PORTFOLIO:

Analiza zastosowania systemów hurtowni danych w odniesieniu do oczekiwań operatorów systemów dystrybucyjnych (OSD)

Autor: Krzysztof Osiński, Jacek Popow, Wojciech Komnata, Sebastian Ernst

Centrum Inteligentnych Systemów Informatycznych Akademia Górniczo-Hutnicza im. Stanisława Staszica al. Mickiewicza 30, 30-059 Kraków
budynek C-2 pokój 426 tel.: 12 617 44 53 www.isi.agh.edu.pl **isi@agh.edu.pl**

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Celem niniejszego raportu jest określenie wstępnych potrzeb operatorów systemów dystrybucyjnych (OSD) w zakresie wykorzystania technologii hurtowni danych. Wymagania OSD powinny uwzględniać aktualną oraz przewidywaną sytuację całego polskiego sektora energetycznego, który jak powszechnie się uważa stoi przed ważnymi wyzwaniami.

W naszym kraju rośnie zużycie energii elektrycznej, niestety zwiększające się zapotrzebowanie nie idzie w parze z modernizacją infrastruktury zarówno wytwórczej jak i transportowej. Dodatkowo na krajową politykę energetyczną wpływ ma także międzynarodowa polityka ochrony środowiska. Unia Europejska wyznaczyła do roku 2020 tzw. trzy cele ilościowe (3 x 20%) związane ze zmniejszeniem emisji gazów cieplarnianych o 20% (względem roku 1990), zmniejszenie zużycia energii o 20% (względem prognoz na rok 2020) oraz zwiększenie udziału odnawialnych źródeł energii do 20% całkowitego zużycia energii.

Polska czynnie uczestnicząc w opracowywaniu unijnej polityki energetycznej, sukcesywnie wdraża główne wytyczne uwzględniając krajowe uwarunkowania i regionalną specyfikę.

Rada Ministrów 10 listopada 2009 roku przyjęła dokument opracowany przez Ministerstwo Gospodarki pt. „Polityka energetyczna Polski do 2030 roku”[1]. Materiał ów przedstawia najważniejsze obszary polskiej energetyki uwzględniając zarówno rosnące potrzeby jak i nasze zobowiązania względem ochrony środowiska czy też wymagań Unii Europejskiej.

Listę sześciu zagadnień rozpoczyna poprawa efektywności energetycznej Polski będąca punktem krytycznym dla całej strategii. Ministerstwo zakłada, przy utrzymującym się wzroście gospodarczym, brak zwiększonego zużycia energii elektrycznej, co w praktyce oznacza zmniejszenie energochłonności. Punktem odniesienia jest poziom 15-tu krajów Unii Europejskiej (tzw. UE-15). Resort gospodarki przygotował ustawę o efektywności

[1] http://www.mg.gov.pl/files/upload/8134/Polityka%20energetyczna%20ost.pdf. Data odczytu 10 lutego 2015 roku.

energetycznej[2],[3],[4] której celem jest wsparcie systemu opartego na białych certyfikatach[5],[6]. Dodatkowo Ministerstwo Gospodarki będzie stymulowało rozwój i wykorzystanie procesów technologicznych ograniczających zużycie paliwa wykorzystywanego do produkcji energii elektrycznej, a przez to zmniejszenie zanieczyszczenia środowiska.

Kolejnym poruszonym przez resort gospodarki obszarem jest problematyka bezpieczeństwa energetycznego. Istotną kwestią jest wzrost bezpieczeństwa dostaw energii w oparciu o posiadane przez Polskę zasoby, w których najważniejszą rolę odgrywa węgiel kamienny i brunatny. Obok krajowych zasobów mają być prowadzone prace prowadzące do dywersyfikacji dostawców paliw z uwzględnieniem rozwoju współpracy transgranicznej. Równolegle z dywersyfikacją źródeł dostaw paliw mają być prowadzone prace w obszarze dywersyfikacji struktur wytwarzania energii elektrycznej, ze szczególnym zwróceniem uwagi na energetykę jądrową,

Do 2020 roku wykorzystanie odnawialnych źródeł energii (OZE) ma stanowić 15% udział w zużyciu energii – jest to czwarty kierunek omówiony w strategii rozwoju polskiej energetyki. Ostatnie dwa to zwiększenie konkurencji na rynku energii oraz ograniczenie oddziaływania energetyki na środowisko.

---

[2] Ustawa z dnia 15 kwietnia 2011 r. o efektywności energetycznej (Dz.U. 2011 nr 94 poz. 551) http://isap.sejm.gov.pl. Data odczytu 10 lutego 2015 roku.

[3] „Drugi Krajowy Plan Działań dotyczący efektywności energetycznej dla Polski, 2011." Dokument przyjęty przez Radę Ministrów w dniu 17 kwietnia 2012 roku. www.mg.gov.pl. Data odczytu 10 lutego 2015 roku.

[4] „Krajowy Plan Działań dotyczący efektywności energetycznej dla Polski, 2014." Dokument przyjęty przez Radę Ministrów w dniu 20 października 2014 roku. www.mg.gov.pl. Data odczytu 10 lutego 2015 roku.

[5] http://pl.wikipedia.org/wiki/System_białych_certyfikatów : białe certyfikaty to świadectwa potwierdzające zaoszczędzenie określonej ilości energii w wyniku realizacji inwestycji służących poprawie efektywności energetycznej. Data odczytu 10 lutego 2015 roku.

[6] www.oze.pl: w celu usystematyzowania źródeł pochodzenia energii elektrycznej wprowadzone zostały różnokolorowe certyfikaty: (1) zielone – źródła odnawialne, (2) czerwone – wysokosprawna kogeneracja, (3) żółte (wcześniej niebieskie) – małe źródła kogeneracyjne opalane gazem lub o mocy poniżej 1 MW, (4) fioletowe – źródła wykorzystujące gaz z odmetanowanych kopalń lub biogaz, (5) pomarańczowe – wykorzystanie instalacji wychwytywanie i zatłaczania dwutlenku węgla (CCS – Carbon Capture and Storage), (6) błękitne – z nowych wysokosprawnych źródeł, (7) białe – poprawiające efektywność energetyczną, (8) brązowe – produkcja i wprowadzanie do sieci biogazu rolniczego. Data odczytu 10 lutego 2015 roku.

Niezależnie od opracowań resortowych dostępne są analizy publikowane przez wytwórców i dostawców energii. Jednym z przykładów jest studium europejskiego koncernu energetycznego RWE pt. „Scenariusze rozwoju technologii na polskim rynku energii do 2050 roku”[7]. Dokument powstał w oparciu o analizę czterech długookresowych hipotetycznych i holistycznych scenariuszy rozwoju polskiej energetyki: zachowawczy, krajowy, zielony i innowacyjny. Poszczególne scenariusze zależne są od stabilności polityki energetycznej, długookresowości podejmowanych decyzji, wysokości PKB, koncentracji na krajowej niezależności energetycznej, popycie na energię elektryczną, poziomie rozwoju energetyki odnawialnej, unijnej polityki energetyczno-klimatycznej, energetyce słonecznej, wiatrowej, jądrowej oraz intensywności rozwoju innowacji i nowych technologii wytwarzania energii elektrycznej. Przedstawione w raporcie wnioski są dla naszego kraju optymistyczne[8]:

- Polska pozostanie niezależna pod względem produkcji energii elektrycznej,
- Polska może zrealizować ambitne unijne cele redukcji CO_2,
- transformacja systemu elektroenergetycznego nie powinna spowodować wzrostu cen energii,
- węgiel kamienny i brunatny będą odgrywać istotną rolę w strukturze procesu wytwarzania energii,
- zwiększony zostanie udział odnawialnych źródeł energii,
- rozwinie się rynek prosumencki.

Kończąc wstęp niniejszego opracowania należy podkreślić wagę i znaczenie narodowego rynku energii elektrycznej, zarówno w części wytwórczej, jak i przesyłowej. Owa infrastruktura w ramach ustawy z dnia 26 kwietnia 2007 roku o zarządzaniu kryzysowym[9] definiuje infrastrukturę krytyczną jako system powiązanych ze sobą obiektów i usług kluczowych dla bezpieczeństwa państwa i jego obywateli. W ramach

[7] http://www.rwe.pl/web/cms/mediablob/pl/2560216/data/2562680/2/start/dla-mediow/aktualnosci/ Studium RWE: „Scenariusze rozwoju technologii na polskim rynku energetyki do 2050 roku”. Data odczytu 10 lutego 2015.

[8] Tamże.

[9] Ustawa z dnia 26 kwietnia 2007 roku o zarządzaniu kryzysowym. Dz.U. 2007 Nr 89 poz. 590 z późniejszymi zmianami. http://isap.sejm.gov.pl Data odczytu 10 lutego 2015 roku.

Narodowego Programu Ochrony Infrastruktury Krytycznej[10] na pierwszym miejscu wskazywane jest zaopatrzenie w energię, surowce energetyczne i paliwa.

Praktyczne zastosowanie nowoczesnych i efektywnych technologii teleinformatycznych, w szczególności rozwiązań hurtowni danych, staje się jednym z kluczowych i oczekiwanych zagadnień. Współczesny rynek energii elektrycznej charakteryzuje się dążeniem nie tylko do nowoczesnej infrastruktury wytwórczej i przesyłowej, ale także powszechnym wykorzystaniem technologii ITC wspierające zarówno procesy nadzorujące eksploatację, jak i wspomagające podejmowanie decyzji.

[10] Narodowy Program Ochrony Infrastruktury Krytycznej. 2013. http://rcb.gov.pl Data odczytu 10 lutego 2015 roku.

Wprowadzenie w zagadnienia technologii hurtowni danych

Czym jest hurtownia danych, jaką rolę pełni w procesie zarządzania przedsiębiorstwem, czego mogą od niej oczekiwać użytkownicy? Są to ważne zagadnienia, które istotnie wpływają na proces podejmowania decyzji na temat rozpoczęcia wdrożenia zaawansowanych technologicznie rozwiązań zapewniających skuteczny dostęp do zgromadzonych danych.

Warto zwrócić uwagę, iż sam dostęp do zgromadzonych danych nie jest żadnym wyróżnikiem. Systemy baz danych są szeroko stosowanymi rozwiązaniami zapewniającymi gromadzenie, przechowywanie i udostępnianie danych, jednakże są to zwykle dane transakcyjne. Większość typowych baz danych nie ma możliwości cofnięcia się do stanu sprzed godziny, dnia, miesiąca czy roku (nie odnosimy się w tym przypadku do danych zarchiwizowanych lub znajdujących się w kopiach bezpieczeństwa), ograniczając tym samym możliwości analizy predykcyjnej.

W niniejszym rozdziale przedstawiona została koncepcja hurtowni danych wykorzystywanych w procesie podejmowania decyzji. Dalsza część omawia obecnie dominujące architektury hurtowni danych jednocześnie prezentując ich cechy charakterystyczne.

Systemy wspomagające zarządzanie

Podrozdział przybliża proces podejmowania decyzji, koncentrując uwagę czytelnika na wymaganym wsparciu jakie dostarczają dostępne w analizowanym środowisku systemy teleinformatyczne.

Peter Drucker[11], badacz procesów organizacji i zarządzania przedsiębiorstw, w swoich pracach zdefiniował pięć etapów podejmowania decyzji:

- zdefiniowanie problemu, poprzez określenie celu działania, zmieniającego się otoczenia oraz uwarunkowań procesu,

[11] Peter F. Drucker, John Hammond, Ralph Keeney, Howard Raiffa, Alden M. Hayashi: Podejmowanie decyzji. Harvard Business Review. ONEPRESS, Gliwice, 2005

- przeprowadzenie analizy problemu, umożliwiającej dostateczne zakwalifikowanie problemu i tym samym wskazanie wymaganych informacji, umożliwiających podjęcie decyzji,
- opracowanie wariantów rozwiązania, wskazujących nie tylko na ich różnorodność, ale również na możliwość braku podjęcia decyzji,
- wskazanie najlepszej decyzji, z uwzględnieniem wszelkich wymaganych, ale i dostępnych, do realizacji zasobów, określeniem akceptowalnego terminu realizacji oraz poziomu ryzyka,
- podjęcie działań zgodnie ze wskazaną najlepszą decyzją.

Drucker[12] dodaje, że wyłącznie skuteczne działanie udowadnia rzeczywiście podjętą decyzję. Zaniechanie działania, wycofanie się lub wstrzymanie, powoduje, iż proces decyzyjny nie zostaje zakończony. Procesem zakończonym jest jednakże taki w którym nie została podjęta decyzja, jednakże pod warunkiem, iż to był jeden z wcześniej opracowanych wariantów.

Różne są przyjęte podziały decyzji, jeden z nich opiera się na stopniu ryzyka[13], gdzie proces realizowany jest w warunkach:

- pewności – wówczas można z dużym prawdopodobieństwem , wynikającym np. z doświadczenia lub modelu, przewidzieć efekty decyzji,
- niepewności – przy braku możliwości określenia wszystkich potencjalnych konsekwencji, czy też prawdopodobieństwa ich wystąpienia.
- ryzyka – kiedy znane są konsekwencje, wraz z probabilistycznymi przyporządkowanymi im wartościami.

W. Kieżuń[14] dodaje, iż istotą każdego procesu decyzyjnego jest przetworzenie informacji wejściowych i informacji przechowywanej w informację wyjściową, dzieląc decyzje wg roli w procesie zarządzania na strategiczne, operacyjne oraz realizacyjne. Dalsza część opracowania będzie odnosić się do powyższego podziału zwracając uwagę na różnice pomiędzy informacjami leżącymi u podstaw decyzji strategicznych i operacyjnych.

---

[12] Ibidem

[13] Wikipedia – Podejmowanie decyzji. Data odczyt 16 grudnia 2014 roku.

[14] Witold Kieżuń: Sprawne zarządzanie organizacją: zarys teorii i praktyki, Oficyna Wydawnicza SGH, Warszawa, 1997

Ponniah[15] w pięciu punktach charakteryzuje informacje strategiczne, które są:

- zintegrowane – w ramach danej organizacji wszystkie dane muszą mieć jeden, ten sam widok,
- dokładne – względem danych wymagana jest dokładność i zgodność z obowiązującymi regułami biznesowymi,
- dostępne – konieczny jest intuicyjny, łatwy dostęp do systemów analitycznych,
- wiarygodne – każdy mierzony w organizacji parametr musi mieć jedną i tylko jedną wartość,
- terminowe – wymagane, oczekiwane informacje muszą być dostępne w ściśle określonych ramach czasowych.

Niestety dość powszechnym błędem popełnianym podczas procesu decyzyjnego jest mylenie danych strategicznych z danymi operacyjnymi, zasilającymi systemy operacyjne, którymi są np.: przyjmowanie zleceń, reklamacji, fakturowanie, czy rozliczanie należności. W tym samym obszarze mieszczą się analizy bardziej złożone, np.: w którym regionie występują najczęstsze awarie, która brygada wykonała najwięcej napraw, czy też jaka jest średnia awaryjność wskazanej grupy urządzeń elektroenergetycznych.

Warto zwrócić uwagę, że tradycyjny system bazy danych nie da, niestety, odpowiedzi na następujące przykładowe pytania:

- czy jest wpływ warunków atmosferycznych na awaryjność wybranej grupy urządzeń, a jeżeli tak, to jakie są punkty brzegowe,
- które regiony można porównywać pomiędzy sobą ze względu na gęstość zaludnienia, strukturę poboru energii elektrycznej, poziom urbanizacji,
- jak efektywnie rozlokować brygady remontowe, aby uwzględniając predykcję awarii, minimalizować czas naprawy.

Przyjmując, że są one kluczowe, można spróbować wyciągnąć wniosek, iż tradycyjne systemy operacyjne musza być dodatkowo wspomagane systemami przystosowanymi do procesów podejmowania decyzji strategicznych.

[15] Paulraj Ponniah: Data Warehousing fundamentals for IT professionals. A John Wiley and Sons, Inc, Publication, Hoboken, New Jersey, 2nd edition, 2010

Różnice pomiędzy systemami operacyjnymi i strategicznymi zostały zebrane w tabeli[16]:

Parametr	System operacyjny	System strategiczny
Zawartość danych	Aktualne wartości	Dane zarchiwizowane, przetworzone, zagregowane
Struktura danych	Optymalizacja pod kątem realizacji pojedynczej transakcji	Optymalizacja pod kątem złożonych zapytań
Częstość dostępu	Wysoka	Średnia lub niska
Typ dostępu	Odczyt, aktualizacja, zmiana, usunięcie	Odczyt
Użycie	Przewidywalne, powtarzalne	Nieokreślone, ad hoc, przypadkowe
Czas odpowiedzi	Poniżej sekundy	Dziesiątki sekund, minuty
Liczba użytkowników	Duża liczba	Relatywnie mała liczba

W niniejszym raporcie, pozostawiając bez komentarza różne definicje systemów wspomagania podejmowania decyzji (ang. *Decision Support Systems*, DSS), autorzy ograniczą się do jednej, podanej przez Golfarelliego i Rizziego[17]: „*system wspomagania decyzji jest zestawem rozwijanych, interaktywnych technik i narzędzi informatycznych przeznaczonych do przetwarzania i analizowania danych oraz wspierania menadżerów w podejmowaniu decyzji*".

Podstawy hurtowni danych

[16] Dariusz Dymek, Wojciech Komnata, Leszek Kotulski, Piotr Szwed: Architektury Hurtowni Danych. Model referencyjny i formalny opis architektury. Wydawnictwo AGH, Kraków, 2015

[17] Matteo Golfarelli, Stefano Rizzi: Data Warehouse Design, Modern Principles and Methodologies. Tata McGraw Hill Education Private Limited, New Delhi, 2009

Koncepcja hurtowni danych nie jest ideą nową. Pierwsza poważana publikacja pojawiła się w periodyku *IBM Systems Journal*, gdzie Barry Devlin i Paul Murphy w 1988 roku opublikowali artykuł *„An Architecture for a Business and Information Systems”*. Kolejnym krokiem było wydanie w 1991 roku przez Billa Inmona książki pod tytułem „*Building the Data Warehouse*”. Pozycja ta pozwoliła środowisku IT ogłosić Inmona bezapelacyjnym ojcem hurtowni danych. Tytuł ten, pośrednio, zakwestionowała publikacja Ralpha Kimballa „*The Data Warehouse Toolkit*” wydana w 1996 roku rozpoczynając trwający do dnia dzisiejszego spór na temat wyższości poszczególnych podejść do koncepcji hurtowni danych.

Inmon[18] hurtownię danych definiuje następująco:

> *Hurtownia danych jest zorientowaną tematycznie (ang. subject-oriented), zintegrowaną (ang. integrated), nieulotną (ang. nonvolatille), prezentującą wymiar czasowy (ang. time-variant) kolekcją danych wspierającą proces decyzyjny.*

System hurtowni danych poprzez swoją orientację tematyczną gromadzi dane selektywnie, koncentrując się wyłącznie na ściśle zdefiniowanym celu, obszarze działania przedsiębiorstwa. Hurtownia integruje dane, w ramach wskazanego zagadnienia, z różnych systemów źródłowych, przeprowadzając np. w ramach procesu ETL (ang. *Extraction, Transformation, Loading* - proces opisany w dalszej części Raportu) uzgodnienie formatów i zakresów przechowywanych danych. W odróżnieniu od systemów baz danych, gdzie informacje operacyjne ulegają ciągłym zmianom, hurtownia danych po wprowadzeniu informacji udostępnia wyłącznie możliwość jej odczytu bez możliwości usunięcia, czy aktualizacji, zaopatrując równocześnie każdą paczkę w tzw. stempel czasowy.

Podejście Inmona podczas procesu projektowania hurtowni danych zakłada odniesienie do wszystkich obszarów gromadzenia i przetwarzania danych w przedsiębiorstwie, tzw. ujęcie *top-down*. Hurtownia danych stanowi część szeroko rozumianej korporacyjnej fabryki informacji (ang. *Corporate Information Factory*) w skład której wchodzą również dane operacyjne.

Definicja Kimballa[19] odwołuje się do funkcji, szczególnie funkcji analitycznych, jaką hurtownia danych winna spełniać w organizacji, wstępnie abstrahując od charakterystyki gromadzonych danych:

---

[18] William H. Inmon: Building the Data Warehouse, Wiley Publishing, Indianapolis, 4th edition, 2005

[19] Ralph Kimball, Margy Ross: The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling, A John Wiley and Sons, Indianapolis, 3rd edition, 2013

Hurtownia danych jest, strukturalnie przystosowaną do wykonywania efektywnych zapytań i przeprowadzania analiz, kopią danych transakcyjnych.

Ralph Kimball, ponad kwestie budowy hurtowni oraz sposobu pozyskiwania danych, stawia dostarczane, poprzez przyjazny interfejs, użytkownikowi funkcje. Zwraca także uwagę na czas uzyskiwania od systemu odpowiedzi na zadane pytanie. Efektywność pracy środowiska jest nadrzędnym celem, które przyświeca projektantom hurtowni danych wg zasad Kimballa.

W podejściu Kimballa projektowanie najczęściej rozpoczyna się od pojedynczej, zorientowanej działowo hurtowni tematycznej (ang. *data mart*) – jest to spojrzenie na organizację typu *bottom-up*.

Nie ma jednoznacznej odpowiedzi, które z ujęć jest lepsze czy Inmona, czy Kimballa, zależeć to winno od indywidualnie przyjmowanych warunków kryterialnych. Podczas procesu wyboru metodyki dobrze jest zwrócić uwagę na charakterystyczne jej cechy, zebrane w poniższej tabeli[20],[21].

[20] Mary Breslin: Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models, Business Intelligence Journal, s. 6-20, Winter 2004

[21] Dariusz Dymek, Wojciech Komnata, Leszek Kotulski, Piotr Szwed: Architektury Hurtowni Danych…, op. cit.

	Bill Inmon	Ralph Kimball
Metodologia i architektura		
Ogólne podejście do metodologii	Top-down	Bottom-up
Architektura	Korporacyjna hurtownia danych obejmująca bazy departamentowe	Hurtownie tematyczne pojedynczych procesów biznesowych
Złożoność metody	Skomplikowana	Prosta
Odniesienie się do znanych metodologii	Metodyka spiralna	Proces czterech kroków, wywodzący się z metod RDBMS
Projekt fizyczny	Dokładny	Ogólny
Modelowanie danych		
Orientacja danych	Zorientowanie na dane lub na temat	Zorientowanie na proces
Narzędzia	Tradycyjne, związane z DBMS: ERD, DIS	Modelowanie wymiarowe, wywodzące się z modelowania relacyjnego
Zaangażowanie użytkowników	Niskie	Wysokie
Filozofia podejścia		
Główni uczestnicy	Profesjonaliści IT	Końcowi użytkownicy
Miejsce w organizacji	Integralna część korporacyjnej fabryki informacji (CIF)	Transformacja i przejęcie danych operacyjnych
Cel	Dostarczyć należyte rozwiązanie techniczne oparte na sprawdzonych metodach i technologiach baz danych	Dostarczyć rozwiązania, które ułatwiają użytkownikom końcowym bezpośrednie wyszukiwanie danych z zachowaniem rozsądnego czasu reakcji

Kończąc rozważania na temat dwóch podejść warto przeanalizować cechy sprzyjające wyborowi rozwiązania Inmona, czy Kimballa, zebrane poniżej [22],[23]:

Cecha	Sprzyja podejściu Inmona	Sprzyja podejściu Kimballa
Specyfika wymagań systemów DSS	Wymagania strategiczne	Wymagania taktyczne
Integracja danych	Integracja korporacyjna	Integracja na poziomie pojedynczych działów
Struktura danych	Dane niemetryczne i dane które będą wykorzystywane do zaspokojenie wielu różnych potrzeb	Metryki biznesowe, ocena wydajności oraz pojedynczego departamentu
Trwałość danych	Przy częstych zmianach wynikających z danych źródłowych	Gdy system źródłowy jest relatywnie stabilny
Wymagany personel oraz jego kompetencje	Liczny zespół specjalistów	Niewielki zespół o podstawowej wiedzy
Czas dostarczenia rozwiązania	Gdy wymagania organizacji pozwalają na dłuższy okres fazy początkowej projektu	Przy pilnej potrzebie uruchomienia pierwszej hurtowni danych
Koszt wdrożenia	Wysoki koszt początkowy, z niskimi kosztami dalszych prac nad kolejnymi fragmentami korporacyjnej hurtowni danych	Relatywnie niski koszt początkowy, przy czym każdy kolejny etap będzie miał podobny koszt (budowa kolejnej hurtowni tematycznej)

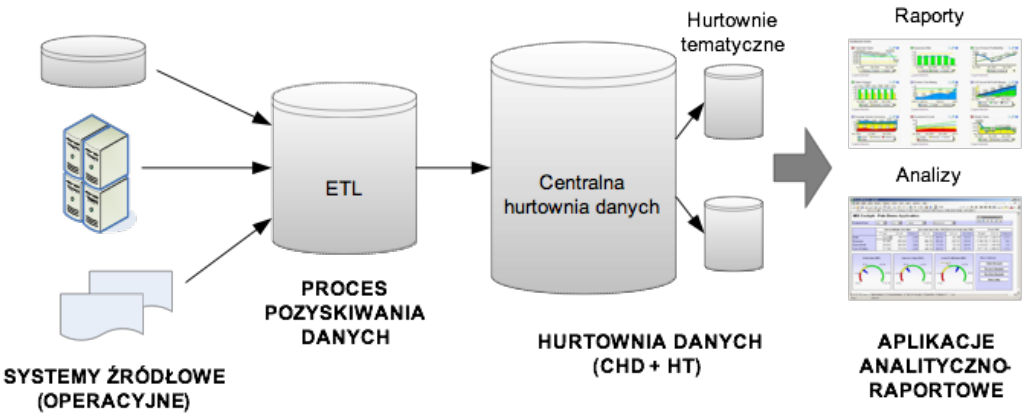
Podstawowa architektura hurtowni danych

[22] Mary Breslin: Data Warehousing Battle ..., op. cit.

[23] Dariusz Dymek, Wojciech Komnata, Leszek Kotulski, Piotr Szwed: Architektury Hurtowni Danych..., op. cit.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Typowa, podstawowa architektura hurtowni danych, niezależnie na razie od ujęcia Inmona i Kimballa, przyjmijmy, że składa się z czterech części.



Rysunek 1. Podstawowa architektura hurtowni danych.

Pierwszym obszarem są tzw. systemy źródłowe, wskazane wewnętrzne i zewnętrzne dane operacyjne, które będą zasilały hurtownię danych. Typowymi wewnętrznymi aplikacjami są: systemy klasy CRM, ERP, MRP, systemy czasu rzeczywistego SCADA, faktury, rozliczenia, wierzytelności, itp. Do typowych danych zewnętrznych można zaliczyć publiczne rejestry administracyjne np. TERYT, jak również systemy GIS, dane z GUS, informacje pogodowe, czy też portale społecznościowe. Praktycznie, nie ma ograniczenia na źródła danych operacyjnych, mogą to być zarówno bazy danych, jak i arkusze kalkulacyjne, czy też pliki tekstowe. Podawane dane winny być w postaci atomowej (podstawowej), aczkolwiek nie ma żadnego ograniczenia, aby wejściowymi były także informacje zagregowane, wstępnie przetworzone, przeanalizowane z wykorzystaniem specjalizowanych systemów informatycznych.

Kolejnym etapem jest proces pozyskiwania danych, określany procesem ETL (ang. *Extraction, Transformation, Loading*). Ów proces odpowiada za pozyskanie danych z systemów źródłowych (ekstrakcja), następnie ich przetworzeniu (transformacja) do wspólnego formatu, a na końcu za załadowanie (ładowanie) do hurtowni danych. Na uwagę zasługuje część odpowiedzialna za transformację, gdzie w przypadku np. różnej długości pola NAZWISKO, zostaje przyjęte jedno wspólne, niezależnie jakie występowało w systemie źródłowym. W praktyce oznacza to, że każdy rekord danych wejściowych poddawany jest procesowi analizy, połączonemu z ewentualną zmianą struktury.

Załadowane dane przez proces ETL trafiają do centralnej hurtowni danych, z której, zgodnie z definicją Inmona, już nie zostaną nigdy systemowo usunięte. Centralna hurtownia danych przechowuje informacje podstawowe, odpowiadające, pomimo, ich ewentualnej transformacji, danym z systemów źródłowych. Niestety często dostęp do informacji atomowych, szczególnie, że wielu użytkowników oczekuje informacji wstępnie przetworzonej powoduje znaczne obciążenie zasobów. Stąd często centralne repozytoria uzupełniane są tzw. hurtowniami tematycznymi (ang. *data mart*). Hurtownie tematyczne najczęściej skorelowane są z określonym departamentem lub procesem biznesowym, zawierając w sobie zagregowane, częściowo przetworzone informacje podstawowe, np. w postaci podsumowań dziennych, tygodniowych, czy miesięcznych, uśrednione wartości zużycia energii elektrycznej w poszczególnych PPE, czy stacjach transformatorowych, lub inne często wykorzystywane agregaty przyspieszające odpowiedź systemu na zapytania użytkowników.

Różnice pomiędzy danymi źródłowymi (klasy OLTP = *OnLine Transaction Processing*) a przechowywanymi w hurtowni danych dobrze ilustruje poniższa tabela[24]:

[24] Thomas Connolly, Carolyn Begg: Database Systems, A Practical Approach to Design, Implementation, and Management, Addison Wesley, imprint Person Education Ltd, London, 4th edition, 2005

System OLTP	System hurtowni danych
Przechowuje aktualne dane	Przechowuje dane historyczne
Gromadzi dane detaliczne	Gromadzi dane detaliczne, ale również dane będące zestawieniami, zarówno niskiego poziomu, jak i wysokiego
Gromadzona informacja jest dynamiczna, jest zmienna	Gromadzona informacja jest głównie statyczna
Powtarzające się przetwarzanie procesowe	Procesy typu ad-hoc, niestrukturalne i heurystyczne
Wysoka wydajność przetwarzania transakcji	Średnia do niskiej wydajność przetwarzania transakcji
Przewidywalny model wykorzystania	Nieprzewidywalny model wykorzystania
System nakierowany na obsługę transakcje	System nakierowany na obsługę analiz
System zorientowany aplikacyjnie	System zorientowany tematycznie
System wspiera codzienne, operacyjne decyzje	System wspiera decyzje strategiczne
Dostępny dla bardzo dużej liczby użytkowników typu urzędniczego i operacyjnego	System dostępny dla relatywnie niewielkiej liczby użytkowników szczebla menadżerskiego

Ostatnia część obejmuje środowisko analityczno-raportowe na które składa się szerokorozumiany system analityki biznesowej (ang. *business intelligence*) wsparty pakietami eksploracji danych (ang. *data mining*). Praktycznie 95% użytkowników hurtowni danych korzysta z modułu raportowego, tablic informacyjnych (ang. *dashboard*), alertów systemowych oraz eksploracji danych.

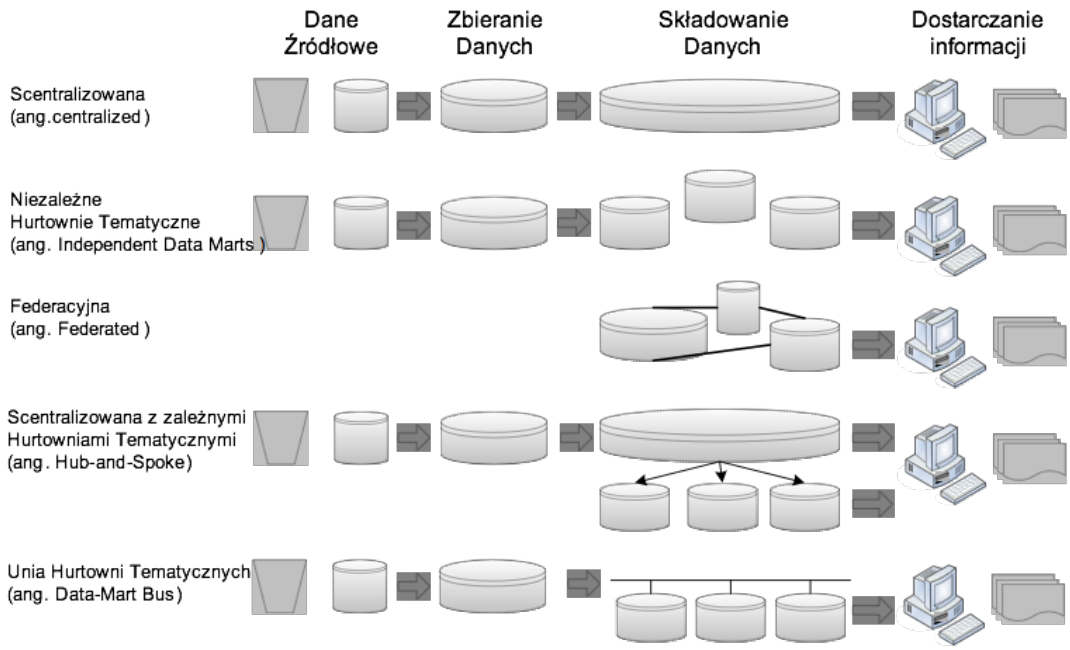
Klasyfikacja architektur hurtowni danych

Generalnie w literaturze występują dwie klasyfikacje architektur hurtowni danych. Pierwsza dzieli hurtownię danych na warstwy, zwykle w modelu 1warstwowym, 2warstowym i 3warstwowym. Druga klasyfikacja odnosi się do organizacji

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

udostępniania danych zgromadzonych w hurtowni dla całego przedsiębiorstwa lub pojedynczych działów, w praktyce do roli poszczególnych elementów, a przede wszystkim centralnej hurtowni danych i hurtowni tematycznych[25].

Niniejszy raport koncentruje się na drugim z kryteriów, czyli na roli jaką pełni hurtownia danych. Kryterium to pozwoliło zdefiniować pięć głównych typów architektury hurtowni danych[26],[27]:



Rysunek 2. Pięć architektur hurtowni danych (wg H.J.Watson i T.Ariyachandra).

1) hurtownia scentralizowana – architektura zakładająca jedno wspólne repozytorium danych, zbudowane zgodnie z zasadami baz relacyjnych (3NF –

[25] Dariusz Dymek, Wojciech Komnata, Leszek Kotulski, Piotr Szwed: Architektury Hurtowni Danych…, op. cit.

[26] Hugh J. Watson, Thilini Ariyachandra: Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures. raport http://www.terry.uga.edu/~hwatson/ DW_Architecture_Report.pdf, July 2005. (data odczytu 1.06.2014)

[27] Thomas Connolly, Carolyn Begg: Database Systems, A Practical Approach to Design…, op. cit.

2) trzecia postać normalna), zapytania za każdy razem przeszukują całą hurtownię, jest to kosztowne czasowo rozwiązanie,

3) hurtownia z niezależnymi hurtowniami tematycznymi – architektura zakładająca wspólny proces pozyskiwania danych, jednakże bez żadnej korelacji, zależności pomiędzy hurtowniami tematycznymi, rozwiązanie dedykowane dla organizacji w której poszczególne departamenty lub procesy biznesowe nie przenikają się,

4) hurtownia federacyjna – podejście typowe dla połączenia systemów informatycznych rozłącznych organizacji, które nie dopuszczają budowy jednego wspólnego repozytorium, podejście wykorzystywane przy połączeniach korporacji, zachowujące specyfikę indywidualnych rozwiązań, jak również możliwe przy pewnych specjalizowanych zastosowaniach[28],

5) hurtownia scentralizowana z zależnymi hurtowniami tematycznymi – architektura uniwersalna, gwarantująca 3NF (ang. *Third Normal Form*), „jedną wersję prawdy" oraz dzięki zależnym hurtowniom tematycznym optymalizując efektywność zapytań użytkowników, rozwiązanie typowe dla podejścia zgodnego z metodą Inmona,

6) unia hurtowni tematycznych – podejście charakterystyczne dla projektów zgodnych z metoda Kimballa, na początku definiowana jest główna hurtownia tematyczna (często nazywana *supermart*) w której będą występowały wspólne dla pozostałych hurtowni tematycznych biznesowe wymiary i metryki, nazywane wzorcowymi.

Tak jak nie ma jednej uniwersalnej metody projektowania hurtowni danych, tak nie ma doskonałej architektury, którą w każdej sytuacji można proponować. Zarówno przed etapem wyboru metodyki pracy, jak i architektury należy przyjąć i przeanalizować warunki kryterialne i dopiero w oparciu o nie dokonać wyboru.

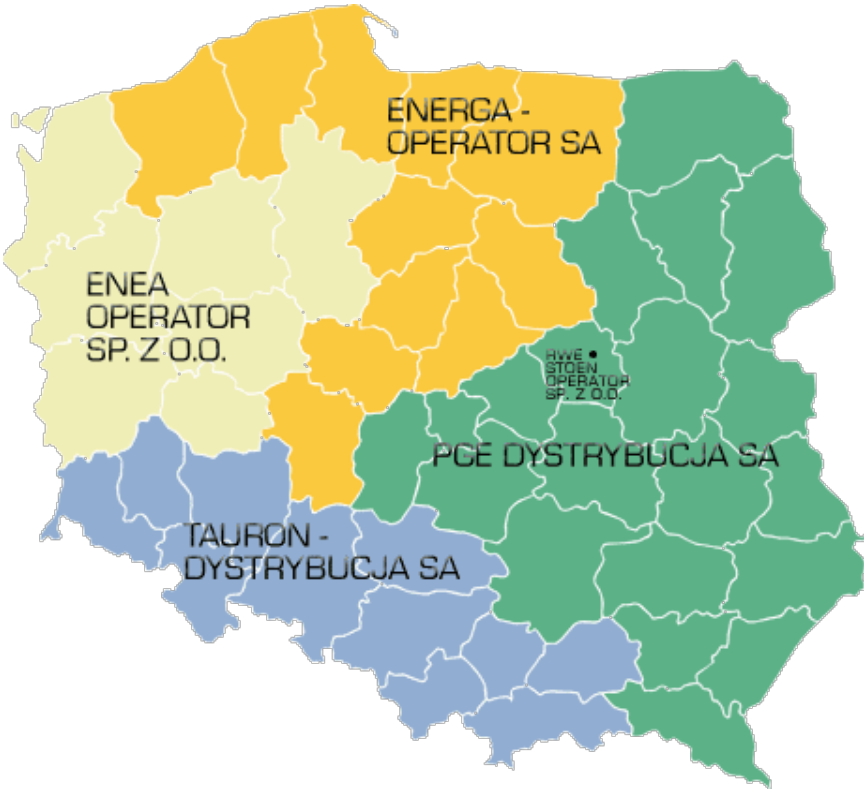
[28] Dariusz Dymek, Wojciech Komnata, Leszek Kotulski: Federacyjna hurtownia danych w dostępie do informacji poufnej. Roczniki Kolegium Analiz Ekonomicznych, zeszyt 33/2014, strony 135-154, Oficyna Wydawnicza SGH, Warszawa, 2014.

Analiza potrzeb Operatorów Systemów Dystrybucyjnych

Opierając się na informacjach na temat pracujących w podmiotach przemysłu energetycznego systemach ITC można przyjąć, iż hurtownie danych są coraz częściej stosowane. Stosowane, co nie oznacza, iż w pełni wykorzystywane w procesie podejmowania decyzji strategicznych. Działania mające na celu ułatwienia w procesie decyzyjnym są wdrażane i powinny wkrótce zacząć przynosić pierwsze efekty.

Istotnym jest jednakże, aby wyraźnie zaznaczyć dwa obszary w których owe rozwiązania znajdują lub mogą znaleźć zastosowanie. Jednym z nich jest obszar biznesowy, drugi eksploatacyjny. Tak jak można przyjąć, iż hurtownie danych zaczynają pracować na rzecz decyzji biznesowych, tak jeszcze nie są stosowane w procesach wspierających analizy eksploatacyjne. Biorąc jednakże pod uwagę zmianę wymagań jakie stosuje regulator[29] ten obszar stanie się krytyczny, gdyż to on zacznie decydować o sukcesie ekonomicznym podmiotów zajmujących się dystrybucją energii elektrycznej w Polsce.

[29] Urząd Regulacji Energetyki - podstawowe zadania: (1) kontrola obowiązku zakupu energii wytworzonej w źródłach odnawialnych oraz udział uczestników rynku w kosztach jej pozyskania, współudział w konstruowaniu i propagowaniu rynku energii elektrycznej pochodzącej z kogeneracji, (2) restrukturyzacja i modernizacja przedsiębiorstw energetycznych, (3) działania przyczyniające się do zmniejszania strat energii, zwłaszcza energii cieplnej. Wikipedia. Data odczytu 16 grudnia 2014 roku.



Rysunek 3. Rynek operatorów systemów dystrybucyjnych (źródło URE).

W naszym kraju obecnych jest na rynku pięciu niezależnych operatorów systemów dystrybucji, są to:

1. TAURON Dystrybucja SA
   Siedziba: ul. Jasnogórska 11, 31-358 Kraków
   Oddziały: Będzin, Bielsko Biała, Częstochowa, Gliwice, Jelenia Góra, Kraków, Legnica, Opole, Tarnów, Wałbrzych, Wrocław
   Liczba odbiorców (tys) 5 300
   Obszar działania (km2) 57 940

2. PGE Dystrybucja SA
   Siedziba: ul. Garbarska 21A, 20-340 Lublin
   Oddziały: Łódź-Teren, Łódź - Miasto, Lublin, Rzeszów, Skarżysko-Kamienna, Zamość, Białystok, Warszawa

3. Enea Operator Sp. z o.o .

Siedziba: ul. Strzeszyńska 58, 60-479 Poznań
Oddziały: Zielona Góra, Gorzów Wielkopolski, Szczecin, Bydgoszcz
Liczba odbiorców (tys) 2 204,74
Obszar działania (km2) 58 192
Długość linii (km) 105 480

4. Energa-Operator SA
   Siedziba: ul. Marynarki Polskiej 130, 80-557 Gdańsk
   Oddziały: Koszalin, Słupsk, Elbląg, Olsztyn, Toruń, Płock, Kalisz
   Obszar działania (km2) 75 000

5. RWE Stoen Operator Sp. z o.o.
   Siedziba: Ul. Rudzka 18, Warszawa
   Liczba odbiorców (tys) 900

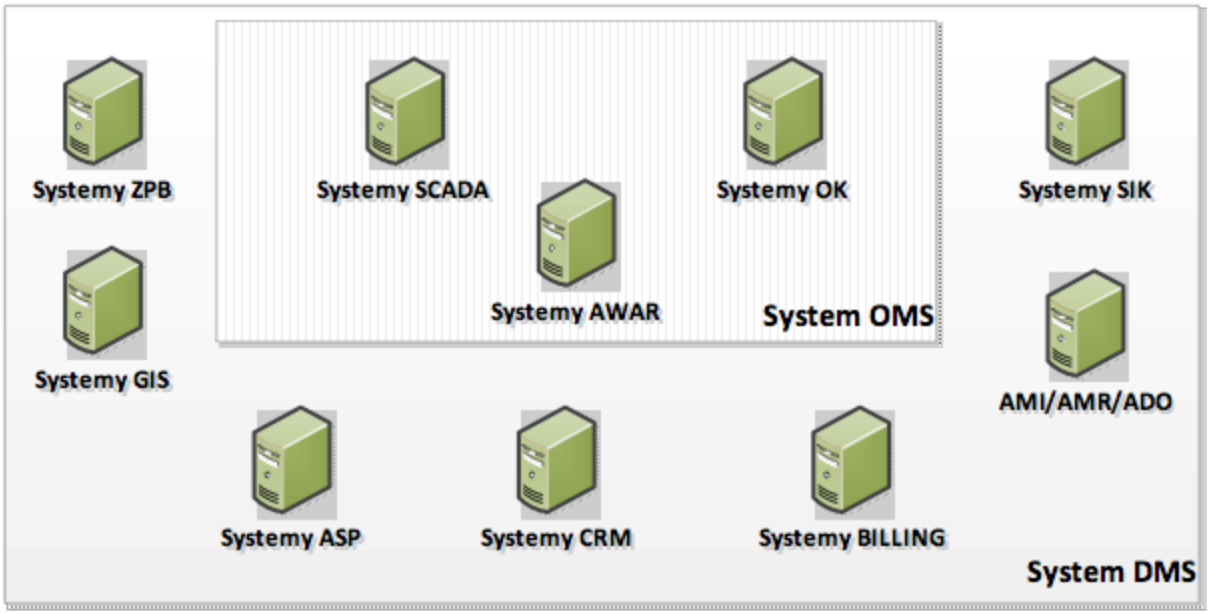
Opis środowiska ITC

OSD od kilku lat wdraża technologię hurtowni danych, preferując tzw. podejście iteracyjne, spiralne. Metoda składa się z czterech etapów z których na początku określany jest cel, następnie przeprowadzana jest szczegółowa analiza, po której wykonywane jest wdrożenie, etap kończy ocena i dalsze planowanie. Obecny system określany jest mianem OMS (*Outage Management Systems*), przy założeniu docelowym doprowadzenia do rozwiązania ADMS (*Advanced Distribution Management Systems*).

Na środowisko ITC w OSD na przykładzie TAURON Dystrybucja S.A. składają się następujące systemy:

- ZPB - zarządzanie pracą brygad,
- GIS - struktura sieci oraz system do zarządzania majątkiem sieciowym,
- ASP (*Asset Strategic Planning*) - zarządzanie inwestycjami,
- AWAR - system zarządzający awariami,
- SCADA - system kontroli i akwizycji danych, nadzór nad ruchem w sieci, przepływem energii elektrycznej,
- OK - system obsługi klienta,

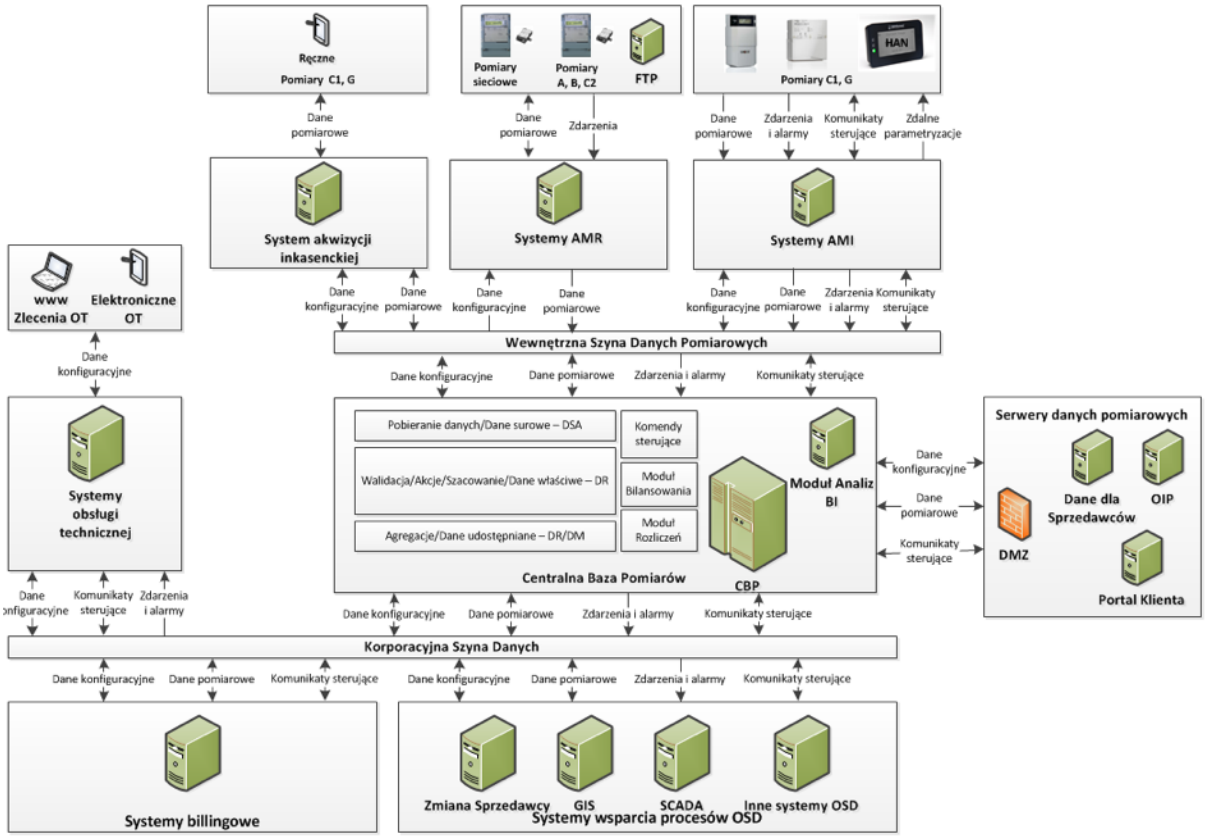
- SIK - system informowania kierownictwa
- AMI - zaawansowane systemy pomiarowe, systemy ADO (system akwizycji i udostępnia danych pomiarowych odbiorców "biznesowych"), AMR (system automatycznego odczytu) oraz MOBOT (mobilne linie pomiarowe),
- BILLING - system rozliczania płatności/ fakturowania,
- CRM - zarządzanie relacjami z klientami instytucjonalnymi i indywidualnymi.



Rysunek 4. System DMS i OMS (na przykładzie TAURON Dystrybucja SA).

Obecnie w TAURON Dystrybucja SA rozwiązania hurtowni danych stosowane są w systemie SCADA oraz w controllingu. Na system OMS składają się rozwiązania SCADA, AWAR oraz OK. Pozostałe wymienione na powyższym rysunku systemu współtworzą rozwiązanie klasy DMS.

Równocześnie wdrażany jest system wspierający projekt pozyskiwania, gromadzenia, przetwarzania i udostępniania danych pomiarowych.



Rysunek 5. System gromadzenia pomiarów (na przykładzie TAURON Dystrybucja SA).

Rozwiązanie zakłada pozyskiwanie danych pomiarowych z trzech źródeł: pomiarów ręcznych, automatycznego pobierania danych (AMR) oraz zaawansowanych systemów pomiarowych (AMI). Pozyskane dane wewnętrzną szyną danych pomiarowych zasilają centralną bazę pomiarów (CBP).

Zmiana wymagań URE

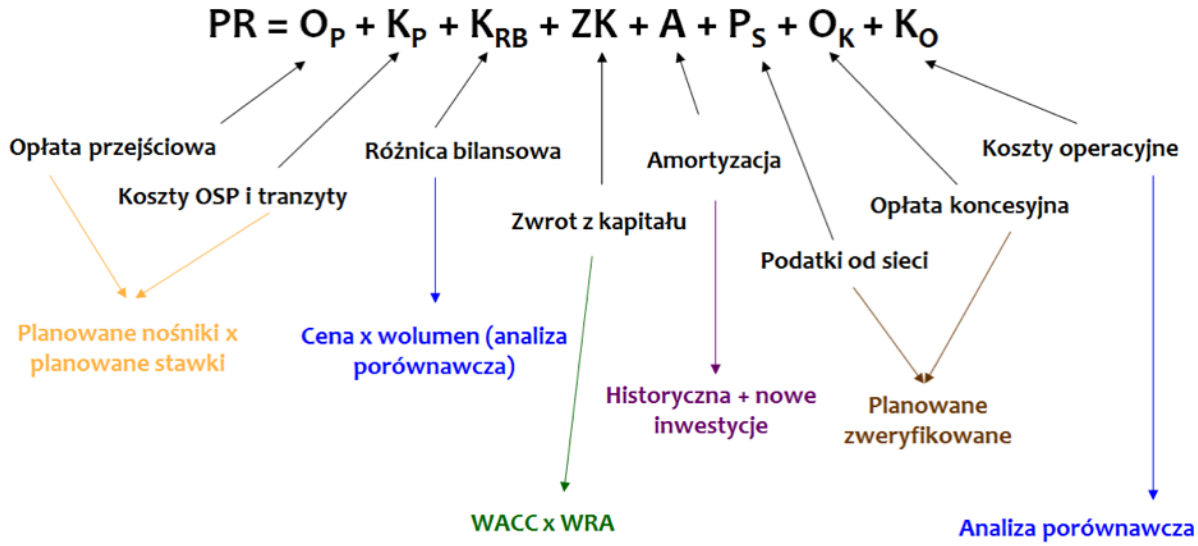
Zgodnie z zapowiedziami Prezesa URE od 01.01.2016 roku nastąpi zmiana dotychczasowego modelu regulacji OSD. Obecnie stosowany przez URE model regulacji, przy wyznaczaniu przychodu regulowanego OSD uwzględnia m.in.:

- koszty ponoszone przez OSD na rzecz OSP (Operatora Sieci Przesyłowej – firmę PSE S.A.),

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

- koszty różnicy bilansowej,
- wartość przyjętego zwrotu z kapitału,
- podatki i opłaty koncesyjne,
- koszty operacyjne.

Aktualnie obowiązujący model regulacji wyrażony jest wzorem[30]:



Gdzie:

- PR – przychód regulowany,
- wolumeny – nośniki do kalkulacji stawek opłat dystrybucyjnych:
  - dostawa energii,
  - moc umowna,
  - liczba odbiorców.

Model ten będzie funkcjonował jedynie do końca 2015 roku. W opinii ekspertów regulacja ta w niewystarczający sposób wspiera efektywną modernizację sieci elektroenergetycznej.

[30] Piotr Ordyna, prezentacja pt. „Regulacja jakościowa z perspektywy Operatora Systemu Dystrybucyjnego" wygłoszona 23 października 2014 roku w Krakowie.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Nowy model regulacji ma uwzględniać zarówno ocenę efektywności OSD w zakresie kosztów objętych obecnym modelem jak i wprowadzenia elementów regulacji jakościowej. Ma on spowodować, że OSD będą promować inwestycje w:

- energetykę prosumencką,
- niezawodność sieci,
- aktywizację strony popytowej,
- efektywność energetyczną.

Regulacja jakościowa ma doprowadzić do nagradzania OSD podwyższających jakość świadczonych usług, poprzez zwiększenie ich przychodu regulowanego. Ma także wprowadzić elementy kary finansowej za brak realizacji wyznaczonych celów jakościowych.

URE na podstawie historycznego wykonania wskaźników jakościowych określi:

- szczegółowe definicje przyjętych wskaźników jakościowych,
- docelowy poziom wskaźników dla wyznaczonych ram czasowych,
- wartość wskaźników w poszczególnych latach z wykorzystaniem ścieżki dojścia,
- parametry modelu regulacji, czyli mechanizm szczegółowy wyznaczania przychodu OSD w zależności od osiągniętego za poprzedni rok poziomu danego wskaźnika.

Nowy model regulacji przychodu OSD będzie zapewne dotyczył następujących „parametrów decyzyjnych":

- wskaźniki SAIDI oraz SAIFI - z obszaru niezawodności dostaw energii elektrycznej,
- wskaźnik czasu przyłączania klientów do sieci - z obszaru jakości obsługi klienta,
- wskaźnik czasu udostępnienia danych pomiarowych - z obszaru jakości obsługi sprzedawców energii.

Dokładną definicję wskaźników SAIDI, SAIFI oraz MAIFI przedstawił Prezes URE w Informacji Nr 16/2012 z dnia 21 czerwca 2012 roku[31], [32]:

- SAIDI (ang. *System Average Interruption Duration Index*) – miara średniego czasu wyłączenia odbiorcy w roku; wskaźnik wyrażony w minutach na odbiorcę na rok jest współczynnikiem niezawodności będącym sumą iloczynów czasu trwania przerwy długiej i bardzo długiej i liczby odbiorców narażonych na skutki danej przerwy w ciągu roku, podzielonej przez łączną liczbę obsługiwanych odbiorców przyłączonych do sieci,
- SAIFI (ang. *System Average Interruption Frequency Index*) – miara częstości wyłączeń odbiorcy w roku, wyznaczana jest jako iloraz liczby odbiorców narażonych na skutki wszystkich przerw długich i bardzo długich w ciągu roku do łącznej liczby obsługiwanych odbiorców,
- MAIFI (ang. *Momentary Average Interruption Frequency Index*) – wskaźnik monitroujący krótkotrwałe wyłączenia, jest ilorazem liczby osób narażonych na skutki wszystkich krótkich przerw zachodzących w ciągu roku do łącznej liczby obsługiwanych odbiorców.

Prezes URE zapowiedział, że poszczególne OSD mogą mieć do osiągnięcia różne cele jakościowe. Będzie to wynikało z analizy danych historycznych poszczególnych wskaźników (benchmarków), gdyż OSD mają różne struktury sieci elektroenergetycznej. Wynika to z rodzaju obsługiwanych obszarów czyli np. proporcji terenów miejskich oraz wiejskich.

Ogólny model działania regulacji jakościowej zaprezentowany został na schemacie[33], [34]:

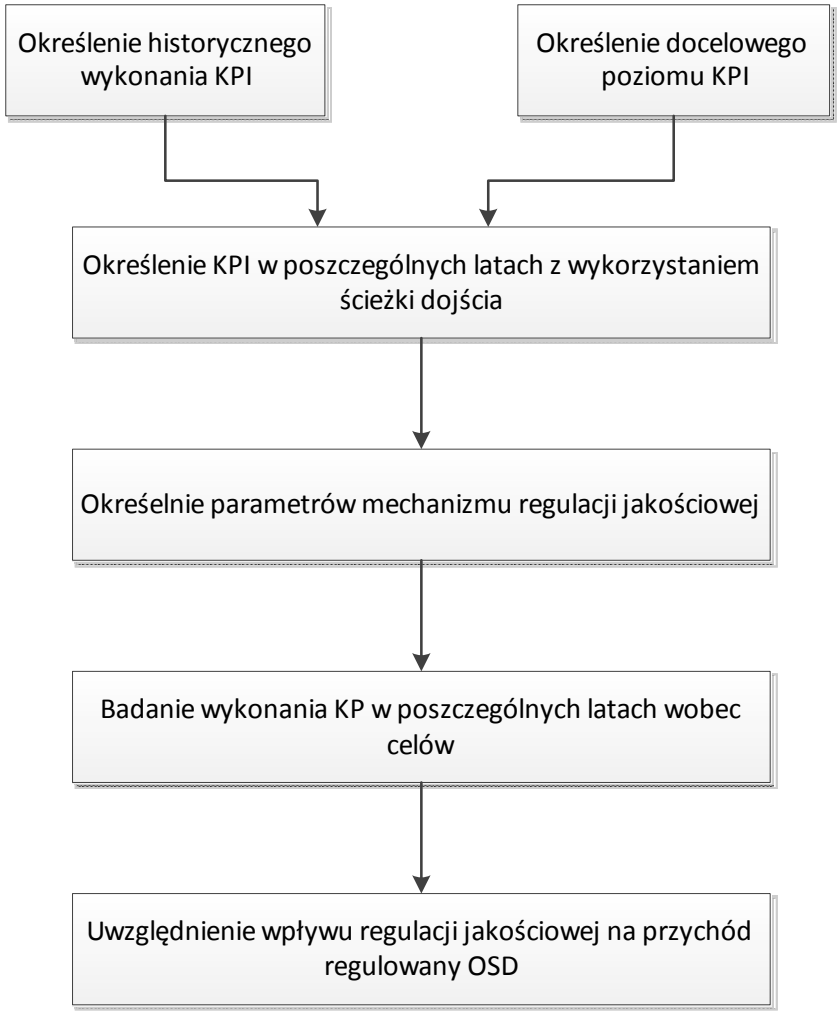
---

[31] Informacja Prezesa URE Nr 16/2012 w sprawie obliczania przez OSD wskaźników SAIDI, SAIFi i MAIFI, o których mowa w rozporządzeniu Ministra Gospodarki z dnia 4 maja 2007r., w sprawie szczegółowych warunków funkcjonowania systemu elektroenergetycznego. Warszawa, 21 czerwca 2012.

[32] Parol Mirosław: Analiza wskaźników dotyczących przerw w dostarczaniu energii elektrycznej na poziomie sieci dystrybucyjnej. Przegląd Elektrotechniczny, R. 90 Nr 8/2014, strony 122-126, 2014.

[33] Piotr Ordyna, prezentacja pt. "Regulacja jakościowa z perspektywy Operatora Systemu …", op. cit.

[34] KPI – Kluczowe wskaźniki efektywności (ang. *Key Performance Indicators*) – finansowe i niefinansowe wskaźniki stosowane jako mierniki w procesach pomiaru stopnia realizacji przyjętych w organizacji celów. Wikipedia. Data odczyt 10 lutego 2015.



Oprócz powyżej wymienionych „parametrów decyzyjnych" regulator wprowadzi „parametry obserwowane" z zakresu:

- niezawodności dostaw energii elektrycznej
  - analiza wskaźników SAIDI oraz SAIFI,
  - monitorowanie obszarów o najniższej jakości dostaw,
- jakości obsługi klienta :
  - mierzenie czasu realizacji przyłączenia (CRP) klienta,
  - satysfakcja klienta z obsługi,

- czas udostępniania danych pomiarowym sprzedawcom,

- racjonalizacji zużycia energii elektrycznej, monitorowanie:
  - edukacji odbiorców oraz
  - aktywizacji odbiorców.

Parametry te nie będą miały wpływu na wielkość przychodów OSD w początkowym okresie regulacji jakościowej. Ale ich rejestrowanie pozwoli w późniejszym czasie na ich włączenie w model regulacji jakościowej.

Wprowadzenie modelu regulacji jakościowej zwiększy przejrzystość procesu regulacji oraz ma doprowadzić do:

- poprawy jakości usług poprzez zwiększenie orientacji OSD na klientów,
- poprawy wizerunku branży poprzez wzrost satysfakcji klientów,
- optymalizacji przychodu regulowanego OSD,
- wzrostu efektywności OSD.

Oczekiwania wobec systemów ITC

Można przyjąć, iż dla operatora systemów dystrybucyjnych najważniejszymi systemami, systemami krytycznymi, są:

- ZPB – zarządzanie pracą brygad,
- SCADA - system kontroli i akwizycji danych, nadzór nad ruchem w sieci, przepływem energii elektrycznej,
- AWAR - system zarządzający awariami,
- ASP (*Asset Strategic Planning*) - zarządzanie inwestycjami.

Równocześnie z powyższymi systemami analizowane są straty energii elektrycznej wynikające z przyczyn technicznych (dane z billingu, systemów AMI) oraz handlowych (np. NPE – nielegalny pobór energii).

Dotychczas nie została podjęta decyzja o ogólnopolskim uruchomieniu systemów Smart Metering dla klientów indywidualnych. Obowiązuje jednakże wymóg, aby do końca 2016 roku większość stacji transformatorowych (w TAURON Dystrybucja SA ponad 55.000 transformatorów) została opomiarowana (częstość pomiarów 15 minut).

Szacunkowe potrzeby, przy założeniu ok. 100B na pojedynczy pomiar, zawierający informacje o energii, liczniku, punkcie poboru energii i kliencie, w okresie 1 roku to ok. 3,4 MB.

Jednym z najważniejszych obszarów jest dbałość o jakość energii elektrycznej, stąd krytycznym staje się system zarządzający ową jakością na który składa się:

- SZMS – system zarządzania majątkiem sieciowym gromadzący różne dane na temat pojedynczego, konkretnego punktu sieci,
- informacje nt. wskaźników SAIDI dla danego obszaru sieciowego,
- ocena stanu technicznego elementu infrastruktury sieciowej,
- wykonane i planowane prace eksploatacyjne,
- dane z systemu nt. planowanych i nieplanowanych wyłączeń,
- informacje nt. zewnętrznych czynników mających bezpośredni lub pośredni wpływ na jakość energii elektrycznej, np. warunki pogodowe.

Operator systemu dystrybucyjnego zwraca szczególną uwagę na pomiar i optymalizację wskaźników SAIDI/SAIFI. Gromadzone dane obejmują:

- nakłady inwestycyjne,
- nakłady eksploatacyjne,
- informacje nt. infrastruktury sieciowej,
- dane nt. planowanych i nieplanowanych wyłączeń,
- informacje z systemu SCADA,
- procedury przyłączeń nowych punktów poboru energii,
- stosowanej technologii (w danym trakcie sieciowym),
- wykorzystywanych agregatów.

Kryteria wyboru

Rozpoczęcie procesu budowy hurtowni danych winno zostać poprzedzone dogłębną analizą procesów biznesowych, które mają zostać wsparte projektowanym rozwiązaniem. Zgodnie z definicją hurtowni danych informacje winny być zorientowane tematycznie, co w praktyce oznacza selekcję tych informacji, które są lub będą wymagane podczas realizacji procesu decyzyjnego. A priori należy przyjąć, że w każdej organizacji, także u OSD, nie ma pojedynczego wymagania, na różnym poziomie występuje wiele potrzeb i wymagań, których zaspokojenie zwiększa prawdopodobieństwo sukcesu wdrożenia hurtowni danych[35]

Wymagania, a tym samym kryteria winny uwzględniać naturalny podział typu „od ogółu do szczegółu", w którym zaczynając od danych strategicznych (wizja i ogólne cele OSD), poprzez identyfikację obszarów biznesowych pozwalających wykonać wgląd w szeroki obszar zagadnień wreszcie do analizy biznesowej skoncentrowanej na analizach i danych biznesowych. Na koniec zwykle okazuje się, że ogrom prac jest bardzo duży i niezbędna jest priorytetyzacji, po której dopiero następuje faktyczne projektowanie hurtowni danych.

Obok powyższych wymagań biznesowych niezbędne jest uwzględnienie wymagań technologicznych. W tym zakresie zwykle określa się oczekiwania względem systemów RDBMS w oparciu o które pracują współczesne hurtownie danych. Poniższa tabela prezentuje przykładowe wymagania technologiczne[36]:

[35] Dariusz Dymek, Wojciech Komnata, Leszek Kotulski, Piotr Szwed: Architektury Hurtowni Danych…, op. cit.

[36] Ibidem

Wymaganie	Opis
Wydajność ładowania (ang. *load performance*)	Hurtownia danych wymaga inkrementalnego ładowania nowych zestawów danych w określonych odcinkach czasu, z wąskim oknem czasowym. Wydajność procesu ładowania powinna być mierzona w setkach milionów wierszy lub gigabajtów na godzinę bez limitów ograniczających prowadzony biznes.
Proces ładowania (ang. *load processing*)	Podczas ładowania nowych lub dodawania kolejnych serii do danych istniejących wykonywanych jest wiele kroków związanych z operacjami na danych: konwersja, filtrowanie, zmiana formatu, weryfikacja integralności, indeksacja i fizyczne przechowania.
Zarządzanie jakością danych (ang. *data quality management*)	Hurtownia danych to zarządzanie faktami, a jeżeli tak, to jakość gromadzonych danych musi być wysoka. Dodatkowo warunkiem wymaganym dla systemu jest zachowanie spójności lokalnej i globalnej, bez względu na jakość danych źródłowych oraz bardzo duży rozmiar bazy. Zdolność do odpowiadania na zapytania użytkownika, coraz bardziej złożone i wyszukane, jest miarą sukcesu systemu.
Wydajność zapytań (ang. *query performance*)	Zarządzanie oparte na analizie faktów, jak również zapytania przygotowywane ad-hoc nie mogą być zbyt wolno obsługiwane lub spowalniane przez system RDBMS.
Terabajtowa skalowalność (ang. *terabyte scalability*)	Należy założyć szybki przyrost hurtowni do rozmiarów terabajtowych (1012 bajtów) i petabajtowych (1015 bajtów). RDBMS nie może mieć ograniczeń względem rozmiaru obsługiwanych danych i powinien dopuszczać modularne i równoległe zarządzanie.
Skalowalność użytkowników (ang. *mass user scalability*)	Obecnie wielu autorów zakłada, że grupa użytkowników hurtowni danych jest niewielka. Zakłada się, że grupa ich będzie rosła i system musi gwarantować poprawną obsługę setek, czy tysięcy równolegle pracujących użytkowników, przy zachowanej jakości obsługi zapytań.
Praca sieciowa (ang. *networked data warehouse*)	System hurtowni danych powinien mieć możliwość współpracy z dużymi sieciami hurtowni danych. System musi obejmować narzędzia wspierające przenoszenie podgrup danych pomiędzy hurtowniami.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Zintegrowana analiza wymiarowa (ang. *integrated dimensional analysis*)	Widoki wielowymiarowe, wsparcie dla modelu wymiarowego to nieodłączne elementy RDBMS dostarczającego najwyższej wydajności analizy relacyjne OLAP (ROLAP).
Zaawansowane funkcje zapytań (ang. *advanced query functionality*)	Klienci wymagają zaawansowanych mechanizmów kalkulacji, sekwencyjnych i porównawczych analiz, dostępu do elementarnych i zagregowanych danych. System powinien udostępniać zestaw zaawansowanych narzędzi wspierających operacje analityczne na danych.

Teoretyczny aspekt wymagań względem warunków kryterialnych kończy podejście prezentujące kryteria i metryki sukcesu integracji danych (na bazie integracji rejestrów)[37]:

[37] Wojciech Komnata, Dariusz Dymek: Integracja rejestrów publicznych na poziomie samorządu terytorialnego. Roczniki Kolegium Analiz Ekonomicznych, zeszyt 33/2014, strony 247-264, Oficyna Wydawnicza SGH, Warszawa, 2014.

Grupa 1 - jakość systemu	
• skalowalność	Stały przyrost danych, etapowe zwiększanie liczby zainteresowanych osób prywatnych podmiotów instytucjonalnych, jak również wzrost ilości zapytań oraz zwiększona ich złożoność nie może wpływać negatywnie na bieżące działanie systemu.
• elastyczność	Rozwiązanie musi uwzględniać pojawiające się nowe potrzeby, nowe procesy czy obszary stosowania dając możliwość łatwego (taniego, w akceptowalnym czasie) adaptowania systemu do potrzeb.
• zdolność do integracji	Zastosowane rozwiązanie winno dawać łatwą (tanią, w akceptowalnym czasie) możliwość dołączenia nowych systemów źródłowych.
• produktywność (wydajność)	Przyrost danych, uczestników ich wymiany, pojawienie się nowych potrzeb, procesów biznesowych, jak również dołączanie nowych systemów źródłowych nie może obniżyć wydajności pracy systemu
Grupa 2 – jakość informacji	
• dokładność	Poziom dokładności gromadzonych danych winien zostać określony podczas fazy projektowej i cały czas monitorowany. Dokładność ma odpowiadać zamierzonym celom do których rozwiązanie jest wykorzystywane.
• długookresowa trwałość	Środowisko źródłowych baz transakcyjnych zmienia się stale w czasie. Zastosowany proces ETL w przypadku HD musi gwarantować niezmienność zgromadzonych danych mimo wielokrotnie powtarzającego się procesu ich pozyskiwania.
• kompletność	Rozwiązanie winno dostarczać 100% wymaganych przez dany proces biznesowy informacji.
• konsystencja	Jednym z ważnych celów jest eliminacja niespójności danych, wynikającej z redundantnego ich przechowywania i aktualizacji (zmiany) w różnych systemach źródłowych. Zastosowane rozwiązanie wymaga opracowania jednej wersji „prawdy”.
Grupa 3 – jakość komunikacji	

• zgodność z SLA	Wdrażany system ma przyspieszyć, ujednoznacznić i ułatwić proces analizy danych, owa analiza musi się jednak odbywać w założonym, nieprzekraczalnym czasie.
• powtarzalność	Wielokrotne zapytania od tego samego lub każdego innego użytkownika, przy tych samych danych wejściowych, muszą gwarantować takie same odpowiedzi.
• eskalacja	Rozwiązanie informatyczne musi dawać analogiczną, do tradycyjnego systemu komunikacji, możliwość eskalacji ze zwróceniem uwagi na niedotrzymywanie parametrów SLA, czy błędne jego funkcjonowanie.
• dostępność	System będzie działał poprawnie, jeżeli wszyscy uczestnicy wymiany, szczególnie dostawcy danych źródłowych będą dostępni on-line.

Hurtownia danych z modułem analitycznym – wymagania

W ostatnim rozdziale opracowania przedstawiona została lista wymagań funkcjonalnych i technicznych względem środowiska bazy danych, systemu pozyskiwania danych do hurtowni, systemu kostek OLAP, eksploracji informacji oraz modułu business intelligence.

Baza danych:

Lista wymagań wobec systemu bazy danych:

1. Dostępność oprogramowania na współczesne 64-bitowe platformy Unix (HP-UX dla Itanium, Solaris dla procesorów SPARC/x86-64, IBM AIX), Intel Linux 64-bit, MS Windows 64-bit. Identyczna funkcjonalność serwera bazy danych na ww. platformach w zakresie przedstawionych wymagań.
2. Niezależność platformy systemowej dla oprogramowania klienckiego / serwera aplikacyjnego od platformy systemowej bazy danych.

3. Możliwość przeniesienia (migracji) struktur bazy danych i danych pomiędzy ww. platformami bez konieczności rekompilacji aplikacji bądź migracji środowiska aplikacyjnego.

4. Przetwarzanie transakcyjne wg reguł ACID (Atomicity, Consistency, Independency, Durability) z zachowaniem spójności i maksymalnego możliwego stopnia współbieżności. Mechanizm izolowania transakcji musi pozwalać na spójny odczyt modyfikowanego obszaru danych bez wprowadzania blokad, z kolei spójny odczyt nie może blokować prawa wykonywania zmian.

5. Wsparcie dla wielu ustawień narodowych i wielu zestawów znaków (włącznie z Unicode).

6. Możliwość migracji 8-bitowego zestawu znaków bazy danych (np MS Windows CP 1252, ISO 8859-2) do Unicode.

7. Skalowanie rozwiązań opartych o architekturę trójwarstwową: możliwość uruchomienia wielu sesji bazy danych przy wykorzystaniu jednego połączenia z serwera aplikacyjnego do serwera bazy danych.

8. Brak formalnych ograniczeń na liczbę tabel i indeksów w bazie danych oraz na ich rozmiar (liczbę wierszy), nie mniej aniżeli 10^{12} rekordów.

9. Wsparcie dla procedur i funkcji składowanych w bazie danych. Język programowania ma być językiem proceduralnym, blokowym (umożliwiającym deklarowanie zmiennych wewnątrz bloku), oraz wspierającym obsługę wyjątków.

10. Możliwość kompilacji procedur składowanych w bazie danych do postaci kodu binarnego.

11. Baza danych musi mieć możliwość deklarowania wyzwalaczy (triggerów) na poziomie instrukcji DML (INSERT, UPDATE, DELETE) wykonywanej na tabeli, poziomie każdego wiersza modyfikowanego przez instrukcję DML.

12. W przypadku, gdy w wyzwalaczu na poziomie instrukcji DML wystąpi błąd zgłoszony przez motor bazy danych bądź ustawiony wyjątek w kodzie wyzwalacza, wykonywana instrukcja DML musi być automatycznie wycofana przez serwer bazy danych, zaś stan transakcji po wycofaniu musi odzwierciedlać chwilę przed rozpoczęciem instrukcji w której wystąpił ww. błąd lub wyjątek.

13. Rozwiązanie musi umożliwiać wymuszanie złożoności hasła użytkownika, czasu życia hasła, sprawdzanie historii haseł, blokowanie konta przez administratora bądź w przypadku przekroczenia limitu nieudanych logowań.

14. Przywileje użytkowników bazy danych mają być określane za pomocą przywilejów systemowych oraz przywilejów dostępu do obiektów aplikacyjnych.

15. Baza danych ma umożliwiać nadawanie ww. przywilejów za pośrednictwem mechanizmu grup użytkowników / ról bazodanowych. W danej chwili użytkownik może mieć aktywny dowolny podzbiór nadanych ról bazodanowych.

16. Baza danych ma mieć możliwość wykonywania i katalogowania kopii bezpieczeństwa bezpośrednio przez serwer bazy danych oraz możliwość zautomatyzowanego usuwania zbędnych kopii bezpieczeństwa przy zachowaniu odpowiedniej liczby kopii nadmiarowych - stosownie do założonej polityki nadmiarowości backup'ów.

17. Baza danych ma mieć możliwość integracji z powszechnie stosowanymi systemami backupu (Legato, Veritas, Tivoli, Data Protector, itd).

18. Baza danych ma mieć możliwość wykonywania kopii bezpieczeństwa w trybie online (hot backup).

19. Odtwarzanie bazy danych z kopii ma umożliwiać odzyskanie stanu danych z chwili wystąpienia awarii bądź cofnięcia stanu bazy danych do wskazanego punktu w czasie.

20. Odtwarzanie musi dawać możliwość objęcia całej bazy danych i, wg decyzji administratora bazy danych, pojedynczych plików danych lub obszarów (tj. logicznie pogrupowanych plików lub nośników). W przypadku, gdy odtwarzaniu podlegają pojedyncze pliki bazy danych lub obszary, pozostałe pliki baz danych lub obszary mają być dostępne dla użytkowników.

21. Baza danych ma mieć możliwość zaimplementowania polityki bezpieczeństwa regulującej dostęp do danych na poziomie pojedynczych wierszy w tabelach. Mechanizm ten ma być realizowany za pomocą mechanizmów motoru bazy danych i musi być przezroczysty dla aplikacji.

22. Motor bazy danych ma udostępniać możliwość zrównoleglenia operacji SQL (zapytania, instrukcje DML, ładowanie danych, tworzenie indeksów, przenoszenie tabel/indeksów pomiędzy przestrzeniami danych) oraz procesów wykonywania kopii bezpieczeństwa bądź odtwarzania.

23. Motor bazy danych ma umożliwiać wykonywanie niektórych operacji związanych z utrzymaniem bazy danych bez konieczności pozbawienia dostępu użytkowników do danych. W szczególności dotyczy to tworzenia/ przebudowywania indeksów oraz reorganizacji bądź redefinicji tabel.

24. Baza danych ma dawać możliwość zakładania/przebudowywania indeksów online bez konieczności odłączenia użytkowników operujących (zapytania, operacje insert, update, delete) na tabelach podlegających indeksowaniu.

25. Motor bazy danych ma umożliwiać zarządzanie przydziałem zasobów obliczeniowych dla użytkowników bazy danych.

26. Oprogramowanie musi mieć zapewnione wsparcie techniczne, umożliwiające otrzymywanie aktualizacji oprogramowania oraz zgłaszanie i rozwiązywanie defektów oprogramowania.

27. Oprogramowanie bazy danych musi zapewniać efektywne metody kompresji danych, pozwalające na zmniejszenie zajmowanej przestrzeni dyskowej. Kompresja danych musi być przezroczysta dla aplikacji, tzn. włączenie bądź wyłącznie kompresji nie może wymuszać modyfikacji aplikacji. Dane muszą być kompresowane zarówno podczas ładowania dużych porcji danych (ładowanie blokowe), jak i podczas wykonywania pojedynczych operacji wstawiania, bądź aktualizacji danych (instrukcje SQL INSERT, UPDATE).

28. Oferowana technologia kompresji musi zapewnić, dla nieskorelowanych danych tekstowych, zmniejszenie wymagań na przestrzeń dyskową dla tabel bazy danych i indeksów. Dostarczone oprogramowanie baz danych musi zawierać odpowiednie licencje na wykorzystanie kompresji.

29. Oprogramowanie baz danych musi posiadać mechanizmy przyspieszające dostęp do danych oraz operacje wykonywane na danych - wykorzystywane podczas przetwarzania danych, analizy informacji oraz udostępniania danych: indeksowanie, równoległe wykonywanie zapytań i procesów, partycjonowanie danych oraz widoki zmaterializowane lub ich funkcjonalne odpowiedniki.

30. Oprogramowanie bazy danych musi zapewnić kontrolę dostępu na poziomie pojedynczego wiersza w taki sposób, by użytkownicy widzieli tylko te rekordy, do których mają uprawnienia. Musi istnieć także możliwość ukrywania wartości kolumn tabel, zależnie od nadanych uprawnień. Tzn. użytkownik może odczytać wartości tylko tych kolumn, dla których ma uprawnienia. Dla pozostałych baza musi wygenerować zamaskowane wartości lub ukryć ich wartość.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

31. Oprogramowanie bazy danych musi zapewniać mechanizm pozwalający na odpytywanie zawartości tabeli dla zadanego w zapytaniu punktu w czasie (np. agregacja wartości kolumn dla stanu tabeli sprzed miesiąca). Mechanizm nie może ograniczać składni zapytań języka SQL, ani okresu czasu, dla którego będą wykonywane tego typu operacje.

32. Oprogramowanie bazy danych musi udostępniać mechanizmy zarządzania obciążeniem, w szczególności musi dawać możliwość automatycznego określania priorytetów wykonywanych w bazie danych zadań, np. na podstawie nazwy użytkownika, nazwy aplikacji, adresu IP komputera klienckiego, ilości czytanych wierszy, czy czasu wykorzystanego procesora. Dla bardzo długo wykonujących się zadań musi istnieć możliwość automatycznego obniżenia priorytetu oraz zatrzymania zadania po upływie określonego czasu.

33. Obok klasycznych danych relacyjnych system musi zapewniać możliwość wydajnego składowania i obsługi danych w formacie XML. Baza musi obsługiwać natywny typ danych XML.

34. Baza danych musi wspierać języki zapytań XQuery, XPath oraz SQL/XML.

35. Baza danych musi wspierać efektywny mechanizm składowania dokumentów XML pozwalający na kompresję zawartości dokumentów XML.

36. System musi być wyposażony w mechanizmy monitorujące, ułatwiające wykrywanie potencjalnych źródeł awarii.

37. System musi umożliwiać wielu użytkownikom równoległy dostęp do tych samych danych.

38. Oprogramowanie baz danych musi posiadać mechanizmy zarządzające optymalnym rozłożeniem danych na dyskach, w celu uzyskania lepszej wydajności rozwiązania.

39. Oprogramowanie baz danych ma dawać opcjonalnie możliwość zapewnienia przechowywania wszystkich danych zastrzeżonych (dane osobowe) w bazach danych w sposób zaszyfrowany przy użyciu mechanizmów wbudowanych w silnik baz danych.

40. Oprogramowani baz danych musi zapewnić na poziomie bazy danych oraz warstwy prezentacji dostępność operacji transformacji i przekształcenia danych zgodnych ze standardem ISO/ANSI SQL92.

System ETL/ ELT

Lista wymagań jaką powinien spełniać system ETL:

1. System musi obsługiwać ładowanie danych z relacyjnych baz danych (minimum Oracle DB, Microsoft SQL Server, IBM DB2, źródła JDBC) oraz plików płaskich.
2. System musi zapewnić integrację z hurtownią danych w taki sposób, by transformacje na danych nie wymuszały kopiowania danych poza hurtownię danych.
3. System musi dostarczać narzędzie do graficznego projektowania transformacji danych.
4. System musi dostarczać narzędzie do graficznego zarządzania procesami transformacji danych (uruchamianie, harmonogramowanie zadań, śledzenie statusu).
5. Dostarczone narzędzie musi umożliwiać przy pomocy graficznego interfejsu deklaratywne projektowanie następujących operacji:
    - import danych z pliku,
    - export danych do pliku,
    - pobieranie danych z tabeli bazy danych,
    - zapisywanie danych do tabeli bazy danych,
    - szybkie, blokowe ładowanie danych,
    - doczytywanie wartości kluczy (key lookup),
    - obsługa wolno-zmieniających się wymiarów (slow changing dimensions),
    - scalanie danych (merge),
    - grupowanie danych (group by),
    - sortowanie (order by),
    - filtrowanie danych po kolumnach i wierszach,
    - łączenie tabel (join),

- wywoływanie funkcji i procedur składowanych,
- uruchamianie zapytań SQL,
- uruchamianie zapytań XQuery,
- warunkowe wykonywanie transformat,
- wykonywanie transformat w pętli,
- równoległe wykonywanie zbioru transformat.

6. Zaprojektowane transformacje danych muszą być automatycznie optymalizowane dla oferowanej hurtowni danych.
7. System musi zapewnić możliwość uruchamiania i harmonogramowania zewnętrznych skryptów oraz programów wykonywalnych.
8. System musi umożliwiać obsługę błędów.
9. System musi umożliwiać powiadamianie o błędnym zakończeniu procesu transformacji / ładowania danych pocztą elektroniczną.
10. System musi posiadać graficzne narzędzie do zarządzania użytkownikami i ich przywilejami uruchamianych zadań.
11. System musi umożliwiać odseparowanie środowiska deweloperskiego oraz środowiska produkcyjnego.
12. System musi zapewnić możliwość debuggowania zaprojektowanych transformat.
13. Narzędzie ma umożliwiać wydajne ładowanie danych do hurtowni danych, „data martów", kostek OLAP i innego rodzaju docelowych składnic danych.
14. System musi obsługiwać w sposób przeźroczysty ładowanie przyrostowe, wymiary wolnozmienne, jednocześnie zapewniając integralność i spójność danych – wystarczy metoda SQL.
15. System ma działać w architekturze SOA i musi posiadać usługi związane z integracją i transformacją danych, które mogą być swobodnie wykorzystywane w różnego rodzaju procesach biznesowych.
16. System ma umożliwiać deklaratywne modelowanie transformacji danych.

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI spin. AGH Małopolska UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

17. System musi posiadać graficzne narzędzie do zarządzania użytkownikami i ich przywilejami.

18. System musi posiadać graficzne narzędzie do monitorowania wykonania zadań.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

System eksploracji danych (data mining)

Lista wymagań względem systemu eksploracji danych:

1. Rozwiązanie ma mieć możliwość wykonywania analiz data mining bezpośrednio w bazie danych (bez potrzeby eksportowania danych do zewnętrznych narzędzi) w tym m.in.:
    - budowanie modelu,
    - zastosowanie modelu,
    - wykorzystanie funkcji statystycznych.
2. Rozwiązanie ma mieć obsługę algorytmów: klasyfikacji, regresji, klastrowania, istotności atrybutu, drzewa decyzyjne, asocjacje, wykrywanie anomalii.
3. Rozwiązanie ma mieć możliwość uruchomienia algorytmów języka R bezpośrednio na bazie danych bez potrzeby eksportu danych na zewnątrz.
4. Rozwiązanie ma mieć dostęp do algorytmów statystycznych bezpośrednio poprzez API SQL i R, w tym m.in dostęp do skryptów R dostępnych razem w CRAN R packages.
5. Zapewnienie warstwy pośredniej tłumaczącej skrypty R na natywny język SQL bazy danych. W sytuacji w której zapytanie R nie będzie mogło być obsłużone przez natywny SQL należy zapewnić obsługę przez zintegrowany silnik analityczny R.
6. Zarządzanie, przygotowanie i transformacje danych, budowanie modelu i scoring ze względów wydajnościowych muszą być uruchamiane bezpośrednio jako proces bazy danych.
7. Rozwiązanie musi dostarczać poniższe funkcje i algorytmy:
    - Anomaly Detection / np. One-Class Support Vector Machine,
    - Association Rules / Apriori,
    - Attribute Importance / Minimum Descriptor Length lub Chi-square,

- Classification / Decision Tree,
- Classification / Decision Tree (cross validation), dopuszczalny jest tryb automatyczny,
- Classification / Logistic Regression,
- Classification / Naive Bayes,
- Classification / Support Vector Machine,
- Clustering / k-Means,
- Regression / Linear Regression,
- Regression / Support Vector Machine,
- Text Mining – podejście statystyczne lub lingwistyczne,
- Principal Components Analysis (PCA),
- Neural Networks,
- Scoring tabel bazy danych z wykorzystaniem modeli open-source R, sampling wykonany bezpośrednio na bazie danych.

8. Rozwiązanie musi posiadać graficzny interfejs do wykonywania części analiz pozwalający na budowanie i ewaluację modeli, zastosowanie modelu, analitycznego workflow i opublikowania rezultatów.
9. Rozwiązania musi mieć możliwość budowania modeli powinno zapewniać możliwość przetwarzania równoległego, agregacji.
10. Rozwiązanie musi mieć możliwość wykorzystania zaawansowanych funkcji statystycznych języka R dostępnych razem z R CRAN packages.
11. Platforma analityczna data mining ma posiadać możliwość uruchomienia w trybie wysokiej dostępności.
12. Rozwiązanie musi mieć możliwość integracji z kodem SQL oraz PL/SQL.
13. Rozwiązanie musi mieć możliwość pobierania danych z baz i plików csv.
14. Platforma analityczna data mining ma posiadać możliwość wykorzystania funkcji statystycznych takich jak:

- rankingi: ranking, seryjny ranking, kumulacyjne wartości w grupie wartości,
- agregacje w oparciu o okna (ruchome oraz skumulowane),
- funkcje operujące na offsetach względem odwołania do wierszy,
- agregacje służące raportowaniu sum, avg, min, max, count,
- agregacje statystyczne: korelacja, regresje liniowe, kowariancje,
- regresja liniowa, kowariancja populacji, kowariancja próbki, współczynnik korelacji zbioru par liczb,
- statystyka opisowa: średnie, odchylenie standardowe, wariancja, min, max, mediana, group-by,
- aregacje oraz operacje na wierszach tabeli: min, max, range, mean, , variance, standard deviation, median, kwantyle, +/- n sigma values, top/ bottom 5 values,
- korelacje Współczynniki korelacji Pearson'a oraz Spearman'a,
- analiza tablicowa chi squared, phi coefficient, contingency coefficient,
- testowanie hipotez: Student t-test , F-test, Binomial test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA,
- dystrybucje Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential,
- analiza Pareto oraz cumulative results table.

System OLAP

Lista wymagań wobec systemu budowy i zarządzania kostkami OLAP:

1. System musi zapewniać możliwość wirtualnego łączenia wielu kostek OLAP w jedną logiczną kostkę OLAP.
2. Silnik OLAP musi wspierać zadawanie zapytań w języku SQL lub MDX.

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI spin. AGH Małopolska UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

3. Kostki OLAP muszą mieć możliwość wykorzystania mechanizmów bezpieczeństwa bazy danych (wykorzystanie ról bazodanowych).
4. System musi udostępniać narzędzie graficzne do projektowania kostek wielowymiarowych.
5. System musi dostarczać narzędzie graficzne umożliwiające:
    - definiowanie faktów i miar,
    - definiowanie wymiarów oraz hierarchii wymiarów.
6. System OLAP musi zapewnić mechanizmy przyspieszające przetwarzanie zapytań wielowymiarowych.
7. System musi zapewniać możliwość budowania kostek OLAP bez potrzeby przenoszenia danych między środowiskiem bazy danych a silnikiem OLAP.
8. Rozwiązanie ma umożliwiać dostęp do kostek i wymiarów OLAP bezpośrednio przez SQL.
9. Rozwiązanie ma umożliwiać dostęp do kostek OLAP z poziomu API JAVA.
10. Rozwiązanie ma umożliwiać transparentną poprawę wydajności zapytań OLAP agregujących dane poprzez wykorzystanie widoków zmaterializowanych (materialized views).
11. Funkcje wymagane w ramach obsługi kostki OLAP:
    a. funkcje arytmetyczne:
        - dodawanie, odejmowanie, mnożenie, dzielenie lub stopniowanie, porównanie procentowe,
    b. funkcje analityczne
        - indeksowanie,
        - operacje na poprzednich i przyszłych okresach,
        - operacje typu Period to Date,
        - stosunki/ współczynniki,
        - rankingi,

- kalkulacje typu “moving”,
- kalkulacje łączne,
- kalkulacje zagnieżdżone,

c. funkcje na pojedynczym wierszu kostki:

- funkcje numeryczne,
- funkcje na datach/ czasie,
- funkcje na znakach,
- funkcje porównujące,
- funkcje konwertujące.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Business Intelligence

Wymagania dotyczące architektury systemu BI:

1. System musi zapewniać bezpośredni (bez potrzeby ładowania danych do pośredniego silnika bazodanowego lub silnika in-memory) dostęp do różnych typów źródeł danych: np. XML, Web Services, procedur składowanych, plików płaskich, baz relacyjnych, baz wielowymiarowych, systemów transakcyjnych, hurtowni danych, hurtowni tematycznych.
2. Rozwiązanie musi być oparte o jedne spójny interfejs użytkownika, oparty o jeden model metadanych. Interfejs musi zapewnić możliwość publikacji raportów na kokpicie, raportów ad-hoc, raportowania operacyjnego, bezpośrednich zapytań do źródeł fizycznych podpiętych do serwera analitycznego.
3. System powinien obsługiwać m.in. następujące źródła danych: Baza Oracle, Baza DB2, Baza Microsoft SQL Server, Microsoft Analysis Services (MDX), ESSBASE.
4. Rozwiązanie musi być oparte o jeden łatwo zarządzalny, spójny model metadanych wykorzystywany przez wszystkie elementy interfejsu użytkownika.
5. Rozwiązanie musi pozwalać na definiowanie metadanych serwera BI poprzez intuicyjny interfejs graficzny a nie w oparciu o skrypty i kodowanie.
6. Rozwiązanie musi posiadać możliwość łączenia na poziomie modelu metadanych informacji pochodzących z różnych źródeł. Musi pozwalać na proste wykonywanie raportów opartych o fragmentację danych pochodzących z wielu źródeł, drążenie poprzez różne źródła danych w ramach jednego raportu.
7. System powinien umożliwiać użytkownikowi/administratorowi zmianę nazw elementów warstwy fizycznej na pojęcia biznesowe, przyjazne użytkownikowi końcowemu.
8. System musi natywnie wspierać wielojęzyczność przez mechanizmy wbudowane w rozwiązanie. Wielojęzyczność musi być wspierana w obrębie jednej warstwy metadanych i nie może wymagać dla każdego języka instalacji odrębnej warstwy metadanych lub ich części.

9. W celu osiągnięcia skalowania systemu powinien być wykorzystywany mechanizm puli połączeń (“connection pooling”). Oznacza to, że pojedyncze połączenie do bazy danych jest wykorzystywane do wykonywania wielu zapytań.

10. Użytkownik musi mieć dostęp do informacji biznesowej w sposób on-line (raporty) wyłącznie przez standardową przeglądarkę sieci Web za pomocą języka DHTML (technologia AJAX). Wykorzystanie przeglądarki internetowej jako interfejsu użytkownika nie może wymuszać instalacji dodatkowych komponentów typu ActiveX lub Applet Java. Wymagane jest wsparcie przynajmniej dwóch następujących przeglądarek internetowych – Internet Explorer, Mozilla Firefox, Chrome.

11. System powinien udostępniać biblioteki API do warstwy modelu biznesowego metadanych.

12. Rozwiązanie musi być oparte o architekturę trójwarstwową.

13. System musi mieć możliwość instalacji na platformie systemu operacyjnego MS Windows oraz Linux.

14. System musi zapewniać możliwość tworzenia agregatów w relacyjnym źródle danych na podstawie logiki biznesowej warstwy metadanych serwera analitycznego. W rezultacie musi istnieć możliwość wygenerowania skryptu fizycznego, który będzie uruchomiony po stronie bazy danych i pozwoli na utworzenie odpowiednich agregatów. Jednocześnie system zapewni automatycznie obsługę tych agregatów po stronie modelu metadanych serwera analitycznego.

15. System musi zapewniać natywną możliwość wizualizacji informacji na mapie.

16. System musi zapewnić możliwość klastrowania środowiska BI.

17. Kokpity menedżerskie BI (management dashboard) muszą oferować możliwość układu wielozakładkowego (tzw. taby) w ramach każdego pojedynczego kokpitu z możliwością ręcznego przechodzenia między tymi zakładkami jak również mechanizm łatwego dodawania prostych przycisków nawigacyjnych służących do przełączania się między poszczególnymi tabami.

18. Wymaga się by kokpity były zrealizowane w klasycznej technologii przeglądarki internetowej (D)HTML i zapewniały możliwość łatwej, dynamicznej zmiany sposobu prezentacji danych bezpośrednio na danym komponencie – np. z tabeli

19. na wykres słupkowy, kołowy, etc. bez konieczności edytowania źródłowej zawartości komponentu/raportu/analizy.

Wymagania analityczno-raportowe:

1. System musi zapewnić możliwość samodzielnego tworzenia raportów przez użytkowników końcowych inaczej niż w sposób ściśle programistyczny.
2. System musi pozwolić na prezentowanie informacji z wielu źródeł danych na jednym raporcie (nie na kokpicie menedżerskim).
3. System musi wykorzystywać tabele agregatów w sposób transparentny dla użytkownika końcowego.
4. System musi umożliwiać prezentację danych na raportach z wykorzystaniem takich wizualizacji jak: Waterfall, Map, 100 Percent Stacked Chart, zamrożonych nagłówków tabeli i tabeli przestawnych.
5. System musi rekomendować użytkownikowi odpowiedni sposób wizualizacji danych w zależności od danych zdefiniowanych w kryterium raportu tzw. data-driven insight.
6. System musi potrafić dynamicznie udostępniać użytkownikom listy wartości wykorzystywane do filtrowania danych na raporcie.
7. System musi wspierać kaskadowe podpowiedzi (prompts) np. 2-ga podpowiedź wyświetla tylko wyfiltrowane ważne wartości dla niej bazując na wartościach zwróconych w 1-ej podpowiedzi.
8. System powinien potrafić wyeksportować dane w formacie HTML, PDF, Excel (*.xls), Excel (*.xsls), MHTML.
9. System powinien potrafić wizualizować graficznie tzw. wyjątki tzn. wartości przekraczające wartości oczekiwane, nie mieszczące się w pewnych zakresach.
10. System musi umożliwiać wykonywanie kalkulacji: matematycznych, statystycznych, znakowych, konwersji.
11. System musi wspierać tworzenie warunków wyliczanych, wykorzystywanych do filtrowania danych.

12. System powinien umożliwiać wizualizację danych aktualnych, historycznych oraz trendu.

13. System musi umożliwiać użytkownikowi budowę nowego raportu tylko i wyłącznie za pomocą standardowej przeglądarki internetowej np. Internet Explorerze ,Mozilla Firefox, Chrome.

14. System powinien pozwalać użytkownikowi na sortowanie danych dowolnego wymiaru w porządku rosnącym lub malejącym w przeglądarce internetowej.

15. System powinien pozwalać użytkownikowi na sortowanie wyników raportu w postaci tabeli przestawnej.

16. System powinien pozwalać użytkownikom ustawiać warunki potrzebne do filtrowania danych w przeglądarce internetowej.

17. System powinien pozwalać użytkownikom na wykonywanie operacji drążenia danych do danych bardziej szczegółowych (drill down) w przeglądarce internetowej.

18. System ma pozwalać użytkownikowi na drążenie hierarchii wymiaru.

19. Powinna istnieć możliwość definiowania na raporcie nowych obiektów wyliczalnych oraz grup, wykorzystując zarówno elementy z hierarchii danego wymiaru oraz korzystając z atrybutów wymiaru.

20. System musi posiadać możliwość drążenia informacji pochodzących z kilku źródeł danych bez potrzeby tworzenia dodatkowych raportów (tzn. bez potrzeby łączenia kilku raportów zawierających informacje z różnych źródeł danych).

21. Powinna istnieć możliwość wykorzystania na raporcie kilku hierarchii wymiarów jednocześnie oraz możliwość umieszczenia w raporcie jednocześnie hierarchii wymiarów wraz z atrybutami wymiarów.

22. System musi zapewniać możliwość tworzenia nowych grup wyliczalnych z uwzględnieniem struktury hierarchicznej wymiaru.

23. Tworzenie każdego dodatkowego widoku danych nie może wymagać osobnego, nowego zapytania SQL.

24. System musi umożliwiać użytkownikom dodawanie logicznych kolumn, wyrażeń, obliczeń na raporcie uruchomionym w przeglądarce internetowej.

25. System musi umożliwiać użytkownikom na zmianę nazw kolumn na raporcie uruchomionym w przeglądarce internetowej, na dowolnie wybrane przez użytkownika nagłówki i etykiety .

26. System musi umożliwiać tworzenie raportów operacyjnych o dokładnie określonym układzie (tzw. pixel-perfect formatting).

27. Dostęp do kokpitów, tworzenie raportów ad-hoc i tworzenie raportów operacyjnych (tzw. pixel-perfect) musi być realizowane poprzez jeden spójny interfejs oparty o przeglądarkę internetową.

28. System musi umożliwiać tworzenie poprzez przeglądarkę internetową firmowego stylu (template) który raz stworzony może być dziedziczony przez wszystkie raporty.

29. Strony portalu informacyjnego muszą mieć możliwość personalizacji na poziomie użytkownika.

30. Kokpity informacyjne muszą mieć możliwość osadzenia w nich treści z zewnętrznego serwisu internetowego.

31. Powinna istnieć możliwość udostępnienia raportu i kokpitu w postaci adresu URL z zachowaniem praw dostępu odnośnie zawartej tam treści.

32. Portal (kokpit) informacyjny musi mieć możliwość osadzenia w nim dowolnej zawartości DHTML (HTML oraz Java Script).

33. System powinien umożliwiać bezpośrednie połączenie jednego raportu z kilkoma innymi w ramach jednego kokpitu menadżerskiego, tak aby kliknięcie na atrybucie raportu powodowało automatyczne filtrowanie danych na pozostałych raportach (tzw. master-detail).

34. System musi mieć możliwość tworzenia zapytań analitycznych opartych o SQL bazujących na modelu logicznym metadanych serwera BI i bezpośrednich strukturach fizycznych podpiętych do serwera BI.

35. System musi zapewniać obsługę funkcji szeregów czasowych jak np. SQL PERIODROLLING, AGGREGATE AT, AGO, TODATE

36. System musi wspierać formatowanie warunkowe dla tabeli przestawnej.

37. System powinien wspierać obsługę hierarchii niezbalansowanych.

38. System musi zapewniać możliwość zaimportowania raportów do programu MS Excel i MS Power Point z możliwością automatycznego odświeżania zawartości raportów. Zawartość raportów powinna być automatycznie filtrowana zgodnie z uprawnieniami użytkownika.

39. System musi zapewniać możliwość podglądu rezultatu/ układu wygenerowanego raportu na etapie jego tworzenia bez potrzeby wcześniejszego zapisywania raportu.

40. System musi umożliwić generowanie raportów operacyjnych w oparciu o bezpośrednie zapytania fizyczne do źródła danych oraz w oparciu o model metadanych serwera BI.

41. System musi umożliwić importowanie danych zawartymi w raportach systemu BI oraz zawartych w metadanych serwera BI z poziomu MS Excel.

42. Użytkownik musi mieć możliwość przywrócenia kokpitu z raportami do stanu wyjściowego po wykonaniu drążenia, filtrowania lub zmiany układu tabeli przestawnej bezpośrednio na kokpicie.

43. System musi zawierać kontekstową Pomoc dla użytkowników/ administratorów.

44. System musi zapewnić możliwość automatycznej weryfikacji modelu metadanych pod kątem potencjalnych błędów w projektowaniu metadanych.

45. System musi umożliwiać użytkownikom planowanie wykonywania raportów o określonym czasie, cykliczności lub jednorazowo.

46. System musi pozwalać użytkownikom końcowym na samodzielne ustawianie harmonogramów wykonania ich zadań/ raportów oraz zapytań.

47. System musi dostarczać mechanizmy do zmiany układu kolumn i wierszy raportów umieszczonych w tabeli lub tabeli przestawnej, poprzez prosty mechanizm „Drag&Drop".

48. System musi ukrywać złożoność struktur danych fizycznych oraz wszystkich aspektów związanych z ich dostarczeniem. Użytkownik musi posługiwać się pojęciami i elementami posiadającymi nazwy biznesowe oraz nie musi znać lokalizacji danych na których pracuje.

49. System musi pozwalać na ukrywanie kolumn na raporcie.

50. System musi zawierać stronę domową na której użytkownik będzie miał dostęp do najczęściej używanych raportów, listy folderów z raportami.

51. System powinien umożliwiać łatwe wyszukiwanie raportów po słowach kluczowych w opisie raportu z możliwością wykorzystania mechanizmu „full text search".

52. System musi pozwalać użytkownikowi na zdefiniowanie procesu obserwowania wyników zwracanych przez raport lub raporty w sposób cykliczny i ostrzegania użytkownika jeżeli wartości progowe zostaną przekroczone. System musi zapewniać możliwość sekwencyjnego obserwowania zdarzeń tzn. jeżeli jeden proces obserwacji wyników zwróci oczekiwane rezultaty może zostać uruchomiony kolejny proces lub procesy sprawdzające kolejne obszary. W rezultacie użytkownik powinien dostać stosowne powiadomienie mailem.

53. System musi zapewniać możliwość szybkiego generowania analiz poprzez umieszczenia wybranego raportu w pamięci cache serwera analitycznego, na podstawie ustalonego wcześniej harmonogramu oraz zdefiniowanego wcześniej warunku lub grupy warunków.

54. System musi zapewnić możliwość podglądu i drukowania schematu fizycznego i biznesowego modelu metadanych.

55. Możliwość zastosowania filtrów bez konieczności wykorzystania dodatkowych przycisków typu „zastosuj" („Apply") lub „wyczyść"(„Reset").

Wymagania w obszarze bezpieczeństwa i administracji:

1. System musi umożliwiać proces zewnętrznej identyfikacji użytkowników. Wśród wspieranych sposobów identyfikacji wymagane jest co najmniej identyfikacja przez wykorzystanie serwera LDAP.
2. System musi wspierać wielopoziomowy model bezpieczeństwa jak użytkownik, rola.
3. System musi dynamicznie przypisywać użytkownikom poziom bezpieczeństwa bazując na atrybutach przypisanych użytkownikowi w procesie identyfikacji.
4. System musi w sposób natywny wspierać śledzenie aktywności użytkowników poprzez identyfikator użytkownika, dostarczając informacje m.in. o czasie

5. wykonania raportu, nazwie raportu, statusie raportu (zakończony/ nie zakończony).

6. System musi pozwalać na definiowanie autoryzacji dostępu do danych na poziomie metadanych biznesowych serwera BI.

7. System musi zapewniać przezroczystość zmian atrybutów fizycznych obiektów bazy danych w stosunku do raportów.

8. System musi dostarczać graficzne narzędzie administracyjne które pozwoli na zdefiniowanie metadanych serwera analitycznego bez potrzeby ręcznego pisania SQL.

9. System musi zapewniać konsolę do zarządzania systemem umożliwiającą min. uruchomienie/ zatrzymanie poszczególnych komponentów systemu, konfigurację, mierzenie wydajności, diagnostykę systemu BI.

10. System musi zapewniać możliwość lokalizacji struktury metadanych jak i warstwy prezentacji.

11. System musi dostarczać inteligentnego, wielożytkownikowego mechanizmu cache.

12. System musi wspierać funkcjonalność klastrowania do operacji równoważenia obciążenia (load balancing) oraz operacji przełączania podczas awarii dla wielu instancji serwerów aplikacyjnych.

13. System musi wspierać realizację wielu równoległych zapytań SQL.

14. System musi wykorzystywać zalety architektury SMP (Symmetric Multi-Processing).

15. System musi wspierać wielowątkowość.

16. System musi wspierać możliwość wcześniejszego buforowania wyników i wyliczeń niezbędnych do szybkiego dostarczenia raportu użytkownikowi końcowemu. Mechanizm musi posiadać możliwość ustalenia harmonogramu zasilania pamięci cache serwera analitycznego żądanymi wynikami.

17. System musi automatycznie optymalizować zapytania analityczne tzn. obliczenia zawarte w logicznym zapytaniu po stronie systemu BI mogą być w ramach optymalizacji całkowicie wykonane po stronie serwera BI, częściowo wykonane

18. po stronie serwera BI i na bazie danych, całkowicie wykonane po stronie bazy danych.

Wymagania dotyczące dystrybucji informacji:

1. System musi zapewnić możliwość samodzielnej subskrypcji użytkownika końcowego do rozsyłanej informacji.
2. Musi istnieć możliwość dystrybucji wybranych raportów.
3. System musi zapewniać możliwość dystrybucji informacji na podstawie listy dystrybucyjnej.
4. System musi pozwolić na generowanie wielu wersji raportów z automatycznym podziałem informacji na podstawie jednego szablonu raportu.
5. System musi zapewnić dystrybucję informacji poprzez email, drukarka, fax.
6. System musi umożliwiać tworzenie raportów operacyjnych o dokładnie określonym układzie (tzw. pixel-perfect formatting).

Wykaz rysunków

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Bibliografia

Breslin Mary: Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models, Business Intelligence Journal, s. 6-20, Winter 2004

Connolly Thomas, Begg Carolyn: Database Systems, A Practical Approach to Design, Implementation, and Management, Addison Wesley, imprint Person Education Ltd, London, 4th edition, 2005

Drucker Peter F., Hammond John, Keeney Ralph, Raiffa Howard, Hayashi Alden M.: Podejmowanie decyzji. Harvard Business Review. ONEPRESS, Gliwice, 2005

Dymek Dariusz, Komnata Wojciech, Kotulski Leszek: Federacyjna hurtownia danych w dostępie do informacji poufnej. Roczniki Kolegium Analiz Ekonomicznych, zeszyt 33/2014, strony 135-154, Oficyna Wydawnicza SGH, Warszawa, 2014

Dymek Dariusz, Komnata Wojciech, Kotulski Leszek, Szwed Piotr: Architektury Hurtowni Danych. Model referencyjny i formalny opis architektury. Wydawnictwo AGH, Kraków, 2015

Golfarelli Matteo, Rizzi Stefano: Data Warehouse Design, Modern Principles and Methodologies. Tata McGraw Hill Education Private Limited, New Delhi, 2009

Inmon William H.: Building the Data Warehouse, Wiley Publishing, Indianapolis, 4th edition, 2005

Kieżuń Witold: Sprawne zarządzanie organizacją: zarys teorii i praktyki, Oficyna Wydawnicza SGH, Warszawa, 1997

Kimball Ralph, Ross Margy: The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling, A John Wiley and Sons, Indianapolis, 3rd edition, 2013

Komnata Wojciech, Dymek Dariusz: Integracja rejestrów publicznych na poziomie samorządu terytorialnego. Roczniki Kolegium Analiz Ekonomicznych, zeszyt 33/2014, strony 247-264, Oficyna Wydawnicza SGH, Warszawa, 2014

Ponniah Paulraj: Data Warehousing fundamentals for IT professionals. A John Wiley and Sons, Inc, Publication, Hoboken, New Jersey, 2nd edition, 2010

Watson Hugh J., Ariyachandra Thilini: Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures. raport http:// www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf, July 2005. Data odczytu 1 czerwca 2014 roku