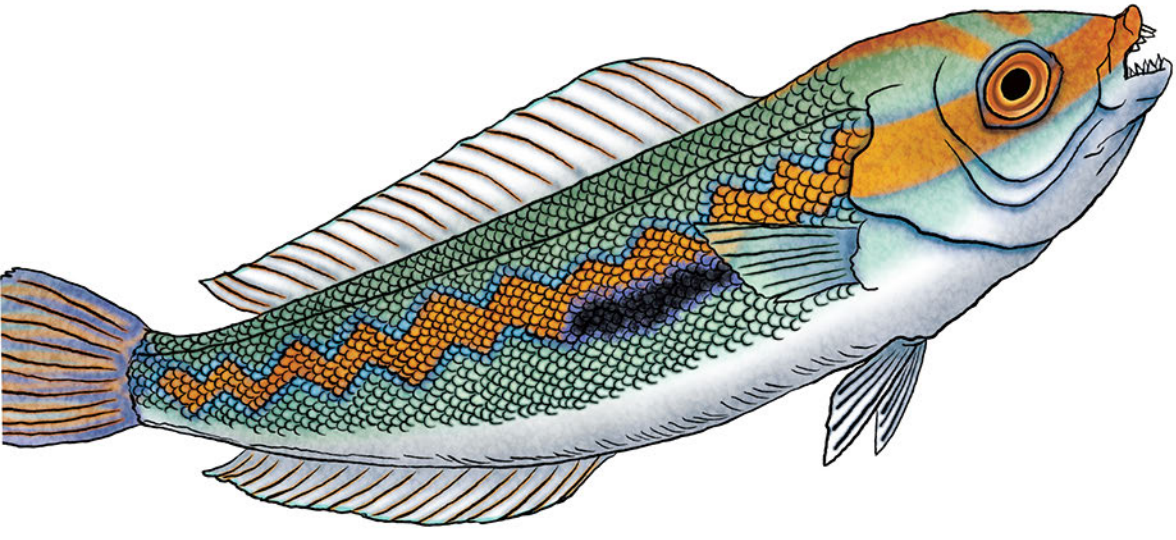
O'REILLY®

Dane grafowe w praktyce

Jak technologie grafowe ułatwiają rozwiązywanie złożonych problemów



Denise Gosnell
Matthias Broecheler

Helion

Tytuł oryginału: The Practitioner's Guide to Graph Data: Applying Graph Thinking and Graph Technologies to Solve Complex Problems

Tłumaczenie: Joanna Zatorska

ISBN: 978-83-283-7460-7

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)

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
https://ftp.helion.pl/przyklady/dagraf.zip

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

Printed in Poland.

Spis treści

ROZDZIAŁ 1.

Myślenie grafowe

Przypomnij sobie, kiedy po raz pierwszy dowiedziałeś się o technologii wykorzystującej grafy.

Prawdopodobnie znajdowałeś się w pomieszczeniu wyposażonym w białą tablicę, a zespół dyrektorów, architektów, naukowców i inżynierów dyskutował o problemach związanych z danymi. W pewnym momencie ktoś zaczął rysować połączenia między różnymi elementami danych. Po chwili ktoś inny zauważył, że linie łączące poszczególne elementy układają się w graf.

Tak zaczęła się podróż zespołu do świata grafów. Grupa zauważyła, że na podstawie relacji między elementami danych można wysnuć nowe i znaczące wnioski związane z działalnością biznesową. Prawdopodobnie zlecono jednej osobie lub małej grupie zapoznanie się z dostępnymi technikami i narzędziami służącymi do przechowywania, analizy i (lub) pobierania danych uporządkowanych w formie grafu.

Prawdopodobnie następnym ważnym odkryciem poczynionym przez zespół była łatwość objaśnienia danych w postaci grafu. Okazało się również, że dane w tym formacie są trudne w użyciu.

Brzmi znajomo?

Różne zespoły podobnie jak podczas opisanego spotkania przy białej tablicy odkrywały połączenia między elementami posiadanych danych i na tej podstawie opracowały aplikacje, z których korzystamy na co dzień. Wystarczy wspomnieć o Netfliksie, LinkedInie i GitHubie. W tych produktach wykorzystuje się relacje między danymi, tworząc integralny zasób wykorzystywany przez miliony ludzi na świecie.

W naszej książce pokazujemy, jak osiągnąć podobny cel.

Jako twórcy i użytkownicy narzędzi mieliśmy setki okazji, aby uczestniczyć w dyskusjach przy białej tablicy jako przedstawiciele obydwu stron. Zdobyliśmy doświadczenie, które ułatwia dokonywanie różnych wyborów i podejmowanie decyzji o technologiach ułatwiających wędrówkę po świecie technologii grafowych.

Nasza książka będzie przewodnikiem, dzięki któremu zrozumiesz dane o strukturze grafu i zaczniesz ich używać.

Dlaczego teraz? Kontekst technologii bazodanowych

Grafy są znane już od wielu wieków. Dlaczego więc nabrały znaczenia właśnie *teraz*?

Zanim zdecydujesz się pominąć ten podrozdział, spróbuj poświęcić mu chwilę uwagi. Zawiera on zarys historii, ale niezbyt szczegółowy, ponieważ nie chcieliśmy zbytnio przedłużać wywodu. Znajomość aspektów historycznych jest jednak kluczowa, ponieważ wzloty i upadki z przeszłości wyjaśniają, dlaczego technologia grafowa ponownie zyskała na znaczeniu.

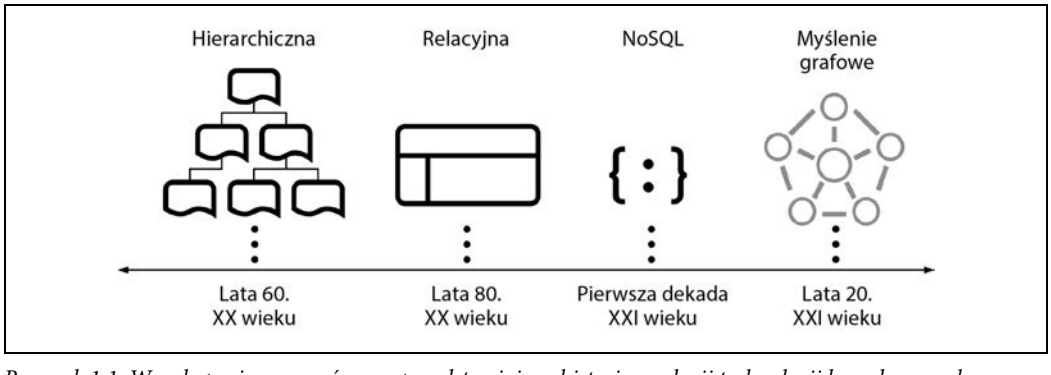
Duże znaczenie grafów wynika ze znacznych zmian technologicznych, jakie dokonały się w ciągu kilku ostatnich dekad. Wykorzystywane wcześniej technologie i bazy danych koncentrowały się na uzyskaniu możliwie *najwyższej wydajności* przechowywania danych. Na czoło wysunęły się technologie relacyjne, które umożliwiały osiągnięcie tego celu. Obecnie zależy nam, aby na podstawie danych uzyskać możliwie *najwyższą wartość*.

Powszechnie uważa się, że wartość danych rośnie dzięki występującym w nich połączeniom.

Kontekst historyczny ewolucji technologii bazodanowych rzuca sporo światła na obecny stan rzeczy, a może nawet wyjaśnia, dlaczego wybrałeś tę książkę. Historię rozwoju technologii bazodanowych można podzielić mniej więcej na trzy okresy: hierarchiczny, relacyjny i NoSQL. Opiszemy je teraz w skrócie, koncentrując się na ich znaczeniu dla tematyki tej książki.

W dalszej części tego rozdziału poznasz skróconą wersję ewolucji technologii grafowych. Skupimy się na najistotniejszych aspektach bogatej historii naszej branży. Dzięki nam zaoszczędzisz swój cenny czas, który musiałbyś przeznaczyć na samodzielne przeglądanie kolejnych odsyłaczy z Wikipedii, chociaż należy przyznać, że samodzielne odkrywanie nowej wiedzy byłoby wędrówką przez najbardziej dostępny graf wiedzy.

Ten krótki rys historyczny przeprowadzi nas od lat 60. XX wieku do dziś. Nasza wędrówka zakończy się na pukającej do naszych drzwi czwartej erze myślenia grafowego, przedstawionej na rysunku 1.1. Zapraszamy Cię na tę krótką wyprawę, ponieważ uważamy, że znajomość historii jest kluczowa, aby zacząć korzystanie z technologii grafowych w naszej branży.



Rysunek 1.1. Wysokopoziomowa oś czasu przedstawiająca historię ewolucji technologii bazodanowych i ilustrująca powstanie myślenia grafowego

Okres od lat 60. do lat 80. XX wieku — dane hierarchiczne

W literaturze technicznej technologie bazodanowe, wykorzystywane od lat 60. po lata 80. XX wieku, są określane naprzemiennie mianem hierarchicznych lub nawigacyjnych. Niezależnie jak je będziemy nazywać, w tym okresie skupiano się na organizacji danych w strukturach drzewiastych.

Technologie bazodanowe przechowywały dane w postaci połączonych ze sobą rekordów. Architekci tych systemów zakładali, że każdy rekord w takich strukturach powinien być dostępny poprzez klucz, skanowanie systemu lub przechodzenie przez połączenia między elementami drzewa danych.

Na początku lat 60. XX wieku utworzono grupę Database Task Group, działającą w ramach CODASYL, czyli Conference/Committee on Data Systems Languages. Jej celem było utworzenie pierwszego zbioru standardów branżowych. Grupa Database Task Group opracowała standard pobierania rekordów ze struktur drzewiastych. Ten wczesny standard jest znany jako „podejście CODASYL" i opiera się na następujących trzech zasadach pobierania rekordów z systemów zarządzania bazami danych[1]:

1. Używanie klucza głównego.
2. Skanowanie wszystkich rekordów po kolei.
3. Nawigacja poprzez połączenia między rekordami.

Konsorcjum CODASYL założono w 1959 roku. To ono utworzyło i ustandaryzowało język COBOL.

Pomijając aspekt historyczny, warto przytoczyć pewną ciekawostkę. Przyjmując opisane podejście, członkowie konsorcjum CODASYL zakładali pobieranie danych za pomocą kluczy, skanowania i połączeń. Dotychczas byliśmy świadkami istotnych innowacji i wdrożeń związanych z dwoma ze wspomnianych trzech standardów: z kluczami i skanowaniem.

Co się stało z trzecim standardem pobierania danych opracowanym przez CODASYL, czyli nawigacją przez połączenia między jednym rekordem a drugim? Magazynowanie, nawigowanie i pobieranie rekordów na podstawie połączeń między nimi jest cechą współczesnej technologii grafowej. Jak już wspomnieliśmy, grafy nie są nową koncepcją; technolodzy korzystają z nich od wielu lat.

Podsumowując ten okres historii, trzeba wspomnieć, że opracowane przez CODASYL technologie nawigacji wykorzystujące połączenia były zbyt trudne i powolne. Najbardziej innowacyjne rozwiązanie, jakie udało się opracować w tamtym czasie, polegało na wprowadzeniu B-drzew, czyli samoorganizujących się drzewiastych struktur danych, które służyły do strukturalnej optymalizacji problemów z wydajnością. W tym kontekście B-drzewa przyspieszały pobieranie rekordów, ponieważ dostarczały alternatywnych ścieżek dostępu do połączonych rekordów[2].

[1] T. William Olle, *The CODASYL Approach to Data Base Management* (Chichester, England: Wiley-Interscience, 1978). Nr 04; QA76. 9. D3, O5.

[2] Rudolph Bayer i Edward McCreight, *Organization and Maintenance of Large Ordered Indexes* w „Software Pioneers", pod redakcją Manfreda Broya i Ernsta Denerta (Berlin: Springer-Verlag, 2002), 245 – 262.

Ostatecznie ze względu na brak równowagi między kosztami implementacji, dojrzałością sprzętową i uzyskiwanymi korzyściami zarzucono to rozwiązanie i zaczęto używać szybszych technologii, czyli systemów relacyjnych. W efekcie CODASYL już nie istnieje, chociaż niektóre z wchodzących w jego skład komitetów kontynuują pracę.

Okres od lat 80. XX wieku do pierwszej dekady XXI wieku — encja-relacja

Pomysł Edgara F. Codda na rozdzielenie uporządkowania danych od systemu ich pobierania zapoczątkował nową falę innowacji w technologiach zarządzania danymi[3]. Jego prace stały się podwaliną okresu encja-relacja w bazach danych.

Okres encja-relacja obejmuje dekady, w ciągu których nasza branża dopracowała pobieranie danych na podstawie klucza, czyli jedną z metod, jakie proponowały grupy robocze w latach 60. XX wieku. W trakcie tego okresu opracowano technologię, która była i jest nadal niesłychanie skuteczna w przechowywaniu i pobieraniu danych z tablic oraz zarządzaniu nimi. Techniki opracowane w tych dziesięcioleciach nadal są w rozkwicie, ponieważ są przetestowane, udokumentowane i zrozumiałe.

Systemy z tego okresu wprowadziły i spopularyzowały specyficzny sposób myślenia o danych. Po pierwsze systemy relacyjne opierają się na solidnej matematycznej teorii algebry relacyjnej. W systemach relacyjnych dane są uporządkowane w zbiorach, które mają na celu magazynowanie i pobieranie encji znanych ze świata rzeczywistego, takich jak ludzie, miejsca i rzeczy. Podobne encje, na przykład ludzie, są zgrupowane w tabeli. Każdy rekord tworzy wiersz tabeli. Określony rekord można pobrać z tabeli na podstawie jego klucza głównego.

W systemach relacyjnych encje można ze sobą łączyć. Aby utworzyć połączenia między encjami, trzeba utworzyć więcej tabel. Tabela łącząca zawiera klucze główne każdej encji i przechowuje je w osobnych wierszach. Pracujący w tym czasie innowatorzy opracowali rozwiązanie dla danych tabelarycznych, które sprawdza się doskonale do dziś.

Na rynku dostępnych jest tyle książek i innych zasobów dotyczących systemów relacyjnych, że nie sposób ich wszystkich wymienić. Nasza książka do nich nie należy. Chcemy się w niej skupić na procesach i zasadach projektowych, które obecnie zyskują popularność.

W tym okresie pojawił się i zakorzenił pogląd, że wszystkie dane można przedstawić w tabeli.

Jeśli dane trzeba zapisać w tabeli i je z niej pobierać, preferowanym rozwiązaniem pozostaną technologie relacyjne. Niezależnie jednak od swej dominującej pozycji technologie relacyjne nie są idealnym rozwiązaniem wszystkich problemów.

Pod koniec lat 90. XX wieku pojawiły się wczesne sygnały ery informacyjnej w postaci rosnącej popularności internetu. W tym czasie zaczęły się pojawiać takie ilości i struktury danych, których wcześniej nie planowano i nie używano. Na tym etapie innowacji technologii bazodanowych aplikacje zostały zalane niewytłumaczalną ilością danych o różnorodnej postaci. Na tej podstawie

[3] Edgar F. Codd, *A Relational Model of Data for Large Shared Data Banks*, „Communications of the ACM 13", nr 6 (1970): 377 – 387.

wysnuto kluczowy wniosek, że model relacyjny się nie sprawdza, ponieważ nie uwzględnia przewidywanego użycia danych. Branża dysponowała szczegółowym modelem magazynowania, ale brakowało jej sposobów analizowania i inteligentnego wykorzystania danych.

Docieramy teraz do trzeciej i najnowszej fali innowacji w technologiach bazodanowych.

Od początku XXI wieku do lat 20. XXI wieku — NoSQL

Rozwój technologii bazodanowych od początku do lat 20. XXI wieku można sprowadzić do powstania ruchu NoSQL (non-SQL, czyli „nie tylko SQL"). W tym czasie obrano za cel utworzenie skalowalnych technologii magazynujących dane, zarządzających nimi i pobierających dane o dowolnym kształcie.

Okres NoSQL można opisać, porównując innowacje w technologiach bazodanowych do szybkiego rozkwitu rynku piwa rzemieślniczego w Stanach Zjednoczonych. Proces fermentowania piwa nie uległ zmianie, ale zaczęto stosować nowe dodatki i dbać o wyższą jakość oraz o świeżość składników. Zawiązały się bliższe relacje między warzelnikami a klientami, których opinie prowadziły do natychmiastowych zmian w produkcji. Obecnie zamiast trzech marek piwa w supermarkecie mamy często do wyboru ponad 30.

W branży bazodanowej rozwój przebiegał inaczej. Nie szukano nowych dodatków wzbogacających ten sam proces fermentacji, ale w tempie wykładniczym rozwijano technologie zarządzania danymi. Architekci potrzebowali skalowalnych technologii, umożliwiających zarządzanie różnymi formami, rozmiarami i wymaganiami swoich szybko rozrastających się aplikacji. Podczas tego procesu pojawiły się takie popularne formaty danych jak klucz-wartość, bazy kolumnowe, dokumenty, strumienie i grafy.

Przesłanie ery NoSQL było dość jasne: skalowalne mechanizmy magazynowania i przeglądania danych, a także zarządzania danymi zapisanymi w tabelach nie zawsze się sprawdzały, podobnie jak nie każdemu smakuje piwo typu pilzner.

Ruch NoSQL rozwinął się w wyniku kilku czynników. Są one kluczowe w zrozumieniu obecnego etapu krzywej popularności technologii grafowych. Chcemy zwrócić uwagę na trzy czynniki, a mianowicie na potrzebę opracowania standardów serializacji danych, wyspecjalizowanych narzędzi oraz skalowalności poziomej.

Wzrost popularności aplikacji internetowych doprowadził do powstania naturalnych kanałów przekazywania danych między aplikacjami. Na podstawie tych kanałów innowatorzy opracowali nowe i różne standardy serializacji danych, takie jak formaty XML, JSON i YAML.

Nowe standardy wymagały oczywiście wyspecjalizowanych narzędzi, czyli drugiego czynnika. Protokoły wymiany danych w internecie wykorzystywały struktury, które z natury *nie* były tabelaryczne. To doprowadziło do nowych innowacji i wzrostu popularności danych w formacie klucz-wartość, dokument, graf i innych wyspecjalizowanych baz danych.

Ponadto do wspomnianych aplikacji nowego typu napływały dane, które obciążały mechanizmy skalowania systemów w niespotykanym dotychczas stopniu. Zastosowania i pochodne praw Moore'a dawały pewną nadzieję, ponieważ koszt sprzętu, a co za tym idzie koszt przechowywania danych,

stale malał. W efekcie działania prawa Moore'a można było sobie pozwolić na duplikację danych i powstanie wyspecjalizowanych systemów. Spadł również koszt ogólnych mocy obliczeniowych[4].

Innowacje i wymagania ery NoSQL wytyczyły wspólną drogę do migracji branży od systemów skalowalnych pionowo do systemów skalowalnych poziomo. W systemie skalowalnym poziomo dodaje się maszyny fizyczne lub wirtualne, aby zwiększyć ogólne możliwości obliczeniowe. System skalowalny poziomo, ogólnie określany mianem „klastra", wydaje się użytkownikowi pojedynczą platformą; użytkownik nie wie, że zadania są wykonywane przez kilka serwerów. Natomiast system skalowalny pionowo wymaga potężniejszych komputerów. Brakuje miejsca? Znajdźmy większe pudło, które jest droższe; aż okaże się, że większych pudeł już nie ma.

Skalowanie poziome oznacza dodanie większej ilości zasobów w celu rozłożenia obciążenia, zwykle w równoległy sposób. Skalowanie pionowe oznacza powiększenie lub przyspieszenie zasobów w celu obsługi większego obciążenia.

Uwzględniając opisane trzy czynniki, uniwersalny zbiór narzędzi do budowania skalowalnej architektury danych nietabelarycznych okazał się najważniejszym osiągnięciem ery NoSQL. Obecnie zespoły programistyczne mogą wybierać najlepsze rozwiązanie dla swoich nowych aplikacji. Mają do wyboru wiele technologii obsługujących różne formaty, szybkość i skalowalność, dostosowane do swoich danych. Dostępne są narzędzia do zarządzania danymi, przechowywania, przeszukiwania i pobierania danych w postaci dokumentów, par klucz-wartość, kolumn i (lub) grafów w dowolnej skali. Za pomocą tych narzędzi zaczęliśmy wykorzystywać wiele formatów danych na sposoby, które wcześniej były niemożliwe do osiągnięcia.

Co można zrobić z tą unikalną kolekcją narzędzi i danych? Możemy rozwiązywać bardziej skomplikowane problemy szybciej i na znacznie większą skalę.

Lata 20. XXI wieku do ? — grafy

Obiecaliśmy, że rys historyczny będzie krótki i skoncentrowany na określonym celu. W tym podrozdziale spełniamy tę obietnicę i łączymy ze sobą ważne wydarzenia z krótkiego opisu historycznego. Te relacje między zdarzeniami historycznymi stanowią fundament czwartej ery innowacji bazodanowych: fali myślenia grafowego.

Podczas tej ery innowacji dokonuje się przesunięcie od wydajności systemów magazynowania do ekstrakcji wartości na podstawie danych przechowywanych w systemach magazynowych.

Dlaczego lata 20. XXI wieku?

Zanim poznasz nasze spojrzenie na erę grafów, prawdopodobnie chciałbyś się dowiedzieć, dlaczego uważamy, że zaczęła się w roku 2020. Poświęćmy chwilę na wyjaśnienie wyboru tej daty jako znaczącej dla rynku grafów.

[4] Clair Brown i Greg Linden, *Chips and Change: How Crisis Reshapes the Semiconductor Industry* (Cambridge: MIT Press, 2011).

Wybór roku 2020 wynika ze skrzyżowania dwóch strumieni myślowych. Na tym skrzyżowaniu spotyka się popularny model przyjmowania innowacji[5] Geoffreya Moore'a z czasem zaobserwowanym w ciągu trzech minionych okresów innowacji bazodanowych.

Podobnie jak CODASYL, cykl upowszechniania technologii przypisywany Moore'owi datuje się na lata 50. XX wieku. Więcej informacji na ten temat można znaleźć w książce *Diffusion of Innovations*[6] Everetta Rogera z 1962 roku.

Zaobserwowano, że od wczesnego wykorzystania nowych technologii do ich szerokiego rozpowszechnienia upływa zwykle sporo czasu. Przykład jest opisany w podrozdziale „Okres od lat 80. XX wieku do pierwszej dekady XXI wieku — encja-relacja", w którym omawiamy relacyjne bazy danych w latach 70. XX wieku. Między pierwszą publikacją a rzeczywistą implementacją technologii relacyjnej upłynęło 10 lat. Przykłady takiego opóźnienia można znaleźć, analizując dowolną technologię.

Historia uczy nas, że w każdym okresie przed nastaniem ery grafów można wydzielić okres rozwoju niszowej technologii, która doczekała się popularyzacji wiele lat później. Analizując lata 20. XXI wieku, przyjmujemy to samo założenie dotyczące rynku grafów. Historia uczy nas również, że nie oznacza to, że istniejące narzędzia odejdą w zapomnienie.

Niezależnie od przyjętego sposobu pomiaru nie są to analizy trendów na rynku akcji, które koncentrują się na określonej dacie. Nasze przewidywania dotyczą nowej ery we wdrażaniu technologii, napędzanej ewolucją wartości. Ocena wartości coraz rzadziej opiera się na wydajności, a coraz częściej na połączonych gęstą siecią powiązaniach w zasobach danych. Te zmiany są czasochłonne i nie mają ustalonego harmonogramu.

Łączenie kropek

Przypomnijmy trzy wzorce dostępu do danych, przewidywane przez konsorcjum COSASYL w latach 60. XX wieku. Dotyczyły one pobierania danych za pomocą kluczy, skanowania i połączeń. Wyodrębnianie elementu danych za pomocą klucza pozostaje najwydajniejszym sposobem dostępu niezależnie od formatu danych. Tę wydajność uzyskano podczas ery encja-relacja. Nadal jest to popularne rozwiązanie.

Drugim celem konsorcjum CODASYL było pobieranie danych poprzez skanowanie, co stało się możliwe w erze NoSQL, kiedy to powstały technologie umożliwiające skanowanie ogromnych ilości danych. Dysponujemy obecnie oprogramowaniem i sprzętem, które umożliwiają przetwarzanie i pobieranie wartości z olbrzymich zbiorów danych na ogromną skalę. Oznacza to, że udało się osiągnąć dwa cele konsorcjum.

Ostatnim celem na liście jest pobieranie danych poprzez nawigację przez łącza. Nasza branża zatoczyła koło.

[5] Geoffrey A. Moore i Regis McKenna, *Crossing the Chasm* (New York: HarperBusiness, 1999).

[6] Everett M. Rogers, *Diffusion of Innovations* (New York: Simon and Schuster, 2010).

Powrót branży do technologii grafowych idzie w parze z odchodzeniem od wydajnego zarządzania danymi i potrzebą wyodrębnienia wartości na ich podstawie. Ta zmiana nie oznacza, że nie musimy już wydajnie zarządzać danymi; oznacza natomiast, że udało się nam rozwiązać jeden problem i możemy się zająć trudniejszym. Nasza branża kładzie obecnie nacisk nie tylko na szybkość i koszty, ale również na wartość.

Wartość można wyodrębnić z danych, jeśli uda się połączyć ze sobą różne informacje i wyciągnąć na tej podstawie nowe wnioski. Ekstrakcja wartości z danych wymaga zrozumienia złożonej sieci powiązań między elementami danych.

Jest to synonim rozpoznawania złożonych problemów i systemów, jakie można zaobserwować w naturalnej sieci danych.

Nasza branża i ta książka koncentrują się na rozwoju i wdrażaniu technologii, które wyodrębnią wartość z danych. Podobnie jak w erze relacyjnej, trzeba opracować nowy sposób myślenia, umożliwiający zrozumienie, wdrożenie i zastosowanie tych technologii.

Aby dostrzec wartość, o której tu piszemy, niezbędna jest zmiana myślenia. Trzeba przestać postrzegać dane jako tabele i na pierwszym miejscu postawić relacje, jakie w nich występują. Właśnie to nazywamy myśleniem grafowym.

Czym jest myślenie grafowe?

Nie napisaliśmy tego jasno, ale przedstawiliśmy już przykład myślenia grafowego podczas opisu spotkania przy białej tablicy na początku tego rozdziału.

Opisując sytuację, w której okazało się, że dane można przedstawić w postaci grafu, odtworzyliśmy możliwości myślenia grafowego. Jest ono bardzo proste: myślenie grafowe obejmuje doświadczenie i wnioski, jakie nasuwają się, gdy dostrzegasz wartość w relacjach istniejących w zbiorze danych.

Myślenie grafowe polega na zrozumieniu domeny problemu w postaci połączonego grafu i użyciu technik grafowych w celu opisu dynamiki domeny, co ma doprowadzić do rozwiązania problemów domenowych.

Możliwość zauważenia grafów w danych można porównać do rozpoznawania złożonej sieci powiązań w domenie. W złożonej sieci można znaleźć najbardziej złożone problemy do rozwiązania. A większość problemów i możliwości biznesowych o najwyższej wartości można uznać za złożone.

To dlatego w następnym etapie innowacji technologii bazodanowych następuje przesunięcie od wydajności na rzecz znajdowania wartości za pośrednictwem technologii grafowych.

Złożone problemy i złożone systemy

Dotychczas kilkakrotnie użyliśmy terminu *problem złożony*, nie podaliśmy jednak jego definicji. Rozważając problemy złożone, mamy na myśli sieci w obrębie systemów złożonych.

Problemy złożone

Problemy złożone są problemami, które można zaobserwować i zmierzyć w obrębie systemów złożonych.

Systemy złożone

Systemy złożone składają się z wielu pojedynczych komponentów połączonych wzajemnie w taki sposób, że zachowanie całego systemu nie stanowi prostego zbioru zachowań poszczególnych komponentów (nazywamy je „zachowaniem emergentnym").

Systemy złożone opisują relacje, wpływy, zależności i interakcje między poszczególnymi komponentami rzeczywistych obiektów. Upraszczając, w systemie złożonym wiele obiektów wchodzi ze sobą w interakcje. Do przykładów systemów złożonych należy ludzka wiedza, łańcuchy dostaw, systemy transportu lub komunikacji, organizacje społeczne, globalny klimat na Ziemi i cały wszechświat.

Większość problemów biznesowych o wysokiej wartości można uznać za problemy złożone, wymagające myślenia grafowego. W tej książce poznasz cztery główne wzorce — sąsiedztwa, hierarchii, ścieżek i rekomendacji — wykorzystywane do rozwiązywania złożonych problemów za pomocą technologii grafowych dla firm z całego świata.

Problemy złożone w biznesie

Dane nie są już produktem ubocznym działalności biznesowej. Dane są coraz częściej zasobem strategicznym w gospodarce. Wcześniej dane wymagały możliwie wygodnych i tanich sposobów zarządzania, umożliwiających działalność biznesową. Obecnie traktuje się je jako inwestycję, która powinna przynieść korzyści. Wymaga to zmiany sposobu myślenia o obsłudze i wykorzystaniu danych.

Przykładowo: pod koniec ery NoSQL firma Microsoft przejęła portale LinkedIn i GitHub. Te przejęcia umożliwiły pomiar wartości danych rozwiązujących skomplikowane problemy. Microsoft kupił LinkedIn za 26 miliardów dolarów, spodziewając się 1 miliarda dolarów przychodów. Natomiast GitHub kosztował 7,8 miliarda dolarów, przy spodziewanych przychodach 300 milionów dolarów.

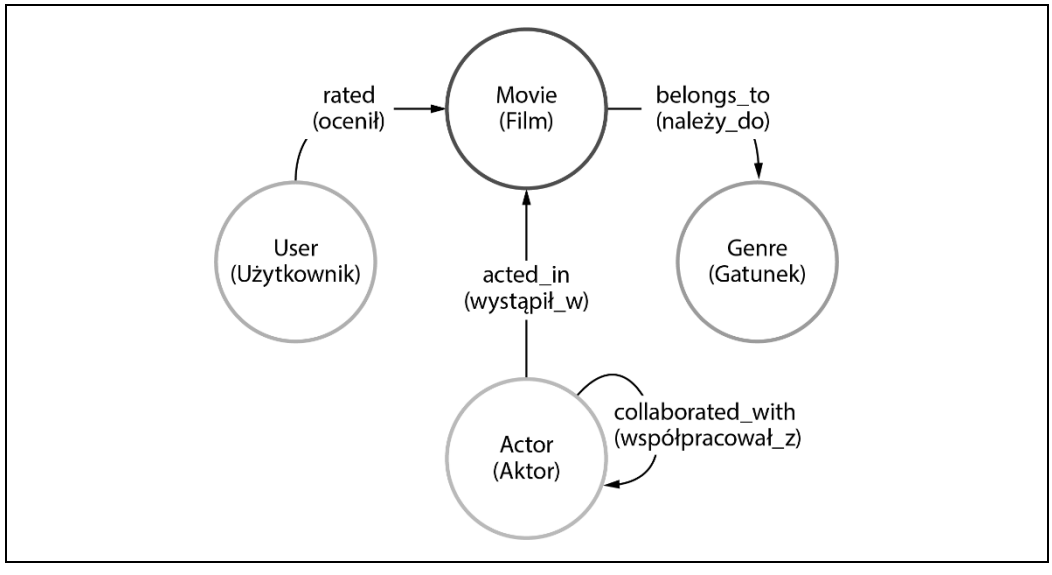
Dane przetwarzane przez LinkedIn oraz GitHub można przedstawić w postaci grafu dzięki sieciom utworzonym w ramach tych serwisów. Są to sieci ekspertów i programistów. Na podstawie przytoczonych wartości można zauważyć, że podczas szacunków zastosowano mnożnik 26, wyznaczając wartość danych modelujących złożony system domenowy. Te dwa przejęcia ilustrują strategiczną wartość danych modelujących graf domenowy. Posiadanie grafu domenowego okazuje się istotnym aspektem w ocenie wartości firmy.

Przytaczając te informacje, nie chcemy narzucać błędnej interpretacji naszych intencji. To, że wartość szybko rozwijających się start-upów rośnie, nie jest niczym nowym. Opisaliśmy te dwa konkretne przykłady, ponieważ właściciele GitHuba i LinkedIna dostrzegli i zdołali spieniężyć wartość wynikającą z danych. Dzięki zasobom danych wzrost przychodów jest wyższy niż w przypadku innych start-upów o podobnej wielkości i dynamice rozwoju.

Stosując myślenie grafowe, firmy te mogą zaprezentować i zrozumieć najbardziej złożone problemy w swojej domenie i uzyskać do nich dostęp. W skrócie: te firmy stworzyły rozwiązania dla największych i najtrudniejszych systemów złożonych.

Firmy, które jako pierwsze zaczęły zmieniać podejście do strategii danych, zbudowały technologie modelujące najbardziej złożone problemy w swoich domenach. Co mają ze sobą wspólnego Google, Amazon, FedEx, Verizon, Netflix i Facebook? Pomijając fakt, że należą do najwyżej wycenianych firm na świecie, każda z nich posiada dane modelujące najważniejsze i najbardziej złożone problemy domenowe. Każda z nich posiada dane tworzące graf jej domeny.

Pomyśl tylko. Google ma graf całej wiedzy ludzkiej. Amazon i FedEx mają grafy globalnego łańcucha dostaw i transportu. Dane Verizona tworzą największy na świecie graf telekomunikacyjny. Facebook ma graf całej światowej sieci społecznościowej. Netflix ma dostęp do grafu rozrywki, przedstawionego na rysunku 1.2 i zaimplementowanego w ostatnich rozdziałach tej książki.



Rysunek 1.2. Jeden ze sposobów przedstawienia danych Netfliksa na grafie i ostatni przykład, jaki zaimplementujesz w tej książce: skalowalne filtrowanie kolaboratywne

W przyszłości firmy inwestujące w architektury danych w celu modelowania złożonych systemów domenowych dołączą do szeregu tych gigantów. Inwestycje w technologie modelowania złożonych systemów można porównać do priorytetyzacji wyodrębniania wartości na podstawie danych.

Jeśli ktoś chce uzyskać wartość ze swoich danych, najpierw powinien się przyjrzeć istniejącym w nich wzajemnym połączeniom. Należy szukać systemu złożonego, opisywanego przez dane. Następnie należy podjąć decyzje dotyczące odpowiednich technologii magazynowania, zarządzania i wyodrębniania tych wzajemnych połączeń.

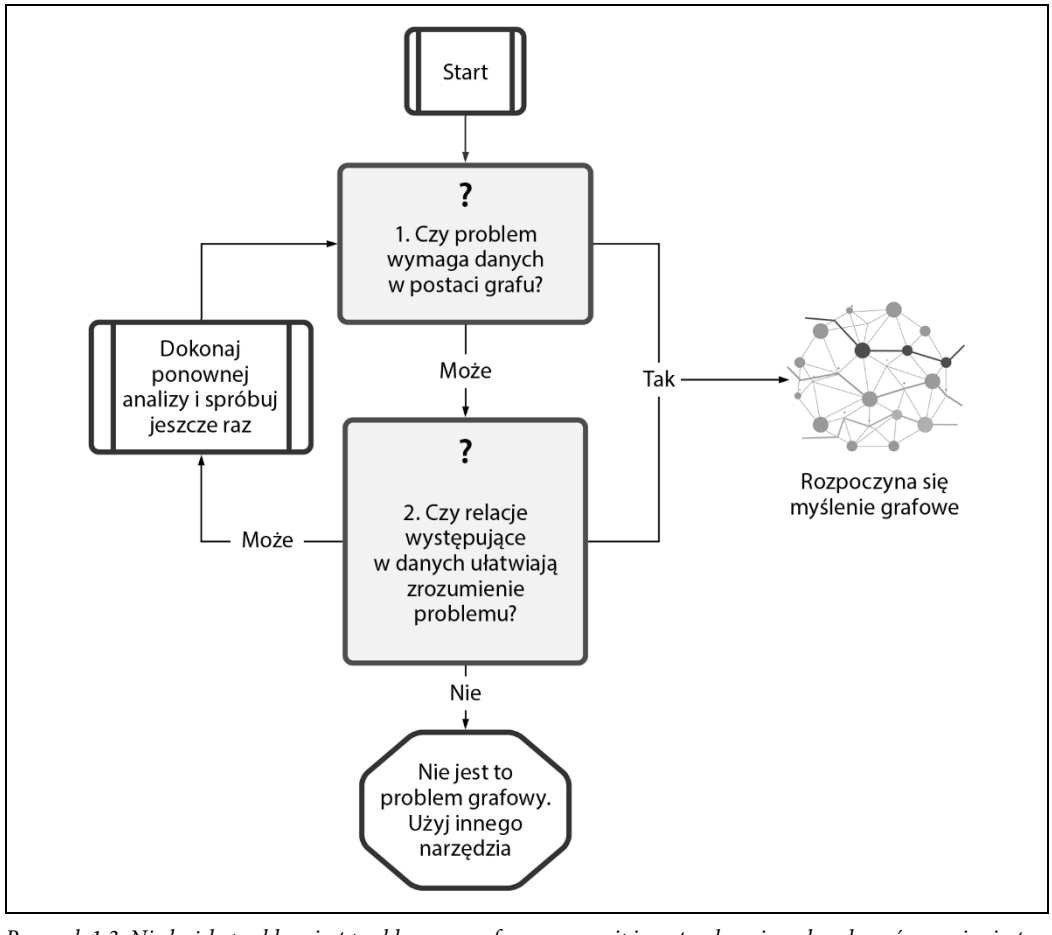
Podejmowanie decyzji o technologii rozwiązywania złożonych problemów

Niezależnie od tego, czy pracujesz w jednej ze wspomnianych firm, czy nie, możesz się nauczyć myślenia grafowego o danych ze swojej domeny.

Od czego zacząć?

Trudności w nauce i stosowaniu myślenia grafowego zaczynają się od rozpoznania obszarów, w których relacje zwiększają wartość danych, i tych, w których jej nie zwiększają. Skorzystamy z dwóch rysunków, na których w uproszczony sposób prezentujemy poszczególne kroki, jakie należy wykonać, oraz wyzwania, z którymi trzeba się będzie zmierzyć.

Chociaż rysunek 1.3 jest prosty, wymaga odpowiedzi na zasadnicze pytania o dane. Aby podjąć pierwszą decyzję, zespół musi wiedzieć, jakich danych wymaga aplikacja. Zaczynamy od tego pytania, ponieważ jest ono często ignorowane.



Rysunek 1.3. Nie każdy problem jest problemem grafowym — najpierw trzeba więc zdecydować, czy nim jest

Inne zespoły pominęły wcześniej aspekty przedstawione na rysunku 1.3, ponieważ blask nowości odwrócił ich uwagę od podążania ustalonymi ścieżkami budowania aplikacji produkcyjnych. To napięcie między nowym a ustalonym sposobem postępowania spowodowało, że pierwsze zespoły zbyt szybko przeprowadzały analizę krytycznych celów aplikacji. Z tego powodu wiele projektów wykorzystujących grafy zakończyło się porażką.

Przeanalizujmy diagram z rysunku 1.3, aby uniknąć powtarzania typowych błędów, jakie popełnili początkowi użytkownicy technologii grafowych.

Pytanie 1. Czy problem wymaga danych w postaci grafu?

Dane można postrzegać na różne sposoby. Aby odpowiedzieć na pierwsze pytanie na drzewie decyzyjnym, trzeba zrozumieć format danych wymagany przez aplikację. Dobrym przykładem danych, dla których odpowiedź na pytanie 1. brzmi „tak", jest sekcja wspólnych kontaktów w serwisie LinkedIn. LinkedIn wykorzystuje relacje między kontaktami, dzięki czemu można nawigować po swojej sieci zawodowej i zrozumieć wspólne kontakty. Prezentowanie użytkownikowi wspólnych kontaktów jest popularnym sposobem wykorzystania danych w kształcie grafu na Twitterze, w Facebooku i w innych sieciach społecznościowych.

Pisząc o „kształcie danych", mamy na myśli strukturę wartościowych informacji, jakie chcemy wydobyć z danych. Chcesz poznać imię i wiek osoby? Dane te można przedstawić jako wiersz danych z tabeli. Chcesz się dowiedzieć, w którym rozdziale, podrozdziale, na której stronie i w którym przykładzie znajdziesz informacje o dodawaniu wierzchołka do grafu? Informację tę można zaprezentować jako zagnieżdżone dane, a je umieścić w dokumencie lub hierarchii. Chcesz wiedzieć, którzy znajomi Twoich znajomych znają Elona Muska? Jest to pytanie o serię relacji, które można najlepiej przedstawić na grafie.

Przy zastosowaniu myślenia od ogółu do szczegółu zalecamy, aby to kształt danych dyktował decyzję o wyborze bazy danych i technologii. W tabeli 1.1 przedstawiono typy danych często wykorzystywane w nowoczesnych aplikacjach.

Tabela 1.1. Krótkie podsumowanie popularnych typów danych, ich kształtów i zalecanych baz danych

Opis danych	Kształt danych	Użycie	Zalecana baza danych
Arkusze kalkulacyjne lub tabele	Relacyjne	Pobieranie za pomocą klucza głównego	Systemy RDBMS
Zbiór plików lub dokumentów	Hierarchiczne lub zagnieżdżone	Węzeł główny identyfikowany za pomocą parametru ID	Bazy danych dokumentów
Relacje lub połączenia	Graf	Przeszukiwanie za pomocą wzorca	Grafowe bazy danych

Aby rozwiązać najciekawsze współczesne problemy związane z danymi, trzeba skorzystać ze wszystkich trzech sposobów postrzegania danych. Musisz osiągnąć biegłość w stosowaniu każdej metody do rozwiązywania problemów z danymi i wynikających z nich drobniejszych problemów. W przypadku każdego aspektu problemu trzeba rozumieć kształt danych przychodzących, przechowywanych i generowanych przez aplikację. Każdy z tych punktów oraz każdy moment transferu danych wpływa na wymagania dotyczące wyboru technologii dla aplikacji.

Jeśli nie jesteś pewien, jakiego kształtu danych wymaga Twój problem, zadaj sobie kolejne pytanie z rysunku 1.3, które wymaga rozważenia ważności relacji występujących w danych.

Pytanie 2. Czy relacje w danych ułatwiają zrozumienie problemu?

Bardziej istotne pytanie z rysunku 1.3 dotyczy istnienia relacji w danych oraz znaczenia, jakie mają one dla problemu biznesowego. Skuteczne wykorzystanie technologii grafów zależy od odpowiedzi na drugie pytanie z drzewa decyzyjnego. Dla nas istnieją tylko trzy odpowiedzi na to pytanie: tak, nie lub może.

Jeśli możesz z całą pewnością odpowiedzieć tak lub nie, droga jest prosta. Na przykład w przypadku sekcji wspólnych kontaktów w serwisie LinkedIn z łatwością odpowiemy „tak" na korzyść danych w postaci grafu. Natomiast pole wyszukiwania w tym samym portalu wymaga nawigacji fasetowej i odpowiedź na pytanie brzmi „nie". Wybór odpowiedzi jest łatwy dzięki temu, że rozumiemy kształt danych wymaganych do rozwiązania problemu biznesowego.

Jeśli relacje występujące w danych ułatwiają rozwiązanie problemu biznesowego, trzeba skorzystać w aplikacji z technologii grafowych. W innej sytuacji trzeba znaleźć inne narzędzie. Może wybór z tabeli 1.1 pomoże rozwiązać aktualny problem.

Trudności pojawiają się, gdy nie ma pewności, czy relacje są istotne dla problemu biznesowego. Tę sytuację przedstawia ścieżka „Może?" z lewej strony rysunku 1.3. Doświadczenie podpowiada nam, że jeśli ktoś znajduje się w tym punkcie, próbuje rozwiązać zbyt duży problem. Zalecamy podzielenie problemu na mniejsze elementy i powrót do górnej części rysunku 1.3. Ponowną analizę najczęściej zalecamy zespołom w przypadku problemu *łączenia encji* lub jeśli trzeba ustalić, kto jest kim w danych. W rozdziale 11. opisany jest przykład użycia struktury grafu w procesie łączenia encji.

Typowe błędy w rozumieniu danych

Czasem dostrzeżenie grafu w danych może przesłonić znaczenie dwóch innych kształtów danych: zagnieżdżonych i tabelarycznych. Zespoły zwykle błędnie interpretują ten fałszywy trop.

Nawet jeśli uważasz, że problem jest złożony i jego rozwiązanie wymaga zastosowania myślenia grafowego, nie oznacza to, że technologie grafowe trzeba zastosować we wszystkich komponentach danych. W rzeczywistości czasem korzystniej jest przedstawić niektóre komponenty lub mniejsze problemy w postaci tabeli lub dokumentów zagnieżdżonych.

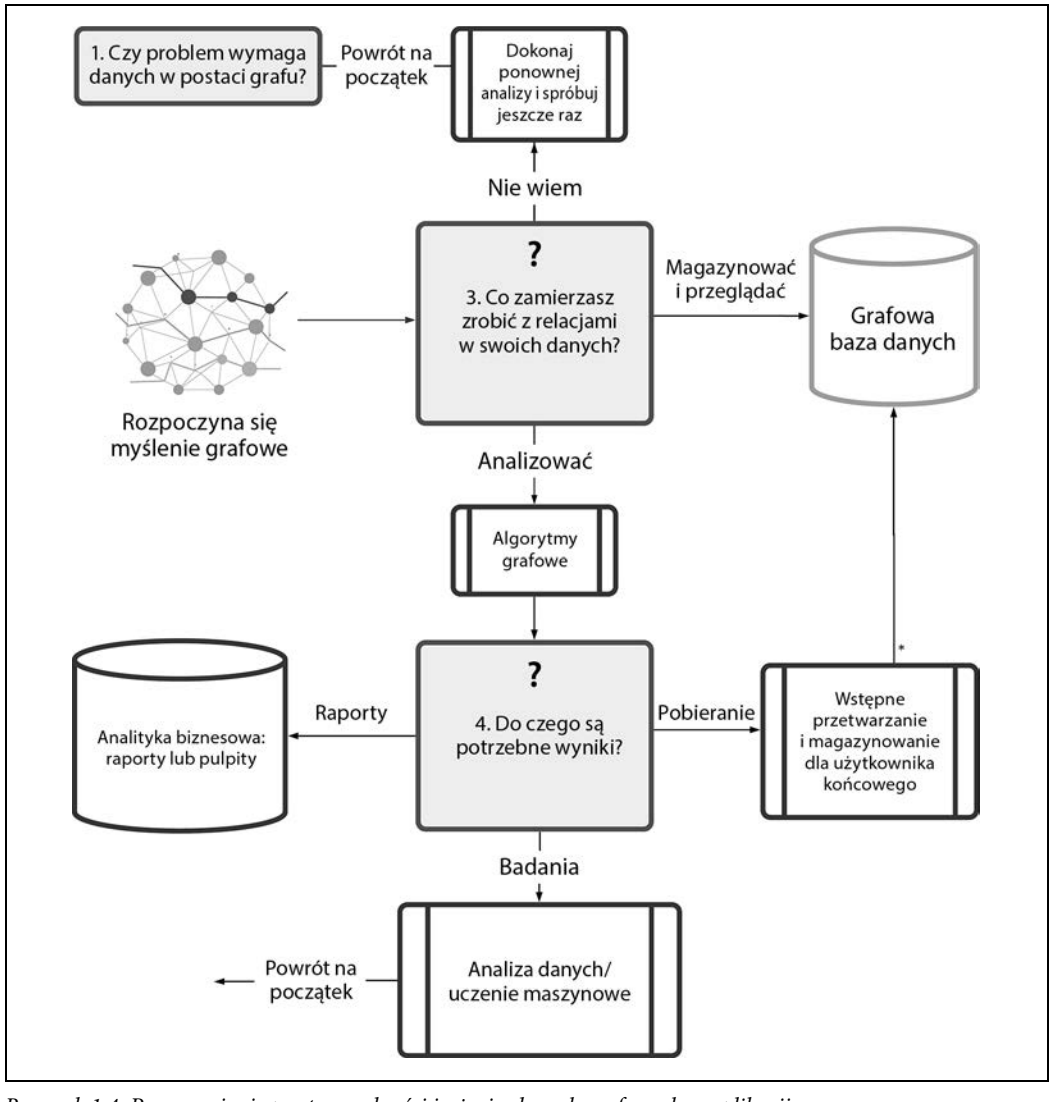
Zawsze warto „myśleć odwzorowaniami" (na pliki lub tabele). Zatem ćwiczenie myślowe z rysunku 1.3 wymaga raczej odpowiedzi na pytanie „Jaki jest najlepszy sposób myślenia o danych?". Wykracza ono poza zwinniejszy proces myślowy polegający na podziale złożonych problemów na mniejsze komponenty. Zachęcamy zatem do znalezienia najlepszego sposobu myślenia o danych z *bieżącego problemu*.

Oto najkrótsze podsumowanie rysunku 1.3: użyj najlepszego narzędzia do rozwiązania bieżącego problemu. „Narzędzie" w tym przypadku jest bardzo ogólnym określeniem. Niekoniecznie dotyczy ono wyboru bazy danych; mamy w tym przypadku na myśli wybór sposobu reprezentacji danych.

Twoje dane są grafem. Co teraz?

Pierwsze pytanie z rysunku 1.3 wymaga podjęcia decyzji o reprezentacji danych poprzez zastosowanie projektowania opartego na zapytaniach. Możliwe, że pewne aspekty złożonego problemu najlepiej prezentują się w postaci tabel lub zagnieżdżonych dokumentów. Takie są oczekiwania.

A jeśli masz dane grafowe i musisz z nich skorzystać? To prowadzi nas do drugiej części procesu myślenia grafowego, przedstawionej na rysunku 1.4.



Rysunek 1.4. Poruszanie się po stosowalności i użyciu danych grafowych w aplikacji

Idąc dalej, zakładamy, że dla rozwoju aplikacji korzystne jest zrozumienie, modelowanie i użycie relacji występujących w danych.

Pytanie 3. Co zamierzasz zrobić z relacjami występującymi w danych?

W świecie technologii grafowych dane można potraktować na dwa najważniejsze sposoby: można je analizować lub przeszukiwać. Kontynuując przykład LinkedIna, sekcja wspólnych kontaktów jest przykładem przeszukiwania i wczytywania danych grafowych do widoku. Zespół zajmujący się badaniami w LinkedInie prawdopodobnie śledzi średnią liczbę połączeń między dowolną parą osób, co stanowi przykład *analizy* danych grafowych.

Odpowiedź na trzecie pytanie dzieli decyzje dotyczące technologii grafowej na dwie grupy: analizę danych i zarządzanie danymi. To pytanie zostało przedstawione na środku rysunku 1.4 wraz z przepływem decyzji dla każdej opcji.

Pisząc o *analizie*, mamy na myśli dokładne przeglądanie danych. Zespoły zwykle poświęcają czas na studiowanie relacji występujących w danych, aby sprawdzić, które z nich są ważne. Proces ten różni się od *przeszukiwania* danych grafowych. *Przeszukiwanie* wymaga pobrania danych z systemu. W tym przypadku wiadomo, jakie pytanie trzeba zadać oraz jakie relacje są potrzebne, aby odpowiedzieć na to pytanie.

Zacznijmy od opcji, która kieruje nas w prawo: przypadków, gdy wiadomo, że aplikacja musi magazynować i przeszukiwać relacje w danych. Co prawda obecnie jest to najmniej prawdopodobna ścieżka ze względu na obecny stan rozwoju branży grafowej, jednak w tych przypadkach jesteśmy gotowi do zmiany polegającej na rozpoczęciu używania grafowej bazy danych w aplikacji.

Współpracując z różnymi firmami, odkryliśmy typowy zbiór przypadków, w których bazy danych muszą zarządzać danymi grafowymi. Te przypadki są tematem następnych rozdziałów, dlatego nie będziemy ich teraz opisywać.

Najczęściej jednak zespoły wiedzą, że ich problem wymaga danych grafowych, ale nie wiedzą dokładnie, jak odpowiedzieć na swoje pytania lub które relacje są ważne. To oznacza, że trzeba przeanalizować dane grafowe.

Chcemy zmobilizować Ciebie i Twój zespół do wykonania jeszcze jednego kroku w tej podróży. Zastanówcie się nad wynikami analizy danych grafowych. Zdefiniowanie struktury i celu analizy grafu ułatwi zespołowi podejmowanie bardziej świadomych decyzji dotyczących infrastruktury i narzędzi. Jest to ostatnie pytanie widoczne na rysunku 1.4.

Pytanie 4. Do czego są potrzebne wyniki?

Tematyka analizy danych może dotyczyć wielu aspektów, począwszy od zrozumienia określonego rozkładu w relacjach, po uruchamianie algorytmów przetwarzających całą strukturę. Jest to obszar algorytmów takich jak połączone komponenty, wykrywanie klik, liczenie trójkątów, obliczanie rozkładu stopnia grafu, obliczanie rankingu strony, wnioskowanie, filtrowanie kolaboratywne i wiele innych. W następnych rozdziałach zdefiniujemy wiele tych terminów.

Najczęściej spotkaliśmy się z trzema różnymi celami wykorzystania wyników algorytmów grafowych: raportami, badaniami lub pobieraniem danych. Opiszemy teraz, jak rozumiemy każdy z tych celów.

Opiszemy szczegółowo wszystkie trzy opcje (raporty, badania i pobieranie danych), ponieważ w ten sposób najczęściej wykorzystuje się obecnie dane grafowe. Pozostałe przykłady techniczne i opisy w naszej książce dotyczą przede wszystkim sytuacji, gdy decyzja o użyciu grafowej bazy danych została już podjęta.

Najpierw rozważmy raportowanie. Nasze użycie słowa „raporty" dotyczy tradycyjnej analityki i wglądu w dane biznesowe. Najczęściej te działania określa się mianem analityki biznesowej (*business intelligence* — BI). Chociaż w wielu wczesnych projektach błędnie stosowano grafy, starano się, aby uzyskane wyniki dostarczały metryk lub danych wejściowych dla ustalonych procesów BI. Narzędzia i infrastruktura potrzebne do rozbudowy lub opracowania procesów analityki biznesowej wykorzystujących dane grafowe zasługują na osobną książkę i dokładne omówienie. Nasza książka nie koncentruje się ani na architekturze, ani na sposobach rozwiązywania problemów BI.

Kolejne typowe zastosowania dla algorytmów grafowych można znaleźć na polu analizy danych i uczenia maszynowego. Są to ogólne badania i rozwój. Firmy inwestują w badania i rozwój w celu uzyskania korzyści z danych w postaci grafów. Na rynku istnieje kilka książek opisujących narzędzia i infrastrukturę potrzebne do badania danych o strukturze grafowej; ta książka do nich nie należy.

Dochodzimy do ostatniej ścieżki, oznaczonej etykietą „pobieranie". Na rysunku 1.4 prezentujemy te zastosowania, które dostarczają usług dla użytkowników końcowych. Mamy na myśli produkty przetwarzające dane i wykorzystywane przez klientów. Użytkownicy mają pewne oczekiwania dotyczące opóźnień, dostępności, personalizacji i innych cech wspomnianych aplikacji. Aplikacje te mają inne wymagania architektoniczne niż aplikacje dostarczające metryk dla użytkowników wewnętrznych. Te zagadnienia oraz przypadki użycia są opisane w kolejnych rozdziałach tej książki.

Powróćmy do przypadku LinkedIna. Jeśli korzystasz z tego serwisu, prawdopodobnie miałeś do czynienia z najlepszym przykładem implementacji ścieżki „pobierania" z rysunku 1.4. W LinkedInie zdefiniowana jest funkcja opisująca połączenie użytkownika z każdą inną osobą w sieci. Jeśli przyjrzymy się profilowi innego użytkownika, znajdziemy informację, czy jest on kontaktem 1., 2. czy 3. stopnia. Odległość między użytkownikami na LinkedInie jest przydatną informacją o sieci zawodowej. Ta funkcja LinkedIna jest przykładem produktu danych, który znajduje się na końcu ścieżki pobierania z rysunku 1.4 i dostarcza kontekstowych metryk grafowych użytkownikom końcowym.

Granice między tymi ścieżkami mogą być płynne. Różnica polega na tworzeniu produktu opartego na danych lub produktu dostarczającego wnioski na podstawie danych. Produkty oparte na danych dostarczają użytkownikom unikalnych wartości. Następny etap innowacji tych produktów będzie obejmować wykorzystanie danych grafowych w celu uzyskania trafniejszych i sensownych doświadczeń. Są to ciekawe problemy i architektury, które chcemy przeanalizować w tej książce.

Dokonaj ponownej analizy i spróbuj jeszcze raz

Od czasu do czasu okaże się, że odpowiedź na pytania z rysunków 1.3 i 1.4 brzmi „nie wiem". Nie trzeba się tym przejmować.

Ostatecznie prawdopodobnie czytasz tę książkę, ponieważ pracujesz w firmie przetwarzającej dane i borykającej się ze złożonymi problemami. Tego typu problemy są obszerne i zależne od siebie. Spoglądając na najwyższy poziom problemu, możesz uznać, że proces myślowy przedstawiony na rysunkach1.3 i 1.4 nie ma żadnego związku z Twoimi złożonymi danymi.

Jednak na podstawie naszego doświadczenia, jakie zdobyliśmy, pomagając setkom zespołów na całym świecie, zalecamy podzielić problem na mniejsze elementy i ponownie przeprowadzić proces decyzyjny.

Próba zrównoważenia wymagań interesariuszy, umiejętności programistów i branży jest niesłychanie trudna. Trzeba zacząć od małych kroków i zbudować fundament na podstawie znanych i sprawdzonych wartości, aby stopniowo zbliżyć się do rozwiązania złożonego problemu.

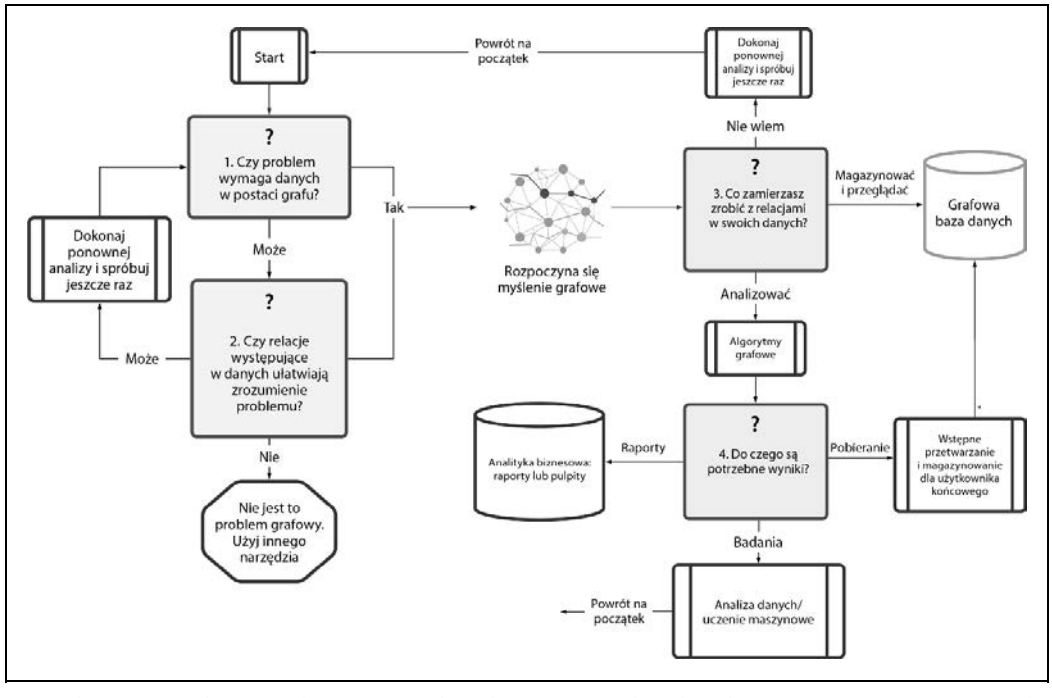
Co się stanie, jeśli zignorujesz podejmowanie decyzji? Zbyt często obserwowaliśmy, jak wspaniałe pomysły nie przechodzą od fazy badania i rozwoju do aplikacji produkcyjnej. Jest to przykład paraliżu analitycznego. Celem użycia algorytmów grafowych jest określenie, w jaki sposób relacje przynoszą wartość aplikacji opartej na danych. Trzeba podjąć pewne trudne decyzje dotyczące czasu i zasobów, jakie można na to poświęcić.

Spojrzenie z szerszej perspektywy

Droga do zrozumienia strategicznego znaczenia danych biznesowych wymaga sprawdzenia, gdzie (i czy) technologia grafowa pasuje do aplikacji. Aby ułatwić sobie wyznaczenie strategicznego znaczenia danych grafowych dla działalności biznesowej, przeprowadziliśmy Cię przez cztery bardzo ważne pytania dotyczące rozwoju aplikacji:

1. Czy problem wymaga danych grafowych?
2. Czy relacje w danych ułatwiają zrozumienie problemu?
3. Co zamierzasz zrobić z relacjami w swoich danych?
4. Co musisz zrobić z wynikami algorytmu grafowego?

Na rysunku 1.5 przedstawiono zbiorczy diagram zawierający wszystkie powyższe pytania.



Rysunek 1.5. Proces decyzyjny, który zapoczątkował powstanie tej książki: jak poruszać się po zastosowaniach i użyciu technologii grafowej w aplikacji

Istnieją dwa powody, dla których poświęciliśmy czas na przeprowadzenie Cię przez drzewo decyzyjne. Po pierwsze drzewo decyzyjne pokazuje cały proces myślowy, za pomocą którego doradzamy, budujemy i stosujemy technologie grafowe. Po drugie drzewo decyzyjne prezentuje, w jaki sposób cel tej książki pasuje do przestrzeni myślenia grafowego.

Oznacza to, że nasza książka jest przewodnikiem po myśleniu grafowym wzdłuż ścieżek przedstawionych na rysunku 1.5, które prowadzą do powstania potrzeby skorzystania z grafowej bazy danych.

Ruszamy na wyprawę z myśleniem grafowym

Jeśli odpowiednio wykorzysta się dane dostępne w firmie, mogą one stanowić strategiczny zasób i inwestycję, które przyniosą nam korzyści. Grafy mają w tym obszarze szczególne znaczenie, ponieważ efekty sieciowe są potężnym bodźcem będącym źródłem wyjątkowej przewagi konkurencyjnej. Ponadto współczesne myślenie projektowe zachęca architektów do postrzegania danych biznesowych jako czegoś, co wymaga maksymalnie wygodnego i minimalnie kosztownego zarządzania.

Takie nastawienie wymaga zmiany myślenia o obsłudze i pracy z danymi.

Zmiana nastawienia jest długą wyprawą, a każda wyprawa zaczyna się od jednego kroku. Postawmy ten krok razem i nauczmy się nowej terminologii, której będziemy używać w trakcie podróży.

Grafy: przełomowa koncepcja w analizie danych!

Komputer do pracy potrzebuje liczb i danych. Człowiek chętniej wysnuwa wnioski i wyodrębnia kontekst na podstawie relacji. Te dwa sposoby myślenia są tak odmienne, że komputery do niedawna z trudem wykonywały zadania związane z operowaniem na relacjach. Obecnie może się to zmienić dzięki grafom. Technologie grafowe łączą ludzkie postrzeganie świata i liniową pamięć komputerów. Ich wdrożenie na szerszą skalę będzie stanowić przełom i pozwoli osiągnąć nieznany dziś poziom. Ale najpierw trzeba się nauczyć stosować myślenie grafowe w rozwiązywaniu problemów technicznych.

Dzięki tej książce opanujesz podstawy myślenia grafowego. Zapoznasz się z elementarnymi koncepcjami grafowymi: teorią grafów, schematami baz danych, systemami rozproszonymi, a także analizą danych. Dowiesz się również, jak wyglądają typowe wzorce wykorzystania danych grafowych w aplikacjach produkcyjnych. Poznasz sposób, w jaki można te wzorce stosować w praktyce. Pokazano tu, jak używać technik programowania funkcyjnego oraz systemów rozproszonych do tworzenia zapytań i analizowania danych grafowych. Opisano też podstawowe podejścia do proceduralnego przechodzenia przez dane grafowe i ich wykorzystanie za pomocą narzędzi grafowych.

W książce:

- nowy paradygmat rozwiązywania problemów: dane grafowe
- wzorce wykorzystania danych grafowych
- przykładowa architektura aplikacji w technologiach relacyjnych i grafowych
- technologie grafowe a przewidywanie preferencji i zaufania użytkowników
- filtrowanie kolaboratywne i jego zastosowanie

Dr Denise Gosnell bada i wdraża dane grafowe. Obecnie jest dyrektorką do spraw danych w DataStax, wcześniej zajmowała się łańcuchami bloków, uczeniem maszynowym i analizą wykresów. Opatentowała wiele zastosowań grafów i algorytmów grafowych.

Dr Matthias Broecheler pełni funkcję dyrektora technicznego w DataStax. Jest ekspertem w zakresie grafowych baz danych, relacyjnego uczenia maszynowego i analizy dużych zbiorów danych, a także twórcą bazy danych Titan.

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

KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-283-7460-7
9 788328 374607
INFORMATYKA W NAJLEPSZYM WYDANIU
Cena: 89,00 zł