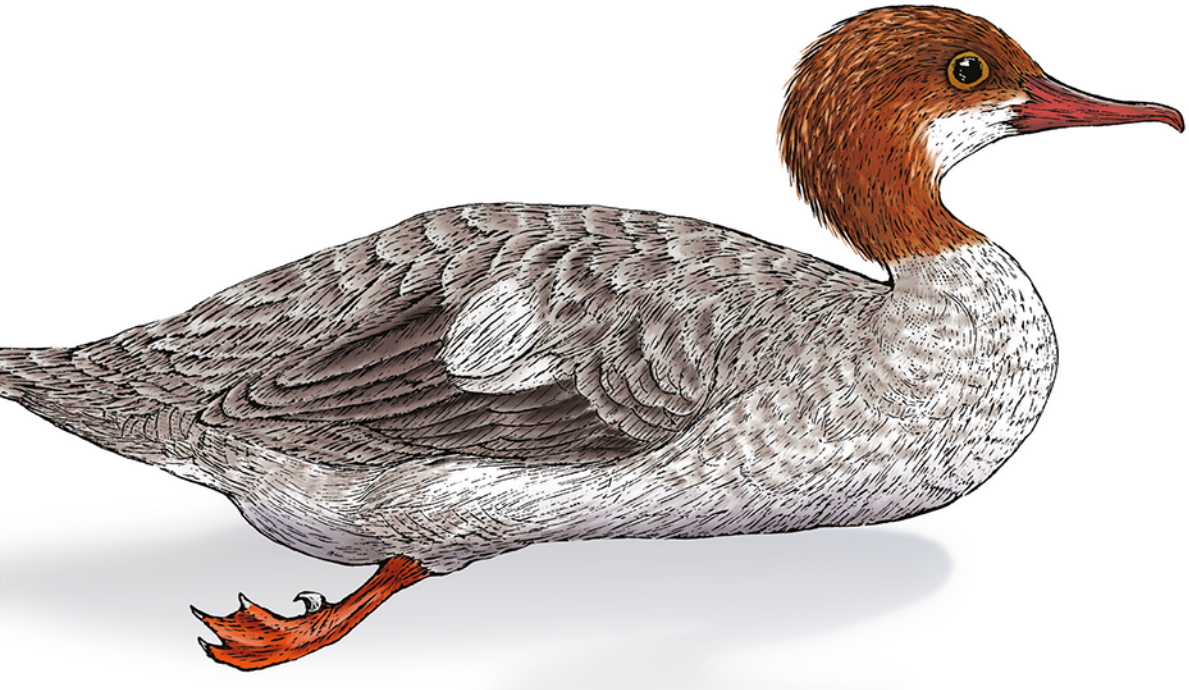
O'REILLY®

Korporacyjne jezioro danych

Wykorzystaj potencjał big data w swojej organizacji



Helion

Alex Gorelik

Tytuł oryginału: The Enterprise Big Data Lake: Delivering on the Promise of Hadoop and Data Science in the Enterprise

Tłumaczenie: Lech Lachowski

ISBN: 978-83-283-5078-6

Helion SA
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 231 22 19, 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *http://helion.pl* (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
http://helion.pl/user/opinie/kojeda
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

Spis treści

ROZDZIAŁ 1.

Wprowadzenie do jezior danych

Podejmowanie decyzji na podstawie danych zmienia nasz sposób pracy i życia. Osoby uprawnione do podejmowania decyzji wymagają danych dostarczanych przez naukę o danych, uczenie maszynowe i zaawansowane analizy wyświetlane na pulpitach w czasie rzeczywistym. Firmy, takie jak Google, Amazon i Facebook, to napędzane danymi molochy, które przejmują tradycyjne firmy, przy użyciu w tym celu właśnie pozyskanych danych. Organizacje świadczące usługi finansowe i firmy ubezpieczeniowe zawsze wykorzystywały dane, analitycy giełdowi i automatyczny trading też. Internet rzeczy (ang. *Internet of Things* — IoT) zmienia produkcję, transport, rolnictwo i opiekę zdrowotną. Od rządów i korporacji w każdej branży po organizacje non profit i instytucje edukacyjne, wszędzie dane są postrzegane jako czynnik zmieniający zasady gry. Sztuczna inteligencja i uczenie maszynowe przenikają do wszystkich aspektów naszego życia. Świat zachłysnął się danymi ze względu na potencjał, jaki reprezentują. Istnieje nawet specjalny termin dla tego zachłyśnięcia się — to ***big data***. Pojęcie to zostało zdefiniowane przez Douga Laneya z firmy Gartner w kategoriach trzech *V* (ang. *volume, variety, velocity*), czyli objętości, różnorodności i prędkości przetwarzania (danych), do której później dodał czwarte i — moim zdaniem — najważniejsze *V* (ang. *veracity*), czyli wiarygodność (weryfikacja) danych.

Przy tak dużej różnorodności, objętości i prędkości przetwarzania danych stare systemy i procesy nie są w stanie dłużej obsługiwać przedsiębiorstw w zakresie operowania danymi. Wiarygodność danych jest nawet jeszcze większym problemem dla zaawansowanych systemów analiz i sztucznej inteligencji, gdzie zasada „GIGO" (ang. *garbage in = garbage out*), czyli w wolnym tłumaczeniu „śmieci na wejściu — śmieci na wyjściu", jest jeszcze bardziej kluczowa, ponieważ w modelach statystycznych i uczenia maszynowego praktycznie nie można stwierdzić, czy złe dane spowodowały podejmowanie złych decyzji, czy zły był sam model.

Aby wesprzeć te wysiłki i sprostać tym wyzwaniom, w zarządzaniu danymi następuje rewolucja dotycząca sposobu przechowywania i przetwarzania danych oraz zarządzania nimi i udostępniania ich osobom podejmującym decyzję. Technologia *big data* umożliwia skalowalność i opłacalność większego rzędu wielkości, niż jest to możliwe przy tradycyjnej infrastrukturze zarządzania danymi. Samoobsługa zaczyna wypierać starannie opracowane i pracochłonne podejścia z przeszłości, w których armie specjalistów IT tworzyły dobrze zarządzane hurtownie i składnice danych, ale wprowadzenie zmian zajmowało miesiące.

Jezioro danych (ang. *data lake*) to śmiałe nowe podejście, które wykorzystuje moc technologii *big data* i łączy ją ze zwinnością samoobsługi. Obecnie większość dużych przedsiębiorstw już wdrożyła lub jest w trakcie wdrażania jezior danych.

Książka jest oparta na wywiadach przeprowadzonych z ponad stu organizacjami, od opartych na nowej organizacji danych, takich jak Google, LinkedIn i Facebook, po administrację rządową i samorządową oraz tradycyjne przedsiębiorstwa korporacyjne. Wywiady dotyczyły inicjatyw związanych z jeziorami danych, projektów analitycznych, doświadczeń i najlepszych praktyk. Jest ona przeznaczona dla dyrektorów IT i praktykujących informatyków, którzy rozważają budowę jeziora danych, są w trakcie budowy albo już je mają, ale walczą o zwiększenie jego produktywności i stopnia zaadaptowania.

Czym jest jezioro danych? Dlaczego go potrzebujemy? Czym się różni od tego, co już mamy? Ten rozdział zawiera krótki przegląd, który zostanie szczegółowo rozwinięty w kolejnych rozdziałach. Aby to streszczenie pozostało zwięzłe, nie zamierzam tutaj szczegółowo wyjaśniać ani badać poszczególnych pojęć i koncepcji; ich szersze omówienie znajdziesz na odpowiednich etapach książki.

Podejmowanie decyzji na podstawie danych to ostatni krzyk mody. Jak już pisałem, osoby uprawnione do podejmowania decyzji wymagają danych dostarczanych przez naukę o danych, uczenie maszynowe i zaawansowane analizy wyświetlane na pulpitach w czasie rzeczywistym. Te dane wymagają „domu", a jezioro danych jest preferowanym rozwiązaniem tworzenia tego domu. Termin ten został wymyślony i po raz pierwszy opisany przez Jamesa Dixona, dyrektora ds. technicznych Pentaho, który na swoim blogu (*https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/*) napisał: „Składnicę danych można potraktować jako magazyn wody butelkowanej, oczyszczonej, zapakowanej i umożliwiającej łatwe spożycie. Natomiast jezioro danych jest dużym zbiornikiem wodnym w bardziej *naturalnym* stanie. Zawartość jeziora danych płynie ze źródła, aby wypełnić to jezioro, a jego *różni użytkownicy* mogą sprawdzać i pobierać próbki zawartości z tego jeziora lub zanurzać się w nim". Niżej pogrubioną czcionką zaznaczyłem kluczowe kwestie.

- Dane są przechowywane w oryginalnej formie i formacie (dane **naturalne** lub surowe).
- Dane są wykorzystywane przez **różnych użytkowników** (oznacza to, że są dostępne dla dużej społeczności użytkowników).

Książka jest o tym, jak zbudować jezioro danych, które udostępnia surowe (a także przetworzone) dane dużej społeczności analityków biznesowych, a nie tylko do wykorzystania w projektach IT. Powodem udostępnienia analitykom surowych danych jest potrzeba umożliwienia im wykonywania analityki samoobsługowej. Samoobsługa była ważnym megatrendem w kierunku demokratyzacji danych. Zaczęło się to od używania narzędzi wizualizacji samoobsługowej, takich jak Tableau i Qlik (czasami nazywanych **narzędziami do wykrywania danych**), które pozwalają analitykom analizować dane bez pomocy informatyków. Trend samoobsługowy jest kontynuowany przez narzędzia do przygotowywania danych, które pomagają analitykom kształtować dane dla instrumentów analitycznych, narzędzia do katalogowania, pomagające znaleźć potrzebne dane, oraz narzędzia nauki danych, które pomagają przeprowadzać zaawansowane analizy. W jeszcze bardziej zaawansowanej analizie, ogólnie określanej jako nauka o danych, jezioro danych jest podstawowym źródłem informacji również dla nowej klasy użytkowników, zwanych naukowcami zajmującymi się danymi.

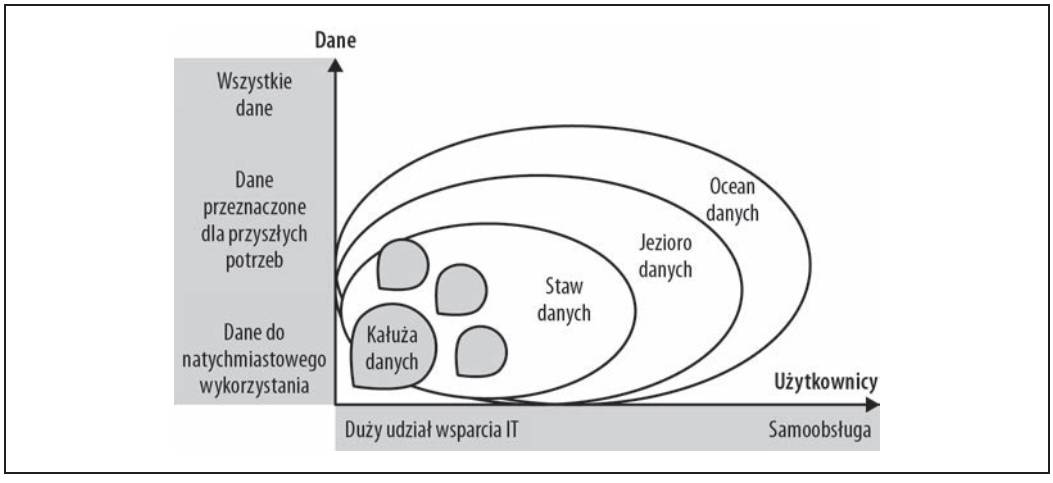
Oczywiście, dużym wyzwaniem związanym z samoobsługą jest zarządzanie danymi i ich bezpieczeństwo. Każdy zgadza się, że dane muszą być przechowywane bezpiecznie, ale w wielu regulowanych branżach istnieją opisane zasady bezpieczeństwa danych, które muszą zostać wdrożone, i nielegalne jest zapewnianie analitykom dostępu do wszystkich danych. Jest to uważane za zły pomysł nawet w niektórych branżach nieregulowanych. Powstaje więc pytanie, w jaki sposób udostępnić dane analitykom bez naruszania wewnętrznych i zewnętrznych przepisów dotyczących zgodności danych? Nazywa się to czasem demokratyzacją danych i zostanie szczegółowo omówione w kolejnych rozdziałach.

Dojrzewanie jeziora danych

Jezioro danych jest stosunkowo nową koncepcją, dlatego przydatne jest zdefiniowanie niektórych etapów dojrzewania, które można obserwować i wyraźnie wyróżnić.

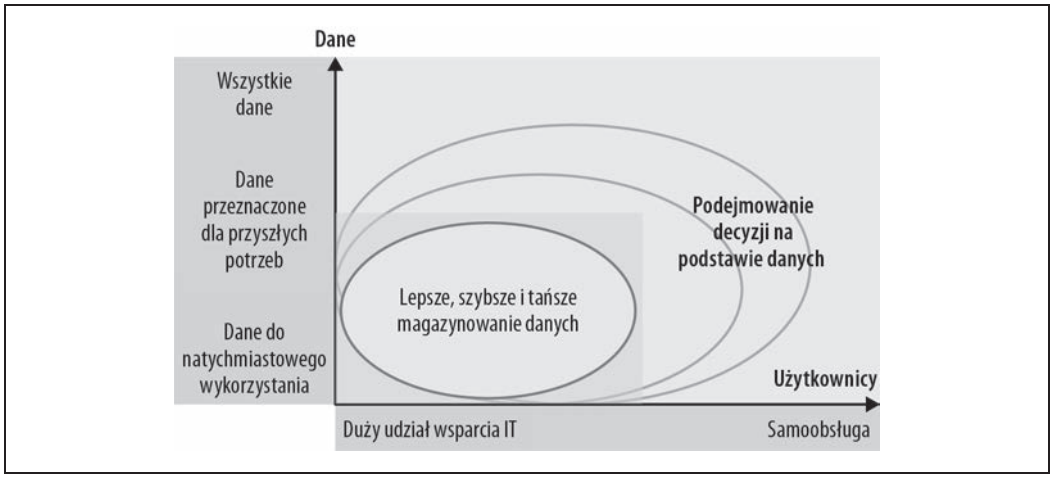
- **Kałuża danych** (ang. *data puddle*) jest w zasadzie składnicą danych służącą pojedynczemu celowi lub projektowi, zbudowaną przy użyciu technologii *big data*. Zwykle jest to pierwszy krok w przyjęciu technologii *big data*. Dane w kałuży danych są ładowane dla celów pojedynczego projektu lub zespołu. Zwykle są dobrze znane i rozumiane, a powodem użycia technologii *big data* zamiast tradycyjnej hurtowni danych jest obniżenie kosztów i zapewnienie lepszej wydajności.
- **Staw danych** (ang. *data pond*) to zbiór kałuż danych. Może przypominać źle zaprojektowaną hurtownię danych, która w rzeczywistości jest zbiorem kolokowanych składnic danych lub może być odciążeniem istniejącej hurtowni danych. Chociaż niższe koszty technologiczne i lepsza skalowalność zapewniają oczywiste i atrakcyjne korzyści, te konstrukcje wciąż wymagają wysokiego poziom uczestnictwa działu IT. Ponadto stawy danych ograniczają dane tylko do potrzeb określonego projektu i wykorzystują je tylko do projektu, który ich wymaga. Biorąc pod uwagę wysokie koszty IT i ograniczoną dostępność danych, stawy danych zdecydowanie nie pomagają w osiągnięciu założeń demokratyzacji wykorzystania danych ani nie zapewniają użytkownikom biznesowym wsparcia dla samoobsługi i podejmowania decyzji na podstawie danych.
- **Jezioro danych** (ang. *data lake*) różni się od stawu danych pod dwoma istotnymi względami. Po pierwsze, oferuje samoobsługę, w której użytkownicy biznesowi mogą znajdować i wykorzystywać żądane zbiory danych bez konieczności wzywania na pomoc działu IT. Po drugie, ma zawierać dane, których potrzebować mogą użytkownicy biznesowi, nawet jeśli nie realizują w danym momencie żadnego konkretnego projektu.
- **Ocean danych** (ang. *data ocean*) rozszerza samoobsługę danych i podejmowanie decyzji na podstawie danych na całe dane korporacyjne, gdziekolwiek się znajdują, niezależnie od tego, czy zostały załadowane do jeziora danych, czy nie.

Rysunek 1.1 ilustruje różnice między tymi pojęciami. Wraz z dojrzewaniem i rozwojem — od kałuży, przez staw i jezioro, aż do oceanu — rośnie ilość danych i liczba użytkowników, czasami dość znacznie. Wzorzec wykorzystania zmienia się od wysokiego poziomu zaangażowania działu IT do samoobsługi, a dane wykraczają poza to, co jest natychmiast potrzebne do realizowanych projektów.



Rysunek 1.1. Cztery etapy dojrzałości

Kluczową różnicą między stawem danych a jeziorem danych jest to, na czym się koncentrują. Stawy danych mają zapewnić tańszą i bardziej skalowalną technologiczną alternatywę dla istniejących relacyjnych hurtowni i składnic danych. Natomiast te drugie koncentrują się na umożliwianiu wykonywania rutynowych zapytań zwracających wyniki gotowe do wykorzystania w środowisku produkcyjnym, umożliwiając użytkownikom biznesowym wykorzystywanie danych do podejmowania własnych decyzji, przeprowadzania analiz ad hoc i eksperymentowania z różnymi nowymi rodzajami danych i narzędzi, tak jak pokazano na rysunku 1.2.



Rysunek 1.2. Obietnica korzyści zapewnianych przez jeziora danych

Zanim przejdziemy do tego, co jest potrzebne do utworzenia udanego jeziora danych, przyjrzyjmy się bliżej dwóm etapom dojrzewania, które prowadzą do jego powstania.

Kałuże danych

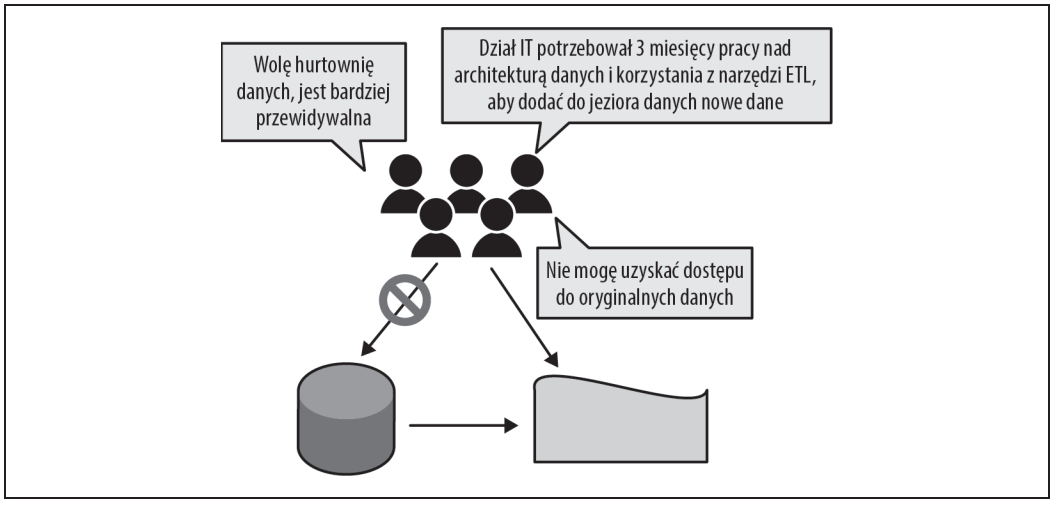
Kałuże danych buduje się zwykle dla małego, skupionego na celu zespołu lub specjalistycznego przypadku użycia. Te „kałuże" to skromnej wielkości zbiory danych, których właścicielem jest często jeden zespół, i najczęściej są budowane w chmurze przez jednostki biznesowe korzystające z shadow IT. W dobie hurtowni danych każdy zespół z reguły budował relacyjną bazę danych dla każdego ze swoich projektów. Proces budowania kałuży danych jest bardzo podobny, z tym wyjątkiem, że wykorzystuje technologię *big data*. Najczęściej kałuże danych są budowane dla projektów, które wymagają dużej mocy i skali *big data*. Do tej kategorii zalicza się wiele zaawansowanych projektów analitycznych, takich jak skupiające się na odpływie klientów lub utrzymywaniu predykcyjnym.

Czasami kałuże danych buduje się, by wesprzeć dział IT w automatyzowaniu intensywnych obliczeniowo i wymagających dużej ilości danych procesów, takich odciążenie narzędzi wyodrębniania, przekształcania i ładowania (ang. *Extract*, *Transform*, *Load* — ETL) — zostaną one omówione szczegółowo w kolejnych rozdziałach — w których wszystkie prace związane z transformacją są przenoszone z hurtowni danych lub drogich narzędzi ETL na platformę *big data*. Innym powszechnym zastosowaniem kałuży danych jest obsługa pojedynczego zespołu poprzez udostępnienie obszaru roboczego, zwanego piaskownicą (ang. *sandbox*), gdzie mogą eksperymentować naukowcy zajmującymi się danymi.

Kałuże danych zwykle mają niewielki zakres i zawierają ograniczoną różnorodność danych. Są zapełniane przez małe, dedykowane strumienie danych, a ich konstruowanie i utrzymanie wymaga wysoce wyspecjalizowanego zespołu technicznego lub dużego zaangażowania ze strony IT.

Stawy danych

Staw danych to zbiór kałuż danych. Tak jak kałuże danych można traktować jak składnice danych zbudowane przy użyciu technologii *big data*, tak staw danych można uznać za hurtownię zbudowaną przy użyciu tej technologii. Może on powstać w sposób naturalny wraz z dodawaniem do platformy *big data* kolejnych kałuż. Innym popularnym podejściem do tworzenia stawów danych jest odciążenie hurtowni danych. W przeciwieństwie do odciążania narzędzi ETL, które wykorzystuje technologię *big data* do wykonania niektórych operacji przetwarzania wymaganych do zapełnienia hurtowni danych, ta koncepcja polega na umieszczeniu wszystkich danych w hurtowni danych i załadowaniu ich na platformę *big data*. Ten pomysł często związany jest z ostatecznym pozbyciem się hurtowni danych, aby obniżyć koszty i poprawić wydajność, ponieważ platformy *big data* są znacznie tańsze i bardziej skalowalne niż relacyjne bazy danych. Jednak samo odciążenie hurtowni danych nie zapewnia analitykom dostępu do surowych danych. Ponieważ rygorystyczna architektura i formalistyczne zarządzanie stosowane w hurtowni danych są nadal utrzymywane, dana organizacja nie może sprostać wszystkim wyzwaniom hurtowni danych, dopóki podstawą wszystkich raportów pozostają długie i kosztowne cykle zmian, złożone transformacje i ręczne kodowanie. Ponadto analitycy często nie lubią przechodzić od precyzyjnie wyregulowanej hurtowni danych z błyskawicznymi zapytaniami do znacznie mniej przewidywalnej platformy *big data*, gdzie duże zapytania wsadowe mogą działać szybciej niż w hurtowni danych, ale obsługa bardziej typowych mniejszych zapytań może zajmować kilka minut. Rysunek 1.3 ilustruje niektóre typowe ograniczenia stawów danych, takie jak brak przewidywalności, zwinności i dostępu do oryginalnych nieprzetworzonych danych.



Rysunek 1.3. Wady odciążania hurtowni danych

Udane tworzenie jeziora danych

Czego więc potrzeba, aby z powodzeniem utworzyć jezioro danych? Jak w przypadku każdego projektu, niezbędne wymagania to zgodność ze strategią biznesową firmy, posiadanie sponsoringu wykonawczego i szerokie poparcie. Ponadto na podstawie wywiadów z dziesiątkami firm wdrażających jeziora danych z różnym poziomem sukcesu można zidentyfikować trzy kluczowe warunki wstępne; są to:

- właściwa platforma,
- właściwe dane,
- właściwe interfejsy.

Właściwa platforma

Technologie *big data* (np. Hadoop) i rozwiązania chmurowe, takie jak Amazon Web Services (AWS), Microsoft Azure i Google Cloud Platform, to najbardziej popularne platformy dla jeziora danych. Technologie te mają kilka ważnych wspólnych zalet. Oto one.

Objętość danych
: Platformy te zostały zaprojektowane do skalowania w poziomie — innymi słowy, do skalowania w nieskończoność bez znaczącej degradacji wydajności.

Koszt
: Zawsze mieliśmy możliwość przechowywania dużych ilości danych na dość niedrogich nośnikach, takich jak taśmy, dyski WORM i dyski twarde. Jednak dopiero w technologii *big data* zyskaliśmy możliwość zarówno przechowywania, jak i przetwarzania ogromnych ilości danych przy tak niewielkich kosztach — zwykle na poziomie od jednej dziesiątej do jednej setnej kosztu komercyjnej relacyjnej bazy danych.

Różnorodność

Platformy te używają systemów plików lub obiektowych baz danych, które pozwalają im przechowywać wszystkie rodzaje plików: Hadoop HDFS, MapR FS, Simple Storage Service (S3) AWS itd. W przeciwieństwie do relacyjnej bazy danych, która wymaga predefiniowanej struktury danych (**schematu zapisu**), dla systemu plików lub obiektowej bazy danych nie ma znaczenia, co zapisujesz. Oczywiście, aby sensownie przetwarzać dane, musisz znać ich schemat, ale tylko wtedy, gdy używasz tych danych. Takie podejście nazywa się schematem odczytu i jest to jedna z ważnych zalet platform *big data*, umożliwiająca tzw. „bezproblemowe spożywanie" (ang. *frictionless ingestion*). Innymi słowy, dane mogą być ładowane absolutnie bez żadnego przetwarzania, w przeciwieństwie do relacyjnej bazy danych, gdzie dane nie mogą zostać załadowane, dopóki nie zostaną przekonwertowane na schemat i format oczekiwany przez bazę danych.

Przyszłościowość rozwiązania

Ponieważ nasze wymagania i świat, w którym żyjemy, podlegają zmianom, bardzo ważne jest zagwarantowanie, żeby posiadane przez nas dane mogły być wykorzystywane do naszych przyszłych potrzeb. Jeśli dzisiaj dane są przechowywane w relacyjnej bazie danych, można uzyskać do nich dostęp tylko przy użyciu relacyjnej bazy danych. Natomiast Hadoop i inne platformy *big data* są bardzo modularne. Ten sam plik może być używany przez różne silniki przetwarzania i programy — od zapytań Hive (Hive zapewnia interfejs SQL do plików Hadoop) po skrypty Pig dla silnika Spark i niestandardowe zadania MapReduce — dostęp do tych samych plików i korzystanie z nich mogą zapewniać różne narzędzia i systemy. Ponieważ technologia *big data* szybko się rozwija, daje to ludziom pewność, że wszelkie przyszłe projekty będą nadal miały dostęp do danych przechowywanych w jeziorze danych.

Właściwe dane

Większość danych gromadzona dziś przez przedsiębiorstwa jest wyrzucana. Niewielki procent jest agregowany i przechowywany w hurtowniach danych przez kilka lat, ale najbardziej szczegółowe dane operacyjne, generowane maszynowo i stare dane historyczne są agregowane lub całkowicie wyrzucane. Utrudnia to prowadzenie analiz. Jeżeli analityk uzna np. wartość pewnych danych, które tradycyjnie zostały wyrzucone, zgromadzenie wystarczającej historii tych danych, aby przeprowadzić znaczącą analizę, może wymagać miesięcy, a nawet lat. Funkcjonalność jeziora danych polega więc na tym, że można przechowywać możliwie jak najwięcej danych do wykorzystania w przyszłości.

Tak więc jezioro danych jest jak świnka skarbonka (zobacz rysunek 1.4) — często nie wiesz, po co zapisujesz dane, ale chcesz je mieć, gdybyś ich pewnego dnia potrzebował. Co więcej, ponieważ nie wiesz, w jaki sposób skorzystasz z tych danych, nie ma sensu przedwcześnie ich konwertować ani przetwarzać. Możesz potraktować to jak podróżowanie ze świnką skarbonką po różnych krajach, dorzucanie do niej pieniędzy w walucie kraju, w którym aktualnie się znajdujesz, i przechowywanie ich w tej walucie, dopóki nie zdecydujesz, w którym kraju chcesz wydać pieniądze. Będziesz mógł wtedy wymienić je wszystkie na walutę tego kraju, zamiast niepotrzebnie wymieniać je (i ponosić koszty) za każdym razem, gdy przekraczasz granicę kolejnego kraju. Możemy podsumować, że celem jest *zapisanie jak największej ilości danych w ich natywnym formacie*.

Rysunek 1.4. Jezioro danych jest jak skarbonka, gdyż pozwala przechowywać dane w ich natywnym lub surowym formacie

Kolejnym wyzwaniem związanym z pozyskaniem właściwych danych są tzw. silosy danych. Różne działy firmy mogą gromadzić swoje własne dane, zarówno dlatego, że ich zapewnianie jest kosztowne i trudne, jak i z tego powodu, że często istnieje polityczna i wewnątrzorganizacyjna niechęć do dzielenia się tymi danymi. Jeżeli w typowym przedsiębiorstwie jedna grupa potrzebuje danych od innej grupy, musi wyjaśnić, jakich danych sobie życzy, a następnie grupa będąca właścicielem danych musi wdrożyć zadania ETL, które wyodrębnią i spakują wymagane dane. Jest to kosztowne, trudne i czasochłonne, dlatego zespoły mogą w miarę możliwości odsuwać w czasie żądania danych, a następnie, tak długo jak to możliwe, unikać ich dostarczenia. Ta dodatkowa praca jest często używana jako wymówka, aby nie udostępniać danych.

Te wyzwania (i wymówki) znikają w przypadku jeziora danych, ponieważ jezioro konsumuje surowe dane poprzez bezproblemowe ich spożywanie (w zasadzie jest to spożywanie bez przetwarzania). Dobrze zarządzane jezioro danych jest również scentralizowane i oferuje przejrzysty proces pozyskiwania dla ludzi w całej organizacji, więc własność staje się znacznie mniejszą barierą.

Właściwy interfejs

Kiedy już mamy odpowiednią platformę i załadowaliśmy dane, docieramy do bardziej skomplikowanych aspektów jeziora danych. Właśnie na tym polu, podczas doboru odpowiedniego interfejsu, zawodzi większość firm. Aby zyskać szeroką adaptację i czerpać korzyści z pomagania użytkownikom biznesowym w podejmowaniu decyzji na podstawie danych, rozwiązania oferowane przez firmy muszą być samoobsługowe, a ich użytkownicy muszą znaleźć, zrozumieć i wykorzystać dane bez konieczności pomocy ze strony działu IT. Ten dział nie będzie po prostu w stanie obsłużyć tak dużej społeczności użytkowników i tak ogromnej różnorodności danych.

Istnieją dwa aspekty umożliwiające samoobsługę; są to dostarczanie użytkownikom danych przy wykorzystaniu odpowiedniego poziomu wiedzy i doświadczenia oraz zagwarantowanie, że użytkownicy będą w stanie znaleźć właściwe dane.

Dostarczanie danych przy odpowiednim poziomie wiedzy specjalistycznej

Aby jezioro danych zostało szeroko zaakceptowane, powinni używać go wszyscy, od naukowców zajmujących się danymi po analityków biznesowych. Jednak przy rozważaniu tak różnorodnych odbiorców z różnymi potrzebami i poziomami umiejętności musimy być ostrożni, aby udostępnić właściwe dane odpowiedniej populacji użytkowników.

Analitycy często nie mają np. umiejętności używania surowych danych. Surowe dane zwykle zawierają wiele szczegółów i są zbyt „drobnoziarniste", a także powodują wiele problemów z jakością, aby można było łatwo z nich korzystać. Jeżeli przykładowo gromadzimy dane o sprzedaży z różnych krajów, które używają różnych aplikacji, dane będą dostarczane w odmiennych formatach z różnymi polami (np. jeden kraj może mieć podatek VAT, a inny nie), jednostkami miary i walutami (takimi jak funty lub kilogramy, dolary i euro itp.).

Aby analitycy mogli korzystać z tych danych, należy je **zharmonizować**, czyli umieścić w ujednoliconym schemacie o tych samych nazwach pól i jednostkach miary, a często również zagregować do postaci dziennej sprzedaży poszczególnych produktów i do konkretnych klientów. Innymi słowy, analitycy chcą „ugotowanych" gotowych posiłków, a nie surowych danych.

Naukowcy zajmujący się danymi stanowią całkowite przeciwieństwo. Dla nich „ugotowane" dane często tracą wartość, której szukają. Jeżeli chcą np. sprawdzić, jak często kupowane są razem dwa określone produkty, ale jedyne dostępne informacje to dzienne sumy sprzedaży według poszczególnych produktów, utkną w martwym punkcie. Naukowcy zajmujący się danymi są jak szefowie kuchni, którzy potrzebują surowych ingrediencji, aby tworzyć swoje kulinarne lub analityczne arcydzieła.

Podczas lektury tej książki dowiesz się, jak zaspokoić rozbieżne potrzeby, konfigurując wiele **stref** lub obszarów zawierających dane spełniające określone wymagania. Przykładowo strefa lądowania (lub inaczej surowa) zawiera oryginalne dane dostarczone do jeziora, podczas gdy strefa produkcyjna (zwana też złotą) zawiera wysokiej jakości, zarządzane dane. Przyjrzymy się pokrótce tym strefom w punkcie „Organizowanie jeziora danych", w kolejnym podrozdziale, a szersze omówienie tych kwestii znajdziesz w rozdziale 7.

Dotarcie do danych

Większość firm, z którymi przeprowadzałem wywiady, ustala model zwany „kupowaniem danych", w którym analitycy używają interfejsu w stylu Amazon.com, aby znaleźć, zrozumieć, ocenić, adnotować i skonsumować dane. Zalety tego podejścia są różnorodne.

Znajomy interfejs
: Większość osób bez problemu potrafi poruszać się w sklepach internetowych i czują się komfortowo, korzystając ze słów kluczowych, nawigacji fasetowej, ocen i komentarzy, więc nie jest wymagane żadne szkolenie.

Wyszukiwanie fasetowe
: Wyszukiwarki są zoptymalizowane pod kątem wyszukiwania fasetowego. Jest ono bardzo pomocne, gdy liczba możliwych wyników wyszukiwania jest duża i użytkownik próbuje zredukować ją do właściwego wyniku. Gdybyś przeszukiwał np. Amazon pod kątem tosterów (zobacz rysunek 1.5), fasety wymieniałyby producentów, pozwalały określić, czy toster powinien umożliwiać opiekanie

Rysunek 1.5. Interfejs sklepu internetowego

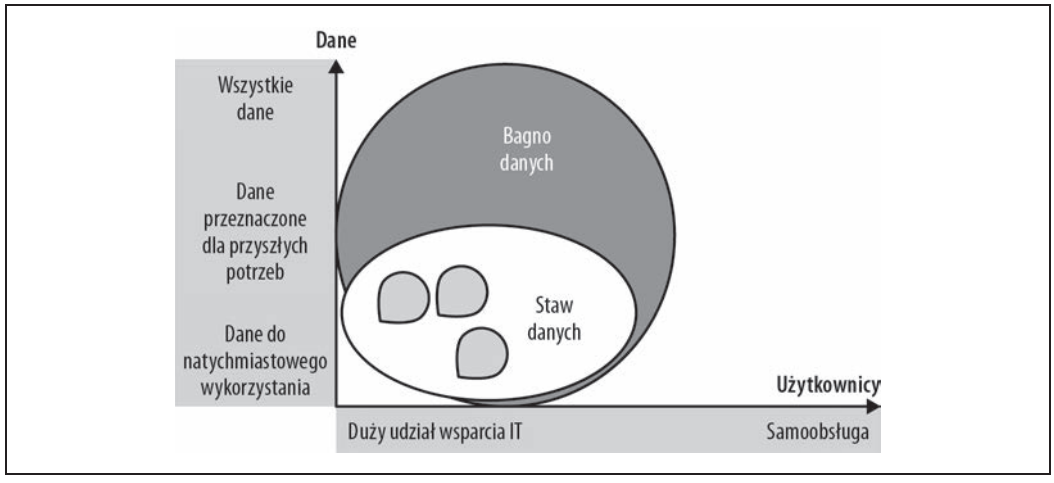
bajgli, ile plasterków pieczywa powinien jednocześnie opiekać itd. Podobnie jest z użytkownikami wyszukującymi właściwe zbiory danych — fasety mogą pomóc im określić wymagane atrybuty zbioru danych, typ i format zbioru danych, system, który je przechowuje, rozmiar i wiek danych, dział, który jest jego właścicielem, jakie posiada uprawnienia i dowolną liczbę innych przydatnych cech.

Ranking i sortowanie
: Możliwość prezentowania i sortowania zasobów danych, szeroko obsługiwana przez wyszukiwarki, to ważne cechy przy wyborze odpowiedniego składnika aktywów na podstawie określonych kryteriów.

Wyszukiwanie kontekstowe
: W miarę, jak katalogi będą stawać się inteligentniejsze, będzie rosnąć znaczenie możliwości wyszukiwania zasobów danych przy użyciu semantycznego zrozumienia tego, czego szukają analitycy. Przykładowo sprzedawca szukający klientów może poszukiwać określonych perspektyw, podczas gdy osoba z działu pomocy technicznej szukająca klientów może poszukiwać istniejących klientów.

Bagno danych

Chociaż jeziora danych zawsze zaczynają się od dobrych intencji, czasami sprawy przybierają zły obrót i w rezultacie otrzymujemy **bagno danych** (ang. *data swamp*). Bagno danych to staw danych, który urósł do rozmiaru jeziora danych, ale nie udało mu się przyciągnąć szerokiej społeczności analityków, zazwyczaj z powodu braku ułatwień związanych z samoobsługą i zarządzaniem danych. W najlepszym przypadku bagno danych jest używane jako staw danych, a w najgorszym nie jest w ogóle używane. Chociaż różne zespoły używają niewielkich obszarów jeziora dla swoich projektów (biały obszar stawu danych na rysunku 1.6), większość danych (zaznaczona na rysunku ciemnym kolorem) pozostaje nieudokumentowana i nie nadaje się do użytku.



Rysunek 1.6. Bagno danych

Gdy pojawiły się jeziora danych, wiele firm kupowało klastry Hadoop i wypełniało je surowymi danymi, bez jasnego zrozumienia, w jaki sposób będą wykorzystywane. Doprowadziło to do powstania ogromnych bagien danych z milionami plików zawierającymi petabajty danych bez żadnego sposobu umożliwiającego ich zrozumienie.

Tylko najbardziej wyrafinowani użytkownicy byli w stanie poruszać się po bagnach, zwykle wyodrębniając z nich małe kałuże, z których mogli korzystać oni i ich zespoły. Co więcej, regulacje prawne uniemożliwiły otwieranie tych bagien dla szerokich grup odbiorców bez ochrony poufnych danych. Ponieważ nikt nie był w stanie określić, gdzie znajdują się wrażliwe dane, użytkownicy nie mogli uzyskać dostępu, a dane w dużej mierze pozostały bezużyteczne i nieużywane. Jeden z naukowców zajmujących się danymi podzielił się ze mną swoim doświadczeniem na temat tego, jak jego firma zbudowała jezioro danych, szyfrując dla ochrony wszystkie dane w jeziorze. Przed odszyfrowaniem i pozwoleniem ich użycia wymagała od naukowców udowodnienia, że dane, których szukali, nie były wrażliwe. Okazało się to być jak paragraf 22: ponieważ wszystko było zaszyfrowane, wspomniany przeze mnie naukowiec zajmujący się danymi nie tylko nie mógł niczego znaleźć, nie mógł też udowodnić, że nie są to dane wrażliwe. W rezultacie nikt nie korzystał z tego jeziora danych (lub, jak sam je nazywał, bagna).

Wskazówki dotyczące sukcesu w budowaniu jezior danych

Ponieważ wiemy już, co jest potrzebne, aby jezioro danych odniosło sukces, i na jakie pułapki uważać, zastanówmy się, jak zabrać się do budowania takiego jeziora danych. Zazwyczaj firmy wykorzystują następujący proces.

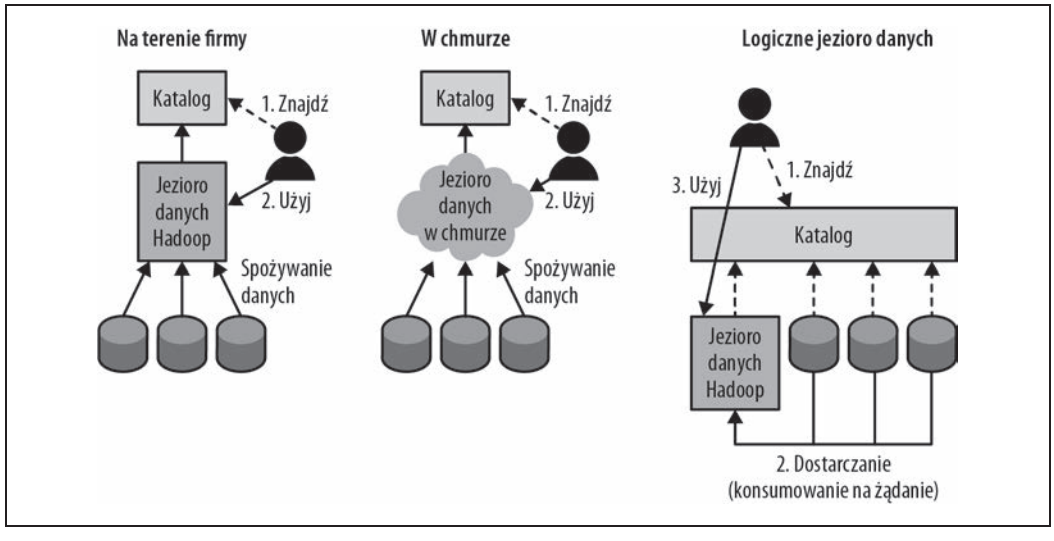
1. Utworzenie infrastruktury (skonfigurowanie i uruchomienie klastra Hadoop).
2. Uporządkowanie jeziora danych (utworzenie stref do użytku przez różne społeczności użytkowników i pobieranie danych).

3. Skonfigurowanie jeziora danych pod kątem samoobsługi (utworzenie katalogu zasobów danych, ustawienie uprawnień i udostępnienie narzędzi dla analityków).
4. Otwarcie jeziora danych dla użytkowników.

Tworzenie jeziora danych

Kiedy zacząłem pisać tę książkę w roku 2015, większość przedsiębiorstw budowała jeziora danych w swoich siedzibach z wykorzystaniem zarówno rozwiązań typu open source, jak i komercyjnych dystrybucji Hadoop. Do roku 2018 co najmniej połowa przedsiębiorstw budowała swoje jeziora danych w całości w chmurze lub tworzyła hybrydowe jeziora danych, mieszczące się i w siedzibie firmy, i w chmurze. Sporo firm wytworzyło również wiele jezior danych. Cała ta różnorodność każe przedsiębiorstwom przedefiniować to, czym jest jezioro danych. Widzimy teraz koncepcję *logicznego jeziora danych* — wirtualnego jeziora danych rozmieszczonego na wielu heterogenicznych systemach. Leżące u jego podstaw systemy mogą być bazami danych Hadoop, relacyjnymi lub NoSQL, znajdującymi się na terenie firmy lub w chmurze.

Na rysunku 1.7 przedstawiono porównanie tych trzech podejść. Wszystkie oferują katalog, w którym użytkownicy przeprowadzają kwerendy, aby znaleźć potrzebne dane. Te zasoby danych już znajdują się w jeziorze danych Hadoop lub są do niego dostarczane, by analitycy mogli z nich korzystać.



Rysunek 1.7. Różne architektury jeziora danych

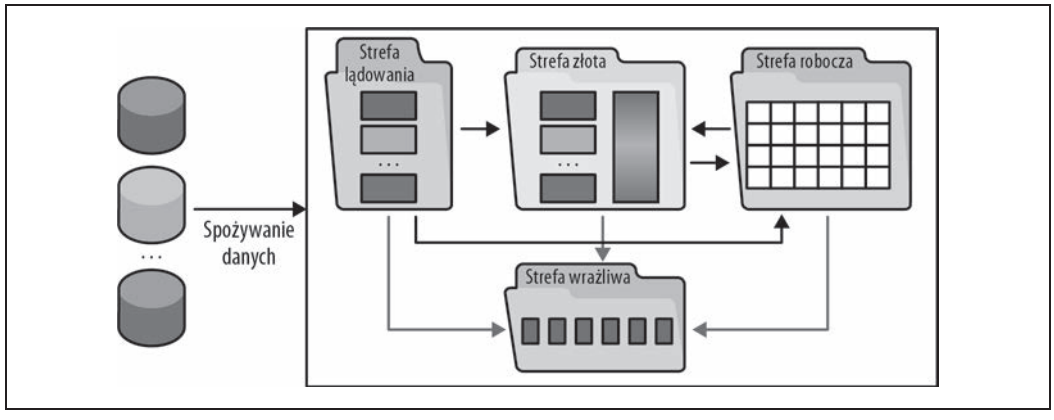
Organizowanie jeziora danych

Większość jezior danych, które napotkałem, jest zorganizowana mniej więcej w ten sam sposób, czyli posiada różne strefy.

- Strefa **surowa** lub **lądowania**, w której dane są pobierane i przechowywane w jak najbardziej pierwotnym stanie (w miarę możliwości).

- Strefa **złota** lub **produkcyjna**, w której przechowywane są czyste, przetworzone dane.
- Strefa **robocza** lub **programistyczna** (ang. *dev*), w której swoje zadania wykonują bardziej zaawansowani technicznie użytkownicy, tacy jak naukowcy zajmujący się danymi i inżynierowie danych. Ta strefa może być zorganizowana przez użytkownika, według projektu, tematu lub na wiele innych sposobów. Gdy praca analityczna wykonana w strefie programistycznej staje się pracą produkcyjną, zostaje przeniesiona do strefy złotej.
- Strefa **wrażliwa**, która zawiera poufne dane.

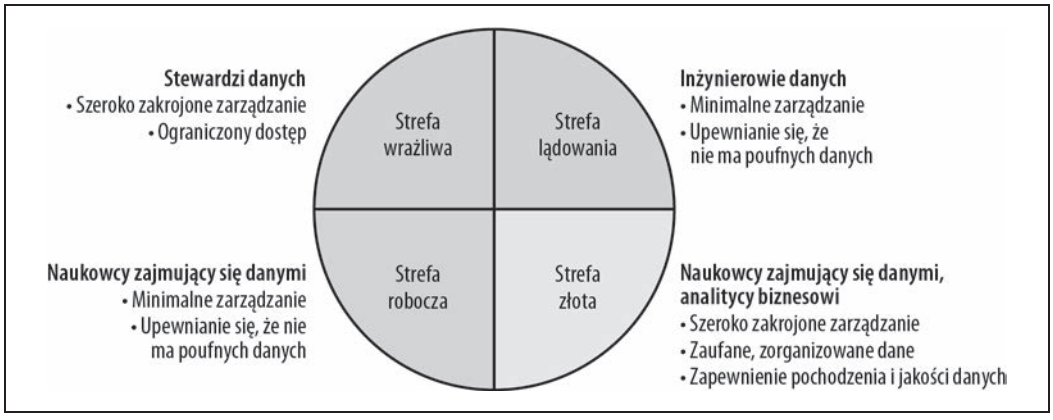
Tę organizację jezior danych można zobaczyć na rysunku 1.8.



Rysunek 1.8. Strefy typowego jeziora danych

Przez wiele lat wśród zespołów zarządzających danymi przeważało przeświadczenie, że dane powinny podlegać tym samym regulacjom bez względu na ich lokalizację lub cel. Jednak od kilku lat analitycy branżowi z firmy Gartner promują koncepcję **multimodalnego IT**, która zasadniczo polega na tym, że zarządzanie powinno odzwierciedlać wykorzystywanie danych i wymagania społeczności użytkowników. Podejście to zostało powszechnie przyjęte przez zespoły jezior danych, z różnymi strefami, o różnych poziomach zarządzania i odmiennych umowach dotyczących poziomu usług (ang. *Service-Level Agreement* — SLA). Dane w strefie złotej są np. zwykle silnie zarządzane, dobrze zorganizowane i udokumentowane oraz zapewniają jakość i świeżość umów SLA, podczas gdy dane w strefie roboczej mają minimalny zakres zarządzania (głównie polega to na upewnieniu się, że nie ma w niej danych wrażliwych) i umowy SLA, które mogą się różnić w zależności od projektu.

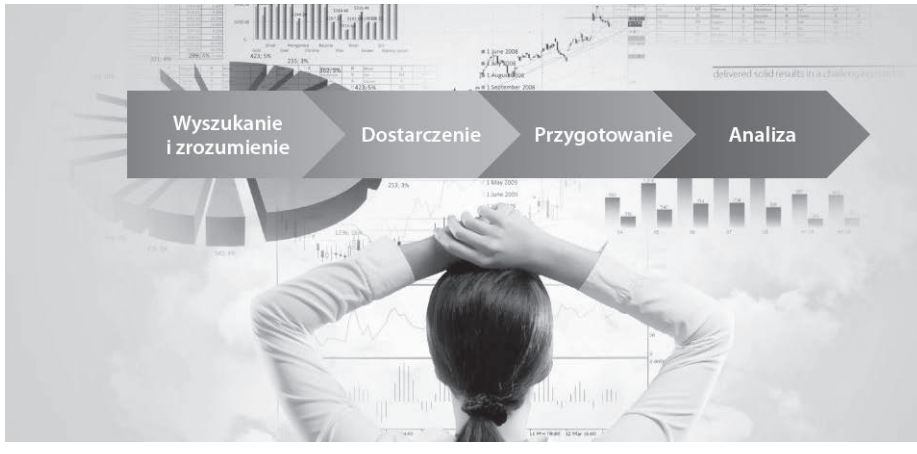
Różne społeczności użytkowników naturalnie kierują się do różnych stref. Analitycy biznesowi wykorzystują dane głównie w strefie złotej, inżynierowie danych pracują na danych w strefie surowej (przekształcając je w dane produkcyjne przeznaczone dla strefy złotej), a naukowcy zajmujący się danymi przeprowadzają swoje eksperymenty w strefie roboczej. Chociaż dla każdej strefy wymagane jest pewne zarządzanie, aby upewnić się, że poufne dane zostaną wykryte i zabezpieczone, stewardzi danych głównie skupiają się na danych w strefach wrażliwych i złotych, aby zagwarantować zgodność z regulacjami prawnymi i firmowymi. Rysunek 1.9 ilustruje różne poziomy zarządzania i odmienne społeczności użytkowników dla różnych stref.



Rysunek 1.9. Oczekiwania dotyczące zarządzania według stref

Konfiguracja jeziora danych pod kątem samoobsługi

Zazwyczaj analitycy, czy to analitycy biznesowi, czy analitycy danych, lub naukowcy zajmujący się danymi, aby wykonać swoją pracę, przechodzą cztery kroki. Kroki te przedstawiłem na rysunku 1.10.



Rysunek 1.10. Cztery etapy analizy

Pierwszym krokiem jest **znalezienie i zrozumienie danych**. Po znalezieniu odpowiednich zbiorów danych trzeba je **dostarczyć**, czyli uzyskać do nich dostęp. Następnie dane często trzeba **przygotować** — tzn. oczyścić je i przekonwertować na format odpowiedni do analizy. Na koniec danych używa się do znalezienia odpowiedzi na określone pytania lub do tworzenia wizualizacji i raportów.

Pierwsze trzy kroki teoretycznie są opcjonalne: jeśli dane są dobrze znane i zrozumiane przez analityka, który ma już do nich dostęp, i mają kształt odpowiedni do przeprowadzenia analizy, analityk może wykonać tylko ostatni krok. W rzeczywistości wiele badań wykazało, że pierwsze trzy kroki zajmują do 80% czasu typowego analityka, a największy udział w tym (60%) ma pierwszy

etap poszukiwania i zrozumienia danych (zapoznaj się np. z raportem Forrester Research *Boost Your Business Insights By Converging Big Data And BI* Borisa Evelsona, z 25 marca 2015 r.: *https://www.forrester.com/report/Boost+Your+Business+Insights+By+Converging+Big+Data+And+BI/-/E-RES115633*).

Przeanalizuję poszczególne etapy, aby dać Ci lepsze wyobrażenie o tym, co dzieje się na każdym z nich.

Znajdowanie i zrozumienie danych

Dlaczego tak trudno znaleźć dane w przedsiębiorstwie? Ponieważ różnorodność i złożoność dostępnych danych znacznie przekraczają ludzkie możliwości zapamiętywania. Wyobraź sobie bardzo małą bazę danych, zawierającą tylko sto tabel (niektóre bazy danych mają tysiące lub nawet dziesiątki tysięcy tabel, więc jest to naprawdę bardzo mała baza danych). Teraz wyobraź sobie, że każda tabela ma sto pól — rozsądne założenie dla większości baz danych, zwłaszcza tych analitycznych, w których dane mają tendencję do denormalizacji. To nam daje 10 000 pól. Czy realnie możliwe jest zapamiętanie wartości 10 000 pól i tego, w których tabelach znajdują się te pola, a następnie prześledzenie ich przy każdym użyciu tych danych do czegoś nowego?

Teraz wyobraź sobie przedsiębiorstwo, które ma kilka tysięcy (lub kilkaset tysięcy) baz danych, z których każda jest o rząd wielkości większa niż nasza hipotetyczna baza danych z 10 000 pól. Kiedyś pracowałem z małym bankiem, który zatrudniał tylko 5000 pracowników, ale zdążył utworzyć 13 000 baz danych. Mogę sobie tylko wyobrazić, ile baz danych może mieć duży bank z setkami tysięcy pracowników. Napisałem, że mogę sobie to „tylko wyobrazić", ponieważ żadne z setek dużych przedsiębiorstw, z którymi współpracowałem w czasie mojej 30-letniej kariery, nie było w stanie mi powiedzieć, ile posiadali baz danych, a tym bardziej, jak wiele tabel lub pól.

Mam nadzieję, że da Ci to wyobrażenie o tym, przed jakimi wyzwaniami stają codziennie analitycy danych.

Typowy projekt polega na tym, że analitycy rozpytują, czy ktokolwiek kiedykolwiek używał określonego rodzaju danych. Są kierowani od osoby do osoby, dopóki nie natkną się na zbiór danych, który ktoś wykorzystał w jednym ze swoich projektów. Zazwyczaj nie mają pojęcia, czy jest to najlepszy zbiór danych do zastosowania, jak ten zbiór danych został wygenerowany lub nawet, czy dane są wiarygodne. Następnie stają przed strasznym wyborem, czy wykorzystać ten zbiór danych, czy rozpytywać dalej, być może nie znajdując niczego lepszego.

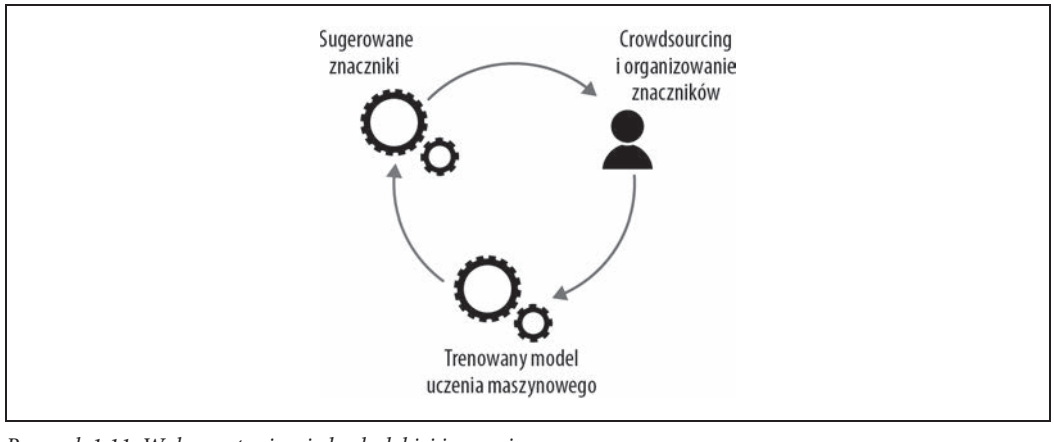
Gdy zdecydują się na użycie jakiegoś zbioru danych, spędzają dużo czasu, próbując rozszyfrować, co oznaczają zawarte w nim dane. Niektóre dane są dość oczywiste (np. nazwy klientów lub numery kont), podczas gdy inne są dość tajemnicze (co np. oznacza kod klienta 1126?). Dlatego analitycy nadal spędzają większość czasu na szukaniu osób, które pomogą im zrozumieć dane. Nazywamy te informacje „wiedzą plemienną". Innymi słowy, wiedza ta zwykle istnieje, ale każdy z członków plemienia posiada tylko jej część i trzeba ją poskładać w całość w bolesnym, długim i podatnym na błędy procesie odkrywania.

Na szczęście, istnieją nowe narzędzia zwane **crowdsourcingiem analitycznym**, które rozwiązują ten problem, zbierając wiedzę plemienną w procesie, który pozwala analitykom dokumentować zbiory danych przy użyciu prostych opisów składających się z terminów biznesowych i tworzyć indeksy

wyszukiwania, aby pomóc im znaleźć to, czego szukają. Takie narzędzia zostały opracowane na zamówienie w nowoczesnych firmach opartych na danych, takich jak Google i LinkedIn. Ponieważ w tych firmach dane są tak ważne i „każdy jest analitykiem", świadomość problemu i chęć przyczynienia się do jego rozwiązania są znacznie wyższe niż w tradycyjnych przedsiębiorstwach. Znacznie łatwiej dokumentować dopiero co utworzone zbiory danych, ponieważ informacje są świeże. Niemniej jednak, nawet w Google, chociaż niektóre popularne zbiory danych są dobrze udokumentowane, nadal istnieje duża ilość nieodkrytych lub nieudokumentowanych danych.

W tradycyjnych przedsiębiorstwach sytuacja jest znacznie gorsza. Istnieją miliony zbiorów danych (plików i tabel), które nigdy nie będą udokumentowane przez analityków, chyba że zostaną użyte — ale też takie, które nigdy nie zostaną znalezione i użyte, chyba że zostaną udokumentowane. Jedynym praktycznym rozwiązaniem jest połączenie crowdsourcingu z automatyzacją. Aby zapewnić takie rozwiązanie, wraz z moim zespołem opracowaliśmy narzędzie Waterline Data. Pomaga ono zebrać za pomocą crowdsourcingu informacje od analityków pracujących z ich zbiorami danych i stosuje je do wszystkich pozostałych nieodkrytych zbiorów danych. Proces ten nazywany jest fingerprintingiem (nadawaniem elektronicznego odcisku palca): narzędzie indeksuje wszystkie ustrukturyzowane dane w przedsiębiorstwie, dodając unikalny identyfikator do każdego pola, a ponieważ pola są adnotowane lub oznaczane przez analityków, szuka podobnych pól i sugeruje dla nich znaczniki. Gdy analitycy szukają zbiorów danych, widzą oba zbiory danych: oznaczone przez innych analityków i oznaczone automatycznie przez nasze narzędzie. Mają wtedy szansę zaakceptować lub odrzucić te sugerowane znaczniki. Następnie narzędzie stosuje uczenie maszynowe (ang. *Machine Learning* — ML), aby ulepszyć swoje automatyczne znakowanie na podstawie opinii zebranych od użytkowników.

Podstawowa koncepcja polega na tym, że sama ludzka adnotacja nie wystarczy, biorąc pod uwagę zakres i złożoność danych, podczas gdy czyste zautomatyzowane znakowanie jest niezależne, gdyż analizuje unikatowe i nieprzewidywalne cechy danych — dlatego w celu osiągnięcia najlepszych rezultatów należy stosować obie te techniki jednocześnie. Ten doskonały cykl zilustrowano na rysunku 1.11.



Rysunek 1.11. Wykorzystanie wiedzy ludzkiej i uczenia maszynowego

Uzyskiwanie dostępu do danych i ich dostarczanie

Po zidentyfikowaniu odpowiednich zbiorów danych analitycy muszą poradzić sobie z ich wykorzystaniem. Tradycyjnie dostęp jest przyznawany analitykom, gdy rozpoczynają projekt lub dołączają do niego. Później jest on już rzadko odbierany. Dlatego weterani mają dostęp do praktycznie wszystkich danych w przedsiębiorstwie, które mogą być nawet w niewielkim stopniu przydatne, podczas gdy początkujący nie mają praktycznie żadnego dostępu i dlatego nie mogą niczego znaleźć ani użyć. Aby rozwiązać problem dostępu do jeziora danych, przedsiębiorstwa zazwyczaj wybierają jedną z dwóch skrajności: przyznają każdemu pełny dostęp do wszystkich danych lub ograniczają cały dostęp, chyba że analityk może udowodnić, że go potrzebuje. Udzielanie pełnego dostępu działa w niektórych przypadkach, ale nie w branżach regulowanych. Aby sprawić, że będzie to bardziej akceptowalne, przedsiębiorstwa czasami nadają anonimowość poufnym danym (anonimizują je), ale to oznacza, że muszą wykonywać pracę, pobierając dane, których nikt nie potrzebuje. Ponadto wraz ze zmianami regulacji coraz większą ilość danych trzeba anonimizować (ten temat zostanie omówiony szerzej w kolejnych rozdziałach).

Bardziej praktycznym podejściem jest publikowanie informacji o wszystkich zbiorach danych w katalogu metadanych, żeby analitycy mogli znaleźć przydatne zbiory danych, a następnie zażądać do nich dostępu, jeśli będzie trzeba. Takie żądania zazwyczaj zawierają uzasadnienie dostępu, informację o projekcie, który wymaga danych, i potrzebnym czasie dostępu. Te żądania są kierowane do stewarda danych. Jeśli zatwierdzi dostęp, zostaje on przyznany na pewien przedział czasu. Ten okres może zostać przedłużony, ale nie na czas nieokreślony, co eliminuje wcześniejszy problem dostępu. Przychodzące żądanie może również zainicjować pracę w celu zidentyfikowania poufnych danych, ale teraz robi się to tylko wtedy, gdy trzeba.

Dostarczanie danych lub fizyczny dostęp mogą zostać przyznane na wiele sposobów.

- Użytkownicy mogą uzyskać dostęp do odczytu całego zbioru danych.
- Jeśli należy przyznać tylko częściowy dostęp, użytkownikowi może zostać przyznana kopia pliku zawierającego tylko wymagane (i aktualizowane) dane albo użytkownik może otrzymać tabelę Hive lub widok zawierający tylko pola i wiersze, które powinien posiadać analityk.
- Jeśli trzeba, można wygenerować zanonimizowaną wersję zbioru danych, która zastępuje wrażliwe informacje z losowo generowanymi równoważnymi informacjami, więc wszystkie aplikacje nadal działają, ale żadne poufne dane nie wyciekają.

Przygotowanie danych

Czasami dane są idealnie oczyszczone i gotowe do analizy. Niestety, w większości przypadków dane muszą zostać właściwie opracowane, aby były odpowiednie dla analityków. Przygotowanie danych zazwyczaj obejmuje następujące czynności.

Kształtowanie
: Wybór podzbioru pól i wierszy do pracy, połączenie wielu plików i tabel (łączenie), przekształcanie i agregowanie, „szufladkowanie” (np. przejście od wartości dyskretnych do zakresów lub segmentów, czyli umieszczanie osób od 0 do 18 roku życia w kategorii „niepełnoletni”, osób w wieku 19 – 25 lat w kategorii „młodzież” itp.), konwertowanie zmiennych w funkcje (np. przekształcanie wieku na funkcję, która ma wartość 0, jeśli osoba ma ponad 65 lat, a 1, jeśli nie) i wiele innych możliwych kroków.

Oczyszczanie
: Wypełnianie brakujących wartości (np. odgadywanie brakującej płci na podstawie imienia lub wyszukiwanie adresu w bazie danych adresów), poprawianie złych wartości, rozwiązywanie sprzecznych danych, normalizowanie jednostek miary i kodów na wspólne jednostki itp.

Ujednolicanie
: Harmonizacja różnych zbiorów danych do tego samego schematu, tych samych jednostek miar, tych samych kodów itd.

Jak widać na tych kilku przykładach, przygotowanie danych wymaga wykonania wielu wyrafinowanych prac i sporo myślenia. Aby skorzystać z lekcji, której nauczyliśmy się dzięki transformacjom, kluczowa jest automatyzacja w celu uniknięcia powtarzania tych samych żmudnych kroków przy każdej tabeli i zbiorze danych, których mogą być tysiące.

Najpopularniejszym narzędziem do przygotowywania danych jest Excel. Niestety, Excel nie skaluje się do rozmiarów jeziora danych, ale mnóstwo nowych narzędzi zapewnia podobne do Excela funkcjonalności dla dużych zbiorów danych. Niektóre, takie jak Trifacta, stosują zaawansowane techniki uczenia maszynowego, aby sugerować transformacje i pomóc analitykom w przygotowywaniu danych. Wielu dużych dostawców także zadebiutowało w branży narzędzi do przygotowywania danych, a ponadto dostawcy narzędzi analitycznych, tacy jak Tableau i Qlik, również zaczynają rozszerzać swoje produkty o funkcjonalności przygotowywania danych.

Analiza i wizualizacja

Po przygotowaniu danych można je przeanalizować. Zakres analizy obejmuje zarówno tworzenie prostych raportów i wizualizacji, jak i zaawansowanych analiz i uczenia maszynowego. To bardzo dojrzała dziedzina, w której setki dostawców zapewniają rozwiązania dla każdego rodzaju analiz. Specjalnie dla jezior danych Hadoop narzędzia do analizy i wizualizacji zaprojektowane do uruchamiania natywnego i wykorzystywania mocy obliczeniowej Hadoop oferują tacy dostawcy jak Arcadia Data, AtScale i wielu innych.

Architektury jeziora danych

Większość firm, w których przeprowadzałem wywiady, początkowo sądziła, że będą miały na miejscu jedno ogromne jezioro danych, które będzie zawierało wszystkie ich dane. Gdy zarządzający nieco lepiej zrozumieli tę technologię i rozwinęły się najlepsze praktyki, większość zdała sobie sprawę, że pojedynczy punkt odniesienia nie był idealnym rozwiązaniem. Z uwagi na przepisy dotyczące suwerenności danych (np. nie wolno Ci pobierać danych z Niemiec) i naciski organizacyjne większa liczba jezior danych zazwyczaj okazywała się lepszym podejściem. Co więcej, ponieważ firmy zdały sobie sprawę ze złożoności obsługi masowo równoległego klastra i doświadczyły frustracji z powodu niezdolności do znalezienia i zatrudnienia doświadczonych administratorów platformy Hadoop oraz innych platform *big data*, zaczęły wybierać jeziora danych oparte na chmurze, w których większość komponentów sprzętowych i platformowych jest zarządzana przez ekspertów pracujących dla takich gigantów jak m.in. Amazon, Microsoft czy Google.

Jeziora danych w chmurze publicznej

Oprócz korzyści wynikających z dostępu do specjalistycznej wiedzy w zakresie technologii *big data* i krótkiego czasu wdrożenia, niski koszt przechowywania i elastyczny charakter chmury obliczeniowej sprawiły, że ta opcja stała się niezwykle atrakcyjna dla implementacji jezior danych. Ponieważ dużo danych jest przechowywanych w celu wykorzystania w przyszłości, sensowne jest zagwarantowanie jak najniższego kosztu ich przechowywania. Sprawdza się to dobrze za sprawą możliwości optymalizacji kosztów obsługiwanych przez różne warstwy przechowywania danych dostarczane przez Amazon i inne firmy: zakresy dostępu wahają się od bardzo szybkiego do dość wolnego, przy czym media o wolniejszym dostępie są znacznie tańsze.

Ponadto elastyczność przetwarzania w chmurze pozwala na budowanie bardzo dużych klastrów na żądanie, jeśli tak trzeba. Możemy porównać to z lokalnym klastrem, który ma stały rozmiar i przechowuje swoje dane w podłączanej bezpośrednio pamięci masowej (chociaż badane są nowe architektury z sieciową pamięcią masową). Oznacza to, że w miarę wypełniania węzłów danymi trzeba będzie dodawać nowe węzły w celu samego przechowywania danych. Co więcej, jeśli obciążenia analityczne są intensywne obliczeniowo i potrzebujesz większej mocy obliczeniowej, musisz dodawać kolejne węzły nawet wtedy, kiedy będziesz ich używać przez krótki czas.

W chmurze płacisz tylko za potrzebne miejsce do przechowywania danych (oznacza to, że nie musisz kupować dodatkowych węzłów obliczeniowych tylko po to, aby uzyskać więcej pamięci) i możesz uruchamiać ogromne klastry na krótkie przedziały czasu. Jeżeli masz np. lokalny klaster 100-węzłowy i zadanie, którego wykonanie zajmuje 50 godzin, nie jest praktyczne kupowanie i instalowanie 1000 węzłów tylko po to, aby wykonać to jedno zadanie szybciej. Jednak w chmurze zapłaciłbyś mniej więcej tyle samo za moc obliczeniową 100 węzłów przez 50 godzin, ile w przypadku 1000 węzłów przez 5 godzin. To ogromna zaleta elastycznych obliczeń.

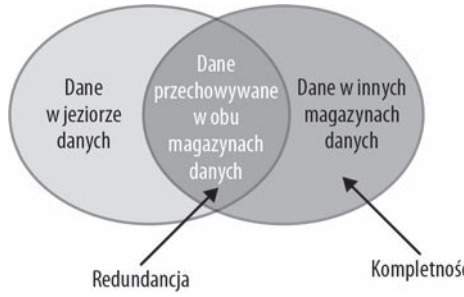
Logiczne jeziora danych

Kiedy przedsiębiorstwa zdały sobie sprawę, że posiadanie jednego scentralizowanego jeziora danych nie było dobrym rozwiązaniem, zakorzeniła się **idea logicznego jeziora danych**. Dzięki takiemu podejściu, zamiast ładować wszystkie dane do jeziora danych tylko na wypadek, gdyby ktoś mógł ich kiedyś potrzebować, dane są udostępniane analitykom za pośrednictwem centralnego katalogu lub oprogramowania do wirtualizacji danych.

Logiczne jeziora danych rozwiązują kwestie kompletności i redundancji, tak jak pokazano na rysunku 1.12.

Te problemy można podsumować w następujący sposób.

Kompletność
: Jak analitycy znajdują najlepszy zbiór danych? Jeśli analitycy mogą znaleźć tylko dane, które istnieją już w jeziorze danych, inne dane, które nie były pobrane do jeziora danych, nie zostaną znalezione ani użyte (obszar półksiężyca po prawej stronie na rysunku 1.12).



Rysunek 1.12. Kwestie kompletności i nadmiarowości

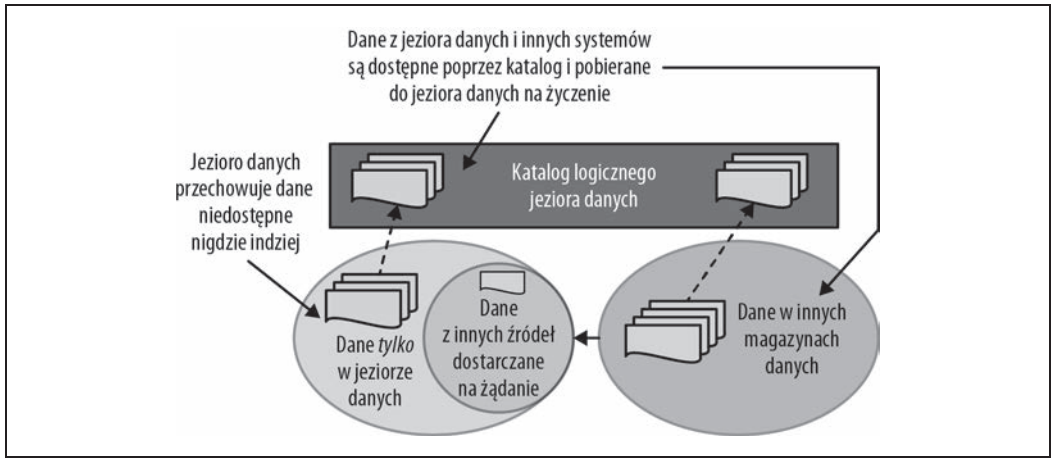
Redundancja

Jeśli do jeziora danych pobierzemy wszystkie dane, będziemy mieli redundancję między źródłami danych i jeziorem danych (część wspólna dwóch okręgów na rysunku 1.12). Aby osiągnąć kompletność w przypadku wielu jezior danych, musielibyśmy pobrać te same dane do każdego jeziora danych.

Co gorsza, w przedsiębiorstwach jest już wiele redundancji. Tradycyjnie, gdy rozpoczynany jest nowy projekt, najbardziej wygodnym i z politycznego punktu widzenia prostym podejściem jest uruchomienie przez zespół projektowy nowej bazy danych, skopiowanie danych z innych źródeł lub hurtowni danych oraz dodanie własnych unikatowych danych. Jest to znacznie łatwiejsze niż studiowanie istniejących składnic danych i negocjowanie ich współużytkowania z aktualnymi właścicielami i użytkownikami. W rezultacie dochodzi do namnożenia się w większości takich samych baz danych. Jeśli na ślepo załadujemy wszystkie dane z tych składnic danych do jeziora danych, będziemy mieli w naszym jeziorze niezwykle wysoki poziom redundancji.

Najlepsze podejście do wyzwań dotyczących kompletności i redundancji, jakie widziałem, obejmuje kilka prostych zasad.

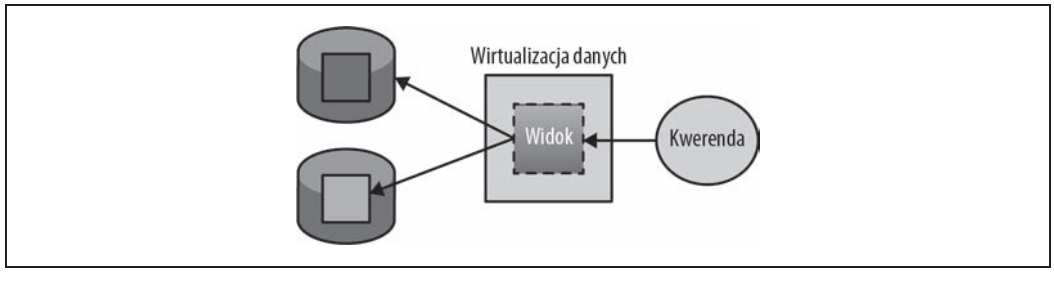
- Aby rozwiązać problem kompletności, tworzymy katalog wszystkich zasobów danych, żeby analitycy mogli znaleźć dowolny zbiór danych, który jest dostępny w przedsiębiorstwie, i zażądać dostępu do niego.
- Aby rozwiązać problem redundancji, należy zastosować proces zilustrowany na rysunku 1.13.
  - W jeziorze danych przechowuj dane, które nie są przechowywane w żadnym innym miejscu.
  - Jeśli trzeba, pobieraj do jeziora danych dane przechowywane w innych systemach; gdy jest to wymagane i potrzebne, utrzymuj ich synchronizację.
  - Pobieraj każdy zbiór danych tylko raz dla wszystkich użytkowników.



Rysunek 1.13. Zarządzanie danymi w logicznym jeziorze danych

Porównanie wirtualizacji i logicznego jeziora danych opartego na katalogu

Wirtualizacja (czasami nazywana również **federacją** lub **integracją informacji w przedsiębiorstwie**, ang. *Enterprise Information Integration* — EII) to technologia opracowana w latach 80. ubiegłego wieku i ulepszana przez kilka pokoleń do roku 2010. W zasadzie tworzy ona wirtualny widok lub tabelę, które ukrywają lokalizację i implementację fizycznych tabel. Na rysunku 1.14 widok jest tworzony przez połączenie dwóch tabel z różnych baz danych. Zapytanie następnie przeprowadza kwerendę w tym widoku, a systemowi wirtualizacji danych pozostawia znalezienie sposobu, jak uzyskać dostęp do tych dwóch baz danych i je połączyć.



Rysunek 1.14. Tworzenie niestandardowego zbioru danych za pomocą widoku

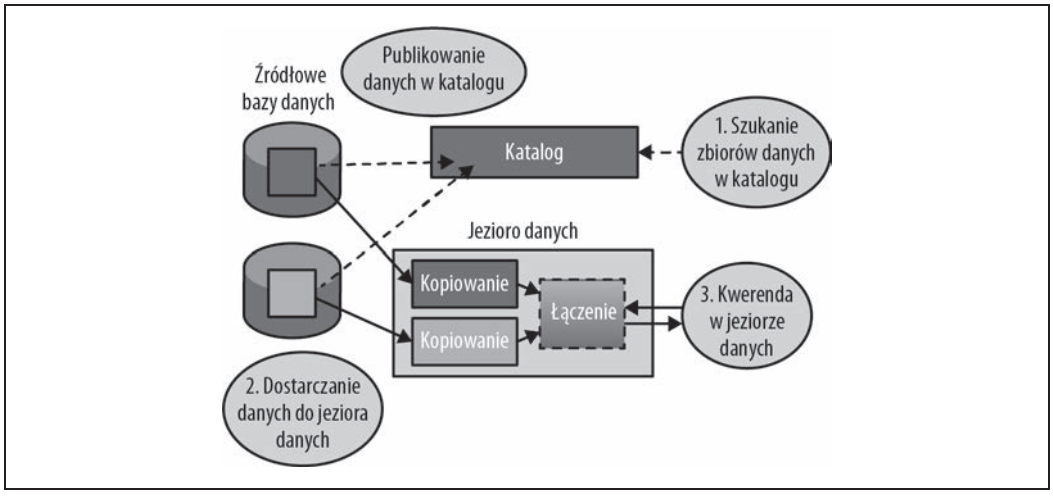
Chociaż ta technologia dobrze się sprawdza w niektórych przypadkach użycia, w logicznym jeziorze danych osiągnięcie kompletności wymagałoby opublikowania każdego zbioru danych jako wirtualnej tabeli i aktualizowania tych tabel, gdy będą zmieniać się bazowe schematy tabel.

Jeśli nawet początkowy problem publikowania każdego zasobu danych zostałby rozwiązany, widoki nadal będą generować poważne problemy. Oto one.

- Tworzenie wirtualnego widoku nie ułatwia znalezienia danych.

- Łączenie danych z wielu heterogenicznych systemów to skomplikowany i intensywny obliczeniowo proces, często powodujący ogromne obciążenia systemów i długie cykle wykonywania. Te tzw. **rozproszone łączenia** (ang. *distributed join*) tabel, które nie mieszczą się w pamięci, zwykle notorycznie wykorzystują zasoby.

Natomiast w podejściu opartym na katalogu publikowane są tylko metadane dotyczące każdego zbioru danych, umożliwiające jego znalezienie. Zbiory danych są następnie dostarczane do tego samego systemu (np. klastra Hadoop) do lokalnego przetwarzania, tak jak pokazano na rysunku 1.15.

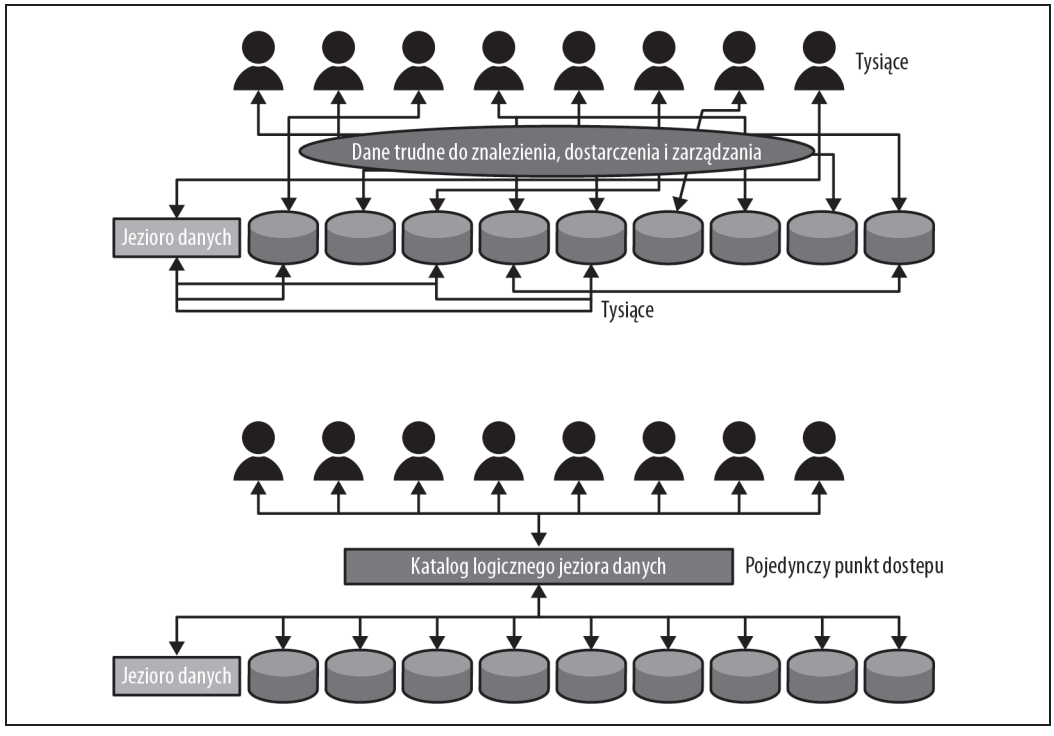


Rysunek 1.15. Dostarczanie metadanych za pośrednictwem katalogu

Oprócz tego, że analitycy mogą znaleźć wszystkie dane i uzyskać do nich dostęp, katalog przedsiębiorstwa może służyć jako pojedynczy punkt dostępu, zarządzania i audytu, tak jak pokazano na rysunku 1.16. Na górze bez scentralizowanego katalogu uzyskanie dostępu do zasobów danych jest możliwe w każdym miejscu i trudne do zarządzania oraz śledzenia. Na dole ze scentralizowanym katalogiem wszystkie żądania dostępu przechodzą przez katalog. Dostęp jest przyznawany na żądanie na określony czas i jest kontrolowany przez system.

Podsumowanie

Można stwierdzić, że wybór odpowiedniej platformy, załadowanie jej właściwymi danymi oraz zorganizowanie i skonfigurowanie samoobsługi z interfejsem odpowiednim do umiejętności i potrzeb użytkowników to klucz do utworzenia udanego jeziora danych. Dalej w książce napiszę, jak wykonać te zadania.



Rysunek 1.16. Dostarczanie danych i zarządzanie przy użyciu katalogu

Skorowidz

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

W

Y

Z

Jeziora danych i big data — zanurz się w ocean możliwości!

Koncepcja *big data*, nauka o danych i analityka danych wspomagają dziś procesy decyzyjne w przedsiębiorstwach w niespotykanym wcześniej zakresie. Zwiększają poziom efektywności pracy w wielu różnych branżach. Korporacje zaczęły więc eksperymenty z wykorzystaniem *big data* i technologii chmury, aby budować jeziora danych oraz tworzyć oparte na nich systemy podejmowania decyzji. Niejeden z tych projektów się nie powiódł, gdyż nie został dostosowany do kultury i potrzeb przedsiębiorstwa. Najwyraźniej zabrakło wiedzy, w jaki sposób skutecznie przeprowadzać tak radykalną transformację.

Ta książka jest praktycznym przewodnikiem, który ułatwia wdrażanie architektury jeziora danych (ang. *data lake*) w przedsiębiorstwie. Omówiono tu różne podejścia do jej uruchamiania i rozwijania, w tym kałuże danych (analityczne piaskownice) i stawy danych (hurtownie danych), a także budowanie jezior danych od podstaw. Opisano konfigurowanie różnych stref, co pozwala na odpowiednie rozmieszczenie zarówno surowych, jak i starannie zarządzanych i przetworzonych danych. Wyjaśniono znaczenie zarządzania dostępem do stref. Zawarto tu również wskazówki umożliwiające zachowanie zgodności z regułami zarządzania danymi przedsiębiorstwa.

W tej książce:

- wprowadzenie do hurtowni danych, *big data* i nauki o danych
- praktyczne techniki budowania jezior danych
- najlepsze praktyki dostarczania analitykom dostępu do danych
- projektowanie architektury jeziora danych oraz różne techniki implementacji
- zalety i wady różnych podejść do budowania magazynów danych i zarządzania nimi

Alex Gorelik jest CEO i założycielem firmy Waterline Data. Wcześniej zakładał startupy, zajmował się marketingiem oraz badaniami produktów, zarządzał zespołem kilkuset inżynierów i pracował nad integracją danych w firmie IBM. Jego kariera jest nieodłącznie związana z nowoczesnymi technologiami przetwarzania danych i wdrażaniem ich dla potrzeb biznesu.

Helion
helion.pl
HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
INFORMATYKA W NAJLEPSZYM WYDANIU

ISBN 978-83-283-5078-6
9 788328 350786
Cena: 49,00 zł