	LIKELY
97	"Hex Nut 21	Hex Nut 21	0.9875	0.9007731	0.9875	LIKELY
98	"LL Bottom Bracket	LL Bottom Bracket	0.9875	0.5996314	0.9875	LIKELY
99	"HL Mountain Rim	HL Mountain Rim	0.9875	0.681129	0.9875	LIKELY
100	"Hex Nut 2	Hex Nut 2	0.9875	0.9713477	0.9875	LIKELY
101	"Lock Washer 11	Lock Washe				
102	"Lock Washer 5	Lock Washe				
103	"Thin-Jam Lock Nut 13	Thin-Jam Loc				
104	"Lock Washar 3	Lock Washe				
105	"ML Grip Tpe	ML Grip Tap				
106	"External Lck Washer 8	External Lock				

pejska
Społeczny

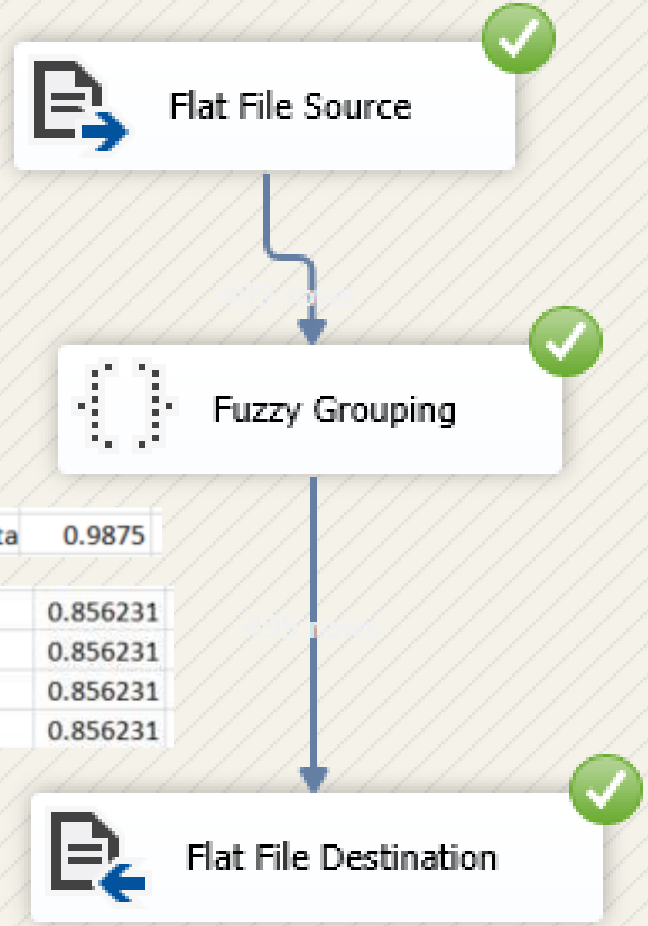
j"

```
"Bearing Ball,Bearing Ball,0.98750001,0.5,0.98750001,LIKELY
"External Lck Washer 8,External Lock Washer 8,0.92961943,0.56844395,0.92961943,NON-MATCH
"External Lock Washer 1,External Lock Washer 1,0.98750001,0.98534936,0.98750001,LIKELY
"External Lock Washer 7,External Lock Washer 7,0.98750001,0.56334531,0.98750001,LIKELY
"External Lock Washer 9,External Lock Washer 9,0.98750001,0.56254739,0.98750001,LIKELY
"Guide Pulley,Guide Pulley,0.98750001,0.56722081,0.98750001,LIKELY
"Headset Ball Bearings,Headset Ball Bearings,0.98750001,0.52441591,0.98750001,LIKELY
"Hex Nut 1,Hex Nut 1,0.98750001,0.97134769,0.98750001,LIKELY
"Hex Nut 10,Hex Nut 10,0.98750001,0.94661856,0.98750001,LIKELY
"Hex Nut 11,Hex Nut 11,0.98750001,0.93934381,0.98750001,LIKELY
"Hex Nut 12,Hex Nut 12,0.98750001,0.93934381,0.98750001,LIKELY
"Hex Nut 13,Hex Nut 13,0.98750001,0.93934381,0.98750001,LIKELY
"Hex Nut 16,Hex Nut 16,0.98750001,0.93017125,0.98750001,LIKELY
"Hex Nut 17,Hex Nut 17,0.98750001,0.90077311,0.98750001,LIKELY
"Hex Nut 2,Hex Nut 2,0.98750001,0.97134769,0.98750001,LIKELY
"Hex Nut 20,Hex Nut 20,0.98750001,0.90077311,0.98750001,LIKELY
"Hex Nut 21,Hex Nut 21,0.98750001,0.90077311,0.98750001,LIKELY
"Hex Nut 22,Hex Nut 22,0.98750001,0.90077311,0.98750001,LIKELY
"Hex Nut 23,Hex Nut 23,0.98750001,0.90077311,0.98750001,LIKELY
"Hex Nut 3,Hex Nut 3,0.98750001,0.9654057,0.98750001,LIKELY
"Hex Nut 5,Hex Nut 5,0.98750001,0.96176714,0.98750001,LIKELY
"Hex Nut 7,Hex Nut 7,0.98750001,0.96176714,0.98750001,LIKELY
"Hex Nut 8,Hex Nut 8,0.98750001,0.96176714,0.98750001,LIKELY
"Hex Nut 9,Hex Nut 9,0.98750001,0.95754081,0.98750001,LIKELY
```

Fuzzy Grouping



486	246	0.9875	490	346	0.824277	United-Sta	South Ame	Sherman	M	31-10-11	United Sta	0.9875	
326	366	0.856231	330	981	2.022023	Iceland	Africa	Bryan Phi	F	31-08-11	Ireland	0.856231	
177	366	0.856231	181	534	10	Iceland	Europe	Sherri Ho	M	28-06-11	Ireland	0.856231	
113	366	0.856231	117	768	4.09318	Iceland	South Ame	Guillermo	F	09-06-11	Ireland	0.856231	
21	366	0.856231	25	713	5.871027	Iceland	South Ame	Geraldine	F	07-05-11	Ireland	0.856231	



377	187	0.764914	381	792	2.603801	South Korea	South America	Kristi Brown	M	24-09-11	North Korea	0.764914

Tworzenie partycji

- fizyczny podział tabeli faktów na mniejsze tabele
- cel: poprawa wydajności zapytań

- zazwyczaj podział względem dat
- uwzględnienie tych części wymiarów, które są potrzebne

Dziennik zmian

- nadmiarowe, zbędne

- wszystkie dane są wprowadzane procesem ETL
- dane ładowane są luzem
- w przypadku niepowodzenia, proces można powtórzyć
- różne systemy bazodanowe korzystają z różnych dzienników
  - jak je zintegrować?

Hurtownie danych

Dziękuję za uwagę

dr inż. Bernadetta Maleszka