ZASTOSOWANIE NARZĘDZI STATYSTYCZNYCH DO ANALIZY DUŻYCH WOLUMENÓW DANYCH W ADMINISTRACJI RZĄDOWEJ

Mirosław Błażej
Ministerstwo Finansów, Departament Polityki Finansowej, Analiz i Statystyki

Wstęp

W obecnym świecie biznesu i administracji mamy do czynienia z ogromnym przyrostem gromadzonej informacji dotyczącej wielu aspektów zachowania się procesów technologicznych, konsumentów, klientów, gospodarstw domowych czy podatników.

Rozwój technologii informatycznych stwarza równolegle możliwości wykorzystania powyższych informacji. Dostępne stają się tak nośniki pamięci o odpowiedniej pojemności, procesory o odpowiedniej wydajności oraz systemy przesyłu informacji o odpowiedniej przepustowości. Ostatnie lata to także okres intensywnego rozwoju narzędzi informatycznych oferujących rozwiązania zapewniające dostęp do danych, coraz powszechniej stosowane są narzędzia i programy oferujące coraz bardziej efektywne algorytmy przetwarzające dane – algorytmy statystyczne, ekonometryczne i data miningowe.

O ile powyżej wymienione zostały elementy „podaży" rozwiązań w zakresie zaawansowanego przetwarzania danych, to głównym elementem skutkującym pojawieniem się „popytu" jest proces narastającej konkurencji pomiędzy podlegającymi koncentracji podmiotami gospodarczymi. Proces koncentracji jest tu istotny, gdy w jego wyniku konsolidowane są bazy danych[1], wzrastają możliwości firm w zakresie stosowania techniki IT i zatrudniania specjalistów - silnie oddziałuje tzw. efekt skali.

[1] Chodzi tu nie tylko o integrację i wzrost baz w sensie ilości obiektów zarejestrowanych w bazie (np. w wyniku połączenia baz danych o klientach dwóch banków), ale także łączenie i wzrost ilości użytecznych danych o obiektach, np. opisujących różne obszary zachowania się obiektów. Proces taki pozwala na znaczące zwiększenie użyteczności danych w bazie dla analiz, czyli użytkowej pojemności informacyjnej danych. Poprzez to pojęcie (niedefiniowane w sposób ścisły) użytkowej pojemności informacyjnej danych należy rozumieć zdolność grupy danych do identyfikowania nowych zależności i relacji (po zastosowaniu odpowiednich technik statystycznych, ekonometrycznych i data miningowych) istotnych z punktu widzenia identyfikowania nowej, użytecznej wiedzy. Pojecie to byłoby analogonem funkcjonującego w ekonometrii pojęcia (definiowanego w sposób ścisły) indywidualnej pojemności informacyjnej zmiennej lub integralnej pojemności informacyjnej danego zbioru zmiennych objaśniających (porównaj np. „Ekonometria", S. Dorosiewicz i in., SGH 1995, str. 19-20).

Przedstawienie zagadnienia

Powyższe procesy dotyczą oczywiście także administracji publicznej, szczególnie duże potencjalne zastosowania narzędzi analizy dużych wolumenów danych otwierają się w odniesieniu do ministerstw finansów, z uwagi na dużą ilość informacji gromadzonej w ramach systemów IT wspierających działalność służb podatkowych, celnych czy budżetowych, informacji dostępnej w formie sprzyjającej stosowaniu tego typu analiz. Instytucje te dysponują ponadto możliwością zgromadzenia odpowiednich zasobów, a potencjalnie duże korzyści z inwestycji w tego typu systemy i procedury analityczne skłaniają do podjęcia wysiłku ich rozwoju.

Poniżej chciałbym odnieść się do jednego z obszarów analitycznych obecnych w pracach Ministerstwa Finansów, a dotyczącego analiz w odniesieniu do systemu podatkowego, a zwłaszcza przewidywania skutków zmian w systemie podatkowym. Analizy wykonywane w ramach rozwijanego aktualnie systemu Sindbad dotyczą głównie właśnie tego wymiaru, ale także analizy sytuacji gospodarczej i społecznej na poziomie makro i mikroekonomicznym.

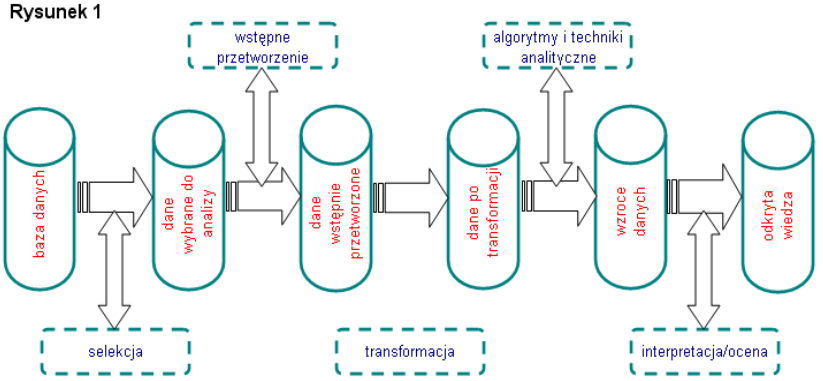
Analizy w zakresie systemu podatkowego, a zwłaszcza związane z szacunkami skutków budżetowych, konsekwencji zmian dla różnych grup podatników, oszacowanie wielkości zmiany efektywnej stawki podatkowej do wykorzystania jako szoku dla ekonometrycznego modelu gospodarki w celu określenia wpływu zmiany na dynamikę podstawowych kategorii ekonomicznych są częstymi zadaniami Departamentu Polityki Finansowej, Analiz i Statystyki czy departamentów podatkowych Ministerstwa. Wykonywane są one w dość specyficznych warunkach, które muszą być uwzględnione przy konstrukcji systemu analitycznego. Wśród tych warunków należy wymienić:

- czas: większość zadań jest poddana bardzo wymagającemu i bardzo silnemu ograniczeniu czasowemu, które jest pochodną wymogów kalendarza politycznego lub prac parlamentarnych bądź rządowych,
- zmienność: analizy dotyczą szerokiego wachlarza zagadnień, począwszy od polityki podatkowej, jak i zagadnień mikro-, makroekonomicznych, polityki społecznej czy budżetowej. Wiąże się także ze stosowaniem najrozmaitszych technik badawczych i analitycznych,
- sposób ich formułowania: zagadnienia analityczne formułowane są w języku politycznym i prawnym. Oznacza to konieczność „przekształcenia" ich na zagadnienia związane z analizą danych oraz ekonomiczne,
- koszt decyzji: choćby z uwagi na skalę (finansową, ilości podmiotów dla jakich analizowane zmiany niosą konsekwencję) koszt popełnienia błędu przy podejmowaniu decyzji jest bardzo wysoki.

Zastosowanie narzędzi IT do wyżej zarysowanych analiz można umiejscawiać w ramach *knowledge discovering in databases* (KDD), które jest nazwą ogólną procesu obejmującego uzyskanie użytecznej i nietrywialnej informacji ze zgromadzonego zbioru danych (najczęściej w postaci elektronicznej bazy danych); nietrywialnej, a więc niepodlegającej

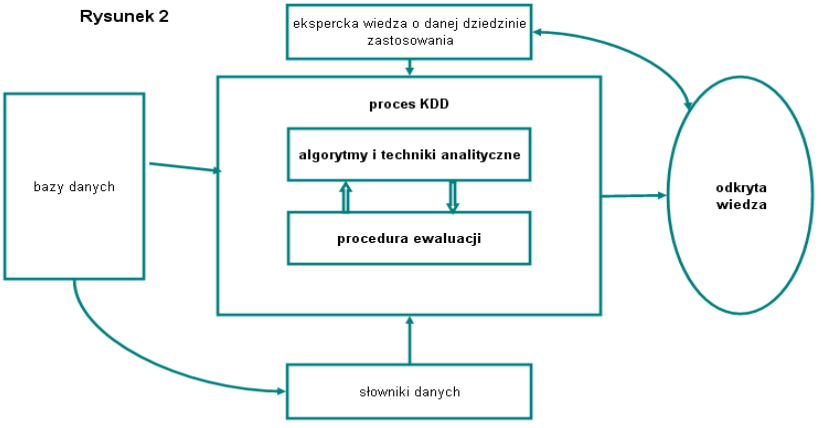
rozpoznaniu za pomocą prostych technik, jak wizualizacja czy podstawowe statystyki (np. średnia). Schemat procesu KDD zamieszczono na rys. 1, a jego kontekst oraz relacje ze źródłami danych na rys. 2, (za U. Fayyad and all 1996).

Rysunek 1



Rysunek: etapy procesu knowledge data discovery
Źródło: U.Fayyad and all 1996

Rysunek 2



Rysunek: proces knowledge data discovery
Źródło: U.Fayaad and all 1996

Rys. 1 i 2. Źródło: U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth „From data mining to knowledge discovery in databases" Artificial Intelligence Magazine, Fall 1996, 37-56 pp.

Jak można się w sposób oczywisty spodziewać (i jest to także zilustrowane na powyższych rysunkach) skuteczne i poprawne analizy bazujące na danych podatkowych wymagają łącznej wiedzy i znajomości szeregu obszarów:

- danych (źródła danych, sposób ich gromadzenia i przekształcania etc.),
- systemu podatkowego,
- zagadnień mikro- i makroekonomicznych,
- metod ilościowych (statystyka, ekonometria),

- finansów publicznych,
- narzędzi IT:
  - bazodanowych (bazy danych, SQL),
  - przekształcania i przetwarzania danych (zapisu algorytmów) (oprogramowanie statystyczno-ekonometryczne, języki programowania strukturalnego),
  - statystyczno-ekonometrycznego.

Opanowanie tak szerokiej wiedzy i umiejętności oraz przekształcenie zespołu osób i narzędzi informatycznych w skuteczne przedsięwzięcie stanowi spore wyzwanie, jednocześnie ma swoje konsekwencje dla budowy systemu IT wspierającego te analizy. Przykładem może być specjalna rola aplikacji do zarządzania metadanymi oraz dostępu do danych w systemie Sindbad.

System Sindbad

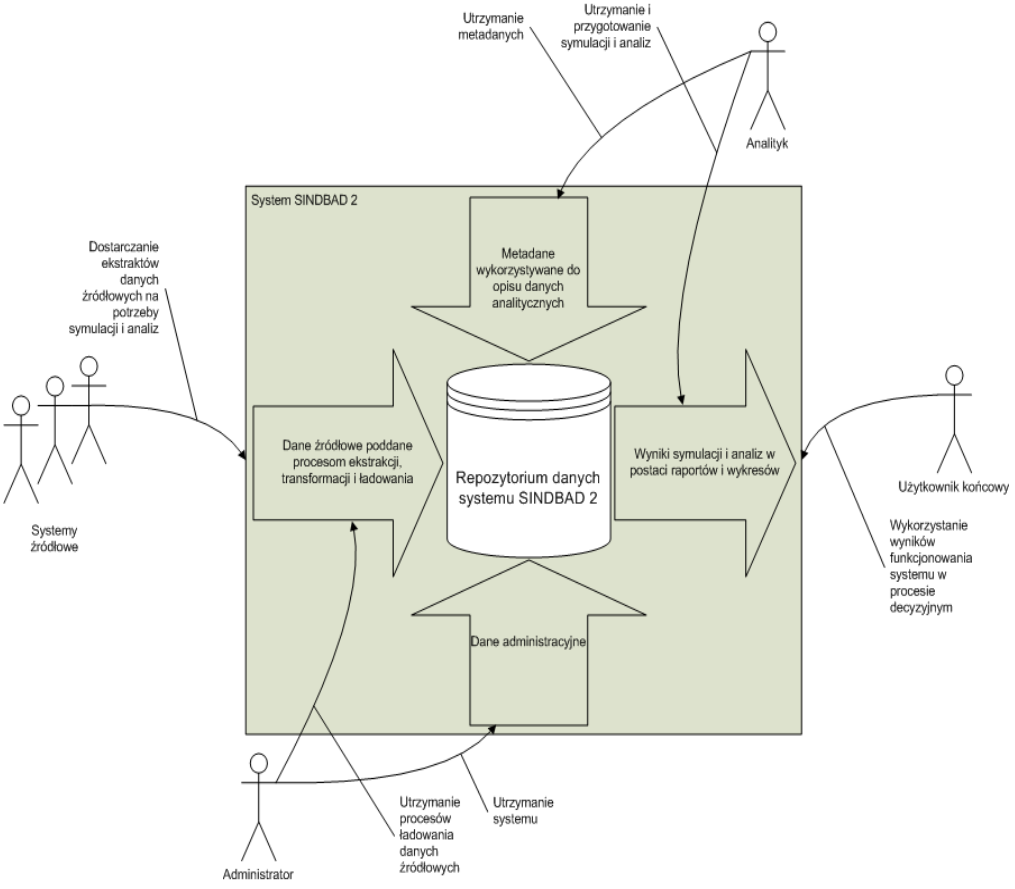
Podstawową przesłanką inicjatywy budowy systemu Sindbad było stworzenie dla analityków Ministerstwa Finansów (odpowiedzialnych za analizy w obszarze podatkowym, makroekonomicznym i dochodowych) zintegrowanego środowiska zapewniającego:

- dane w obszarze koniecznym dla analiz: zasadniczo dane z deklaracji podatkowych, ale także szeregów makroekonomicznych i budżetowych. Poprzez dane należy rozumieć nie dane „surowe", ale raczej w postaci przekształconej do kategorii mających sens z punktu widzenia podatkowo-ekonomicznego oraz zachowujących porównywalność pomiędzy różnymi okresami czasu,
- informacje o danych: podstawa prawna, zmiany zakresu kategorii w poszczególnych latach, relacje z innymi wielkościami etc., najlepiej informacji kontekstowej,
- wygodne narzędzia dostępu do danych i ich przekształcania: dla analityka nie będącego specjalistą w obszarze baz danych i programowania,
- szeroki zestaw narzędzi analitycznych, statystycznych i ekonometrycznych: pozwalających na budowę tak systematycznie tworzonych modeli, analiz i raportów, jak i umożliwiających wykonywanie niewielkich analiz i modeli ad-hoc. Oprogramowanie to powinno zapewniać dostęp do narzędzi analitycznych i modelowych, począwszy do stosunkowo prostych metod statystycznych, aż do wysoce wyspecjalizowanych algorytmów ekonometrycznych, jak analiza kointegracyjna,
- zestaw predefiniowanych modeli podatkowych: pozwalających na szybkie przeanalizowanie różnych alternatywnych rozwiązań podatkowych oraz ich skutków dla budżetu i podatników, we wszystkich wymiarach, jakie mogą mieć sens ekonomiczny lub społeczny, a możliwych do uzyskania na bazie dostępnych informacji (rozkłady, grupowania, przekroje etc.),
- wygodne, funkcjonalne i w miarę proste narzędzia administrowania: danymi i szeregami czasowymi, metadanymi, użytkownikami itd.

Dodatkowym wymogiem jest także, by środowisko to było zintegrowane i jednolite na tyle ile jest to możliwe z uwagi na różnorodność zadań, oraz elastyczne, by można uwzględnić częste zmiany w procesach generujących dane czy potrzebach analitycznych.

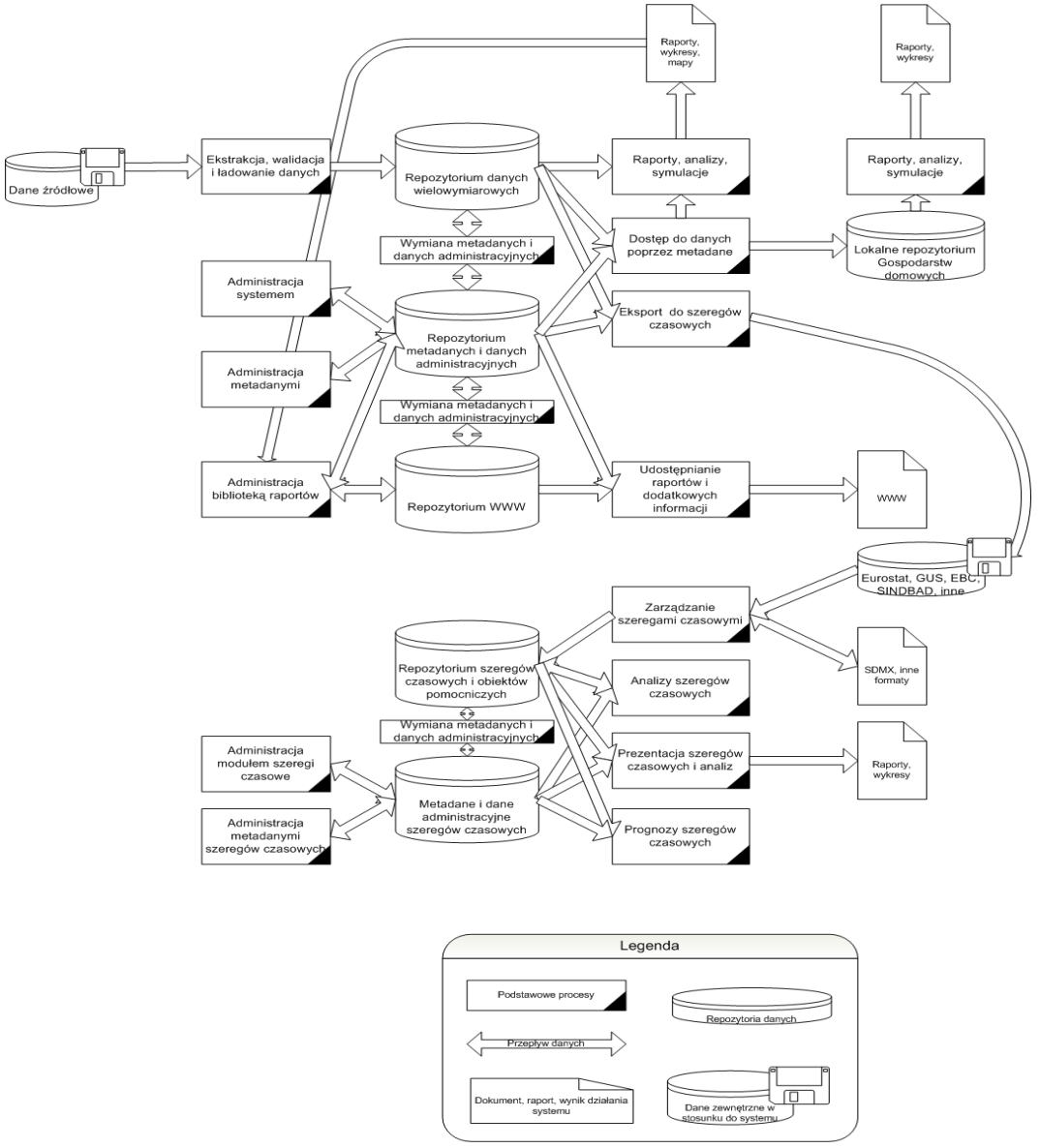
Projekt ten jest obecnie realizowany w Departamencie Polityki Finansowej, Analiz i Statystyki w postaci projektu Transition Facility 2005/017-488.01.04 „Rozwój wiedzy w dziedzinie analiz dochodów budżetowych" we współpracy z firmami ComArch S.A. oraz ITTI Sp. z o.o. Punktem wyjścia dla tego projektu (roboczo określanego jako projekt SINDBAD 2) był uprzednio zrealizowany projekt finansowany w ramach programu PHARE, ale obecne prace stanowią jego znaczące rozwinięcie, zwłaszcza co do możliwości analitycznych i symulacyjnych oraz w zakresie zarządzania szeregami czasowymi danych budżetowych i makroekonomicznych (aktualnie utrzymywanych i wykorzystywanych w pracach analitycznych Departamentu jest ok. 10 tysięcy różnych szeregów lub ich wariantów).

Najbardziej ogólny model funkcjonowania i wykorzystania systemu SINDBAD może zostać zilustrowany poniższym diagramem.



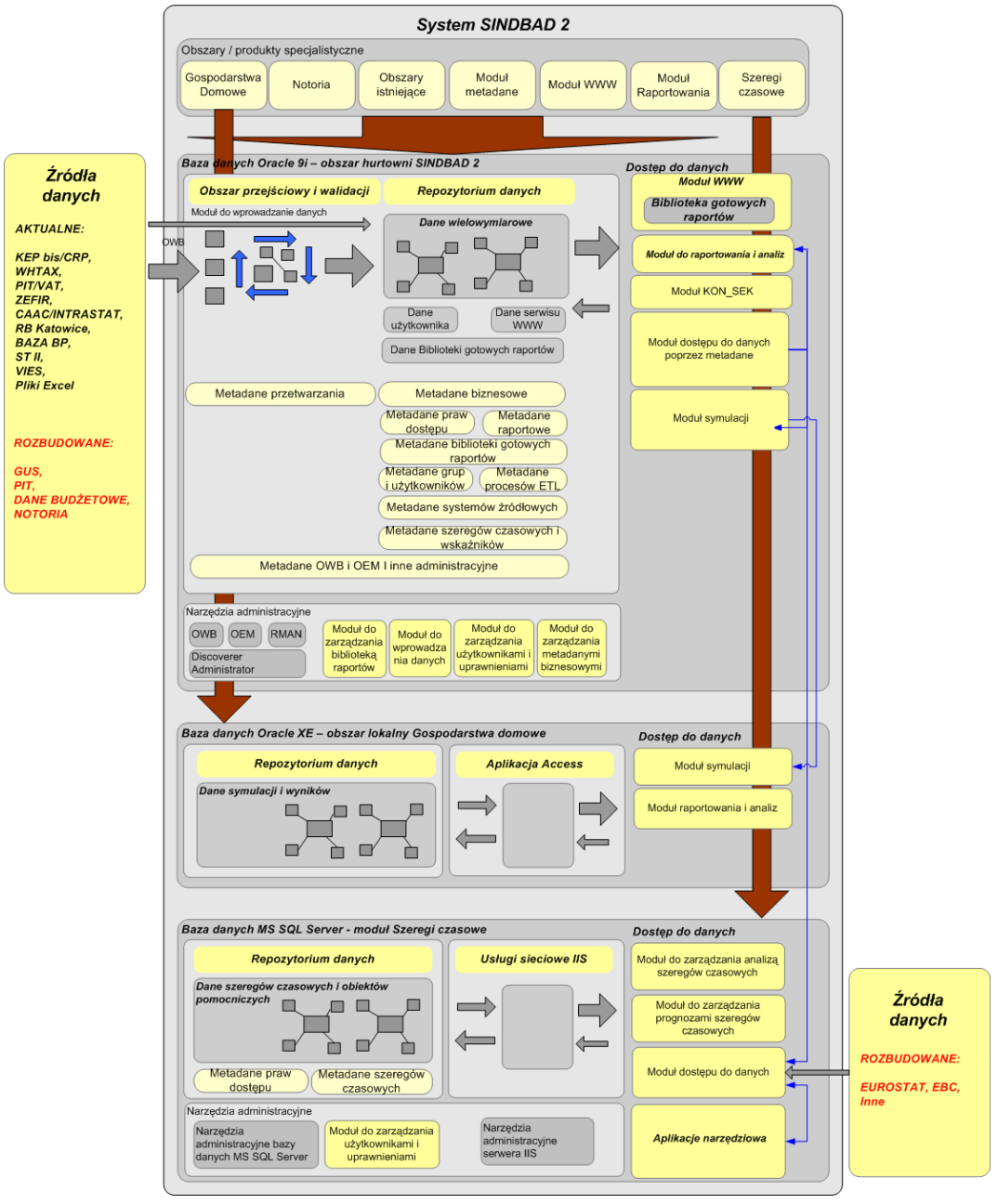
Rys. 3. Ogólny model funkcjonalny systemu SINDBAD i SINDBAD 2 (źródło: dokumentacja techniczna projektu, Projekt Generalny Systemu SINDBAD 2).

Natomiast przechodząc do bardziej szczegółowego opisu systemu, model jego podstawowych procesów i przepływu danych pomiędzy nimi można zilustrować poniższym diagramem.



Rys. 4. Model podstawowych procesów i przepływu danych pomiędzy nimi systemu SINDBAD 2 (źródło: dokumentacja techniczna projektu, Projekt Generalny Systemu SIINDBAD 2).

Poniżej została zaprezentowana architektura logiczna systemu SINDBAD 2.



Rys. 5. Architektura logiczna systemu SINDBAD 2 (źródło: dokumentacja techniczna projektu, Projekt Generalny Systemu SINDBAD 2).

Z uwagi na fakt, że budowa systemu jest w trakcie realizacji, należy podkreślić, że niektóre elementy przedstawione na ww. diagramie podlegają zmianom.

Ponadto na diagramie nie zaznaczono modułów analiz elastyczności podatkowych oraz analiz panelowych realizowanych na danych sytemu SINDBAD, ale w postaci niezależnych podprojektów analitycznych (badawczych). Między innymi w tym obszarze wykorzystana będzie *STATISTICA*.

Wyjaśnienia wymaga także pojęcie „obszarów/produktów specjalistycznych". Budowa systemu jest prowadzona właśnie w kontekście „produktów/obszarów specjalistycznych" wydzielonych ze względu na jednolitość zagadnienia biznesowego, jakiego dotyczą. Podstawowymi przykładami są tu obszary poszczególnych podatków (PIT, CIT, akcyza, VAT itd.), gospodarstwa domowe, elastyczności podatkowe lub analizy panelowe wydzielone z uwagi na jednolitość danych, jednorodność analizowanych tam zagadnień, a czasem także stosowanych metod analitycznych. Pozostałymi obszarami są np. moduły administracyjne, serwis WWW, moduł zarządzania raportami. Osobny obszar stanowi także moduł szeregów czasowych.