Potop Informacyjny, czyli jak podejść do Big Data

29 grudnia 2015

Borys Stokalski

Aby nadążać za potrzebami analitycznym w tempie odpowiadającym szybkości i zmienności współczesnego biznesu konieczna staje się też zdolność do odkrywania i szybkiego wdrażania heurystyk – szybkich metod wnioskowania.

Pojęcia promujące innowacje ze świata szybko zmieniających się technologii IT są często niejednoznaczne, nieprecyzyjne, oraz nadużywane do marketingu istniejących koncepcji i rozwiązań zgodnie z retoryką „przecież nasze produkty robią TO od zawsze". Te nadużycia nie powinny jednak przesłaniać istoty innowacji, mającej nierzadko charakter fundamentalny i długofalowy. Dlatego warto przyjrzeć się jednemu z najważniejszych stosowanych dziś słów-wytrychów, jakim jest Big Data.

Efekt taniejącego przetwarzania danych

Podstawowym motorem przemian w informatyce jest funkcjonujące od wielu dekad wykładniczy wzrost efektywności i spadek jednostkowych kosztów przechowywania, przetwarzania i transmisji danych. Ten prosty wzrost ilościowy prowadzi w rozwoju

technologii regularnie do zmian jakościowych, kiedy to, co dotąd było trudno osiągalne, wyrafinowane i drogie staje się powszechnie dostępne, proste w użyciu i relatywnie tanie. Proces ten doprowadził teleinformatykę do stanu dzisiejszego, gdy liczba nowych procesorów na rynku przekracza 10 mld sztuk rocznie, liczba osób korzystających z mobilnej telefonii wyniosła na początku 2013 roku 6,8 mld, a ślad cyfrowy – dane produkowane przez naszą cywilizację – osiągnęły w 2012 roku blisko 280 exabajtów.

W erze Big Data konieczne staje się tworzenie „fabryk modeli analitycznych", charakteryzujących się dostępnością bardzo dużej liczby predefiniowanych zmiennych i narzędzi pozwalających na szybkie stworzenie i przetestowanie ich.

Podstawowa funkcjonalność tej cyfrowej tkanki świata to zapewnienie dostępu do informacji oraz automatyzacja rutynowych procesów. Widać w związku z tym gdzie postęp w teleinformatyce staje się dziś kluczowy. Musimy nauczy się budować inteligentne rozwiązania, które będą w stanie przetworzyć ten zalew danych w użyteczną, potrzebną w danym miejscu i czasie wiedzę, a na podstawie tej wiedzy wpływać na działania ludzi i urządzeń. W przeciwnym wypadku exabajty generowanych co dzień danych będą w ogromnej mierze jedynie cyfrowym śmieciem bezsensownie obciążającym infostrukturę cywilizacji, pochłaniającym przestrzeń dyskową i energię potrzebną na jej utrzymanie.

Nieaktualne koncepcje analizy danych

W obszarze analizy danych podstawowe koncepcje architektoniczne, technologie, narzędzia i zasady governance dla obszaru pozyskiwania, integracji i analizy danych dla potrzeb procesów decyzyjnych powstały w latach 90. XX wieku. Koncepcje te są również i dzisiaj podstawą do budowy wydajnych rozwiązań Business Intelligence (BI), zapewnienia jakości danych oraz zarządzania relacjami pomiędzy światem zarządzania informacją, a światem automatyzacji działań operacyjnych. Powstały jednak na bazie założeń, które zaczynają dzisiaj coraz bardziej tracić znacznie.

Po pierwsze im bardziej w przedsiębiorstwach skracają się cykle biznesowe i narasta zjawisko hiperkonkurencji, tym większym problemem zaczyna być podstawowy paradygmat klasycznego BI – rozdzielenie procesów podejmowania decyzji od procesów operacyjnych. Przewaga informacyjna – według analityków rynku jeden z kluczowych mechanizmów zapewnienia konkurencyjności przedsiębiorstw w nadchodzących latach – to zdolność do szybszego podejmowania trafnych decyzji na „pierwszej linii" – w procesach sprzedaży, personalizacji usług i produktów, obsługi klienta. Przesuwa to wymagania architektoniczne architektury BI od zapewnienia zdolności do integracji i wielowymiarowej analizy danych historycznych w kierunku zdolności udostępnienia procesom operacyjnym usług analitycznych i predykcyjnych (prognozujących przyszłe zdarzenia). Wymaga to szybkiego dostępu do wysokiej jakości danych oddających aktualny stan ważnych dla biznesu obiektów (klientów, procesów, zasobów organizacji) oraz udostępniania ich w sposób wspierający automatyzację i interoperacyjność procesów biznesowych.

Konieczność zmiany podejścia do analizy danych

Szybkości i zmienności współczesnego biznesu wymaga odkrywania i szybkiego wdrażania heurystyk – szybkich metod wnioskowania. Z jednej strony pozwalają one zastąpić złożone,

pracochłonne przetwarzanie analityczne, z drugiej zaś pozwalają szybko zastępować modele, które z czasem tracą znaczenie gdyż przestają adekwatnie opisywać zmieniającą się rzeczywistość. To zjawisko zostało zauważone w latach 80. XX wieku przez płk. Johna Boyda, amerykańskiego pilota, stratega i „filozofa konfliktu", twórcy modelu cyklu adaptacyjnego (OODA) stanowiącego fundament pojęciowy wielu współczesnych doktryn prowadzenia działań wojennych.

W związku z tymi zmianami pojawia się w architekturze BI nowy wzorzec architektoniczny – można go nazwać „fabryką modeli analitycznych", czy „fabryką heurystyk". Cechą charakterystyczną takiej „fabryki", jest dostępność bardzo dużej liczby predefiniowanych zmiennych i narzędzia pozwalające na szybkie stworzenie i przetestowanie ich z wykorzystaniem modelu analitycznego. Przykładem może być fabryka modeli wykorzystująca kilka tysięcy różnych zmiennych opisujących sieci połączeń telefonicznych, służąca do doskonalenia modeli scoringowych dla kampanii marketingowych operatora.

Kontekst, w którym funkcjonuje klient

Kolejnym aspektem staje się to, że już w niedługim czasie gros informacji kluczowych dla biznesu (takich jak informacje charakteryzujące klientów) stanowić będą informacje, nad którym przedsiębiorstwa będą miały bardzo ograniczoną kontrolę. Personalizacja usług i produktów (ta ostatnia istotnie wzmocniona przez nowe metody produkcji wykorzystujące druk trójwymiarowy) wymaga znajomości szerokiego kontekstu, w którym funkcjonuje klient. Trudno dziś powiedzieć, kto może zostać brokerem takiej informacji. Najpoważniejszym kandydatem są sieci społecznościowe, instytucje finansowe i operatorzy mobilni. Dodatkowo ważną rolę w kształtowaniu się mechanizmów dostępu do takich informacji odegrają regulatorzy działający w obszarze ochrony prywatności.

Musimy nauczy się budować inteligentne rozwiązania, które będą w stanie przetworzyć ten zalew danych w użyteczną wiedzę. W przeciwnym wypadku exabajty generowanych co dzień danych będą w ogromnej mierze jedynie cyfrowym śmieciem.

Tak, czy inaczej tradycyjne architektury i rozwiązania BI – zakładające, że analizujemy dane tworzone i zarządzane wewnątrz przedsiębiorstwa – nieuchronnie odchodzą w przeszłość. Powoli można przestać zakładać, że wszystkie dane poddawane analizie można sprowadzić do jednolitej, wystandaryzowanej postaci. W przypadku informacji pełnotekstowych czy multimedialnych jest to niemożliwe. Dlatego nowa generacja rozwiązań analitycznych musi uwzględniać integrację zewnętrznych źródeł informacji – zapewne raczej w postaci usług niż w postaci dostępu do danych źródłowych, jak ma to miejsce w tradycyjnych architekturach BI. Takie rozwiązania pojawiają się dzisiaj np. w postaci analizatorów treści pełnotekstowych zapisów na fakturach elektronicznych lub opisów transakcji elektronicznych w celu automatycznej ich klasyfikacji.

Czas zacząć pracę u podstaw

Nie dysponujemy dziś przemysłowym, szeroko zweryfikowanym zestawem dobrych praktyk pozwalających w sposób powtarzalny radzić sobie z nietrywialnymi zagadnieniami Big Data – łączącymi w sobie takie elementy, jak ekstremalne wolumeny, zróżnicowanie źródeł i formatów, ekstremalną zmienność. Nie dysponujemy nawet dobrą taksonomią takich

zagadnień, w odróżnieniu od klasycznego BI, które stanowi jeden z lepiej uporządkowanych obszarów inżynierii oprogramowania, w którym wzorce architektoniczne, procesy wytwórcze, technologie, zagadnienia biznesowe i metody analizy tworzą dość jasną i spójną mapę. Widać jednak coraz więcej przykładów potwierdzających, że rozwiązywanie problemów Big Data prowadzić może do zupełnie nowej informatyki. Skoro słynny system Watson ze znanego eksperymentu IBM był w stanie wygrać z ludźmi już nie w szachy, a w teleturnieju w Polsce znanym jako „Va Banque" nic dziwnego, że system wykorzystujący podobne mechanizmy może zacząć pełnić rolę asystenta-researchera skutecznie wspierającego diagnostykę medyczną.

Zautomatyzowana predykacja staje się dziś podstawą do selekcji „playlist" w stacjach radiowych, algorytmy zaczynają też sprawdzać się lepiej od ludzi w procesach naboru i selekcji kandydatów do pracy. Realna staje się wizja auta bez kierowcy, samolotu bez pilota, czy wizja wirtualnych asystentów załatwiających w świecie dostępnych w sieci usług w naszym imieniu. Uprzemysłowienie informatyki w obszarze Big Data – zdefiniowanie metodyk, standaryzacja narzędzi i usług – jest ważnym kamieniem milowym na drodze do znacznie bardziej inteligentnej teleinformatyki niż ta, którą znamy dziś.

Dla twórców narzędzi analitycznych:

- · W niedługim czasie gros informacji kluczowych dla biznesu (takich jak informacje charakteryzujące klientów) stanowić będą informacje, nad którym przedsiębiorstwa będą miały bardzo ograniczoną kontrolę.
- · Tradycyjne architektury i rozwiązania BI – zakładające, że analizujemy dane będące elementem systemu informacyjnego przedsiębiorstwa – nieuchronnie odchodzą w przeszłość.
- · Powoli można przestać zakładać, że wszystkie dane poddawane analizie można sprowadzić do jednolitej, wystandaryzowanej postaci. W przypadku informacji pełnotekstowych czy multimedialnych jest to niemożliwe.
- · Nowa generacja rozwiązań analitycznych musi uwzględniać integrację zewnętrznych źródeł informacji – zapewne raczej w postaci usług niż w postaci dostępu do danych źródłowych.

Cele biznesowe wykorzystania Big Data

- 49% analiza danych i zachowań klientów końcowych
- 18% optymalizacja działań operacyjnych
- 15% zarządzanie ryzykiem i finansami
- 14% przygotowanie nowych modeli biznesowych
- 4% poprawa współpracy wewnątrz firmy

Źródło – Raport IBM „2012 Analytics Study: The real-world use of Big Data" powstały na podstawie badania na 1144 przedstawicielach biznesu i IT w 130 krajach.

Stosunek IT i biznesu do Big Data

- 47% planowane wykorzystanie narzędzi Big Data
- 28% pilotowe wdrożenia Big Data

- 24% brak działań związanych z Big Data
- 1% brak danych

Źródło – Raport IBM „2012 Analytics Study: The real-world use of Big Data" powstały na podstawie badania na 1144 przedstawicielach biznesu i IT w 130 krajach.

---

http://itwiz.pl/potop-informacyjny-czyli-jak-podejsc-big-data/