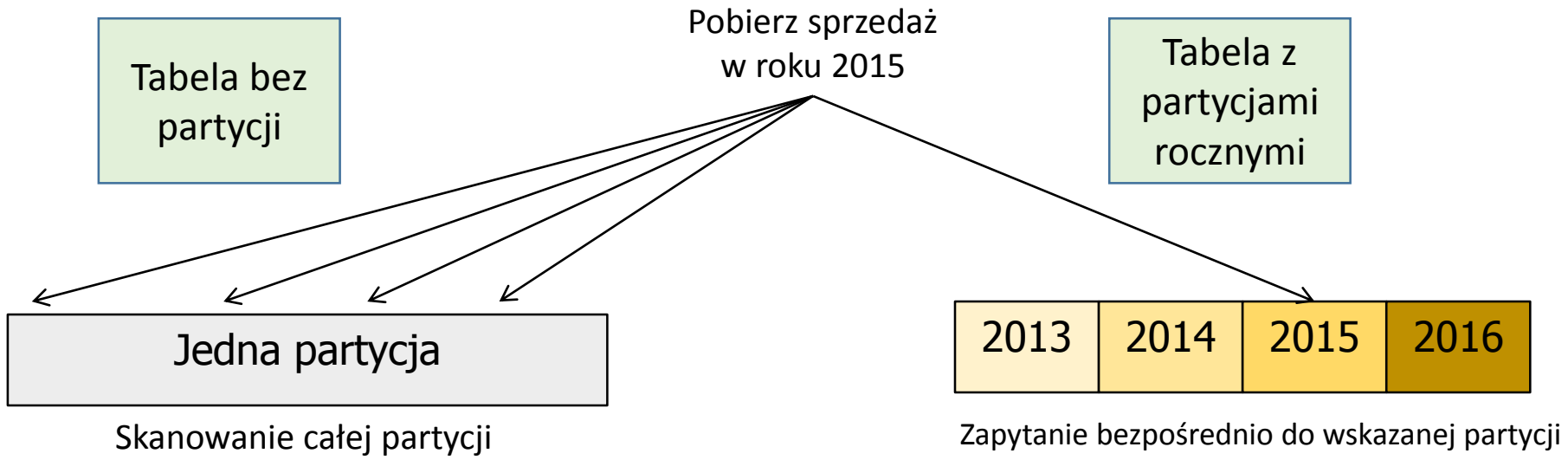
Wprowadzenie do Hurtowni Danych

Mariusz Rafało
mrafalo@sgh.waw.pl

PROJEKTOWANIE WARSTWY DANYCH DETALICZNYCH - ZAGADNIENIA

Partycjonowanie

- Partycja jest wydzielonym miejscem na dysku, w którym przechowywane są określone dane
- Stworzenie partycji pozwala na skrócenie czasu dostępu do danych



Tworzenie partycji

- Dane w różnych partycjach nie mogą się powielać
- Wszystkie dane z tabeli muszą należeć do jakieś partycji
- Dane mogą być wskazane poprzez tabelę lub poprzez zapytanie

WERSJONOWANIE DANYCH

Wersjonowanie

- Wersjonowanie jest powszechnym zagadnieniem w problematyce hurtowni danych. Zdarza się np. że klient zmieni adres swojej siedziby a ta informacja powinna mieć odzwierciedlenie w systemie
- Istnieje kilka podejść do problemu wersjonowania elementów wymiarów
  - Change Data Capture (CDC) – ogół mechanizmów służących do wychwytywania zmian w bazie danych
  - Slowly Changing Dimension (SCD) – wykrycie zmiany w elemencie wymiaru oraz automatyczna reakcja na ta zmianę

SCD – Slowly Changing Dimension (1)

- Nadpisanie wartości
  - Zalety
    - Prosta implementacja (przeliczenie agregatów!)
  - Wady
    - Brak historii (przekłamanie historii!)

Stan przed:

Id Klienta	Imię	Nazwisko	Województwo
101	**Anna**	**Kowalska**	**wielkopolskie**

Stan po:

Id Klienta	Imię	Nazwisko	Województwo
101	**Anna**	**Kowalska**	**mazowieckie**

SCD – Slowly Changing Dimension (2)

- Dodanie nowego rekordu
  - Zalety
    - Prosta implementacja
    - Zachowanie historii
  - Wady
    - Brak możliwości skojarzenia nowej wartości atrybutu z historycznymi faktami

Stan przed:

Id Klienta	PESEL	Imię	Nazwisko	Województwo	Data od
101	**76032465444**	**Anna**	**Kowalska**	**wielkopolskie**	**2016-01-01**

Stan po:

Id Klienta	PESEL	Imię	Nazwisko	Województwo	Data od
101	**76032465444**	**Anna**	**Kowalska**	**wielkopolskie**	**2016-01-01**
123	**76032465444**	**Anna**	**Kowalska**	**mazowieckie**	**2016-04-23**

SCD – Slowly Changing Dimension (3)

- Dodanie nowej kolumny
  - Zalety
    - Zachowanie historii
    - Możliwości skojarzenia nowej wartości atrybutu z historycznymi faktami
  - Wady
    - Nieefektywność dla wielu atrybutów i wielu zmian

Stan przed:

Id Klienta	Imię	Nazwisko	Województwo
101	**Anna**	**Kowalska**	**wielkopolskie**

Stan po:

Id Klienta	Imię	Nazwisko	Województwo	Województwo poprzednie
101	**Anna**	**Kowalska**	**mazowieckie**	**wielkopolskie**

Inne sposoby wersjonowania (1)

- Określenie czasu obowiązywania wartości (bez znacznika)

Stan przed:

Id Klienta	Imię	Nazwisko	Województwo	Data_od	Data_do
101	**Anna**	**Kowalska**	**wielkopolskie**	**2016-01-01**	

Stan po:

Id Klienta	Imię	Nazwisko	Województwo	Data_od	Data_do
101	**Anna**	**Kowalska**	**wielkopolskie**	**2016-01-01**	**2016-04-15**
121	**Anna**	**Kowalska**	**mazowieckie**	**2016-04-16**	

Inne sposoby wersjonowania (2)

- Określenie czasu obowiązywania wartości (ze znacznikiem)

Stan przed:

Id Klienta	Imię	Nazwisko	Województwo	Data_od	Data_do	Aktualny
101	**Anna**	**Kowalska**	**wielkopolskie**	**2016-01-01**	**9999-12-31**	**T**

Stan po:

Id Klienta	Imię	Nazwisko	Województwo	Data_od	Data_do	Aktualny
101	**Anna**	**Kowalska**	**wielkopolskie**	**2016-01-01**	**2016-04-15**	**N**
121	**Anna**	**Kowalska**	**mazowieckie**	**2016-04-16**	**9999-12-31**	**T**

Inne sposoby wersjonowania (3)

- Wersjonowanie przy użyciu identyfikatora

Stan przed:

Id Klienta	Imię	Nazwisko	Województwo	Id Klienta Poprzednie	Aktualny
101	**Anna**	**Kowalska**	**wielkopolskie**		**T**

Stan po:

Id Klienta	Imię	Nazwisko	Województwo	Id Klienta Poprzednie	Aktualny
101	**Anna**	**Kowalska**	**wielkopolskie**		**N**
121	**Anna**	**Kowalska**	**mazowieckie**	**101**	**N**
136	**Anna**	**Kowalska**	**małopolskie**	**121**	**T**

OBSŁUGA BRAKÓW DANYCH

Obsługa braków danych (1)

Transakcje:

Data	Kod produktu	Ilość
2016-04-05	AAA	45
2016-05-26	BBB	12
2016-04-23	CCC	89

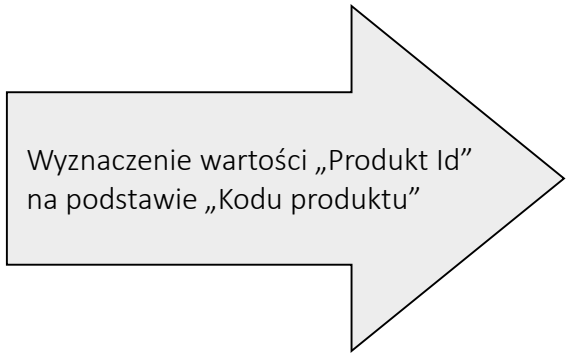


Tabela faktów

Data	Produkt Id	Ilość
2016-04-05	1	45
2016-05-26	**???**	12
2016-05-26	2	89

Słownik Produkt

Produkt Id	Kod produktu	Nazwa produktu
0		Brak danych
1	AAA	Moduł A
2	CCC	Moduł C
3	XXX	Moduł X

Obsługa braków danych (2)

- Pomijamy dane, dla których brak informacji
- Informacja o braku odnotowana w logu

LOG (Transakcje niezasilone)

Data	Kod produktu	Ilość
2016-05-26	BBB	12

Transakcje:

Data	Kod produktu	Ilość
2016-04-05	AAA	45
2016-05-26	BBB	12
2016-04-23	CCC	89

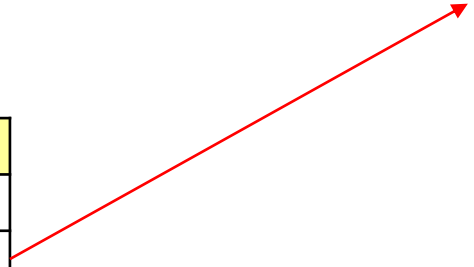


Tabela faktów

Data	Produkt Id	Ilość
2016-04-05	1	45
2016-05-26	2	89

Dalsze działanie:

- Konieczne uzupełnienie słownika
- Ponowne zasilenie danych tym samym zakresem

Obsługa braków danych (3)

Wstawiamy rekord „wartownika”

- Tracimy informację pierwotną
- Możliwa różna interpretacja danych

Tabela faktów

Data	Produkt Id	Ilość
2016-04-05	1	45
2016-05-26	0	12
2016-05-26	2	89

Brak utraty źródłowej informacji

Dalsze działanie

- Uzupełnienie słownika
- Przeliczenie faktów

Tabela faktów

Data	Produkt Id	Kod produktu	Ilość
2016-04-05	1	AAA	45
2016-05-26	0	BBB	12
2016-05-26	2	CCC	89

Obsługa braków danych (4)

Słownik uzupełniany „online”

- Zachowujemy informację
- Możliwy „bałagan”

Tabela faktów

Data	Produkt Id	Ilość
2016-04-05	1	45
2016-05-26	4	12
2016-05-26	2	89

Dalsze działanie

- Zarządzanie zawartością słownika (kontrola jakości danych)

Słownik Produkt

Produkt Id	Kod produktu	Nazwa produktu
0		Brak danych
1	AAA	Moduł A
2	CCC	Moduł C
3	XXX	Moduł X
4	BBB	Moduł B

PROJEKTOWANIE KOSTEK WIELOWYMIAROWYCH

Możliwości w zakresie przechowywania danych

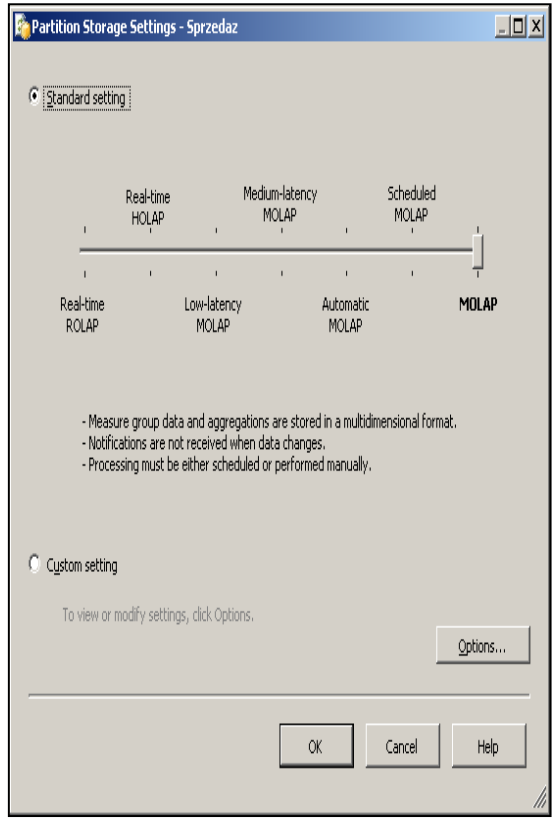
- ROLAP (Relational OLAP)
  - Rozwiązanie bazuje na strukturach relacyjnych
  - Zaletą jest możliwość obsłużenia bardzo dużych ilości danych
  - Wadą są wolne zapytania
- MOLAP (Multidimensional OLAP)
  - Rozwiązanie bazuje na strukturach wielowymiarowych
  - Charakteryzuje się szybką odpowiedzią na zapytania
  - Duże możliwości w zakresie kalkulacji i budowy agregatów
  - Silnik zoptymalizowany pod kątem przetwarzania analitycznego
- HOLAP (Hybrid OLAP)
  - Połączenie obu rozwiązań,
  - Dane detaliczne przechowywane są w strukturach relacyjnych (ROLAP)
  - Agregaty budowane są w oparciu o wielowymiarową bazę danych (MOLAP)

MOLAP vs ROLAP vs OLAP

Model	Opóźnienie	Szybkość zapytań	Szybkość przetwarzania	Rozmiar
MOLAP	Duże	Szybko	Wolno	Duży
ROLAP	Niskie	Wolno	Szybko	Mały
HOLAP	Średnie	Średnio	Szybko	Średni

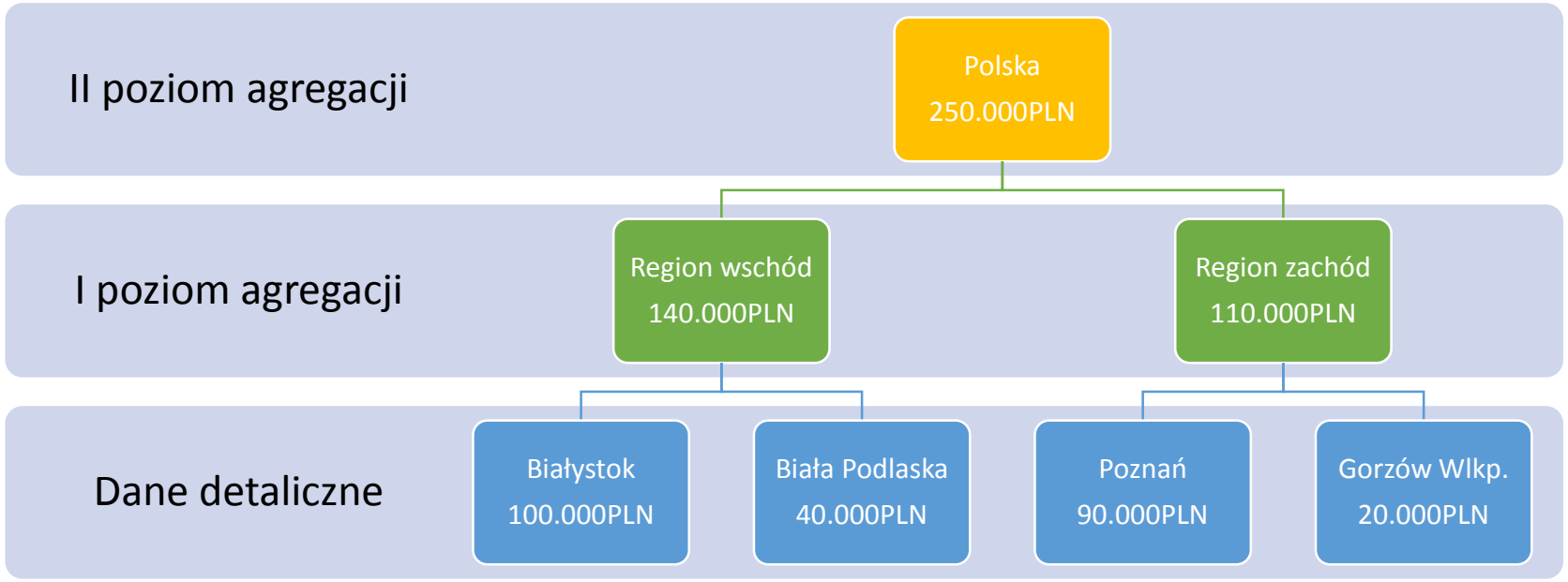
Przykład: Microsoft Analysis Services

- Low-Latency MOLAP
  - Zarówno dane detaliczne jak i agregaty przechowywane są w kostce,
  - Kostka przetwarzana jest co 30min.
- Medium-Latency MOLAP
  - Zarówno dane detaliczne jak i agregaty przechowywane są w kostce,
  - Kostka przetwarzana jest co 120min.
- Automatic MOLAP
  - Kostka przetwarzana jest tylko na Żądanie.
- Scheduled MOLAP
  - Zarówno dane detaliczne jak i agregaty przechowywane są w kostce,
  - Kostka przetwarzana jest zgodnie z harmonogramem.



Agregaty

- Agregat to obliczona i zapisana wartość kalkulowana
- Agregaty mogą znacząco zwiększyć wydajność zapytań MDX



Poziom agregacji

- Duży poziom agregacji
  - Agregaty wyliczone na przecięciach wielu wymiarów
  - Duży rozmiar kostki
  - Wolne przetwarzanie kostki
  - Odciążenie kostki pod kątem kalkulacji(dane są już wyliczone, nie muszą być obliczane ad-hoc)
  - Uzasadniony przy dużych ilościach danych w kostce, gdy kostka przetwarzana jest rzadko
- Mały poziom agregacji
  - Agregaty wyliczone tylko w kluczowych przecięciach wymiarów
  - Mały rozmiar kostki
  - Duże obciążenie kostki podczas agregowania danych ad-hoc
  - Szybkie przetwarzanie kostki
  - Uzasadniony dla małych kostek, często przetwarzanych

Projektowanie agregatów

- System pozwala na automatyczne tworzenie agregatów na podstawie:
  - Przewidywanej objętości na dysku
  - Przewidywanego wzrostu wydajności
- Istnieje także możliwość monitorowania zapytań MDX trafiających na serwer od klientów
- Na podstawie statystyk uzyskanych w ten sposób można trafniej określić agregaty

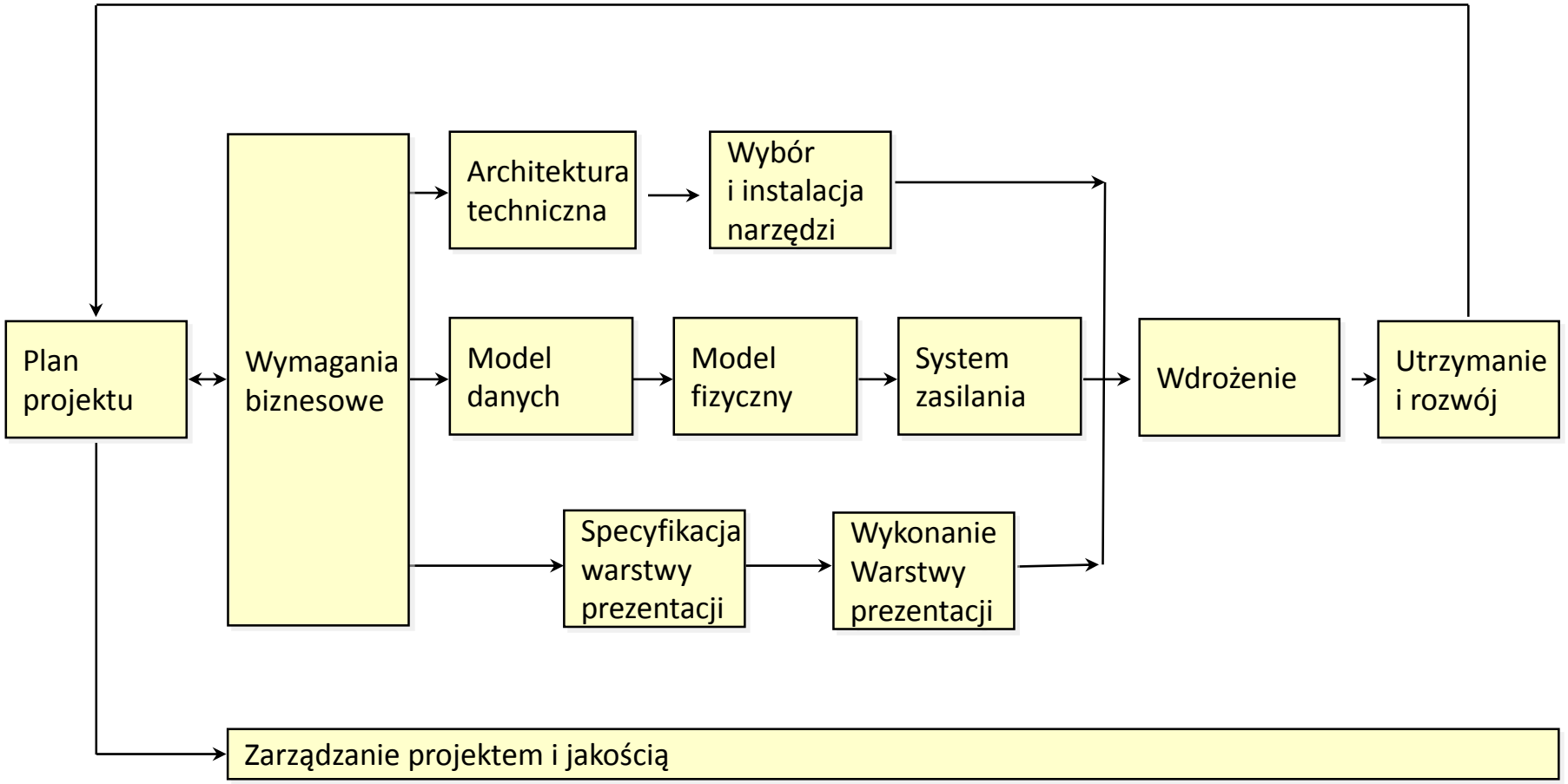
CYKL PROJEKTOWY HURTOWNI DANYCH

Przyczyny niepowodzeń projektów

- 1. Niepełne wymagania - 13.1%
- 2. Brak zaangażowania użytkowników - 12.4%
- 3. Brak zasobów - 10.6%
- 4. Nierealne wymagania - 9.9%
- 5. Brak wsparcia ze strony wyższych szczebli zarządzania - 9.3%
- 6. Zmienne wymagania - 8.7%
- 7. Brak planowania - 8.1%
- 8. System przestał być potrzebny - 7.5%
- 9. Błędy w zarządzaniu - 6.2%
- 10. Nieznajomość technologii - 4.3%
- 11. Inne - 9.9%

źródło: Chaos Report, http://www.standishgroup.com

Hurtownia danych jako proces



Cykl projektowy

- Faza 1: Uruchomienie projektu
  - zdefiniowanie zakresu prac
  - wybór architektury hurtowni danych
  - ustalenie obsady projektu
  - utworzenie harmonogramu projektu
  - określenie wymagań technicznych systemu
  - podział na tematy (określenie przyrostów)

- Faza 2: Wybór tematu
  - w pierwszej kolejności przyrost dający najwięcej korzyści przy najmniejszych kosztach i najmniejszym ryzyku

Cykl projektowy

- Faza 3: Analiza wymagań
  - ścisła współpraca z przyszłymi użytkownikami systemu a projektantami hurtowni
  - cykliczne pojawianie się potrzeb informacyjnych
  - określenie praw dostępu dla poszczególnych grup użytkowników

- Faza 4: Ocena wykonalności
  - określenie dostępności danych
  - określenie jakości danych

Cykl projektowy

- Faza 5: Projekt systemu
  - modelowanie analiz
  - projektowanie struktur danych
  - projektowanie systemu zasilania
  - określenie metadanych

- Faza 6: Implementacja
  - stworzenie struktur danych
  - inicjalne zasilenie systemu
  - skonstruowanie aplikacji i raportów
  - testowanie

Cykl projektowy

- Faza 7: Szkolenia
  - przygotowanie materiałów szkoleniowych
  - szkolenie administratorów
  - szkolenie użytkowników

RYZYKA ZWIĄZANE Z HURTOWNIĄ DANYCH

Ryzyko techniczne

- Nowe narzędzia
  - specjalne tryby pracy baz relacyjnych
  - bazy wielowymiarowe/in memory
  - narzędzia dostępu do danych
  - Cloud computing
  - Big data
- Nowe techniki
  - modelowanie wielowymiarowe
  - czyszczenie danych
  - dystrybucja raportów
  - przetwarzanie rozproszone

Ryzyko biznesowe

- Współpraca
  - wielu wydziałów
  - wielu firm
  - wielu dostawców
- Zakres
  - duża część organizacji
  - założenia powiązane ze strategią
  - trudności z zapewnieniem udziału kluczowych użytkowników

Dziękuję za uwagę