



SYBASE®



Technologie efektywnego zarządzania danymi w systemach rejestrowych na przykładzie IRS

Jacek Barańczuk, Dyrektor ds. Technologii i Usług
Tomasz Piętka, Business Development Manager
General Business & Public Sector

Miedzeszyn, 24 września 2009



Agenda

- Misja IRS
- Skąd pomysł na budowę hurtowni danych?
- Cele jakie postawiono przed hurtownią danych
- Wymagania odnośnie Hurtowni Danych
- Architektura Hurtowni Danych
- Działanie Hurtowni Danych
- ROI - mierzalny sukces
- Wyniesiona lekcja





IRS - wprowadzenie

Co to jest IRS?

IRS to Urząd Podatkowy USA, jest biurem Departamentu Skarbu USA i jedną z najbardziej efektywnych instytucji jeśli chodzi o administrowanie podatkami

- W 2004, IRS zebrał ponad 2 tryliony USD dochodu i rozliczył ponad 224 miliony podatków od dochodów
- IRS wydał tylko 44 centy za każde zebrane 100 USD w 2005r.
- Działa od 1913 na mocy 16 poprawki do Konstytucji USA
- Inicjalna stopa podatkowa?
1% ale w przypadku gdy dochód wynosił ponad 3 tys. USD
6% - dodatkowa opłata przy zarobkach większych niż 500 tys. USD



Misja organizacji

Misja IRS:

- Świadczyć amerykańskim podatnikom usługi na najwyższym poziomie poprzez pomoc w zrozumieniu i przestrzeganiu prawa podatkowego oraz stosowaniu spójnego i sprawiedliwego prawa dla wszystkich obywateli
- Rolą IRS jest pomoc podatnikom, którzy w większości są *uczciwi* w przestrzeganiu prawa podatkowego, oraz wpłynąć na mniejszość, która uchyla się od płacenia podatków

**Cele: IRS Research, Analysis, and Statistics
Organization**

**Dostarczać IRS aktualnych i kompletnych
informacji na temat każdego płatnika.**

SYBASE®



Budowa hurtowni danych – skąd pomysł?

Zrozumienie problemu klienta:

- Istniejący zbiór danych był bardzo ograniczony
- Zapytania umożliwiały limitowaną analizę danych
 - Brak możliwości raportowania ad-hoc
- Czas odpowiedzi na zapytania był mierzony w tygodniach a niekiedy w miesiącach!
 - Każde zaangażowane zapytanie wymaga i transportuje nowy zestaw danych z jednego lub więcej zdalnych systemów baz danych
- Aplikacje OLAP firm trzecich nie mogły bezpośrednio dostać się do istniejących systemów (zarówno polityczne jak i techniczne problemy)



Cele hurtowni danych

Zdefiniowanie celów nowej hurtowni danych:

- Jeden, skonsolidowany projekt bazy danych (3-cia postać normalna) zarówno dla płatników osób fizycznych jak i firm
- Każda informacja o podatku dochodowym od osoby fizycznej czy prawnej przechowywana i gotowa do natychmiastowej analizy
- 10 letnia informacja o podatniku
- Czas odpowiedzi na zapytania w godzinach, minutach, sekundach.
- Możliwość wykorzystania narzędzi Olap firm trzecich bezpośrednio na hurtowni danych

Widoczne korzyści:

- Miliardy dolarów zebrane w nieściągalnych podatków dochodowych
- Wiele nagród i wyróżnień ze strony rządu i przemysłu: TDWI, GCN, ComputerWorld, Excellence.gov, etc.



Wymagania odnośnie hurtowni danych

Przełożenie celów na wymagania projektowe:

- Kierownicy wyższego szczebla w roli wsparcia i sponsora
- Platforma Unix dla serwerów produkcyjnych i deweloperskich
- Zdefiniowanie modelu danych
- „Rozwiązanie” ETL
- Serwer analityczny (nie, tradycyjny serwer bazodanowy)
- Całkowita przestrzeń na dane: 20 TB, wielkość bazy danych: 2-3 TB
- Wsparcie dla 20-50 niezależnych użytkowników
- Procesy walidacji oraz zapewniania jakości danych
- Najlepsze z możliwych wsparcie dla aplikacji SAS

SYBASE®



Serwer dla hurtowni danych

Wybór najlepszych rozwiązań

- Sybase IQ wybrany jako serwer analityczny
 - Lepszy niż konkurencyjne rozwiązania w wielu kluczowych w tym projekcie obszarach, m.in. ładowania danych, wydajności zapytań, wymagań sprzętowych – zweryfikowano podczas testów PoC
 - Możliwość efektywnego wykorzystania modelu w 3rd postaci normalnej
- Standardowa platforma Unix'owa, sieci, pamięci masowe zostały rozbudowane i odpowiednio zaktualizowane
- SAS, BRIO (Hyperion), oraz SPSS Clementine były wiodącymi rozwiązaniami rynkowymi dla celów analitycznych - wymagane przez użytkowników

Sybase IQ – co go charakteryzuje

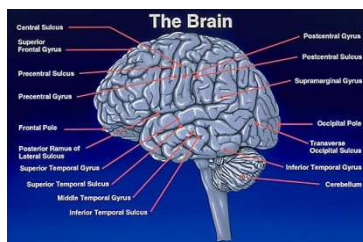


Kolumnowy
zapis danych



SQL

Standardowe
narzędzia



Zaawansowany
optimalizator

Sybase IQ



Bezpieczeństwo



Zasilanie w czasie
rzeczywistym
(24x7)



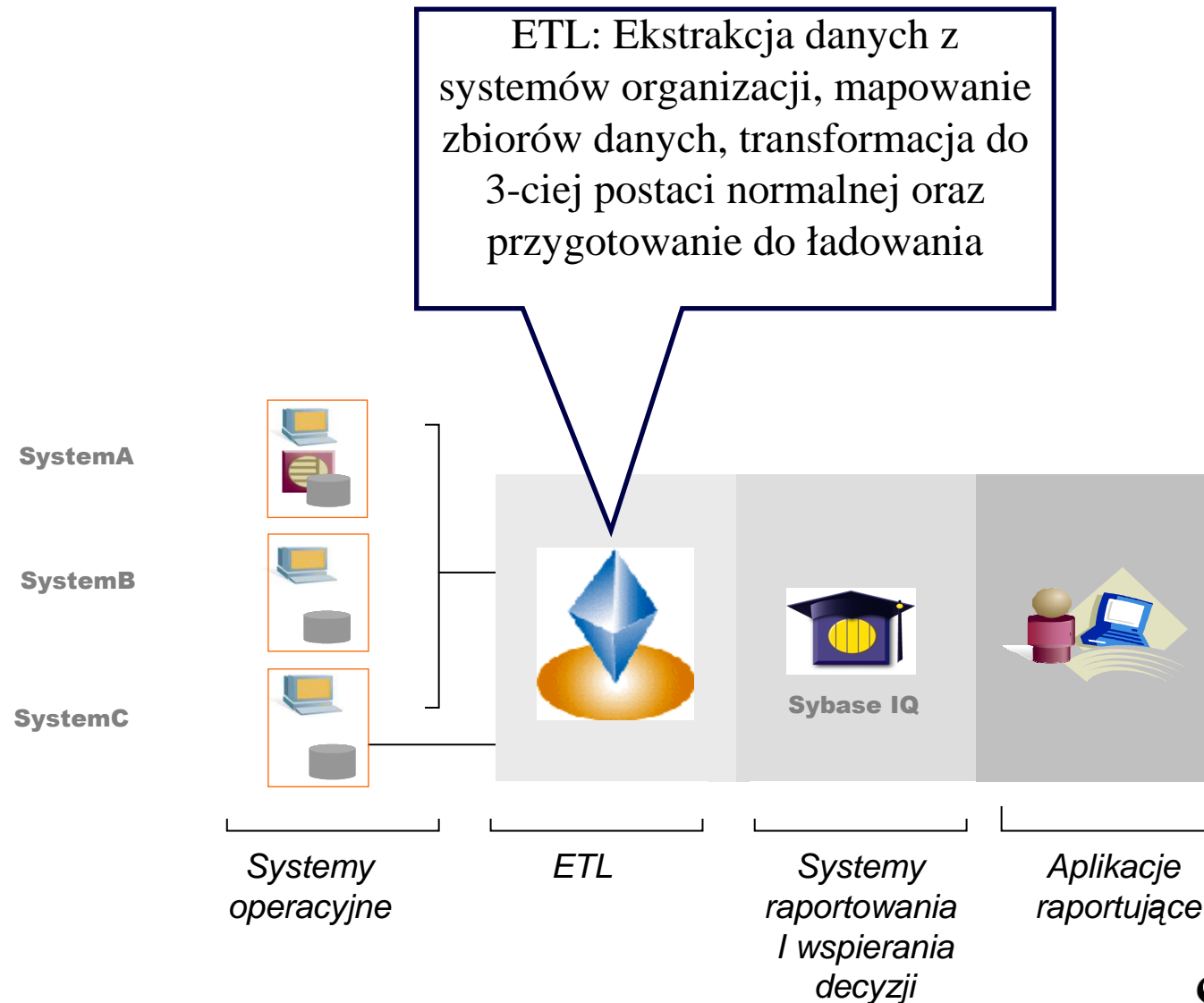
Efektywne wykorzystanie
sprzętu



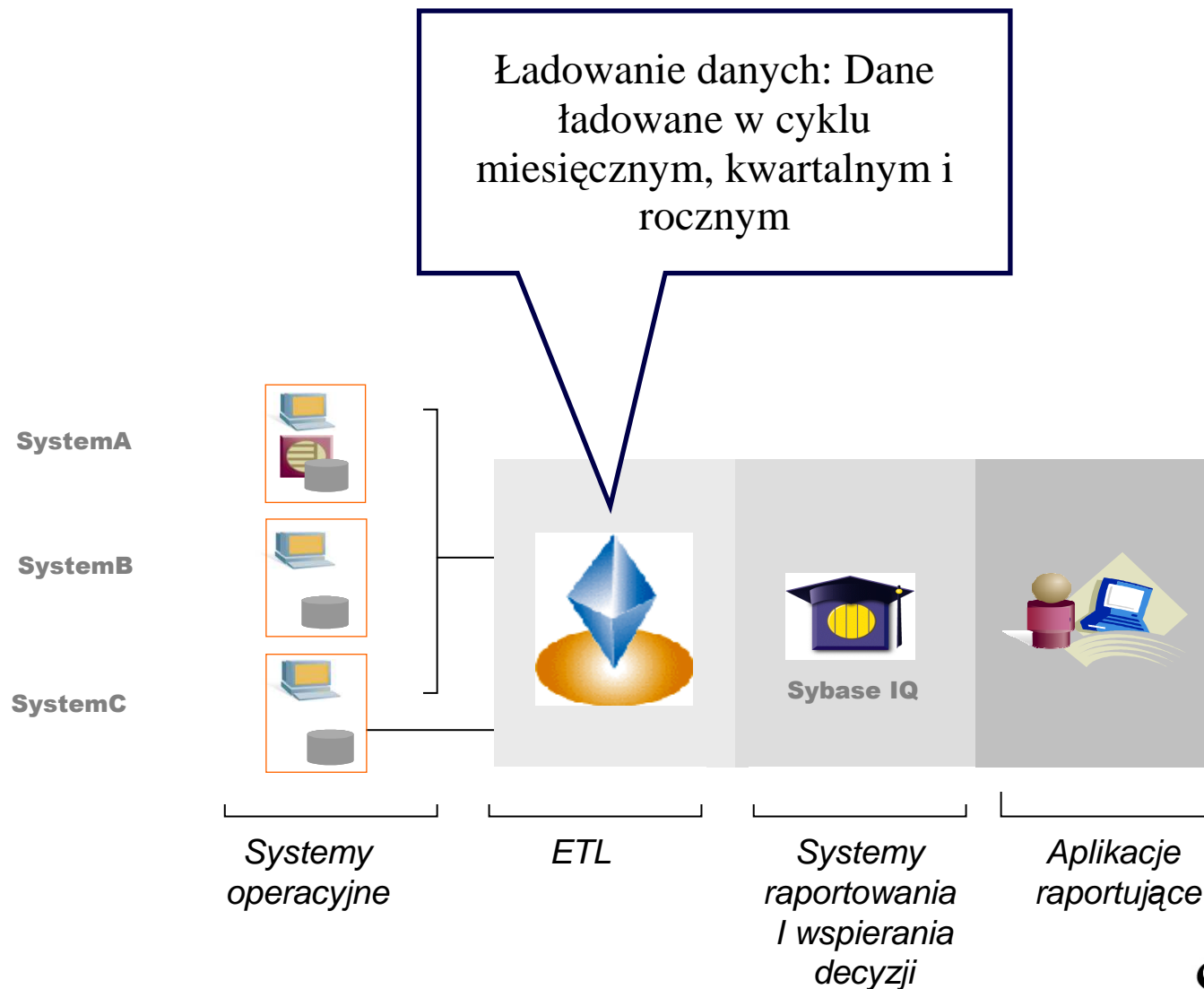
Łatwa administracja,
niski koszt

SYBASE®

Operacje hurtowni danych

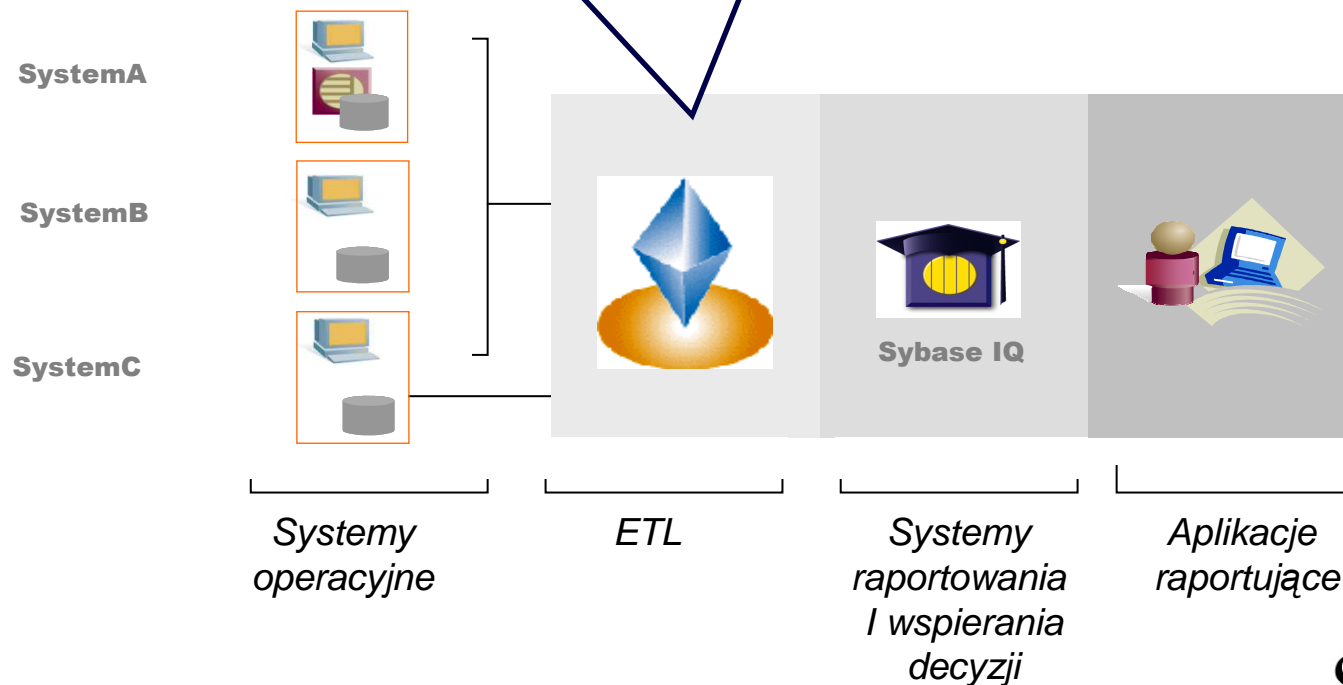


Operacje hurtowni danych



Operacje hurtowni danych

Walidacja danych:
Weryfikacja i walidacja poprawności danych w chwili ich przechwytywania z systemów produkcyjnych
+
Informowanie źródłowego systemu o nieprawidłowości



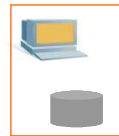
Operacje hurtowni danych

Dostępność: Utrzymanie stałej synchronizacji danych pomiędzy serwerem produkcyjnym oraz zapasowym

SystemA



SystemB



SystemC



Systemy operacyjne



ETL



Sybase IQ

Systemy raportowania i wspierania decyzji



Aplikacje raportujące



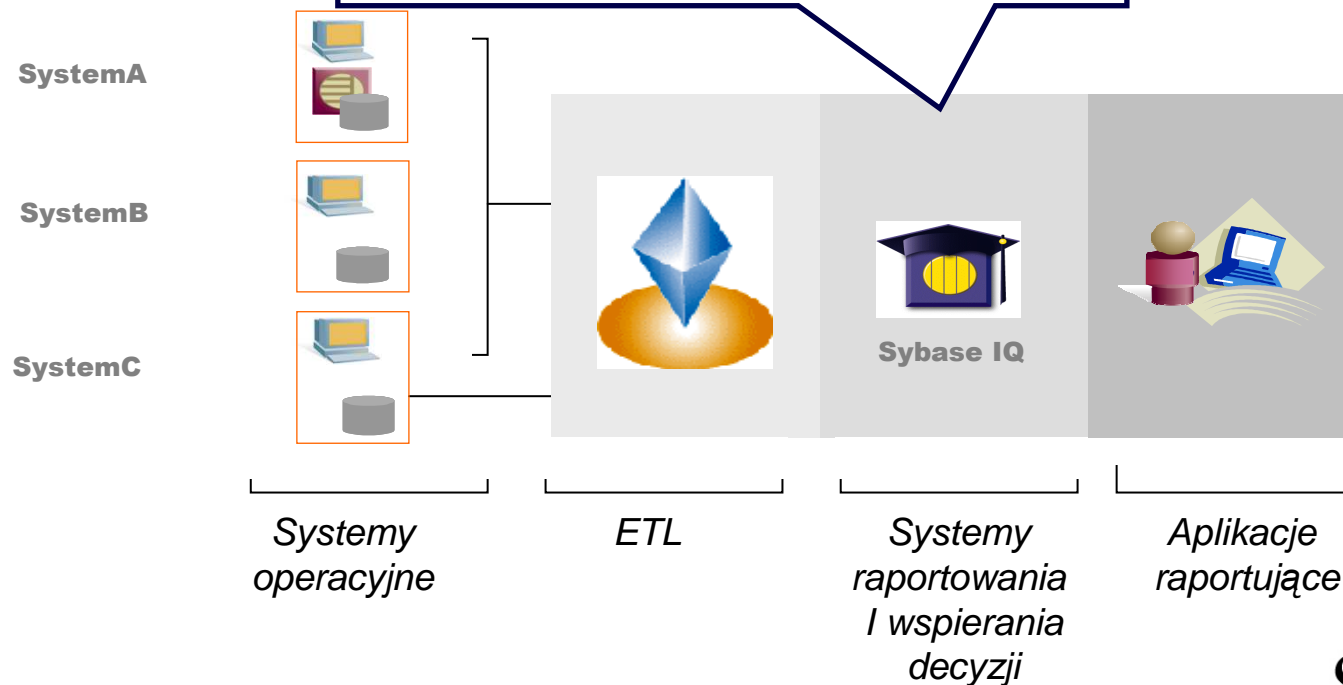
Sybase IQ

Środowisko zapasowe



Operacje hurtowni danych

Programowanie: Modelowanie i analiza nowych zbiorów danych oraz zmian w istniejących zbiorach danych gdy zmieniane jest prawo podatkowe oraz formularze, implementacja zmian
+
„Marketing”





Jak środowisko działa dzisiaj*

- CDW zarządza ponad 150 terabajtami (TB) przestrzeni, dzięki czemu jest największą hurtownią danych dla sektora federalnego
- Bieżący przyrost hurtowni danych na poziomie ponad: 20 TB/rok
- Liczba tabel w CDW: około 700
- Liczba kolumn: około 30 000, indeksów: około 45 000
- Przybliżona liczba użytkowników: ponad 400
- Średni roczny przyrost użytkowników: 100
- Średnia liczba równocześnie pracujących użytkowników systemu: około 60 z maksymalną liczbą ponad 100
- Przybliżona liczba zapytań/miesiąc: ponad 4 000

* Stan na lipiec 2008



Czynniki kluczowe dla sukcesu

- Wydajność zapytań
 - 10-100x szybsze niż tradycyjny RDBMS
- Zasilanie bazy danych
 - Jeden krok, Wysoka wydajność, Jednoczesne przetwarzanie Ładowania/Indeksowania
- Wydajność dyskowych operacji We/Wy
 - Duże blokowe operacje We/Wy znacząco szybsze niż w tradycyjnych DBMS
- Łatwość administrowania
 - Koszt administratora jest ułamkiem kosztu przy tradycyjnym DBMS
 - Dowolny schemat (3NF, wielowymiarowy, etc.)
- Wydajność składowania
 - Typowo konsumuje 30-70% przestrzeni, w porównaniu do danych surowych porównując do 300-900% przy tradycyjnych podejściach



Zwrot z inwestycji dla CDW (ROI)

- Produkcyjność badaczy wzrasta
 - 25% przyrost rocznej produktywności w oparciu o poprawę wydajności dzięki zastosowaniu Sybase IQ oraz jej sposobu indeksowania
 - Możliwość wykonywania skomplikowanych zapytań, zapytań ad-hoc oraz szybkiego wykonywania analiz predykcyjnych
np. zapytanie na bazie konkurencyjnej trwało 14 godzin – na Sybase IQ 40 sekund
- Obniżenie kosztów utrzymania w IT
 - Sybase IQ nie wymaga specjalizowanej optymalizacji
 - Prostszy w utrzymaniu i zarządzaniu
 - Wymaga mniejszej liczby pracowników zaangażowanych w zarządzanie bazą danych, przetwarzanie, ładowanie, konfigurację, strojenie oraz optymalizacje
- Niskie koszty szkoleń
 - Wiele rozwiązań wymaga wielu szkoleń – w tym przypadku tylko jedno rozwiązanie zaangażowane
- Redukcja kosztów sprzętu oraz pamięci
 - Działa na mniej wymagającym sprzęcie: wymagane mniej CPU, serwery średniej klasy
 - Kompresja IQ: Około 70% mniej przestrzeni dyskowej w stosunku do surowych danych przy w pełni poindeksowanych danych

SYBASE®

Czego się nauczono



Biznes

- Uzyskaj pełne wsparcie ze strony menedżerów średniego szczebla
- Upewnij się, że użytkownicy końcowi są w zespole
- Wybierz najlepsze rozwiązania rynkowe (nawet jeżeli jeszcze nie są Ci znane) ale proś o Proof of Concept!!!
- Utwórz odpowiedni zespół
- Twórz rozwiązanie dla Twoich użytkowników
- Mierz swój ROI oraz rzeczywiste korzyści dla organizacji

Strona techniczna

- Zrozum swoje wymagania
- Upewnij się, że aplikacje są w pełni wspierane
- Zainwestuj w swoją infrastrukturę
- Stwórz partnerstwo z wiarygodnymi dostawcami, z którymi możesz mieć długotrwałe relacje
- Upewnij się, że Twoje rozwiązanie jest w pełni skalowalne
- Mierz swój ROI