Zeszyty Naukowe
Wyższej Szkoły Bankowej w Poznaniu
Nr 40/2012

Bogdan Pilawski

Wyższa Szkoła Bankowa w Poznaniu
Bank Zachodni WBK

Narzędzia ETL w zasilaniu repozytoriów danych[1]

Streszczenie. *W artykule omówiono wybrane, podstawowe aspekty stosowania narzędzi ETL (Extract – Transform – Load) do zasilania repozytoriów danych. Na tle ewolucji tych repozytoriów przedstawiono charakterystykę narzędzi ETL oraz wskazano tendencje rozwojowe i formułowane wobec nich nowe wymogi. Całość uzupełniają praktyczne przykłady zastosowań tych narzędzi.*

Słowa kluczowe: *ETL, bazy danych, repozytoria danych, hurtownie danych*

1. Wprowadzenie

Stosowanie rozwiązań informatycznych w celu usprawnienia funkcjonowania przedsiębiorstw, instytucji i administracji w latach 70. i 80. XX w. ograniczało się w zasadzie do bieżącej obsługi. Obejmowało ono stosunkowo proste, ale masowe pod względem ilościowym czynności planowania oraz rejestracji i – wraz z rozwojem

[1] Opracowanie jest rozwinięciem i uzupełnieniem wystąpienia pod tym samym tytułem, jakie miało miejsce w ramach cyklu seminariów „Ku modelowi gospodarki opartej na wiedzy", organizowanego wspólnym staraniem Katedry Informatyki Stosowanej Wyższej Szkoły Bankowej w Poznaniu oraz działającego na tej uczelni Studenckiego Koła Informatyki Stosowanej. Celem wspomnianego wystąpienia było przede wszystkim przedstawienia stosunkowo mało znanych podstaw narzędzi ETL oraz ich miejsca i roli w szerszej dziedzinie hurtowni danych, utożsamianej często z obszarem zbierania danych i przekształcania ich w informacje w wyniku analizy. Przesądziło to o informacyjnym przede wszystkim zakresie tego wystąpienia i jego ograniczonym poziomie szczegółowości. Podobne cechy ma niniejsze opracowanie. W artykule uwzględniono m.in. doświadczenia autora z czynnego udziału w procesie wyboru narzędzi ETL na potrzeby dużej, międzynarodowej korporacji bankowej, zakończonego oceną i decyzją po kilkutygodniowych próbach, z udziałem ówczesnej światowej czołówki producentów takich narzędzi. Autor uczestniczył również w późniejszym opracowaniu strategii wdrożenia wybranych narzędzi.

technik i narzędzi – poszerzyło się z czasem o bieżącą obsługę transakcji. Silnym czynnikiem ograniczającym zakres tego stosowania były wówczas jego wysokie koszty, będące pochodną relatywnie wysokich cen sprzętu komputerowego[2]. Inną przyczyną tego stanu był niedorozwój narzędzi i metod gromadzenia i analizy danych.

Istotną zmianę przyniosło pojawienie się w drugiej połowie lat 70. XX w. tzw. minikomputerów oraz – na początku lat 80 – pierwszych komputerów osobistych. Masowość produkcji tych ostatnich, w połączeniu z rozwojem mikroelektroniki, spowodowała bardzo znaczny spadek cen sprzętu[3], czyniąc opłacalnymi wiele rozwiązań informatycznych, zupełnie nowych albo pozostających wówczas tylko w sferze koncepcji. Stan ten pozwolił m.in. na zwiększanie ilości danych pozostających w bezpośrednim dostępie (czyli – *de facto* – zapisanych w pamięci dyskowej), co pociągnęło też za sobą rozwój w zakresie oprogramowania dostęp ten obsługującego. W efekcie do dyspozycji pozostawały nie tylko bieżące dane transakcyjne, ale również tzw. dane historyczne, nie mające bezpośredniego związku z bieżącą działalnością, pozwalające jednak na ocenę i analizę działań przeszłych oraz – na jej podstawie – planowanie i wyznaczanie strategii na przyszłość.

2. Repozytoria danych i ich zasilanie

Początkowo wspomnianym działaniom analitycznym poddawano pliki i bazy danych wykorzystywane w zastosowaniach transakcyjnych, w czasie od nich wolnym. Typowym przykładem była praktyka jednej z brytyjskich sieci sprzedaży obuwia, która zaprezentowała swe rozwiązanie na początku lat 90., podczas dorocznej konferencji organizacji AMSU[4], odbywającej się na Uniwersytecie w Yorku. Polegało ono na wykonywaniu każdej doby analizy popytu, z podziałem na wzory, rozmiary oraz lokalizację, i kierowanie zaopatrzeniem sklepów według jej wyników. Analizę tę wykonywano na transakcyjnej bazie danych, po zakończeniu obsługi transakcyjnej i tzw. codziennego przetwarzania wsadowego[5].

---

[2] Cechy konstrukcyjne ówczesnego sprzętu komputerowego utrudniają dokładniejszy rachunek. Przykładowo: pojedynczy wymienny pakiet dyskowy o pojemności 30 MB, w systemie komputerowym zakupionym na początku lat 70. XX w. przez Zakłady H. Cegielski w Poznaniu, kosztował 210 GBP i wymagał dla działania napędu kosztującego ok. 10 000 GBP. Dla porównania – tani średniolitrażowy samochód kosztował wtedy w Wielkiej Brytanii (kraju producenta wspomnianego komputera) ok. 1200-1300 GBP (źródło: notatki autora).

[3] Mimo że spadek ten liczył się w rzędach wielkości, nie można go odnosić w równej mierze do sprzętu masowego (np. komputerów osobistych) i – będącego przedmiotem selekcji jakościowej – sprzętu stosowanego profesjonalnie.

[4] Association of Mainframe System Users – organizacja zrzeszająca użytkowników dużych komputerów produkcji brytyjskiej firmy ICL, przejętej później przez japońską firmę Fujitsu; w pracach i działaniach AMSU w latach 80. i 90. XX w. uczestniczyli liczni przedstawiciele użytkowników komputerów firmy ICL w Polsce.

[5] Źródło: notatki autora.

Praktyka taka była wówczas dość powszechna, miała jednak wiele wad. Jedną z nich była konieczność dysponowania odpowiednią ilością czasu, który każdej doby można było przeznaczyć na działania analityczne. Ich nieoczekiwane przedłużenie się stwarzało ryzyko opóźnienia w rozpoczęciu sesji obsługi transakcyjnej kolejnego dnia, co mogło oznaczać nawet brak możliwości obsługi bieżącej sprzedaży. Jej prowadzenie równolegle z działaniami analitycznymi było niemożliwe, gdyż te ostatnie angażowały niemal całość zasobów mocy obliczeniowej. Jeszcze większą przeszkodą okazała się wkrótce sama organizacja zapisów danych w plikach i bazach, dobrze uwzględniająca potrzeby przetwarzania transakcyjnego, ale niezbyt przydatna do złożonych, masowych działań analitycznych.

W efekcie, na przełomie lat 80. i 90. XX w., doprowadziło to do prób wyodrębniania repozytoriów danych przeznaczonych wyłącznie na potrzeby analityczne, na co wpływ miał również dalszy rozwój techniczny (zwiększanie pojemności pamięci dyskowych) i spadek cen sprzętu informatycznego. Repozytoria takie zaczęto określać mianem „hurtowni danych" (*data warehouse*), a za twórców ich podstaw uchodzą Bill Inmon i Ralph Kimball.

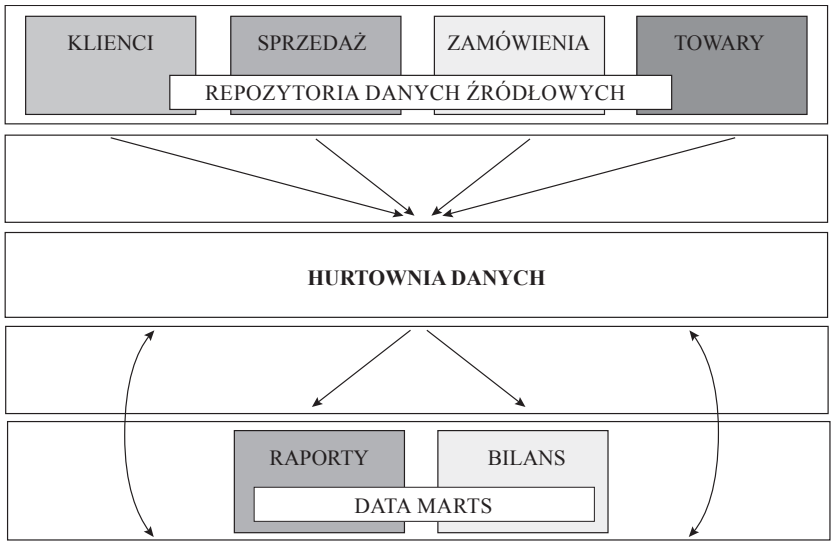
Hurtownie danych cechują się strukturą danych odmienną od znanej z transakcyjnych baz danych, gdzie przeważa „klasyczny" model relacyjny, chociaż – głównie w gałęziach przemysłu opartych na montażu – spotyka się również model hierarchiczny. Ta zasadnicza różnica w strukturze danych przesądza o fizycznej odrębności repozytoriów stanowiących hurtownie danych. Potrzebne do ich analizy znaczne moce obliczeniowe powodują też, że stosuje się do tego celu specjalizowane komputery, pracujące ze specjalistycznym oprogramowaniem.

W przypadku niektórych analiz wykonywanych na danych zgromadzonych w hurtowni zapotrzebowanie na moc obliczeniową jest tak duże, że praktycznie uniemożliwia równoległą realizację więcej niż jednej analizy. Przypadki takie powodują, że – przykładowo – analizy *ad hoc*, wykonywane w tym samym czasie, co zaplanowane, powtarzalne analizy rutynowe, utrudniają lub wręcz uniemożliwiają planowe wykonanie tych ostatnich. W celu zapobiegania takim sytuacjom stosuje się tzw. *data marts*[6], będące w istocie kopiami określonych podzbiorów hurtowni danych, przeznaczonymi do prowadzenia autonomicznych działań analitycznych o określonym zakresie.

Podzbiory takie, dla podkreślenia ich pochodzenia w całości od hurtowni danych, określa się mianem *dependent data marts*. Koncepcję tego rodzaju przedstawiono poglądowo na rys. 1.

---

[6] Termin *data mart* nie ma, jak dotąd, polskiego odpowiednika; samo słowo *mart* w języku angielskim oznacza *targowisko* i jest pewnego rodzaju pochodną, też zapożyczonego z handlu, terminu *hurtownia*, rozumianej jako jednostka umieszczona wyżej niż targowisko w hierarchii pośredników między producentem a konsumentem.



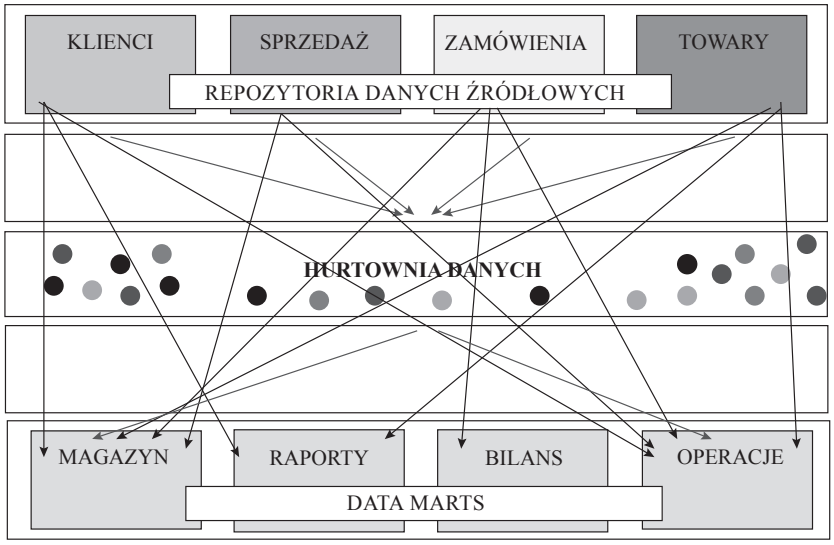
Rys. 1. Koncepcja *dependent data marts*

Ź r ó d ł o: opracowanie własne.

W licznych jednak dużych organizacjach, najczęściej z przyczyn związanych z ich złożoną przeszłością[7], ale również w wyniku sięgania po doraźne, uproszczone rozwiązania, wykształciła się praktyka korzystania z tzw. *independent data marts*. Polega ona na tworzeniu analitycznych *data marts* bezpośrednio z repozytoriów danych źródłowych, z pominięciem hurtowni danych bądź z jej tylko częściowym udziałem. Rozwiązania takie wydają się atrakcyjne dzięki szybkiemu przedstawianiu wyników, jednak – w dłuższej perspektywie – stwarzają ryzyko niespójności takich samych lub podobnych wyników, otrzymywanych z innych repozytoriów danych, również będących *independent data marts*, czy też – samej hurtowni danych. Wyniki takie mogą też być zaprzeczeniem sformułowanej przez B. Inmona zasady „jednej wersji prawdy" (*single version of truth*)[8]. Według tej zasady poleganie wyłącznie na danych pochodzących z ich hurtowni da zawsze takie same wyniki tych samych analiz. Pomijanie natomiast, czy nawet „obchodzenie" hurtowni grozi niespójnościami i rozbieżnościami w wynikach, które – przełożone na decyzje – mogą mieć negatywne skutki. Rozwiązanie z użyciem *independent data marts* przedstawia rysunek 2.

[7] Do takich przyczyn można zaliczyć łączenia się i podziały czy próby stosowania analitycznych rozwiązań informatycznych z okresu przed hurtowniami danych.

[8] Zob. www.b-eye-network.com/view/282.



Rys. 2. Koncepcja *independent data marts*

Ź r ó d ł o: opracowanie własne.

Skoro jednak hurtownia danych stanowi odrębną od pozostałych systemów informatycznych stosowanych w danej organizacji całość, pojawia się kwestia przenoszenia do niej danych z tych innych systemów. Nie stanowi to na ogół większego problemu tam, gdzie dane w hurtowni stanowią kopię danych z transakcyjnej bazy danych, a różnią się jedynie strukturą i organizacją. Przypadki tego rodzaju występują zazwyczaj tylko w niedużych organizacjach. Tam jednak, gdzie są liczne, zróżnicowane źródła danych, a ilość samych danych jest znaczna, terminowe i jakościowo poprawne zasilanie hurtowni danych nabiera szczególnego znaczenia i jest realizowane z udziałem specjalistycznego oprogramowania.

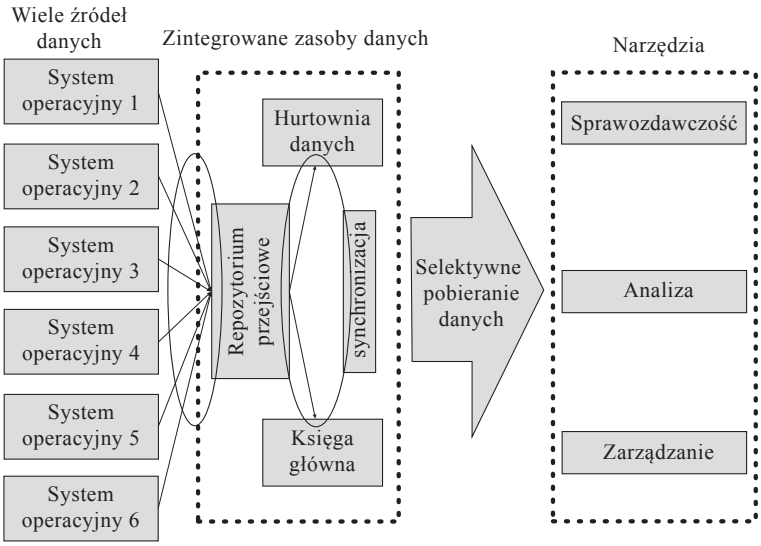
Wspomniane oprogramowanie jest określane symbolem ETL[9], który jest skrótem od angielskich terminów *Extract – Transform – Load*. Terminy te oddają istotę działania tego oprogramowania, które sprowadza się do trzech podstawowych czynności:

– pobrania danych z ich repozytoriów źródłowych (*Extract*),
– dokonania reorganizacji, przekształceń i agregacji danych (*Transform*),
– umieszczenia zreorganizowanych i przekształconych danych w hurtowni danych (*Load*).

[9] Krótki zarys historii narzędzi ETL w: Y. Montcheuil, C. Dupupet, *Third Generation ETL: Delivering the Best Performance*, Sunopsis Inc., Boston 2007.

3. Charakterystyka narzędzi ETL

Przykładowy, całościowy schemat stosowania hurtowni danych w banku przedstawia rys. 3. Po jego lewej stronie występują rozmaite repozytoria-źródła danych, pochodzących z systemów operacyjnych, zapewniających bieżącą obsługę działalności banku. Dane z tych repozytoriów są pobierane według określonych reguł (funkcja „Extract") i umieszczane w repozytorium przejściowym (*data stage*). W celu zagwarantowania, że dane repozytorium źródłowe odzwierciedla stan z określonego momentu (np. na koniec dnia w rozumieniu księgowym), na czas pobierania danych blokuje się możliwość aktualizacji zapisów w takim repozytorium. Blokada taka wyłącza dane repozytorium z działań bieżących, co, w niektórych warunkach, może ograniczać ich zdolność do działania[10]. Powoduje to dążenie do możliwie największego skrócenia operacji pobierania, czego wynikiem jest jej ograniczanie do samego tylko pobrania i przeniesienia danych. Tam, gdzie nie ma takiego ograniczenia, spotyka się rozwiązania, w których w trakcie przenoszenia danych dokonuje się również ich reorganizacji, przekształceń i agregacji. W przypadku wielu źródeł danych działania te mają charakter wstępny, czasem tylko kontrolny, a ich wyniki są podstawą do właściwych przekształceń, wykonywanych w samym już tylko repozytorium przejściowym (por. rys. 3).



Rys. 3. Schemat stosowania hurtowni danych w banku

Ź r ó d ł o: opracowanie własne.

[10] Ograniczenie takie wystąpi wtedy we wszystkich systemach działających w trybie określanym jako 7x24 (24 godziny na dobę, przez wszystkie dni tygodnia), czyli bez żadnych przerw.

Dwa etapy, w których najczęściej znajdują zastosowanie narzędzia ETL, oznaczono na rys. 3 owalami, umieszczonymi po obu stronach repozytorium przejściowego. Spotyka się jednak też rozwiązania, gdzie narzędzie te stosuje się po „drugiej" niejako stronie hurtowni danych, czyli do tworzenia z niej, wspomnianych już wcześniej, *data marts*. Nie jest to jednak zamierzony, główny cel powstania i istnienia tych narzędzi.

Portal internetowy o nazwie ETL Tools[11], zajmujący się różnymi aspektami praktycznymi związanymi z tymi narzędziami, pośród pożądanych cech narzędzi ETL wymienia następujące:

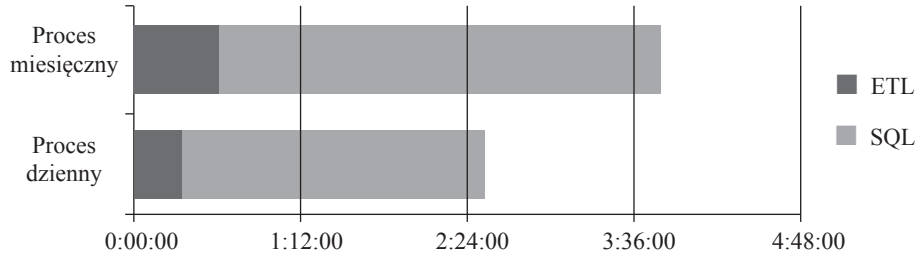
– obsługa podziału dużych tabel danych,
– obsługa bardzo dużych ilości danych,
– wykonywanie kontroli poprawności danych,
– graficzne odwzorowanie związków między repozytoriami i elementami danych,
– działanie z wieloma wersjami systemów operacyjnych i oprogramowania baz danych,
– obsługa metadanych,
– obsługa wykrywania błędów,
– obsługa gwiaździstych schematów organizacji danych,
– działanie wielowątkowe,
– kontrola wersji,
– harmonogramowanie zadań,
– graficzny interfejs użytkownika,
– interfejs przeglądarki internetowej.

Powyższą listę można uznać za dość wyczerpującą i opartą na szerokim doświadczeniu praktycznym, brak w niej jednak jednego szczególnie istotnego kryterium, jakim są tzw. interfejsy własne (*native interfaces*). Istotą tych interfejsów jest ich przygotowanie do współdziałania ze źródłowymi repozytoriami danych oparte na znajomości i wykorzystaniu ich wewnętrznych mechanizmów, zamiast sięgania po rozwiązania znormalizowane, typu język SQL. Żądania wykonania operacji na bazie danych, sformułowane w tym języku, przed właściwym wykonaniem każdorazowo wymagają konwersji i sprowadzenia do poziomu wspomnianych interfejsów własnych. Czynności te wydłużają znacznie czas trwania operacji pobierania danych, wydłużając w ten sposób okres niedostępności danego repozytorium dla innych działań. Przykład różnic między czasem trwania operacji na tych samych danych, raz wykonanych metodami tradycyjnymi (SQL), drugi raz – za pomocą narzędzi ETL z udziałem interfejsu własnego, przedstawia rys. 4[12].

---

[11] Zob. ww.etltools.com.

[12] Przykład z 2006 r. z jednego z polskich banków.



Rys. 4. Czas pobierania tych samych danych metodami SQL i ETL

Ź r ó d ł o: opracowanie własne.

Z danych przedstawionych na rys. 4 wynika, że różnica w czasie przetwarzania jest ponad pięciokrotna na korzyść interfejsów własnych. Po stronie wad tych interfejsów należy jednak wskazać to, że nie zawsze nadążają one za zmianami wprowadzanymi do obsługujących je mechanizmów przez producentów oprogramowania baz danych[13].

Powyższe nie oznacza jednak, że od narzędzi ETL nie oczekuje się „klasycznych" metod operowania danymi. Metody te obejmują nie tylko wspomniany już tu język SQL, ale również metody ODBC/JDBC, pliki jednowymiarowe (*flat files*) oraz usługi ESB[14].

Inną ważną właściwością narzędzi ETL jest ich zdolność do obsługi metadanych – zarówno po stronie systemów źródłowych, z których dane mają być pobierane, jak i po stronie hurtowni danych, gdzie narzędzia te potrafią działać z różnymi tzw. modelami danych. Modele takie to gotowe, wzorcowe struktury danych, przygotowane i udostępniane przez producentów hurtowni danych. Występują one w wielu wersjach, przeznaczonych dla różnych branż, z pewnym zakresem możliwości własnego kształtowania takiego modelu przez użytkownika.

Przygotowanie niektórych dostępnych na rynku narzędzi ETL do obsługi metadanych stanowi znaczne ułatwienie w korzystaniu z tych narzędzi i przyczynia się do ujednolicenia kategorii terminologicznych z zakresu nazewnictwa elementów danych. Ujednolicenie takie ułatwia istotnie porozumiewanie się służb biznesowych i informatycznych działających w danej organizacji. Pełne jednak ujednolicenie w zakresie metadanych jest ciągle bardzo odległą perspektywą, gdyż w praktyce

[13] Producenci ci nie zawsze informują wytwórców narzędzi ETL o wprowadzanych przez siebie zmianach i usprawnieniach, gdyż często sami dostarczają również takie narzędzia i nie leży w ich interesie usprawnianie działania narzędzi konkurentów.

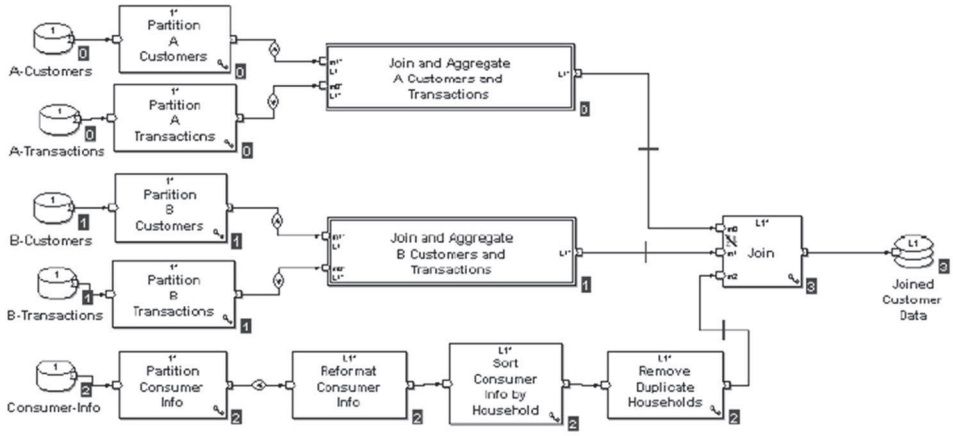
[14] Metody te wymieniane są m.in. w: P. Russom, *How To Evaluate Enterprise ETL*, Forrester Research, Cambridge (US), 2004, s. 8 oraz Y. Montcheuil, C. Dupupet, op. cit., s. 9 (rozdział „Data Access Technologies"); jako wymóg minimum konieczność ich obsługi przez narzędzia ETL wymieniają też autorzy opracowania: W. Eckerson, C. White, *Evaluating ETL and Data Integration Platforms*, The Data Warehousing Institute, Seattle 2003.

omawianej tu dziedziny występują obecnie trzy odmienne główne kategorie metadanych, a mianowicie metadane: biznesowe, operacyjne i techniczne. Pełne ich ujednolicenie dla wszystkich etapów – poczynając od pobierania danych z repozytoriów źródłowych, a kończąc na działaniach analitycznych – długo jeszcze nie będzie możliwe.

Z korzystaniem z metadanych wiąże się inna cecha, występująca pośród przytoczonych tu wcześniej właściwości, jakimi winny cechować się narzędzia ETL. Chodzi o tzw. interfejs graficzny, pozwalający projektować i wyznaczać związki między źródłowymi repozytoriami danych a zasobami hurtowni, w których dane te mają się znaleźć. Projektowanie to i wyznaczanie obejmuje również wskazywanie, posługując się metadanymi, jakie dane podlegają przenoszeniu i jakie transformacje, agregacje i czynności kontrolne mają przy tej okazji być na nich wykonane[15].

Przykład ekranowego interfejsu graficznego narzędzia ETL przedstawia rys. 5. Łączy się tam, przekształca i agreguje w jedną strukturę dane pochodzące z pięciu odrębnych repozytoriów źródłowych[16]. Patrząc od góry – dane z dwóch par repozytoriów są tam łączone wstępnie, podczas gdy dane z piątego repozytorium są poddawane przekształcaniu i reorganizacji, po czym dane z wszystkich pokazanych tam źródeł są łączone i umieszczane w jeszcze innym repozytorium.

Przykład stosowania reguł transformacji i kontroli danych ukazuje rys. 6. Odzwierciedla on jednocześnie przebieg testowy, jaki można wykonać w celach kontrolnych przy projektowaniu reguł pobierania i transformacji danych.

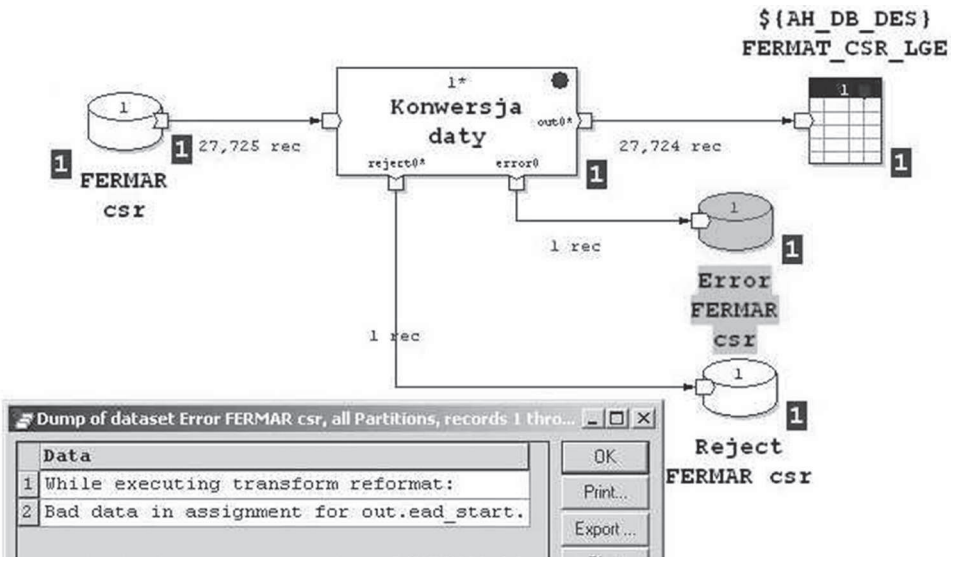


Rys. 5. Narzędzia ETL – graficzny interfejs użytkownika

Ź r ó d ł o: opracowanie własne (z praktyki).

[15] Brak interfejsu graficznego oznaczałby konieczność formułowania zadań w jakimś przeznaczonym do tego języku, co byłoby kłopotliwe w stosowaniu, bardzo pracochłonne i mało elastyczne.

[16] Wszystkie przytoczone tu przykłady pochodzą z systemu ETL o nazwie Co>Operation firmy Ab Initio.



Rys. 6. Narzędzia ETL – przebieg testowy

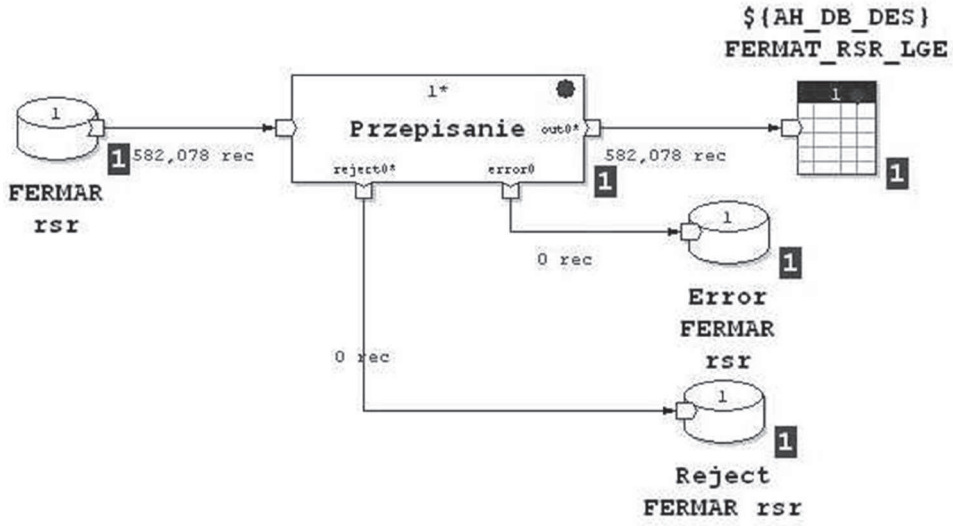
Ź r ó d ł o: opracowanie własne (z praktyki).

Ze schematu na rys. 6 widać, że z repozytorium o nazwie *FERMAR csr* pobrano 27 725 zapisów danych, z których jeden został zakwestionowany podczas wykonywania procedury o nazwie „Konwersja daty" i skierowany jednocześnie do dwóch repozytoriów pomocniczych: gromadzącego zapisy błędne (*Error FERMAR csr*) oraz zawierającego zapisy odrzucone w trakcie transformacji (*Reject FERMAR csr*).

Jeszcze inna istotna właściwość narzędzi ETL to realizacja funkcji kontrolnych z zakresu kompletności procesów przekształcania i przenoszenia danych. Hurtownia danych jest podstawą do sporządzania raportów, analiz oraz przygotowywania danych statystycznych. Wykonywanie wielu z tych czynności jest realizacją ustawowych obowiązków, a przekazanie – w ich ramach – błędnych danych może być nawet ścigane sądownie. Podejmowanie decyzji na podstawie błędnych czy chociażby tylko nieprecyzyjnych danych nie leży też w interesie organizacji, w której ma ono miejsce. Wszystkie przytoczone tu względy przemawiają więc za tym, aby przenoszone i przekształcane dane kontrolować również ilościowo, po to by mieć pewność, że uwzględniono wszystkie dane, które powinny być wzięte pod uwagę[17]. Przykład takiej kontroli przedstawia rys. 7. Przy symbolach repozytoriów: źródłowym (*FERMAR rsr*) i docelowym (*FERMAT_RSR_LGE*) widać liczniki przeniesionych zapisów (w obu przypadkach o stanie 582 078). Jedno-

[17] Ten aspekt praktyki stosowania narzędzi ETL poruszany jest m.in. w: P. Russom, op. cit.

cześnie ten sam schemat pokazuje, że żadne zapisy nie zostały w trakcie tego procesu zakwestionowane (zero zapisów w repozytoriach *Error FERMAR rsr* oraz *Reject FERMAR rsr*).



Rys. 7. Narzędzia ETL – kontrola ilościowa

Ź r ó d ł o: opracowanie własne (z praktyki).

Wymogi praktyczne, jakim winny odpowiadać narzędzia ETL zastosowane w konkretnej organizacji, mogą się różnić w szczegółach, ale można też odnieść do nich kilka podstawowych zasad, istotnych w większości sytuacji[18].

4. Narzędzia ETL a inne metody zasilania

Narzędzia ETL należą do najczęściej stosowanych w zasilaniu hurtowni danych w dane, ale nie są jedynym środkiem do tego celu. Ich wadą jest np. konieczność stosowania repozytorium pośredniego, co nie tylko jest źródłem dodatkowych kosztów, ale stanowi również istotny czynnik wydłużający cały proces zasilania. Narzędzia ETL nie radzą sobie też dobrze z przypadkami, w których potrzebne jest zasilanie ciągłe (systemy działające w trybie *on-line* i *quasi on-line*).

[18] Krótkie omówienie tych zasad w: *Managing Big Data: Building the Foundation for a Scalable ETL Environment*, Knightsbridge Solutions, Chicago 2002.

Rozwiązaniem mającym eliminować niektóre z wymienionych wad są narzędzia określane skrótem EL-T[19], których istota działania zakłada odwróconą, w stosunku do klasycznych narzędzi ETL, kolejność operacji: dane pobrane z repozytoriów źródłowych (funkcja *E-xtract*) najpierw są umieszczane w hurtowni danych (funkcja *L-oad*) i dopiero tam poddawane przekształceniu (funkcja *T-ransform*).

Inny postulat formułowany wobec narzędzi ETL dotyczy wydajności ich działania, gdzie stałym problemem są bardzo duże (i rosnące) ilości przenoszonych danych. Dla przestrzegania swoistej „czystości reguł" i w celu sprawowania właściwej kontroli nad procesami umieszczania danych w hurtowniach danych liczne dane są poddawane związanym z tym operacjom wielokrotnie. Działania takie pochłaniają zasoby infrastruktury informatycznej i pociągają za sobą koszty. Stan ten wpłynął na uzupełnienie narzędzi ETL o kolejną właściwość, określaną mianem *change data capture*[20]. Działanie w tym trybie polega na bieżącej, dokonywanej bezpośrednio w hurtowni danych aktualizacji tylko tych danych, które uległy zmianie. Jeszcze inne oczekiwania wobec narzędzi i metod ETL wiążą się z konceptem tzw. *Big Data* i – wiązanej m.in. z nim – metodyki przetwarzania *Hadoop*[21].

5. Podsumowanie

Mimo stosunkowo krótkiej historii hurtownie danych i metody ich zasilania danymi przeszły już długą ewolucję. Trwa ona nadal, gdyż wobec rozwiązań tych formułuje się coraz to nowe wymogi. Wiele z tych wymogów można spełnić również w wyniku rozwoju technicznego, powodującego, że konkretne rozwiązania, często od dawna przygotowane teoretycznie, znacznie później znajdują swe uzasadnienie ekonomiczne.

Liczne pierwsze zastosowania hurtowni danych zakładały, że wystarczy zebrać odpowiednio dużą ilość możliwie najbardziej szczegółowych danych, by w wyniku odnalezienia ukrytych w nich, często głęboko, prawidłowości uzyskać

[19] Przykład takiego rozwiązania i jego zalety przedstawiono w: *Is ETL Becoming Obsolete? Why a Business-Rules-Driven "E-LT" Architecture is Better*, Sunopsis Inc., Boston 2006.

[20] Metodykę tę bliżej omówiono m.in. w: *Augmenting ETL Systems With Real-Time Change Data Capture*, GoldenGate Software Inc., San Francisco 2007; tamże w związku z tym poddaje się również krytyce samą koncepcję „dnia operacyjnego" jako nieprzystającą do współczesnych potrzeb.

[21] Szeroki przegląd obecnego stanu wymogów wobec narzędzi ETL można znaleźć w: N. Yuhanna, *The Forrester Wave™: Enterprise ETL, Q1, 2012*, Forrester Research, Cambridge (US), 2012, tam też dokonano przeglądu i porównania bieżących możliwości tych narzędzi i zawarto wielokryteriową ocenę przodujących rozwiązań z tego zakresu.

zaskakujący efekt biznesowy, dający skokowy wzrost przewagi nad konkurentami. Przypominało to – w jakimś sensie – działania średniowiecznych alchemików, poszukujących dobrze ukrytego przez naturę sposobu na przemianę żelaza w złoto, o istnieniu którego byli przekonani.

Obecna praktyka hurtowni danych, jak każda inna dziedzina informatyki, rządzi się jednak realnymi i twardymi regułami, co nie oznacza, że nie może, w niektórych przypadkach, być źródłem spektakularnych sukcesów. Codzienność tej dziedziny to jednak żmudne, powtarzalne działania, których realizacja wymaga stałej dbałości o jakość danych i precyzję wyników. Istotną w tym rolę pełnią narzędzia ETL, które – podobnie jak cała dziedzina gromadzenia i analizy danych – podlegają ciągłej ewolucji i doskonaleniu.

Literatura[22]

Augmenting ETL Systems With Real-Time Change Data Capture, GoldenGate Software Inc., San Francisco 2007.

Data Integration: Moving Beyond ETL, DataFlux Corporation, Cary 2010.

Eckerson W., White C., *Evaluating ETL and Data Integration Platforms*, The Data Warehousing Institute, Seattle 2003.

Is ETL Becoming Obsolete? Why a Business-Rules-Driven "E-LT" Architecture is Better, Sunopsis Inc., Boston 2006.

Managing Big Data: Building the Foundation for a Scalable ETL Environment, Knightsbridge Solutions, Chicago 2002.

Montcheuil Y., Dupupet C., *Third Generation ETL: Delivering the Best Performance*, Sunopsis Inc., Boston 2007.

Russom P., *How To Evaluate Enterprise ETL*, Forrester Research, Cambridge (US), 2004.

Yuhanna N., *The Forrester Wave™: Enterprise ETL, Q1, 2012*, Forrester Research, Cambridge (US), 2012.

[22] Wykaz ten należałoby uzupełnić o liczne podręczniki systemu o nazwie Co>Operation firmy Ab Initio, czołowego producenta narzędzi ETL, które, mimo że nie są przytaczane w opracowaniu bezpośrednio, były wykorzystywane przy jego powstawaniu.