INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 2(40) • 2016

ISSN 1507-3858
e-ISSN 2450-0003

Ryszard Zygała
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: ryszard.zygala@ue.wroc.pl

MOŻLIWOŚCI ROZWIJANIA SYSTEMÓW ANALITYCZNYCH DLA MŚP W OPARCIU O ŚRODOWISKO JĘZYKA PYTHON

DEVELOPING ANALYTICAL SYSTEMS IN SMES BASED ON PYTHON ENVIRONMENT

DOI: 10.15611/ie.2016.2.08
JEL Classification: L86, O32, O33

Streszczenie: W artykule podjęta została problematyka różnych aspektów rozwoju systemów analitycznych w sektorze małych i średniej wielkości przedsiębiorstw (MŚP), z wykorzystaniem narzędzi *open source*, a w szczególności bibliotek języka Python. Głównym celem artykułu jest wskazanie na przykładzie środowiska języka Python, że w oprogramowaniu *open source* tkwi duży potencjał, który może być wykorzystany do rozwijania i eksploatacji systemów analitycznych w firmach sektora MŚP. Autor wyraża przekonanie, że firmy sektora MŚP nie tylko mogą, ale powinny rozważać wdrożenie strategii konkurencyjnych opartych na wysokiej jakości danych pozyskiwanych z systemów analitycznych. W tym celu mogą skutecznie rozwijać systemy analityczne w oparciu o oprogramowanie *open source* i metody wypracowane przez naukę o danych. Artykuł zawiera również identyfikację barier, przed którymi stają MŚP, decydując się na inwestowanie w technologie informacyjne. Kluczowym aspektem badawczym w artykule jest analiza funkcjonalności bibliotek języka Python, na podstawie której autor wykazuje, że poszczególne komponenty środowiska Python mogą wspomagać rozwój każdej warstwy systemu analitycznego, a zatem mogą one stanowić kompletną i solidną podstawę realizacji strategii opartej na wysokiej jakości danych pozyskiwanych z tego typu systemu w firmach sektora MŚP.

Słowa kluczowe: MŚP, systemy analityczne, *open source*, Python.

Summary: The paper describes different issues concerning the analytical systems development in small and medium-sized enterprises (SME). The main purpose of the paper is to demonstrate that SMEs not only can but also should consider to develop competitive strategies based on high quality of data gathered from analytical systems. In order to implement these strategies they can develop analytics based on open source software and data science methods. The paper analyses the key Python libraries, useful for data analysis and an analytical system development.

Keywords: SME, analytical systems, open source, Python.

1. Wstęp

W krajach wysoko rozwiniętych gospodarczo sektor małych i średnich przedsiębiorstw w istotnym stopniu wpływa na rozwój danego kraju. Ten sektor przedsiębiorstw stanowi ok. 90 % ogółu przedsiębiorstw w krajach Unii Europejskiej, jest miejscem pracy dla ok. 2/3 zatrudnionych pracowników i wypracowuje blisko 60% dochodu narodowego (por. [Eurostat 2016]). W Polsce struktura udziału omawianego sektora w gospodarce jest podobna jak w całej Unii Europejskiej, przy czym niższy jest udział sektora w dochodzie narodowym, bo wynosi ok. 50% [Łapiński i in. 2014]. Jedną z zasadniczych przyczyn takiego stanu rzeczy może być niższa w Polsce w stosunku do średniej unijnej ogólna produktywność zatrudnionych w tym sektorze. Nie bez znaczenia dla rozwoju tego sektora gospodarki jest ich zdolność do absorbcji nowoczesnych technologii informacyjnych i komunikacyjnych (ICT).

Zarówno małe, jak i duże przedsiębiorstwa potrzebują wysokiej jakości informacji do podejmowania decyzji, jednak charakterystyczną cechą mniejszych przedsiębiorstw jest ich niższa niż w dużych firmach zdolność do wdrażania nowoczesnych rozwiązań informatycznych. Wynika to nie tylko z niższego potencjału finansowego MŚP, ale, co za tym idzie, również z ograniczonej możliwości sięgania po wysoko wykwalifikowaną kadrę specjalistów IT i/lub umiejętności posługiwania się zaawansowanym oprogramowaniem wśród specjalistów spoza działu IT. Bariery natury finansowej często skutkują unikaniem przez MŚP rozwiązań IT (głównie oprogramowania), których koszty zakupu są akceptowalne, ale koszty eksploatacji mogą stanowić barierę.

W artykule zaprezentowano podejście do problemu sugerujące wdrażanie w MŚP strategii opartych na pozyskiwaniu, przetwarzaniu i wykorzystaniu w procesach decyzyjnych wysokiej jakości danych pochodzących z systemów analitycznych opartych na darmowym oprogramowaniu *open source*, a w szczególności na środowisku obiektowego języka programowania Python. Kluczowym aspektem badawczym w artykule jest analiza funkcjonalności bibliotek języka Python, na podstawie której autor wykazuje, że poszczególne komponenty środowiska Python mogą wspomagać rozwój każdej warstwy systemu analitycznego, mogą one zatem stanowić kompletną i solidną podstawę realizacji strategii opartej na wysokiej jakości danych pozyskiwanych z tego typu systemu w firmach sektora MŚP. Nawet pobieżna analiza dostępnych narzędzi *open source* dla systemów analitycznych przekraczałaby ramy niniejszej publikacji, stąd naturalne stało się skoncentrowanie analizy na jednym z przedstawicieli tej kategorii narzędzi. Chociaż w praktyce implementacji rozwiązań *data science* wykorzystuje się wiele języków programowania[1], to w przekonaniu

[1] Z przeprowadzonych badań ankietowych wynika, że w obszarze analizy danych, *business intelligence*, *data science* i *big data* Python jest równie popularny jak język R [http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html]. Siłą Pythona jest jego uniwersalność, co daje możliwość tworzenia kompletnych aplikacji analitycznych.

autora właśnie język Python można uznać za taki, który najbardziej odpowiada potrzebom rozwijania i eksploatacji systemów analitycznych w małych i średniej wielkości firmach. Decydują o tym cechy środowiska Pythona szerzej opisane w pkt. 5.

Koszty rozwijania i użytkowania systemów analitycznych zbudowanych na bazie środowiska języka Python mogą być relatywnie znacznie niższe, niż ma to miejsce w odniesieniu do komercyjnych pakietów analitycznych. Autor artykułu posiada doświadczenie w kształceniu różnych grup studenckich (w tym międzynarodowych) z wykorzystania języka Python w *data science* i na bazie tych doświadczeń można było zaobserwować relatywnie szybkie opanowanie funkcjonalności tego języka w omawianym zakresie. Co więcej, spostrzeżenie to jest również prawdziwe w stosunku do studentów, których wcześniejsze umiejętności programistyczne były bardzo słabe. Wynika to przede wszystkim z właściwości samego języka. W firmach sektora MŚP można wykorzystać potencjał Pythona w bardzo różnym zakresie, od zamiennika arkusza kalkulacyjnego w analizach danych po alternatywę dla komercyjnych pakietów analitycznych dla zaawansowanych algorytmów eksploracji danych, uczenia maszynowego lub sztucznej inteligencji, w tym w środowisku *big data*. Należy przy tym mieć na uwadze, że poziom wiedzy i umiejętności informatycznych we współczesnych firmach MŚP jest wyższy niż 10-20 lat temu, a umiejętności programowania zdobywają obecnie uczniowie na różnych poziomach edukacji powszechnej.

2. Definiowanie strategii opartej na danych w MŚP

Wysoka jakość danych dostępnych w organizacji sprzyja podejmowaniu optymalnych decyzji. Dane i informacja mogą nie tylko być czynnikiem wspomagającym procesy decyzyjne, ale również stanowić podstawę do formułowania strategii. Według Davenporta i Harrisa istnieje wiele firm, które skutecznie oparły swoje strategie na wysoko przetworzonych danych w systemach analitycznych (*analytics*) [Davenport, Harris 2007].

Z przeprowadzonego w 2003 r. badania 1200 specjalistów IT w brytyjskich MŚP wynikało, że połowa firm nie posiadała w ogóle strategii informatyzacji. Dla 76% firm powodem niewdrażania nowych rozwiązań IT były kwestie kosztów. W 30% firm powodem braku rozwoju nowych rozwiązań jest wewnętrzny opór przed zmianą. Aż 27% firm nie wydało na szkolenia IT żadnych pieniędzy w ostatnim roku. Problemem zauważalnym w tym sektorze był niski poziom świadomości menedżerów najwyższego szczebla, w jaki sposób IT może tworzyć wartość dla firmy lub jak istniejące zasoby informatyczne można lepiej wykorzystać do realizacji celów firmy. Patrzenie na infrastrukturę IT bardziej przez pryzmat kosztu niż korzyści stanowiło istotną barierę w inwestowaniu w technologie informacyjne [Beckett 2003]. Adaptacji technologii informacyjnych w MŚP towarzyszy wiele rodzajów ryzyka [Ghobakhloo i in. 2011]:

1. Niewłaściwe związanie wdrożonych rozwiązań IT ze strategiami firmy.
2. Niewłaściwa realizacja problemów organizacyjnych.

3. Nieodpowiednia realizacja potrzeb użytkowników.
4. Brak wymaganych zasobów (wiedzy, umiejętności, finansów i umiejętności menedżerskich).
5. Nieadekwatne szkolenia i przygotowanie użytkowników końcowych.
6. Ograniczenia organizacyjne i finansowe w zatrudnianiu specjalistów IT.
7. Brak kwalifikacji kierownictwa w mocno scentralizowanej strukturze zarządzania firmą.
8. Niewystarczające wsparcie państwa i regulacji prawnych.
9. Brak satysfakcji z tworzenia przewagi konkurencyjnej przez IT.
10. Partykularyzm charakteryzujący kulturę organizacji i koneksje rodzinne.

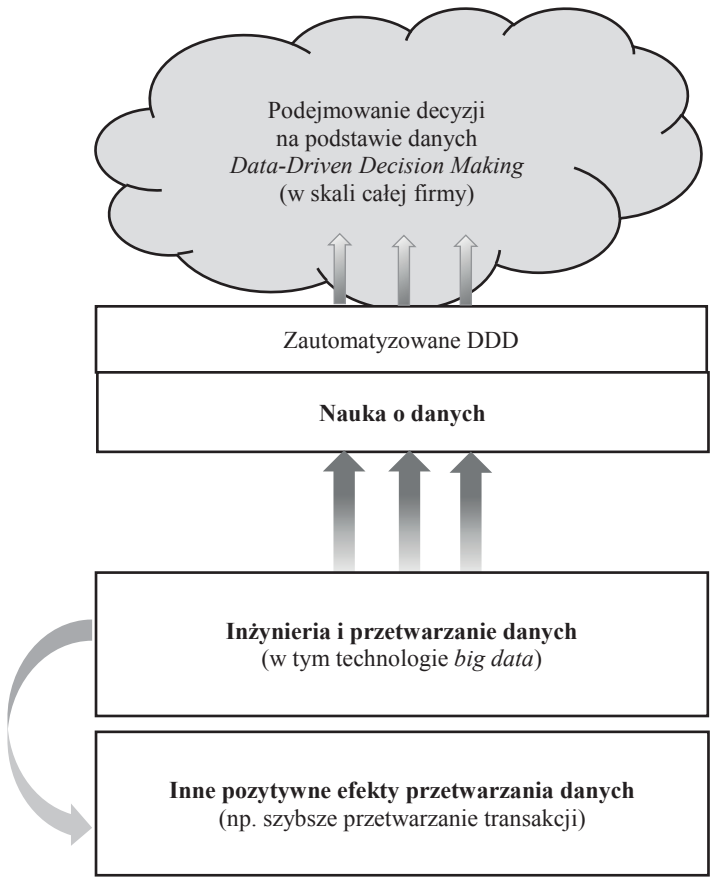
W obszernym studium badawczym Brynjolfsson, Hitt i Kim przeprowadzili badania, które potwierdziły, że efektywność organizacji jest wyższa w firmach opierających swoje decyzje ma danych oraz systemach analitycznych. Sprzyja temu oparcie procesów decyzyjnych na danych pozyskiwanych z systemów analitycznych, określane jako *data-driven decisionmaking approach* (DDD). Zaproponowali oni wskaźnik DDD, który klasyfikuje firmy według kryterium wykorzystania danych w procesach podejmowania decyzji. Używając aparatu statystycznego, wykazali, że efektywność firmy rośnie w zależności od stopnia, w jakim firma wykorzystuje wysokiej jakości dane w procesie decyzyjnym. Uzyskiwana w ten sposób ogólna efektywność jest widoczna w postaci poprawy rentowności aktywów czy wyższej wartości rynkowej firmy [Brynjolfsson i in. 2011].

Według Provosta i Fawcetta zastosowanie podejścia DDD może być skutecznie wzmocnione przy wykorzystaniu technologii dedykowanych przetwarzaniu wielkich zbiorów danych rozproszonych *big data* oraz infrastruktury *data science* (zob. rys. 1). W warunkach działania firm z sektora MŚP oparcie strategii firmy na podejściu DDD wydaje się nie tylko realne, ale również w wielu przypadkach wskazane. W pierwszej kolejności takie podejście powinny rozważyć firmy, które potrafią oszacować wielkość i strukturę udziału informacji w łańcuchu tworzenia wartości, a w konsekwencji w swoich produktach. Mogą się posłużyć pięcioetapową procedurą [Porter 2001]:

1. Ocenić nasilenie informacji.
2. Określić rolę techniki informacyjnej w strukturze sektora.
3. Rozpoznać i uszeregować sposoby uzyskania przewagi konkurencyjnej dzięki technice informacyjnej.
4. Zbadać, w jaki sposób technika informacyjna mogłaby się przyczynić do powstania nowych dziedzin działalności.
5. Opracować plan wykorzystania techniki informacyjnej.

Zaproponowana przez M. Portera procedura może być przydatna do celów strategii w podejściu DDD. Oparcie procesów decyzyjnych na wysoko przetworzonych danych pochodzących z systemów analitycznych można rekomendować firmom sektora MŚP, w których:

- krytyczne procesy gospodarcze generują lub mogą generować duże zbiory jednorodnych lub heterogenicznych danych, zwłaszcza w obszarach sprzedaży i marketingu,
- specyfika działalności umożliwia prowadzenie sklepów internetowych dla masowego klienta,
- pożądane jest mocno spersonalizowane budowanie oferty rynkowej z wykorzystaniem automatyzacji tworzenia profili klientów i w oparciu o nie wykorzystanie systemów rekomendujących,
- analizy predykcyjne mogą usprawnić procesy planowania działań w całym łańcuchu tworzenia wartości dla klienta,
- procesy wytwórcze mogą być wspomagane o zastosowanie algorytmów sztucznej inteligencji,



Rys. 1. Nauka o danych w kontekście różnych związanych z danymi procesów w ramach organizacji

Źródło: adaptacja z [Provost, Fawcett 2013, s. 5].

- podstawowe zasoby, które są konsumowane lub wytwarzane w działalności podstawowej, można zaliczyć do zasobów intelektualnych (np. firmy doradcze lub innowacyjne).

Oczywiście powyższa lista nie wyczerpuje innych sytuacji, w których firmy MŚP nie mogą realizować strategii DDD. Jak wskazuje się w wielu publikacjach poruszających problematykę budowania optymalnego środowiska informacyjnego organizacji, oprócz wielu wymogów stawianych samej technicznej infrastrukturze systemu informacyjnego, to czynniki „miękkie" decydują o finalnym powodzeniu[2]. Innymi słowy, infrastrukturze SI musi towarzyszyć dojrzałość organizacji w absorbcji technologii, aby dane i informacja stały się strategicznym zasobem organizacji. Stwierdzenie to jest aktualne również w odniesieniu do firm sektora MŚP.

3. Bariery implementacji systemów analitycznych w MŚP

Tematyka nauki o danych (*data science*) jest jeszcze na tyle świeża, że wielu autorów podkreśla, iż istota *data science* jest jeszcze przedmiotem dyskusji, stąd brak powszechnie akceptowanej definicji staje się uzasadniony. Provost i Fawcett wskazują, że „nauka o danych obejmuje zasady, procedury i techniki stosowane w celu zrozumienia zjawiska poprzez (zautomatyzowaną) analizę danych" [Provost, Fawcett 2014]. Jednym z kluczowych czynników niezbędnych w kształtowaniu praktyki zgodnej z *data science* jest odpowiednia infrastruktura systemów analitycznych.

Pojęcie systemy analityczne (*analytics*) przyjęło się w międzynarodowym piśmiennictwie jako termin integrujący różne kategorie systemów informacyjnych, których podstawowym celem jest wspomaganie procesów decyzyjnych na różnych szczeblach zarządzania organizacjami. W niniejszym artykule do tej kategorii rozwiązań informatycznych zaliczone zostaną takie rodzaje systemów, jak: systemy wspomagania decyzji (*decision support systems*), systemy informowania kierownictwa (*executive information systems*), ale również różnej kategorii systemy *business intelligence*. Takie uproszczenie ma uzasadnienie przede wszystkim na podstawie analizy realnie wdrażanych rozwiązań w przedsiębiorstwach, gdzie obecnie bardzo popularne są pakiety oprogramowania analitycznego wiodących dostawców światowych (np. IBM, SAS, SAP, Oracle itd.), w których monolitycznych instalacjach trudno jest rozdzielać fragmenty funkcjonalne odpowiadające określonej kategorii systemów analitycznych.

Tematykę poruszającą spektrum problemów i barier w implementacji zaawansowanych rozwiązań informatycznych w MŚP możemy spotkać w bardzo licznych publikacjach badawczych, zarówno w źródłach anglojęzycznych, jak i krajowych. Na podstawie obserwacji praktyki gospodarczej można wskazać, że jedną z kluczowych barier tego typu strategii w MŚP jest brak kapitału potrzebnego do zbudowania i eksploatacji systemów analitycznych.

---

[2] Szerokie omówienie tego zagadnienia można znaleźć m.in. w [Zygała 2007].

Bariery mają przełożenie także na mniejszą niż w dużych (bogatszych) przedsiębiorstwach skalę inwestowania w systemy analityczne. Polityka cenowa wiodących dostawców tej klasy rozwiązań stanowi istotną barierę finansową wykluczającą mniejsze, mniej zasobne finansowo firmy. Problem finansowania omawianej klasy rozwiązań w MŚP odnosi się nie tylko do fazy nabycia i wdrożenia, ale również eksploatacji. Można także dostrzec problem niedopasowania funkcjonalności dużych pakietów do potrzeb mniejszych organizacji. Mówiąc wprost, rosnąca złożoność funkcjonalna i eksploatacyjna komercyjnych rozwiązań analitycznych w wielu przypadkach może być nie do wykorzystania przez MŚP.

Na bazie obserwacji praktyki gospodarczej można zaryzykować tezę, iż potrzeby informacyjne w sektorze MŚP są w głównych nurtach tożsame jak w dużych organizacjach gospodarczych. Problemem może być brak dostępu do danych z rozmaitych przyczyn, w tym takich, które były omawiane powyżej. Alternatywną ścieżką absorbcji systemów analitycznych w sektorze MŚP może być skorzystanie z coraz bogatszej oferty darmowego oprogramowania *open source*, na bazie którego można tworzyć zaawansowane, wydajne, skalowalne i bogate funkcjonalnie rozwiązania.

Przezwyciężanie różnych obiektywnych ograniczeń, które dotyczą firm tego sektora, powinno skutkować ewolucyjnym wykorzystaniem systemów analitycznych, w tym rozwiązań opartych na technologii *big data*. Przemawiają za tym następujące argumenty [Scholz i in. 2010; Brooke 2015; Passerini, El Tarabishy, Patten 2012]:

1. Firmy sektora MŚP zazwyczaj dysponują mniejszymi zespołami specjalistów IT niż firmy duże, ale w takich sytuacjach można rozważyć cykl szkoleń z popularnych narzędzi *data science* albo delegowanie lepiej przygotowanych pracowników na dodatkowe szkolenia zewnętrzne. Można również wspierać własne kadry specjalistami z zewnątrz.

2. Menedżerowie sektora MŚP często mają przekonanie, że ich biznes da się kierować bardziej tradycyjnymi metodami, a te bardziej zaawansowane, naukowo wspomagane narzędzie i metody są dla dużych firm. Problem jednak w tym, że nie zawsze da się zauważyć, kiedy proste metody są niewystarczające, albo widać negatywne zjawiska, ale nie ma na nie diagnozy opartej na faktach. Lepsze wydaje się w takich sytuacjach podejście podglądania dużych firm i stawianie sobie pytania, czy możemy działać podobnie.

3. Firmy MŚP zazwyczaj obsługują mniej klientów, często więc bezpośrednie kontakty z nimi wystarczają do wyrobienia w sobie przekonania o wystarczającej wiedzy na temat potrzeb klienta. Pomijanie w procesach obsługi klienta zaawansowanej analizy danych może prowadzić do utraty pozycji rynkowej lub stagnacji w rozwoju.

4. Mniejsze firmy mają przewagę nad większymi pod względem elastyczności i szybszej reakcji na sygnały rynkowe, nawet w przypadku konieczności przeprowadzenia głębszych zmian w funkcjonowaniu. Systemy analityczne mniejszych firm mogą generować na mniejszym poziomie złożoności informacje dotyczące wartości swoich klientów, motywacji nowych klientów, oceny własnej pozycji rynkowej (np.

analizy SWOT), problemów, z jakimi borykają się ich klienci. Problemy tego typu są co do istoty takie same jak w przypadku dużych firm.

5. Firmy MŚP nie zawsze mogą sobie pozwolić na kosztowne zlecenia badań rynkowych, ale wnikliwa penetracja darmowych zasobów danych w Internecie może być w wielu wypadkach wystarczająca.

6. Segmentacja i personalizacja obsługi rynku przez MŚP nie musi być złożona i wieloelementowa. Często wystarczy ograniczyć swoje analizy do newralgicznych dla firmy segmentów rynku i zbudowania kilku profili klientów.

7. W porównaniu z dużymi firmami, w sektorze MŚP dominują mniej sformalizowane i ujęte w struktury podejścia w stosunku do procesów IT, a także bardziej oportunistyczne podejście do inwestycji w IT. Nieformalne procesy informacyjne muszą więc być poddane procesowi formalizacji, najlepiej ze wspomaganiem IT, wówczas gromadzone dane „odsłonią" wewnętrzną efektywność firmy, do czego mogą wydatnie przyczynić się systemy analityczne.

Ewolucyjnemu podejściu do wykorzystania systemów analitycznych w praktyce MŚP z pewnością sprzyja obecna podaż oprogramowania, które może być wykorzystane do tego celu. Ogólną analizę tego typu oprogramowania zawiera kolejny punkt.

4. Oferta rynku oprogramowania *open source* na potrzeby *data science*

Wdrażanie rozwiązań analitycznych w MŚP nie musi oznaczać funkcjonalności odbiegającej znacznie od tej, którą posiadają rozwiązania dla dużych organizacji. Nawet w odniesieniu do rozwiązań *big data* firmy MŚP powinny rozważać ich implementację, naśladując przy tym duże firmy [Brooke 2015]. Ważnym aspektem dla rozwoju systemów analitycznych w sektorze MŚP jest coraz bardziej rozbudowana funkcjonalnie grupa narzędzi typu *open source*. Poza aspektami czysto funkcjonalnymi nie bez znaczenia jest to, iż dostępne tego typu oprogramowanie ma charakter komplementarny, pozwalając na zbudowanie kompletnych aplikacji analitycznych dedykowanych dużemu spektrum problemów (zob. tab. 1).

Zaprezentowana w tab. 1 lista absolutnie nie wyczerpuje bardzo bogatej oferty oprogramowania *open source*, które może być wykorzystywane w zaawansowanych rozwiązaniach *data science*. Wskazuje ona, że opierając się głównie na kompetencjach pracowników, można tworzyć dzisiaj rozwiązania skalowalne dla różnych problemów analizy danych, od transakcyjnych dobrze ustrukturyzowanych niedużych zbiorów danych po słabo usktrukturyzowane i rozproszone wielkie zasoby danych, typowe dla aplikacji *big data*. Bardziej szczegółową analizę, zorientowaną na język Python zawiera kolejny punkt opracowania.

Tabela 1. Wybrane oprogramowanie *open source* stosowane w *data science*

Nazwa oprogramowania	Ogólna charakterystyka
Systemy bazodanowe	
CUBRID	Darmowe oprogramowanie *open source*, szczególnie przydatne w przetwarzaniu danych aplikacji internetowych. Przystosowane do współpracy z językami PHP, Perl, Ruby i Python
MariaDB	System bazodanowy stworzony przez twórców MySQL może być alternatywą dla tego popularnego serwera bazodanowego. Szczególna uwaga twórców systemu została skupiona na kwestiach bezpieczeństwa danych, stąd jest wykorzystywany m.in. przez Facebook czy Google
MongoDB	Bardzo popularny serwer bazodanowy z kategorii NoSQL. System pracuje na danych przechowywanych w postaci dokumentów (np. XML, JSON). MongoDB bardzo dobrze współpracuje z Pythonem
MySQL	Jest to jeden z najpopularniejszych relacyjnych systemów bazodanowych. Stanowi on poważną alternatywę dla komercyjnych produktów
PostgreSQL	Konkurencyjny dla MySQL serwer, który może pracować na wielu platformach systemowych. W praktyce podkreśla się jego zgodność ze standardem *de facto* ACID (*Atomicity*, *Consistency*, *Isolation*, *Durability*). Dobrze współpracuje z Pythonem
Oprogramowanie dla *big data*	
Apache Hadoop	Framework dedykowany programistom do przetwarzania *big data*
Apache Mahout	Skalowalne algorytmy *machine learning* dla Apache Hadoop
Spark	Framework przetwarzania klastrowego dedykowany problemom analizy danych
Języki programowania dla *data science*	
SQL	Język zapytań powszechnie wykorzystywany w relacyjnych bazach danych. Wykorzystując potencjał tego języka, można rozwiązywać nawet bardzo złożone problemy analityczne *data science*
R	Język skryptowy, który należy współcześnie do najczęściej wykorzystywanych narzędzi *data science*. Do jego zalet należy warta podkreślenia bogata funkcjonalność w analizach statystycznych, *data mining* oraz *machine learning*
Scala	Nowoczesny język obiektowy, w którym napisany został Apache Spark, co uzasadnia jego powszechne wykorzystanie do rozwiązywania złożonych problemów analitycznych *big data*
Python	Język skryptowy ogólnego zastosowania z bardzo obszernymi bibliotekami obsługującymi wszelkie zagadnienia *data science*. Stanowi alternatywę dla języka R, ale zyskuje uznanie praktyków z uwagi na szybkość, ogólne zastosowanie oraz współpracę z innymi technologiami infrastruktury systemów analitycznych (zob. pkt 5).

Źródło: opracowanie własne.

5. Koncepcja komponentów systemu analitycznego dla MŚP na bazie bibliotek Pythona

Język programowania Python stworzony został na początku lat 90. przez G. Van Rossem i zaliczany jest dzisiaj do najważniejszych narzędzi – obok języka R – używanych w problematyce *data science* [Davies 2016]. Najważniejsze cechy Pythona mające wpływ na jego użyteczność dla *data science* to (por. [Thakur 2016, s. 9-10; Boschetti, Massaron 2015]):

1. **Przenośność**. Programy napisane w Pythonie można uruchamiać na różnych platformach systemowych (Linux, Windows, Mac).

2. **Spójność**. Python jest rozwijany w dużym stopniu przez środowisko naukowe. Logiczna składnia Pythona jest szczególnie podkreślana jako przyjazna dla początkujących programistów.

3. **Produktywność programistyczna**. Program napisany w Pythonie może stanowić 1/5 do 1/3 rozmiaru kodu ekwiwalentu napisanego w Javie lub C++, co finalnie oznacza oszczędność czasu i kosztów.

4. **Bogata funkcjonalność bibliotek ogólnego zastosowania**. Standardowe biblioteki Pythona rozwijane są przez coraz liczniejszą społeczność międzynarodową. Do wielu zadań programista może znaleźć propozycje obsługi problemu przez kilka bibliotek.

5. **Bogata funkcjonalność bibliotek do analizy danych, uczenia maszynowego i technologii *big data***. Potencjał analityczny Pythona jest porównywalny do tego, który oferuje język R, ale w przypadku tego pierwszego języka istnieje możliwość zbudowania kompletnych i przenaszalnych aplikacji analitycznych, co daje mu przewagę nad językiem R. Python jest również rekomendowanym narzędziem do współpracy z najnowszymi technologiami *big data* (np. Spark).

6. **Jakość oprogramowania**. Wysoka czytelność kodu napisanego w Pythonie stanowi istotną cechę dla kolektywnego tworzenia oprogramowania. Python oparty jest na wielu paradygmatach nowoczesnych obiektowych języków programowania, może być używany jako interaktywny język skryptowy, ale również jako uniwersalny język do tworzenia złożonych aplikacji.

7. **Integracja oprogramowania**. Python może być rozszerzany i integrowany z wieloma innymi językami (np. Java, C, Fortran), co umożliwia traktowanie Pythona jako integratora pomiędzy różnymi aplikacjami.

Python wszedł do programów dydaktycznych wielu kursów renomowanych uczelni na świecie w obszarach zarządzania, analizy danych i *data science*[3]. Popularność tę można postrzegać jako docenienie potencjału Pythona na potrzeby analizy danych, *data science* oraz aplikacji *big data*. Postrzegając potencjał tego środowiska programistycznego z pozycji wymagań stawianych aplikacjom analitycznym przez *data science*, należy przede wszystkim uwzględnić wymóg obsługi przez system różnych formatów danych wejściowych, które mogą pochodzić z różnych dzia-

[3] Potwierdzenie tej tezy można znaleźć na stronach www.edx.org oraz www.coursera.org.

łań realizowanych w MŚP, ale przede wszystkim nie zawsze generujących dobrze ustrukturyzowane dane typowe dla zintegrowanych systemów transakcyjnych. Potencjał Pythona umożliwia MŚP realizację omawianych wyżej strategii DDD, co w ujęciu operacyjnym powinno sprowadzać się do następujących czynności (por. [Anderson 2015 s. 5-9]):

1. **Gromadzenie danych**. Dane muszą być nie tylko relewantne do rozwiązywanych problemów, ale również dostarczane na czas, obiektywne (*unbiased*) i godne zaufania.

2. **Umożliwienie dostępu do danych**. Oprócz dostępu do danych użytkownik musi również mieć możliwość generowania zapytań (*queryable*). Oznacza to, że dane muszą mieć zdolność łączenia ich z innymi repozytoriami danych. Wewnątrz organizacji musi zafunkcjonować kultura współdzielenia danych (*data-sharing culture*).

3. **Raportowanie**. W organizacji zorientowanej na dane (*data-driven*) tradycyjne raportowanie, które wykonują poszczególne komórki na swoje potrzeby, nie jest wystarczające, konieczne jest również współdzielenie tych raportów w ramach organizacji.

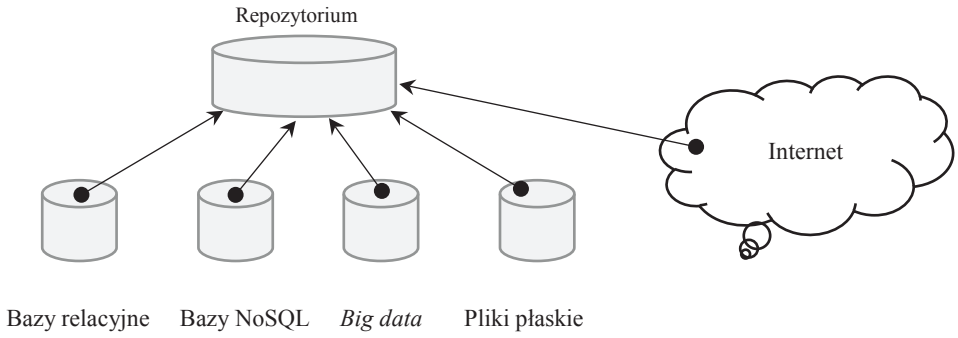
4. **Generowanie alertów (*alerting*)**. Oznacza to raportowanie o zaistnieniu zdarzenia istotnego dla organizacji. Dotyczy to specyficznych danych w postaci dobrze zaprojektowanych wskaźników (np. efektywności, wydajności czy też bezpieczeństwa informacji).

5. **Raporty i alerty stają się podstawą analizy danych**. Na podstawie cyklicznych raportów i alertów organizacja zorientowana na dane przekształca zasoby danych na informację i wiedzę wspomagającą podejmowanie decyzji i funkcjonowanie ludzi, procesów i technologii w firmie.

Wymienione cechy organizacji zorientowanej na dane stanowią bazę do dalszej analizy potencjału języka Python do zdefiniowania architektury systemu analitycznego, który może być budowany na potrzeby MŚP.

Na poziomie gromadzenia danych do repozytorium systemu analitycznego biblioteki Pythona oferują bardzo bogatą funkcjonalność, umożliwiającą zasilanie systemu analitycznego w dane pochodzące z bardzo różnych źródeł i w wielu formatach przechowywania danych (zob. rys. 2). Repozytorium danych można zorganizować w różnych układach i na różnych poziomach integracji.

W postaci niezintegrowanej poszczególne pliki z danymi mogą być przechowywane w oryginalnych formatach. Rola Pythona może polegać na pobieraniu danych na potrzeby konkretnej analizy, przekształceniu ich na wynikowy format wewnętrzny Pythona, a następnie wykonaniu procesu wyodrębnienia tylko tych danych, które są potrzebne do finalnej analizy (*data wrangling*). W postaci zintegrowanej bazy danych dane mogą być konwertowane i integrowane do jednorodnych formatów, a następnie przechowywane w systemach baz danych. Przykładowo, za pośrednictwem biblioteki SQLAlchemy można obsłużyć dwukierunkową obsługę popularnych serwerów relacyjno-obiektowych, takich jak: SQLite, Postgresql, MySQL, Oracle, MS-SQL, Firebird czy Sybase.



Rys. 2. Zasilenia systemu analitycznego MŚP

Źródło: adaptacja z [Provost, Fawcett 2014, s. 5].

Za pośrednictwem biblioteki PyMongo można obsłużyć zasilenie repozytorium w dane oparte na formacie „klucz-wartość" popularnego obecnie formatu JSON. Podstawowe biblioteki Pythona do analizy danych, Pandas i NumPy, umożliwiają użytkownikowi korzystanie z danych konwertowanych nie tylko z formatów bazodanowych, ale również z plików Excela, txt, csv do formatu DataFrame, a następnie przechowywanie tych danych w repozytorium w postaci rozproszonej lub zintegrowanej.

Python umożliwia również wykonywanie analizy danych przechowywanych w technologiach *big data*. W szczególności dobrze wygląda obsługa przez Pythona frameworku Spark, wykorzystywanego do przetwarzania i analizy danych *big data*. Taki model przetwarzania nie wychodzi z ram *open source* i może być skalowany dla bardzo dużych wolumenów danych w rozproszonym układzie klastrów. Do warstwy zasilania systemu analitycznego dla MŚP należy również zaliczyć dane pochodzące z różnych witryn w Internecie. Środowisko języka Python umożliwia również bezpośrednie pozyskiwanie danych z witryn internetowych. Do pobierania danych z plików HTML i XML (*web scrapingu*) analityk danych może posłużyć się wydajnym parserem z biblioteki Beautiful Soup. Pandas-datareader jest biblioteką przydatną do zasilania wspomnianych już struktur DataFrame na zasadzie dostępu zdalnego do takich witryn internetowych, jak: Yahoo! Finance, Google Finance, FRED, World Bank, OECD, Eurostat czy EDGAR Index. Można się spodziewać, że lista ta będzie się powiększać, a korzyści automatycznego zasilania systemu analitycznego w aktualne dane z Internetu wydają się oczywiste.

Analityk danych w MŚP, dysponujący bogatym w dane repozytorium, może wykonywać w zasadzie nieograniczone analizy ekonomiczne, posługując się ogromną funkcjonalnością bibliotek Pythona napisanych na potrzeby analizy danych. Do podstawowych bibliotek analizy danych Pythona można zaliczyć:

1. **NumPy** – powszechnie traktowana jest jako biblioteka bazowa dla analizy danych i obliczeń naukowych. Biblioteka ta przede wszystkim umożliwia tworzenie

struktur tablicowych oraz macierzowych, a następnie wykonywanie na nich wielu typów operacji. Oprócz tego NumPy oferuje zaawansowane funkcje matematyczne i statystyczne do obliczeń na wspomnianych strukturach, ale również obsługę wejścia i wyjścia danych.

2. **SciPy** – jest biblioteką bazującą na NumPy, oferującą algorytmy obliczeń naukowych i technicznych, w szczególności w problemach optymalizacji, algebry liniowej, interpolacji, funkcji specjalnych, szybkiej transformacji Fouriera, przetwarzania sygnałów i obrazów, a także równań różniczkowych.

3. **Pandas** – biblioteka oferująca kluczowe dla analizy danych struktury Series i DataFrame oraz praktyczną funkcjonalność analizy danych w obszarach finansów, statystyce, naukach społecznych i inżynierii. Pandas pozwala analitykowi wykonywać różne operacje na danych, takie jak porządkowanie, selektywne usuwanie, transformacje, uzupełnianie, wyodrębnianie zdefiniowanych zakresów itd. Biblioteka obsługuje również dwukierunkową współpracę z zewnętrznymi zbiorami danych.

4. **IPython** – stanowi rozszerzenie funkcjonalności standardowego Pythona o środowisko dodające wiele funkcji pozwalających na bardziej interaktywny tryb obliczeń i wizualizacji.

W bardziej zaawansowanych analizach biznesowych mogą mieć zastosowanie algorytmy uczenia maszynowego (*machine learning*), *data mining* i przetwarzania tekstu. Python udostępnia analitykom kilka ważnych bibliotek do tego typu analiz:

1. **Scikit-learn** – jest biblioteką powstałą na bazie NumPy oraz SciPy. Oferuje ona zbiór algorytmów dla uczenia maszynowego i *data mining*, które mogą być szczególnie przydatne do filtrowania spamu, rozpoznawania obrazów, analiz giełdowych, segmentacji klientów, grupowania wyników eksperymentów, personalizacji ofert marketingowych (algorytmy rekomendacyjne) itd.

2. **Theano** – używa składnię podobną do NumPy do definiowania, optymalizacji i oceny obliczeń matematycznych na bazie wielowymiarowych tablic w zastosowaniach uczenia głębokiego (*deep learning*).

3. **NLTK** – jest zbiorem bibliotek dedykowanych przetwarzaniu języka naturalnego (*natural language processing*).

4. **Scrapy** – biblioteka przydatna do pozyskiwania danych z Internetu w technice *web scraping*.

Wymienione biblioteki Pythona pozwalają na kompleksową analizę danych zgromadzonych w repozytorium systemu analitycznego – od stosunkowo prostych analiz do metod z obszaru sztucznej inteligencji i uczenia maszynowego. Finalnym efektem procesu analizy w wielu przypadkach jest wizualizacja i dystrybucja informacji. W tym zakresie do dyspozycji analityków danych Python oferuje bardzo bogatą funkcjonalność skupioną w bibliotekach:

1. **Matplotlib** – stanowi standardową bibliotekę do tworzenia dwuwymiarowej grafiki i wizualizacji. Cechą wyróżniającą tę bibliotekę jest różnorodność oferowanych form graficznych i oszczędność kodu, co czyni ją szczególnie przydatną do celów analizy danych.

2. **Seaborn** – jest biblioteką bazującą na matplotlib. Seaborn jest biblioteką wysokiego poziomu, szczególnie przydatną w wizualizacji statystycznej, w tym obliczeń regresji liniowej, analizy szeregów czasowych, macierzy danych, map cieplnych itd.

3. **Bokeh** – pozwala na tworzenie interaktywnych wydruków osadzonych w nowoczesnych przeglądarkach internetowych z generowanym kodem JavaScript, co umożliwia uzyskanie efektów dostępnych w bardzo popularnej w świecie deweloperów bibliotece d3.js. Biblioteka bokeh może być polecana szczególnie do tworzenia pulpitów menedżerskich (*digital dashboards*) w środowisku webowym.

4. **Basemap** i **Folium** – umożliwiają tworzenie wizualizacji z wykorzystaniem interaktywnych map geograficznych w środowisku webowym.

5. **NetworkX** – jest biblioteką do tworzenia i analizy grafów i sieci. Oferuje użytkownikowi tworzenie grafiki z wykorzystaniem zarówno standardowych, jak i niestandardowych formatów danych.

Dla firm sektora MŚP ważną kwestią jest umożliwienie taniego wielodostępu do aplikacji i danych. Atrakcyjnym i ekonomicznym wariantem uzyskania tego celu jest osadzenie interfejsu użytkownika w środowisku przeglądarki internetowej. Do tego mogą być szczególnie przydatne frameworki Pythona: Django, Flask i Pyramid. Pozytywnie można ocenić potencjał Flaska pod kątem przydatności do celów systemu analitycznego firm MŚP. Za pomocą Flaska użytkownik może przedstawiać DataFrame z Pandas bezpośrednio w Excelu lub w postaci pulpitów menedżerskich z warstwą kliencką w przeglądarce internetowej (zob. [Dempsey 2015]). Szczególnie druga opcja może być atrakcyjna dla MŚP, gdyż zapewnia dystrybucję raportów wśród wielu użytkowników rozproszonych terytorialnie.

6. Zakończenie

W artykule zaprezentowany został potencjał współczesnych narzędzi deweloperskich, dystrybuowanych na darmowych licencjach *open source*, które mogą być użyte do stworzenia kompletnych aplikacji analitycznych w firmach sektora MŚP. W zaprezentowanym podejściu do problemu autor wskazuje, że coraz bardziej atrakcyjne dla firm tego sektora są strategie oparte na pozyskiwaniu, przetwarzaniu i wykorzystaniu wysokiej jakości danych w procesach decyzyjnych (*data-driven decisionmaking approach*). Warunkiem krytycznym przy tego typu strategiach jest oparcie ich na praktycznej implementacji dorobku nauki o danych (*data science*) i nowoczesnej infrastruktury systemów analitycznych.

Autor wykazuje, że dobrym wyborem dla takiego podejścia może być środowisko programistycznego Pythona, jako narzędzia z jednej strony łatwego w implementacji, nawet przez średnio zaawansowanych specjalistów, a z drugiej strony bardzo bogatego w specjalizowane biblioteki i frameworki do zbudowania skalowalnych systemów analitycznych dla wielu użytkowników, nawet rozproszonych geograficznie. Zaprezentowaną w artykule tematykę można uznać za rozwojową

pod względem badawczym oraz aplikacyjnym. Narzędzia informatyczne podlegają ustawicznej ewolucji, ale bardziej trwała jest idea, aby zaawansowane algorytmy bazujące na sztucznej inteligencji i statystyce znajdowały swoje zastosowanie również w aplikacjach mniejszych podmiotów gospodarczych. Wokół tej idei można formułować nowe cele badawcze do osiągnięcia w przyszłości.

Literatura

Anderson C., 2015, *Creating Data-Driven Organization*, O'Reilly, Sebastopol.

Beckett H., 2003, *Half of SMEs have no IT Strategy,* Computer Weekly, 1, 34, http://www.computerweekly.com/feature/Half-of-SMEs-have-no-IT-strategy (listopad 2016).

Boschetti A., Massaron L., 2015, *Python Data Science Essentials*, Packt, Birmingham.

Brooke S., 2015, *SMEs should do more to embrace big data and learn from the big players*, http://realbusiness.co.uk/tech-and-innovation/2015/10/30/smes-should-do-more-to-embrace-big-data-and-learn-from-the-big-players/ (listopad 2016).

Brynjolfsson E., Hitt L.M., Kim H.H., 2011, *Strenght in numbers; How does data-driven decision-making affect firm performance?*, https://pdfs.semanticscholar.org/dde1/ 9e960973068e541f634--b1a7054cf30573035.pdf.

Davies J., 2016, *10 Programming Languages and Tools Data Scientists Use Now*, InformationWeek: http://www.informationweek.com/devops/programming-languages/10-programming-languages-and-tools-data-scientists-use-now/d/d-id/1326034?image_number=2 (1.11.2016).

Davenport T.H., Harris J.G., 2007, *Competing on Analytics: The New Science of Winning*, Harvard Business Press.

Dempsey R., 2015, *Python Business Intelligence Cookbook*, Packt, Birmingham.

Eurostat, http://ec.europa.eu/eurostat/web/structural-business-statistics/ structural-business-statistics/sme (listopad 2016).

Ghobakhloo M., Sadegh Sabouri M., Hong T.S., Zulkifli N., 2011, *Information technology adoption in Small and Medium-Sized Enterprises*; *An appraisal of two decades literature*, Interdisciplinary Journal of Research in Business, vol. 1, Issue. 7, July, s. 53-80.

http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html.

Łapiński J., Nieć M., Rzeźnik G., Zakrzewski R., 2014, *Małe i średnie przedsiębiorstwa w Polsce*, [w:] *Raport o stanie sektora małych i średnich przedsiębiorstw w Polsce w latach 2011-2012*, https://badania.parp.gov.pl/images/badania/ROSS_2013_2014.pdf (1.11.2016).

Passerini K., El Tarabishy A., Patten K., 2012, *Information Technology for Small Business: Managing the Digital Enterprise*, Springer Science and Business Media, Heidelberg.

Porter M., 2001, *Porter o konkurencji*, PWE, Warszawa.

Provost F., Fawcett T., 2014, *Data Science for Business*, O'Reilly, Sebastopol.

Scholz P., Schieder Ch., Kurze Ch., Gluchowski P., Boehringer M., 2010, *Benefits and Challenges of Business Intelligence Adoption in Small and Medium-Sized Enterprises*, Proceedings of 18th European Conference on Information Systems: http://ai2-s2-pdfs.s3.amazonaws.com/96ba/be7026fdb-8c558d0c98a940d7ab8df5ae6d7.pdf.

Thakur A., 2016, *Data Science with Python*, Packt, Birmingham.

Zygała R., 2007, *Podstawy zarządzania informacją w przedsiębiorstwie*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.