

# Przetwarzanie danych w dużej skali

---

NIEZAWODNOŚĆ, SKALOWALNOŚĆ  
I ŁATWOŚĆ KONSERWACJI SYSTEMÓW



Tytuł oryginału: Designing Data-Intensive Applications:  
The Big Ideas Behind Reliable, Scalable, and Maintainable Systems

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-283-4065-7

© 2018 Helion S.A.

Authorized Polish translation of the English edition of Designing Data-Intensive Applications ISBN 9781449373320 © 2017 Martin Kleppmann.

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiegokolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz Wydawnictwo HELION dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz Wydawnictwo HELION nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Wydawnictwo HELION  
ul. Kościuszki 1c, 44-100 GLIWICE  
tel. 32 231 22 19, 32 230 98 63  
e-mail: [helion@helion.pl](mailto: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/przdan>  
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

- [Kup książkę](#)
- [Poleć książkę](#)
- [Oceń książkę](#)

- [Księgarnia internetowa](#)
- [Lubię to! » Nasza społeczność](#)

---

# Spis treści

Przedmowa .....	11
<b>I Podstawy systemów danych .....</b>	<b>17</b>
<b>1. Niezawodne, skalowalne i łatwe w konserwacji aplikacje .....</b>	<b>19</b>
Myślenie o systemach danych	20
Niezawodność	22
Skalowalność	26
Łatwość konserwacji	33
Podsumowanie	36
<b>2. Modele danych i języki zapytań .....</b>	<b>41</b>
Model relacyjny a model oparty na dokumentach	42
Język zapytań o dane	54
Modele danych przypominające graf	60
Podsumowanie	72
<b>3. Przechowywanie i pobieranie danych .....</b>	<b>79</b>
Struktury danych używane w bazie	79
Przetwarzanie transakcji czy analityka?	98
Bazy kolumnowe	103
Podsumowanie	110
<b>4. Kodowanie i zmiany .....</b>	<b>119</b>
Formaty kodowania danych	120
Sposoby przepływu danych	134
Podsumowanie	144

<b>II Dane rozproszone .....</b>	<b>151</b>
Skalowanie pod kątem wyższego obciążenia	151
<b>5. Replikacja .....</b>	<b>157</b>
Liderzy i obserwatorzy	158
Problemy z opóźnieniem replikacji	166
Replikacja z wieloma liderami	171
Replikacja bez lidera	180
Podsumowanie	193
<b>6. Podział na partycje .....</b>	<b>201</b>
Podział na partycje i replikacja	202
Podział na partycje danych typu klucz-wartość	203
Podział na partycje a indeksy pomocnicze	207
Równoważenie partycji	210
Trasowanie żądań	214
Podsumowanie	216
<b>7. Transakcje .....</b>	<b>223</b>
Niejasne pojęcie transakcji	224
Niskie poziomy izolacji	233
Sekwencyjność	250
Podsumowanie	263
<b>8. Problemy z systemami rozproszonymi .....</b>	<b>271</b>
Błędy i awarie częściowe	272
Zawodne sieci	274
Zawodne zegary	283
Wiedza, prawda i kłamstwa	295
Podsumowanie	304
<b>9. Spójność i konsensus .....</b>	<b>315</b>
Gwarancje spójności	316
Liniowość	317
Gwarancje uporządkowania	331
Transakcje rozproszone i konsensus	343
Podsumowanie	361

<b>III Dane pochodne .....</b>	<b>375</b>
Systemy zapisu a systemy danych pochodnych	375
Przegląd rozdziałów	376
<b>10. Przetwarzanie wsadowe .....</b>	<b>379</b>
Przetwarzanie wsadowe z użyciem narzędzi uniksowych	380
MapReduce i rozproszone systemy plików	386
Poza model MapReduce	406
Podsumowanie	415
<b>11. Przetwarzanie strumieniowe .....</b>	<b>425</b>
Przesyłanie strumieni zdarzeń	426
Strumienie a bazy danych	436
Przetwarzanie strumieniowe	447
Podsumowanie	462
<b>12. Przyszłość systemów danych .....</b>	<b>473</b>
Integrowanie danych	474
Podział baz danych na komponenty	482
Dążenie do poprawności	497
Robienie tego, co słuszne	513
Podsumowanie	522
<b>Słowniczek .....</b>	<b>533</b>
<b>Skorowidz .....</b>	<b>542</b>



# Niezawodne, skalowalne i łatwe w konserwacji aplikacje

*Internet został opracowany tak dobrze, że większość ludzi traktuje go jak zasoby naturalne takie jak Pacyfik, a nie jak coś zbudowanego przez człowieka.*

*Kiedy po raz ostatni stworzono wolną od usterek technologię na tak dużą skalę?*

— Alan Kay, z wywiadu dla magazynu „Dr Dobbs’s Journal” (2012)

Obecnie wiele aplikacji to rozwiązania *intensywnie przetwarzające dane* (ang. *data-intensive*), a nie *wymagające obliczeniowo* (ang. *compute-intensive*). W tego rodzaju aplikacjach sama moc procesora rzadko jest czynnikiem ograniczającym. Większymi problemami są zwykle: ilość danych, ich złożoność i szybkość, z jaką się zmieniają.

Aplikacje intensywnie przetwarzające dane są zwykle zbudowane ze standardowych cegiełek, które zapewniają często potrzebne mechanizmy. Na przykład wiele aplikacji musi:

- przechowywać dane, aby później dana aplikacja (lub inna) mogła je znaleźć (*baza danych*);
- zapamiętywać wyniki kosztownych operacji, aby przyspieszyć odczyt (*pamięć podręczna*);
- umożliwiać użytkownikom wyszukiwanie danych na podstawie słów kluczowych lub filtrów (*indeksy wyszukiwania*);
- przysyłać do innego procesu komunikat w celu jego asynchronicznej obsługi (*przetwarzanie strumieniowe*);
- okresowo przetwarzać duże ilości zakumulowanych danych (*przetwarzanie wsadowe*).

Jeśli brzmi to aż nazbyt oczywiście, to dlatego, że tego typu *systemy danych* są tak udaną abstrakcją. Posługujemy się nimi przez cały czas, nie myśląc o tym wiele. W trakcie tworzenia aplikacji większość inżynierów nie myśli nawet o pisaniu od podstaw nowego systemu składowania danych, ponieważ bazy doskonale nadają się do obsługi danych.

Jednak rzeczywistość jest bardziej złożona. Istnieje wiele systemów bazodanowych o różnych cechach, ponieważ poszczególne aplikacje mają odmienne wymagania. Istnieje też wiele metod obsługi pamięci podręcznej, kilka sposobów tworzenia indeksów wyszukiwania itd. W trakcie tworzenia aplikacji trzeba określić, które narzędzia i podejścia będą najbardziej odpowiednie do wykonywania danego zadania. Ponadto czasem trudno jest połączyć komponenty, gdy trzeba zrobić coś, z czym pojedyncze narzędzie sobie nie radzi.

Ta książka to podróż po zasadach i praktycznych aspektach systemów danych oraz po wykorzystywaniu ich do budowania aplikacji intensywnie przetwarzających dane. Zobaczysz tu podobieństwa i różnice między poszczególnymi narzędziami. Dowiesz się też, z czego wynikają ich cechy.

Ten rozdział zaczyna się od przeglądu podstawowych cech, jakie programiści starają się uzyskać: niezawodności, skalowalności i łatwości konserwacji systemów danych. Wyjaśniono tu, co oznaczają te cechy. Przedstawione są też sposoby myślenia o nich i podstawy potrzebne w dalszych rozdziałach. W następnych rozdziałach omawiane są kolejne warstwy. Poznasz tam różne decyzje projektowe, które trzeba uwzględnić w trakcie pracy nad aplikacjami intensywnie przetwarzającymi dane.

## Myślenie o systemach danych

Bazy danych, kolejki, pamięć podręczna itd. są zwykle traktowane jako narzędzia należące do zupełnie różnych kategorii. Choć baza danych i kolejka komunikatów pozornie są podobne (oba te narzędzia przechowują dane przez pewien czas), cechują się zdecydowanie odmiennymi wzorcami dostępu. To oznacza inną charakterystykę pracy, a tym samym i bardzo odmienne implementacje.

Dlaczego więc wszystkie takie rozwiązania są łączone pod zbiorczym określeniem *systemy danych*?

W ostatnich latach pojawiło się wiele nowych narzędzi do przechowywania i przetwarzania danych. Są one zoptymalizowane pod kątem różnych przypadków użycia i nie wpasowują się w tradycyjne kategorie [1]. Istnieją np. magazyny danych używane też jako kolejki komunikatów (Redis), a także kolejki komunikatów z gwarancjami trwałości typowymi dla baz danych (Apache Kafka). Granice między kategoriami się zacierają.

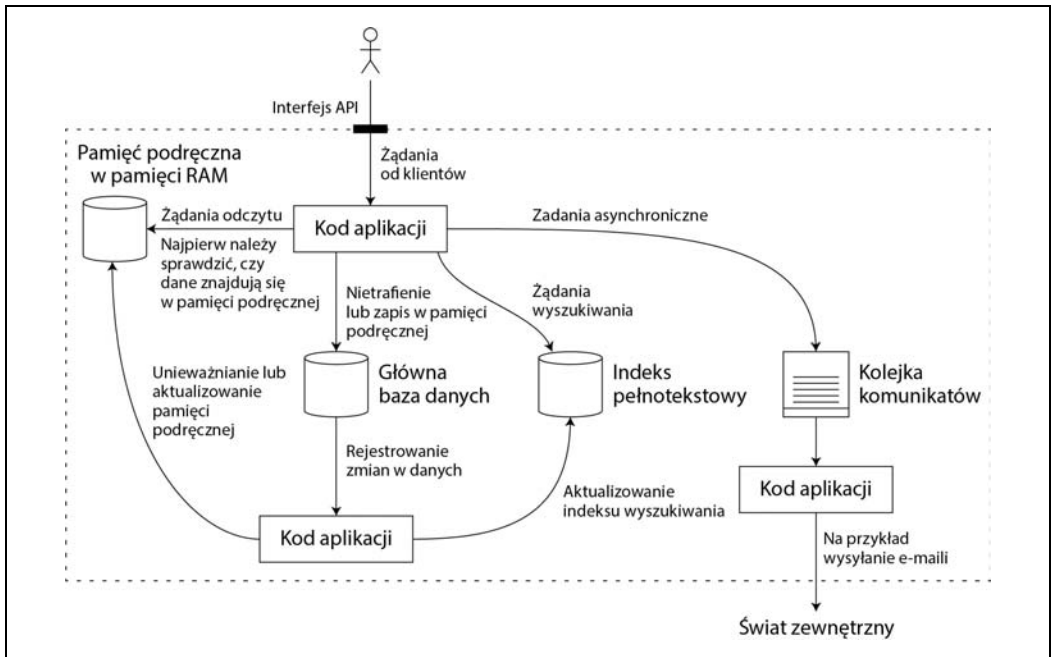
Ponadto coraz większa liczba aplikacji cechuje się tak wysokimi lub szerokimi wymaganiami, że żadne pojedyncze narzędzie nie potrafi zaspokoić wszystkich potrzeb z zakresu przetwarzania i składowania danych. Dlatego pracę dzieli się na zadania, które *mogą* być wydajnie wykonywane przez pojedyncze narzędzie, a różne narzędzia są łączone za pomocą kodu aplikacji.

Na przykład jeśli korzystasz z zarządzanej przez aplikację warstwy pamięci podręcznej (wykorzystującej system Memcached lub podobny) lub serwera wyszukiwania pełnotekstowego (takiego jak Elasticsearch lub Solr) oddzielnie od głównej bazy, zwykle to kod aplikacji odpowiada za synchronizowanie pamięci podręcznej i indeksów z podstawową bazą. Na rysunku 1.1 pokazano prosty przykład tego, jak może to wyglądać (szczegóły znajdziesz w dalszych rozdziałach).

Gdy łączysz kilka narzędzi, aby udostępnić usługę, interfejs usługi lub interfejs **API** (ang. *application programming interface*) zwykle ukrywają szczegóły implementacji przed klientem. Oznacza to utworzenie nowego systemu danych o specjalnym przeznaczeniu na podstawie mniejszych komponentów o przeznaczeniu ogólnym. Złożony system danych może zapewniać określone gwarancje dotyczące np. poprawnego unieważniania lub aktualizowania pamięci podręcznej po zapisie danych, tak aby klienci zewnętrzne otrzymywały spójne wyniki. W ten sposób stajesz się nie tylko programistą aplikacji, ale też projektantem systemu danych.

Jeśli projektujesz system danych lub usługę, musisz odpowiedzieć na wiele skomplikowanych pytań. Jak zapewnisz, że dane pozostaną prawidłowe i kompletne, nawet gdy wystąpią wewnętrzne błędy? Jak zapewnisz klientom stabilnie wysoką wydajność, nawet gdy wystąpią problemy w komponentach systemu? Jak system będzie się skalował, aby obsłużyć wzrost obciążenia? Jak powinien wyglądać właściwy interfejs API usługi?





Rysunek 1.1. Możliwa architektura systemu danych łączącego kilka komponentów

Na projekt systemu danych wpływa wiele czynników, w tym umiejętności i doświadczenie jego twórców, zależności od starszych systemów, termin zakończenia prac, odporność organizacji na różnego rodzaju ryzyka, ograniczenia prawne itd. Te czynniki są wysoce zależne od sytuacji.

W tej książce uwzględniane są głównie trzy kwestie ważne w większości systemów informatycznych:

### Niezawodność

System powinien działać *prawidłowo* (poprawnie wykonywać swoje funkcje z oczekiwaną wydajnością) nawet w obliczu *problemów* (awarii sprzętowych lub programowych, a nawet błędów ludzkich). Zob. punkt „Niezawodność”.

### Skalowalność

Gdy system się *rozrasta* (ze względu na ilość danych lub ruchu albo złożoność), powinny istnieć sensowne sposoby radzenia sobie z tym wzrostem. Zob. punkt „Skalowalność”.

### Łatwość konserwacji

W przyszłości nad systemem pracować będzie wiele osób (inżynierów i pracowników operacyjnych zajmujących się zarówno konserwacją aktualnych funkcji, jak i dostosowywaniem systemu do nowych przypadków użycia). Wszyscy oni powinni móc pracować nad nim w *produktywny* sposób. Zob. punkt „Łatwość konserwacji”.

Te słowa są często używane bez dokładnego zrozumienia ich znaczenia. Reszta rozdziału została przeznaczona na rozważania o niezawodności, skalowalności i łatwości konserwacji. Później, w dalszych rozdziałach, przedstawione są różne techniki, architektury i algorytmy używane do osiągnięcia tych celów.

# Niezawodność

Każdy intuicyjnie rozumie, co oznacza, że coś jest niezawodne lub zawodne. Oto typowe oczekiwania z obszaru oprogramowania:

- Aplikacja wykonuje zadania żądane przez użytkownika.
- Aplikacja jest odporna na pomyłki popełnione przez użytkownika lub korzystanie z niej w niezaplanowany sposób.
- Wydajność aplikacji jest wystarczająca dla oczekiwanego sposobu użytkowania przy spodziewanym obciążeniu i ilości danych.
- System zapobiega nieuprawnionemu dostępowi i nadużyciom.

Jeśli wszystkie te rzeczy razem oznaczają „poprawne działanie”, *niezawodność* można w przybliżeniu rozumieć jako „ciągłe poprawne działanie nawet po wystąpieniu problemów”.

Rzeczy, które mogą pójść źle, są nazywane *błędami*, a systemy przewidujące błędy i radzące sobie z nimi określa się mianem *odpornych na błędy*. To ostatnie pojęcie może być nieco mylące, ponieważ sugeruje, że można zbudować system odporny na błędy wszelkiego rodzaju, co w praktyce jest niewykonalne. Gdyby cała ziemia (wraz z wszystkimi serwerami) została wchłonięta przez czarną dziurę, zapewnienie odporności na taki „błąd” wymagałoby hostingu rozwiązania w kosmosie. Życzę powodzenia w pozyskiwaniu w budżecie środków na taki system! Dlatego sensowne jest mówienie tylko o odporności na błędy *określonych rodzajów*.

Zauważ, że błąd (ang. *fault*) to nie to samo co awaria (ang. *failure*) [2]. Błąd zwykle definiuje się jako odstępstwo jednego komponentu systemu od specyfikacji, natomiast *awaria* następuje, gdy system jako całość przestaje świadczyć oczekiwane usługi użytkownikowi. Nie da się ograniczyć ryzyka wystąpienia błędów do zera. Dlatego zwykle najlepiej jest projektować mechanizmy zapewniania odporności na błędy, aby nie dopuścić do tego, by błąd doprowadził do awarii. W tej książce opisano kilka metod budowania niezawodnych systemów z użyciem zawodnych elementów.

Sprzeczne z intuicją jest to, że w tego typu odpornych na błędy systemach sensowne może być *zwiększanie* liczby błędów przez celowe ich wywoływanie (np. przez losowe zamykanie poszczególnych procesów bez ostrzeżenia). Wiele błędów krytycznych wynika ze złej obsługi błędów [3]. Celowo wprowadzając błędy, gwarantujesz, że mechanizmy zapewniania odporności są stale sprawdzane i testowane. Zwiększa to pewność, że błędy zostaną poprawnie obsłużone, gdy wystąpią z przyczyn naturalnych. Przykładem zastosowania tego podejścia jest mechanizm *Chaos Monkey* [4] opracowany przez firmę Netflix.

Choć zwykle preferuje się zapewnianie odporności na błędy, a nie zapobieganie im, w niektórych sytuacjach zapobieganie okazuje się lepsze niż leczenie (np. wtedy, gdy nie istnieje lekarstwo). Dotyczy to np. kwestii bezpieczeństwa. Jeśli napastnik włamie się do systemu i uzyska dostęp do poufnych danych, nie da się odwrócić skutków tego zdarzenia. Jednak w tej książce opisywane są głównie błędy, których skutkom można zaradzić. Błędy te są opisane w następujących punktach.

## Błędy sprzętowe

Gdy zastanawiasz się nad awariami systemów, na myśl przychodzą błędy sprzętowe. Awarie dysków twardych, błędy w pamięci RAM, przerwy w dostawie prądu, odłączenie niewłaściwego kabla sieciowego. Każdy, kto pracował w dużym centrum danych, potwierdzi, że gdy liczba maszyn jest duża, tego typu sytuacje zdarzają się *nieustannie*.

Średni czas do awarii (ang. *mean time to failure* — **MTTF**) dysku twardego wynosi od ok. 10 do 50 lat [5, 6]. Dlatego w klastrze składowania danych obejmującym 10 tys. dysków należy oczekiwać, że dziennie średnio zepsuje się jeden dysk.

Pierwszą reakcją jest zwykle zapewnienie nadmiarowości poszczególnych komponentów sprzętowych, aby zmniejszyć częstotliwość awarii systemu. Można łączyć dyski w macierze RAID, stosować dwa źródła zasilania i wymienne procesory dla serwerów, a także używać baterii i spalinyowych generatorów prądu, aby zapewnić zasilanie awaryjne centrum danych. Gdy jeden komponent przestanie działać, nadmiarowy zajmie jego miejsce na czas wymiany uszkodzonej jednostki. To podejście nie pozwala w pełni zapobiec awariom z powodu problemów sprzętowych, jednak jest dobrze znane i często sprawia, że maszyna potrafi latami pracować bez przestoju.

Do niedawna w większości aplikacji nadmiarowość komponentów sprzętowych wystarczała, ponieważ sprawiała, że całkowita awaria maszyny zdarzała się stosunkowo rzadko. Jeśli potrafisz szybko odtworzyć kopię zapasową na nowej maszynie, przestój wynikający z awarii w większości systemów nie jest katastrofą. Dlatego nadmiarowość w postaci dodatkowych maszyn może być potrzebna tylko w nielicznych zastosowaniach, w których wysoka dostępność była absolutnie konieczna.

Jednak wraz ze wzrostem ilości danych i wymaganiami obliczeniowymi aplikacji coraz częściej aplikacje potrzebowały większej liczby maszyn, co skutkowało proporcjonalnym wzrostem liczby błędów sprzętowych. Ponadto w niektórych platformach do udostępniania chmury (np. w usłudze Amazon Web Services — AWS) dość często zdarza się, że instancje maszyn wirtualnych stają się niedostępne bez ostrzeżenia [7], ponieważ platformy te są tak projektowane, by stawiały elastyczność i łatwość dostosowywania<sup>1</sup> nad niezawodność pojedynczych maszyn.

Dlatego następuje przechodzenie w kierunku systemów odpornych na utratę całych maszyn. W tym celu programowe metody zapewniania odporności na błędy są stosowane zamiast lub obok nadmiarowości sprzętu. Takie systemy mają też zalety operacyjne. System z jednym serwerem wymaga zaplanowanego przestoju, gdy potrzebujesz ponownie uruchomić maszynę (np. w celu wprowadzenia poprawek bezpieczeństwa systemu operacyjnego), natomiast system, który radzi sobie z awariami maszyn, można aktualizować węzeł po węźle bez powodowania przestoju całego systemu (jest to *aktualizacja stopniowa*; zob. rozdział 4.).

## Błędy programowe

Błędy sprzętowe zwykle są traktowane jako losowe i niezależne od siebie. Awaria dysku w jednej maszynie nie oznacza, że wystąpią problemy także z dyskiem innej maszyny. Mogą pojawić się pewne słabe korelacje (np. z powodu wspólnej przyczyny takiej jak temperatura w szafie serwerowej),

---

<sup>1</sup> Definicję znajdziesz w punkcie „Metody radzenia sobie z obciążeniem”.

jednak zwykle mało prawdopodobne jest, że duża liczba komponentów sprzętowych zawiedzie w tym samym momencie.

Inną kategorią są systematyczne błędy w systemie [8]. Trudniej je przewidzieć, a ponieważ występowanie takich problemów w węzłach jest skorelowane, błędy tego typu powodują znacznie więcej awarii systemu niż nieskorelowane błędy sprzętowe [5]. Oto przykłady takich problemów:

- Usterka oprogramowania powodująca, że każda instancja serwera aplikacji ulega awarii po otrzymaniu określonych niewłaściwych danych wejściowych. Pomyśl np. o sekundzie przestępnej 30 czerwca 2012 r., która sprawiła, że wiele aplikacji jednocześnie przestało działać z powodu usterki w jądrze Linuksa [9].
- Niekontrolowany proces zużywający wspólnie zasoby (czas procesora, pamięć, przestrzeń dyskową lub przepustowość łącza).
- Usługa, od której zależy system, działa powoli, przestaje reagować lub zaczyna zwracać nieprawidłowe odpowiedzi.
- Awaria kaskadowa, w trakcie której niewielka usterka jednego komponentu skutkuje błędem w innym komponentcie, co z kolei prowadzi do dalszych problemów [10].

Usterki, które powodują tego rodzaju błędy programowe, często pozostają w ukryciu przez długi czas — do momentu ujawnienia ich z powodu nietypowego zbiegu okoliczności. Wtedy okazuje się, że w oprogramowaniu przyjęto określone założenie dotyczące środowiska i choć zwykle to założenie jest prawdziwe, z jakiegoś powodu ostatecznie przestało takie być [11].

Nie istnieje proste rozwiązanie problemu błędów systematycznych w oprogramowaniu. Pomocnych może być wiele drobiazgów: staranne przemyślenie założeń dotyczących systemu i występujących interakcji, odizolowanie procesu, umożliwienie procesowi na awarię i restart, pomiar, monitorowanie i analizowanie działania systemu w środowisku produkcyjnym. Jeśli system ma zapewniać określone gwarancje (np. w kolejce komunikatów może to polegać na tym, że liczba komunikatów przychodzących musi być równa liczbie komunikatów wychodzących), może stale sprawdzać swój stan w trakcie pracy i zgłaszać alarm po wykryciu rozbieżności [12].

## Błędy ludzkie

To ludzie projektują i tworzą systemy oprogramowania. Ludźmi są też operatorzy sterujący pracą systemów. Nawet mając najlepsze intencje, ludzie popełniają pomyłki. Na przykład w badaniach nad dużą usługą internetową odkryto, że główną przyczyną przestojów były błędy w konfiguracji spowodowane przez operatorów. Błędy sprzętowe (serwerów lub sieci) odpowiadały tylko za 10 – 25% przestojów [13].

Jak sprawić, by systemy były niezawodne mimo ludzkich pomyłek? W najlepszych systemach stosuje się kilka podejść:

- Systemy są projektowane w taki sposób, aby zminimalizować możliwości spowodowania błędów. Na przykład odpowiednio zaprojektowane abstrakcje, interfejsy API i interfejsy administratora powodują, że łatwo jest robić „właściwe rzeczy”, a zniechęcają do robienia „niewłaściwych”. Jeśli jednak interfejs okaże się zbyt ograniczający, użytkownicy nie będą z niego korzystać, co oznacza utratę korzyści. Trudno więc zachować odpowiednią równowagę w tym obszarze.

- Izolowanie miejsc, w których ludzie popełniają najwięcej pomyłek, od miejsc, gdzie błędy mogą skutkować awariami. Przede wszystkim warto udostępnić kompletne nieprodukcyjne *środowisko izolowane* (ang. *sandbox*), w którym użytkownicy mogą bezpiecznie eksplorować możliwości i eksperymentować na rzeczywistych danych, ale bez wpływu na rzeczywistych użytkowników.
- Prowadzenie dokładnych testów na wszystkich poziomach: od testów jednostkowych po testy integracyjne i ręczne całego systemu [3]. Testy zautomatyzowane są powszechnie stosowane, dobrze zrozumiałe i wartościowe zwłaszcza w kontekście sprawdzania przypadków brzegowych, które rzadko występują w trakcie normalnej eksploatacji systemu.
- Umożliwianie szybkiego i łatwego przywracania stanu po błędach ludzkich w celu zminimalizowania ich wpływu, gdy wydarzy się awaria. Można np. umożliwić szybkie wycofywanie zmian w konfiguracji, stopniowo wprowadzać nowy kod (tak aby nieoczekiwane usterki dotykały tylko małego podzbioru użytkowników) i zapewniać narzędzia do wykonywania ponownych obliczeń na danych (jeśli się okaże, że wcześniejsze obliczenia były błędne).
- Przygotowanie szczegółowego i precyzyjnego systemu monitorowania, np. wskaźników wydajności i współczynników błędów. W innych dziedzinach inżynierii ten obszar nazywa się *telemetrią*. Gdy rakieta już wystartuje, telemetria jest niezbędna do śledzenia tego, co się dzieje, i do zrozumienia awarii [14]. Monitorowanie pozwala wcześniej wykryć sygnały ostrzegawcze i sprawdzać, czy jakieś założenia lub ograniczenia nie zostały naruszone. Gdy wystąpi problem, wskaźniki mogą być nieocenione w diagnozowaniu usterki.
- Wprowadzenie odpowiednich szkoleń i metod zarządzania. To skomplikowany i ważny temat wykraczający poza zakres tej książki.

## Jak ważna jest niezawodność?

Niezawodność jest istotna nie tylko w elektrowniach atomowych i oprogramowaniu do kontroli ruchu lotniczego. Także od bardziej zwyczajnych aplikacji oczekuje się, że będą działały niezawodnie. Usterki w aplikacjach biznesowych powodują utratę wydajności (i problemy prawne, jeśli dane są błędnie prezentowane), a przestoje sklepów elektronicznych mogą prowadzić do poważnych kosztów z powodu utraty dochodów i spadku reputacji.

Nawet w aplikacjach „niekrytycznych” programiści odpowiadają przed użytkownikami. Pomyśl o rodzicu, który przechowuje wszystkie zdjęcia i filmy swoich dzieci w służącym do tego programie [15]. Jak ta osoba się poczuje, jeśli baza danych zostanie nagle uszkodzona? Czy rodzic będzie wiedział, jak ją odzyskać na podstawie kopii zapasowej?

W niektórych sytuacjach zdarza się, że rezygnuje się z niezawodności na rzecz obniżenia kosztów rozwoju (np. w trakcie tworzenia prototypowego produktu na nieznanym rynku) lub kosztów operacyjnych (np. w usłudze o bardzo niskiej marży). Należy jednak zachować dużą ostrożność, wybierając drogę na skróty.

# Skalowalność

Nawet jeśli dziś system działa niezawodnie, nie oznacza to, że będzie tak w przyszłości. Jednym z typowych powodów spadku wydajności jest wzrost obciążenia. Możliwe, że liczba jednoczesnych użytkowników systemu wzrosła z 10 tys. do 100 tys. lub z miliona do 10 mln. Możliwe, że system przetwarza znacznie większą ilość danych niż wcześniej.

*Skalowalność* to pojęcie służące do opisu tego, czy system radzi sobie ze wzrostem obciążenia. Warto jednak zauważyć, że nie jest to jednowymiarowa etykieta, którą można nadać systemowi. Stwierdzenia typu „X się skaluje” lub „Y się nie skaluje” nic nie znaczą. Analizowanie skalowalności wymaga rozważenia pytań takich jak: „Jeśli system rozrośnie się w określony sposób, jakie będą możliwości poradzenia sobie z tym wzrostem?” i „Jak można dodać zasoby obliczeniowe, aby obsłużyć większe obciążenie?”.

## Opisywanie obciążenia

Najpierw należy zwięźle opisać aktualne obciążenie systemu. Dopiero potem można analizować kwestie związane ze wzrostem (co się stanie, jeśli obciążenie wzrośnie dwukrotnie?). Obciążenie można opisać za pomocą kilku liczb nazywanych *parametrami obciążenia*. Odpowiedni dobór tych parametrów zależy od architektury systemu. Może to być liczba żądań do serwera WWW na sekundę, stosunek odczytów do zapisów w bazie, liczba jednocześnie aktywnych użytkowników na czacie, współczynnik trafień pamięci podręcznej lub coś innego. Możliwe, że interesują Cię typowe przypadki. Możliwe też, że przyczyną występowania „wąskiego gardła” jest niewielka liczba skrajnych przypadków.

Aby przyjrzeć się temu w bardziej konkretny sposób, warto rozważyć działanie Twittera na podstawie danych opublikowanych w listopadzie 2012 r. [16]. Dwie podstawowe operacje na Twitterze to:

### *Publikowanie tweetów*

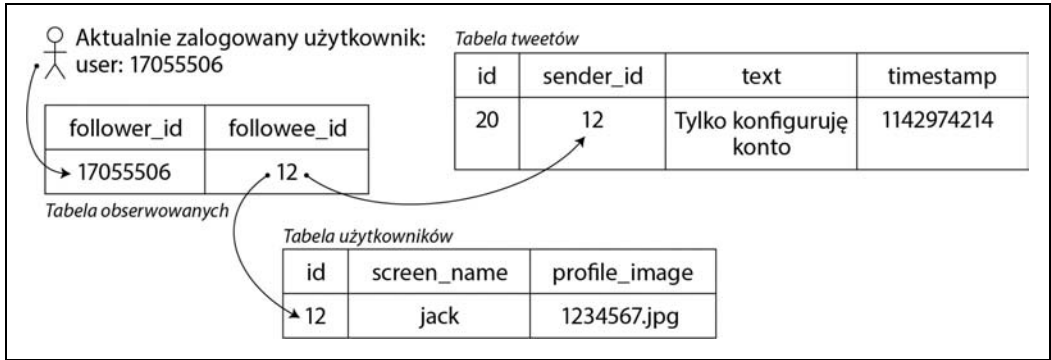
Użytkownik może opublikować nową wiadomość dla swoich obserwatorów (średnio 4,6 tys. żądań na sekundę; szczytowo ponad 12 tys. żądań na sekundę).

### *Wyświetlanie osi czasu*

Użytkownik może wyświetlić wiadomości opublikowane przez obserwowane osoby (300 tys. żądań na sekundę).

Sama obsługa 12 tys. zapisów na sekundę (szczytowa szybkość zamieszczania tweetów) jest stosunkowo prosta. Jednak w przypadku Twittera trudności ze skalowaniem nie wynikają przede wszystkim z liczby tweetów, ale z powodu *rozgałęzień* — każdy użytkownik obserwuje wielu innych i jest obserwowany przez liczne osoby. Są dwie ogólne metody wykonywania dwóch opisanych operacji:

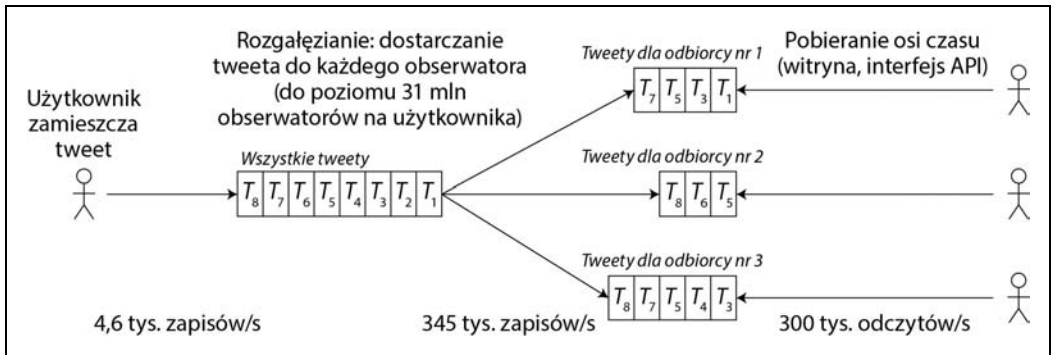
1. Zamieszczenie tweetu powoduje wstawienie go do globalnej kolekcji wiadomości. Gdy użytkownik żąda osi czasu, należy znaleźć wszystkie obserwowane przez niego osoby, odszukać wszystkie tweety każdej z nich i połączyć je (posortowane według czasu). W relacyjnej bazie danej, takiej jak na rysunku 1.2, można napisać zapytanie o następującej postaci:



Rysunek 1.2. Prosty schemat relacyjny służący do obsługi osi czasu na Twitterze

```
SELECT tweets.*, users.* FROM tweets
JOIN users ON tweets.sender_id = users.id
JOIN follows ON follows.followee_id = users.id
WHERE follows.follower_id = current_user
```

- Przechowywanie pamięci podręcznej z osią czasu każdego użytkownika (to coś w rodzaju skrzynki odbiorczej z tweetami każdego odbiorcy; zob. rysunek 1.3). Gdy użytkownik zamieszcza tweet, należy znaleźć wszystkich obserwatorów tej osoby i dodać ten nowy tweet do pamięci podręcznej z osią czasu każdego obserwatora. Żądanie odczytu osi czasu nie jest wtedy kosztowne, ponieważ wynik został wcześniej obliczony.



Rysunek 1.3. Potok danych służący do dostarczania tweetów do obserwatorów na Twitterze z poziomami obciążenia z listopada 2012 r. [16]

W pierwszej wersji Twittera używano podejścia nr 1, jednak systemom trudno było wtedy poradzić sobie z obciążeniem powodowanym przez zapytania pobierające oś czasu, dlatego firma wprowadziła podejście nr 2. Nowe podejście sprawdza się lepiej, ponieważ średnia szybkość publikowania tweetów jest o prawie dwa rzędy wielkości mniejsza niż częstotliwość odczytów osi czasu. Dlatego w tej sytuacji lepiej wykonywać więcej pracy na etapie zapisu i mniej w trakcie odczytu.

Wadę podejścia nr 2 stanowi jednak to, że zamieszczenie tweetu wymaga dużo dodatkowej pracy. Tweet jest średnio dostarczany do ok. 75 obserwatorów, dlatego 4,6 tys. tweetów na sekundę oznacza 345 tys. zapisów na sekundę w osiach czasu w pamięci podręcznej. Ta średnia ukrywa jednak to, że liczba obserwatorów na użytkownika jest bardzo zróżnicowana. Niektóre osoby mają ponad 30 mln



obserwatorów. To oznacza, że jeden tweet może prowadzić do ponad 30 mln zapisów w osiach czasu! Szybkie wykonywanie tej operacji (Twitter stara się dostarczać tweety obserwatorom w czasie 5 s) to poważne wyzwanie.

W przypadku Twittera rozkład liczby obserwatorów na użytkownika (możliwe, że z wagami zależnymi od tego, jak często użytkownicy publikują tweety) to ważny parametr obciążenia w trakcie analizy skalowalności, ponieważ określa on obciążenie wyjściowe. Twoja aplikacja może mieć zupełnie inne cechy, ale możesz zastosować podobne zasady do analizy obciążenia.

Oto zakończenie historyjki o Twitterze: po skutecznym zaimplementowaniu podejścia nr 2 Twitter przechodzi do hybrydowego rozwiązania łączącego obie techniki. Tweety większości użytkowników nadal są zapisywane w modelu rozgałęziania na osiach czasu w momencie publikowania wiadomości, przy czym nie dotyczy to niewielkiej grupy użytkowników (celebrytów) o bardzo dużej liczbie obserwatorów. Tweety celebrytów są pobierane niezależnie i dołączane do osi czasu obserwatora w trakcie jej wczytywania (tak jak w podejściu nr 1). To hybrydowe podejście zapewnia stabilnie wysoką wydajność. Do tego przykładu wrócisz w rozdziale 12. po zapoznaniu się z bardziej technicznymi informacjami.

## Opis wydajności

Po opisaniu obciążenia możesz zbadać, co się stanie po jego wzroście. Możesz na to spojrzeć na dwa sposoby:

- Co się stanie z wydajnością systemu, jeśli obciążenie wzrośnie, a zasoby systemowe (procesor, pamięć, przepustowość sieci itd.) pozostaną niezmienione?
- O ile trzeba zwiększyć ilość zasobów, jeśli chcesz uzyskać tę samą wydajność po wzroście obciążenia?

Oba te pytania wymagają ustalenia liczb określających wydajność. Przyjrzyj się więc pokrótce temu, jak opisywać wydajność systemu.

W systemach przetwarzania wsadowego (np. Hadoop) zwykle ważna jest *przepustowość*, czyli liczba rekordów, jakie można przetwarzać w ciągu sekundy, lub łączny czas potrzebny do wykonania zadania na zbiorze danych o określonej wielkości<sup>2</sup>. W systemach internetowych zwykle ważniejszy jest *czas odpowiedzi* usługi, czyli czas od przesłania żądania przez klienta do otrzymania odpowiedzi.



### Opóźnienie a czas odpowiedzi

Określenia *opóźnienie* (ang. *latency*) i *czas odpowiedzi* (ang. *response time*) są często używane zamiennie, ale nie oznaczają one tego samego. Czas odpowiedzi mierzy się z perspektywy klienta. Obok czasu przetwarzania żądania (*czasu pracy usługi*) obejmuje też czas transferu w sieci i spędzony w kolejkach. Opóźnienie to czas oczekiwania żądania na obsłużenie. W tym okresie żądanie jest *uśpione* (oczekuje na usługę) [17].

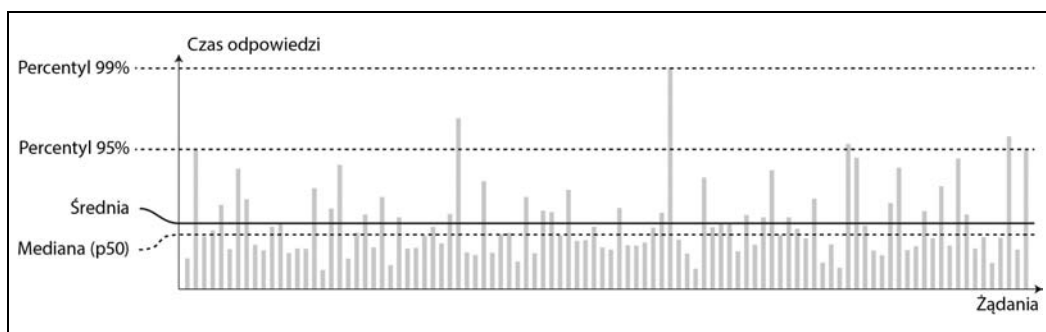
---

<sup>2</sup> W idealnym świecie czas wykonywania zadania wsadowego to wielkość zbioru danych podzielona przez przepustowość. W praktyce ten czas jest często dłuższy. Jest to spowodowane niesymetrycznością (nierównym rozkładem danych między procesy robocze) i koniecznością oczekiwania na zakończenie pracy przez najwolniejsze zadanie.



Nawet jeśli wielokrotnie przesyłasz to samo żądanie, przy każdej próbie uzyskasz nieco odmienny czas odpowiedzi. W praktyce w systemie obsługującym różne żądania czas odpowiedzi może się znacznie wahać. Dlatego o czasie odpowiedzi trzeba myśleć nie jak o jednej liczbie, ale jak o *rozkładzie* wartości, który można zmierzyć.

Na rysunku 1.4 każdy szary słupek reprezentuje żądanie kierowane do usługi. Wysokość słupków pokazuje, ile czasu zajęła obsługa tych żądań. Większość żądań jest obsługiwanych szybko, jednak występują rzadkie *obserwacje odstające* o znacznie dłuższym czasie obsługi. Możliwe, że te wolno obsługiwane żądania są z natury bardziej kosztowne — np. wymagają przetworzenia większej ilości danych. Jednak nawet w sytuacji, gdy uważasz, że wszystkie żądania powinny zajmować tyle samo czasu, i tak pojawiają się różnice. Losowy dodatkowy czas może wynikać z przełączania kontekstu na proces tła, utraty pakietu sieciowego i retransmisji w protokole TCP, przerwy na odzyskiwanie pamięci, błędu stronicowania wymuszającego odczyt z dysku, mechanicznych wibracji szafy serwerowej [18] i wielu innych przyczyn.



Rysunek 1.4. Średnia i percentyle — czasy odpowiedzi dla próbki 100 żądań kierowanych do usługi

Często podaje się *średni* czas odpowiedzi usługi. Ściśle rzecz biorąc, określenie „średnia” nie dotyczy żadnego konkretnego wzoru. W praktyce zwykle rozumie się ją jako *średnią arytmetyczną* (jeśli danych jest  $n$  wartości, należy je zsumować, a następnie podzielić wynik przez  $n$ ). Jednak średnia nie jest bardzo dobrą miarą, gdy chcesz poznać „typowy” czas odpowiedzi, ponieważ nie pokazuje, ile użytkowników dotyczy dany czas odpowiedzi.

Zwykle lepiej posługiwać się *percentylami*. Jeśli przyjrzyysz się liście czasów odpowiedzi i posortujesz ją od najkrótszych do najwolniejszych, *mediana* to punkt środkowy. Na przykład jeśli mediana czasów odpowiedzi wynosi 200 ms, oznacza to, że połowa żądań jest przetwarzana w czasie krótszym niż 200 ms, a połowa żądań zajmuje więcej czasu.

Mediana to więc dobra miara, kiedy chcesz wiedzieć, jak długo użytkownicy zwykle muszą czekać. Połowa użytkowników jest obsługiwana w czasie krótszym niż mediana, a druga połowa w czasie dłuższym. Medianę nazywa się też *percentylem 50%*, co czasem skraca się do postaci *p50*. Zauważ, że mediana dotyczy jednego żądania. Jeśli dany użytkownik zgłasza kilka żądań (w trakcie sesji lub z powodu znajdowania się na jednej stronie wielu zasobów), prawdopodobieństwo, że przynajmniej jedno z nich będzie wymagało więcej czasu niż mediana, jest znacznie wyższe niż 50%.

Aby ustalić, jak kłopotliwe są obserwacje odstające, możesz się przyjrzeć wyższym percentydom. Często wykorzystywane są percentyle 95%, 99% i 99,9% (*p95*, *p99* i *p999*). Są to wartości progowe czasów odpowiedzi, dla których 95%, 99% i 99,9% żądań jest przetwarzanych szybciej niż dana

wartość. Na przykład jeśli czas odpowiedzi dla percentyla 95% wynosi 1,5 s, oznacza to, że 95 ze 100 żądań jest przetwarzanych w czasie krótszym niż 1,5 s, a 5 ze 100 żądań wymaga 1,5 s lub więcej. Przedstawiono to na rysunku 1.4.

Wysokie percentyle czasów odpowiedzi, nazywane też *skrajnymi wartościami opóźnienia* (ang. *tail latencies*), są ważne, ponieważ bezpośrednio wpływają na komfort pracy użytkowników usługi. Na przykład Amazon opisuje wymogi dotyczące czasów odpowiedzi usług wewnętrznych w kategoriach percentyla 99,9%, choć czasy te dotyczą tylko 1 na 1000 żądań. Dzieje się tak, ponieważ klienci generujący najdłuższe żądania to często ci z największą ilością danych na kontach, co wynika z wielu zakupów. Oznacza to najcenniejszych klientów [19]. Ważne jest, aby dbać o zadowolenie tych klientów, zapewniając, że witryna szybko ich obsługuje. W Amazonie zaobserwowano też, że wydłużenie czasu odpowiedzi o 100 ms skutkuje spadkiem sprzedaży o 1% [20]. Z innych badań wynika, że jednosekundowe wydłużenie obsługi zmniejsza wskaźnik satysfakcji klientów o 16% [21, 22].

Optymalizowanie pod kątem percentyla 99,99% (najwolniejsze z 10 tys. żądań) okazało się jednak zbyt kosztowne i nie przyniosło wystarczających korzyści ze względu na cele Amazonu. Skracanie czasu odpowiedzi dla bardzo wysokich percentyli jest trudne, ponieważ wartości te są podatne na losowe zdarzenia pozostające poza naszą kontrolą, a korzyści są coraz mniejsze.

Percentyle są często używane w **SLO** (ang. *service level objective*) i **SLA** (ang. *service level agreement*). Są to kontrakty definiujące oczekiwaną wydajność i dostępność usługi. Kontrakt SLA może określać, że usługę uznaje się jako aktywną, jeśli mediana czasów odpowiedzi wynosi mniej niż 200 ms i percentyl 99% wynosi mniej niż 1 s (jeżeli czas odpowiedzi jest dłuższy, usługa równie dobrze mogłaby być wyłączona). Dodatkowym wymogiem może być to, by usługa była aktywna przez co najmniej 99,9% czasu. Te poziomy wyznaczają oczekiwania klientów usługi i pozwalają im domagać się zwrotu zapłaty, jeśli kontrakt SLA nie zostanie zrealizowany.

W przypadku wysokich percentyli za dużą część czasu odpowiedzi często odpowiadają kolejki. Ponieważ serwer potrafi jednocześnie przetwarzać tylko niewiele elementów (ogranicza to np. liczba rdzeni procesora), wystarczy mała liczba wolnych żądań, by wstrzymać przetwarzania dalszych żądań. To efekt, który można nazwać *blokowaniem na czole kolejki*. Nawet gdy serwer potrafi szybko przetworzyć te dalsze żądania, dla klienta czas odpowiedzi jest długi z powodu czasu oczekiwania na ukończenie wcześniejszych żądań. Z tego powodu ważne jest, by mierzyć czasy odpowiedzi po stronie klienta.

Gdy obciążenie jest generowane sztucznie w celu przetestowania skalowalności systemu, klient generujący obciążenie musi przesyłać żądania niezależnie od czasów odpowiedzi. Jeśli klient będzie oczekiwał na ukończenie wcześniejszego żądania przed przesłaniem następnego, w trakcie testów kolejki będą sztucznie krótsze niż w rzeczywistości, co zakłóca pomiary [23].

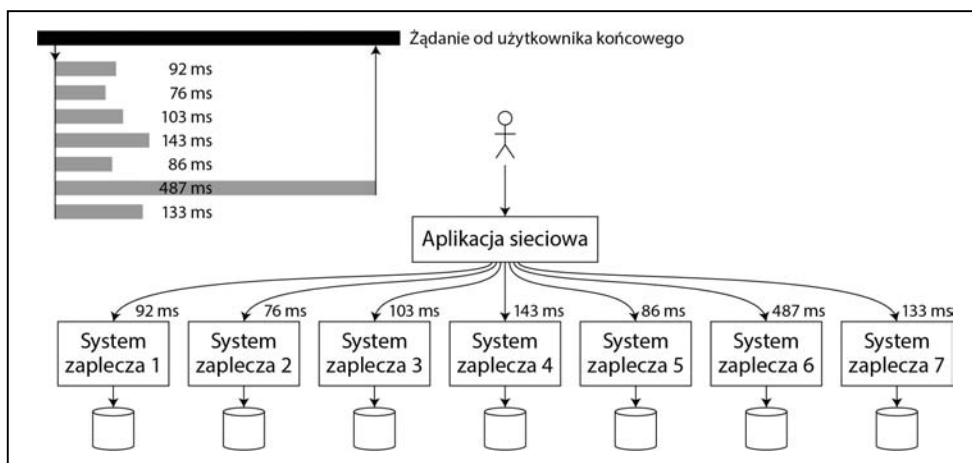
## Sposoby radzenia sobie z obciążeniem

Po omówieniu parametrów opisu obciążenia i wskaźników do pomiaru wydajności można na dobre rozpocząć analizę skalowalności — jak utrzymać wysoką wydajność, gdy obciążenie wzrośnie do określonego poziomu?

Architektura odpowiednia dla danego poziomu obciążenia zapewne nie poradzi sobie z jego 10-krotnym wzrostem. Dlatego jeśli pracujesz nad usługą, której obciążenie szybko rośnie, zapewne będziesz musiał na nowo przemyśleć architekturę po wzroście obciążenia o każdy rząd wielkości, a prawdopodobnie jeszcze częściej.

## Percentyle w praktyce

Wysokie percentyle są ważne zwłaszcza w usługach zaplecza wywoływanych wielokrotnie w ramach obsługi jednego żądania od użytkownika końcowego. Nawet jeśli te wywołania są zgłaszane równoległe, żądanie od użytkownika końcowego nadal musi oczekiwać na przetworzenie najwolniejszego z równoległych wywołań. Wystarczy jedno wolne wywołanie, aby obsługa całego żądania trwała długo. Przedstawia to rysunek 1.5. Nawet gdy tylko niewielki odsetek wywołań na zapleczu jest długo przetwarzany, ryzyko natrafienia na wolne wywołanie rośnie, jeśli żądanie użytkownika końcowego wymaga wielu wywołań. W takiej sytuacji także dla większego odsetka żądań użytkownika końcowego otrzymasz dłuższy czas odpowiedzi (ten efekt to *zwiększenie liczby skrajnych wartości opóźnienia* [24]).



Rysunek 1.5. Gdy obsługa żądania wymaga kilku wywołań do zaplecza, wystarczy jedno długie żądanie do zaplecza, aby wydłużyć przetwarzanie całego żądania od użytkownika końcowego

Jeśli chcesz dodać percentyle czasów odpowiedzi do panelu monitorowania usług, musisz na bieżąco wydajnie je obliczać. Możliwe, że zechcesz utrzymywać okno przesuwne czasów odpowiedzi na żądania z ostatnich 10 minut. Co minutę system może obliczać medianę i różne percentyle na podstawie wartości z tego okna i wyświetlać uzyskane dane na wykresie.

Naiwna implementacja polega na przechowywaniu listy czasów odpowiedzi dla wszystkich żądań z okna czasowego i sortowaniu tej listy co minutę. Jeśli to rozwiązanie jest dla Ciebie za mało wydajne, istnieją algorytmy, które potrafią obliczać dobre przybliżenie percentyli przy minimalnych kosztach procesora i pamięci. Są to algorytmy forward decay [25], t-digest [26] i HdrHistogram [27]. Pamiętaj, że uśrednianie percentyli, np. w celu zmniejszenia rozdzielczości czasowej lub połączenia danych z kilku maszyn, matematycznie nie ma sensu. Poprawną metodę agregowania danych z czasami odpowiedzi stanowi dodawanie histogramów [28].

Ludzie często mówią o rozbieżności między *skalowaniem pionowym* (ang. *scaling up* lub *vertical scaling*; polega to na wymianie maszyny na wydajniejszą) a *skalowaniem poziomym* (ang. *scaling out* lub *horizontal scaling*; jest to rozdzielanie obciążenia między wiele mniejszych maszyn). Rozdzielanie obciążenia między liczne maszyny nazywa się też architekturą *bez zasobów współdzielonych* (ang. *shared-nothing*). System, który potrafi działać na jednej maszynie, jest zwykle prostszy, jednak zaawansowane maszyny bywają bardzo drogie, dlatego przy bardzo wysokim obciążeniu roboczym często nie da się uniknąć skalowania poziomego. W praktyce skuteczne architektury to często pragmatyczne połączenie obu podejść. Na przykład zastosowanie kilku mocnych maszyn może się okazać prostsze i tańsze niż używanie dużej liczby małych maszyn wirtualnych.

Niektóre systemy są *elastyczne*, co oznacza, że potrafią automatycznie dodawać zasoby obliczeniowe po wykryciu wzrostu obciążenia. Inne są skalowane ręcznie (to człowiek analizuje zasoby i decyduje o dodaniu do systemu nowych maszyn). System elastyczny może być bardzo przydatny, gdy obciążenie jest wysoce nieprzewidywalne, jednak systemy skalowane ręcznie są prostsze i powodują mniej niespodzianek w trakcie eksploatacji (zob. punkt „Równoważenie partycji”).

Choć podział usług bezstanowych między wiele maszyn jest stosunkowo prosty, przenoszenie stanowych systemów danych z jednego węzła do środowiska rozproszonego może powodować wiele dodatkowej złożoności. Dlatego do niedawna powszechnie uważano, że bazę należy przechowywać w jednym węźle (ze skalowaniem pionowym) do czasu, gdy koszty skalowania lub wymogi wysokiej dostępności wymuszają przejście do środowiska rozproszonego.

Wraz ze wzrostem jakości narzędzi i abstrakcji z obszaru systemów rozproszonych to powszechne podejście może się zmienić (przynajmniej w niektórych aplikacjach). Możliwe, że w przyszłości rozproszone systemy danych staną się standardem — nawet w sytuacjach, gdy nie trzeba będzie obsługiwać dużych ilości danych lub dużego ruchu. Dalej w książce omówiono wiele rodzajów rozproszonych systemów danych. Dowiesz się nie tylko tego, w jakim stopniu są one skalowalne, ale też tego, jak łatwe jest ich użytkowanie i konserwowanie.

Architektura systemów działających na dużą skalę jest zwykle ściśle powiązana z aplikacją. Nie istnieje nic takiego jak uniwersalna skalowalna architektura na każdą okazję (nieformalnie można ją nazwać *magicznym sosem skalującym*). Problem może stanowić liczba odczytów, liczba zapisów, ilość przechowywanych danych, złożoność danych, wymogi dotyczące czasu odpowiedzi, wzorce dostępu lub (zazwyczaj) połączenie wszystkich tych i wielu innych kwestii.

Na przykład system zaprojektowany do obsługi 100 tys. żądań na sekundę (każde po 1 kB) wygląda zupełnie inaczej niż system zaprojektowany do obsługi trzech żądań na minutę (każde po 2 GB), nawet jeśli ilość danych w obu tych systemach jest podobna.

Architektura dobrze się skalująca w konkretnej aplikacji opiera się na założeniach dotyczących tego, które operacje będą często wykonywane, a które rzadko (są to parametry obciążenia). Jeśli te założenia okażą się błędne, praca inżynierska włożona w skalowanie w najlepszym razie zostanie zmarnowana, a w najgorszym przyniesie skutki odwrotne od zamierzonych. W startupach na wczesnym etapie rozwoju lub w nieprzetestowanych produktach zwykle ważniejsza jest możliwość szybkiego dodawania funkcji produktu niż myślenie o skalowaniu pod kątem hipotetycznego przyszłego obciążenia.

Nawet gdy skalowalną architekturę dostosowuje się do konkretnej aplikacji, zwykle jest zbudowana z komponentów o ogólnym przeznaczeniu łączonych w znane wzorce. W tej książce opisano takie komponenty i wzorce.

## Łatwość konserwacji

Wiadomo, że większość kosztów związanych z oprogramowaniem jest ponoszonych nie w trakcie jego budowania, ale w czasie bieżącej eksploatacji — w związku z naprawianiem błędów, utrzymywaniem zdolności operacyjnej systemów, badaniem awarii, dostosowywaniem do nowych platform, modyfikowaniem na potrzeby nowych przypadków użycia, spłacaniem długu technicznego i dodawaniem nowych funkcji.

Jednak, niestety, wiele osób pracujących nad systemami oprogramowania nie lubi konserwacji *przestarzałych* (ang. *legacy*) systemów. Możliwe, że taka konserwacja wymaga naprawiania cudzych pomyłek albo pracy z nieaktualnymi platformami lub systemami zmuszonymi wykonywać zadania, do których robienia nie są przeznaczone. Każdy przestarzały system jest kłopotliwy w inny sposób. Dlatego tak trudno przedstawić ogólne wskazówki.

Jednak można i należy projektować oprogramowanie tak, aby starać się zminimalizować trudności w trakcie konserwacji i samemu uniknąć tworzenia przestarzałego oprogramowania. W tym celu warto zwrócić uwagę przede wszystkim na trzy zasady projektowania systemów informatycznych:

### *Łatwość eksploatacji*

Należy umożliwić pracownikom operacyjnym łatwe zapewnianie płynnej pracy systemu.

### *Prostota*

Nowi inżynierowie powinni móc łatwo zrozumieć system. W tym celu należy w maksymalnym stopniu wyeliminować z niego złożoność. Zauważ, że nie jest to tym samym co prostota interfejsu użytkownika.

### *Łatwość modyfikowania*

Zadbaj o to, by inżynierowie mogli w przyszłości łatwo wprowadzać zmiany w systemie, dostosowując go do nieprzewidzianych przypadków użycia po zmianie wymagań. Tę cechę można też nazwać *rozszerzalnością*, *modyfikowalnością* lub *plastycznością*.

Podobnie jak w kontekście niezawodności i skalowalności, nie istnieją łatwe rozwiązania pozwalające osiągać te cele. Zamiast tego warto spróbować zastanowić się nad systemami z uwzględnieniem łatwości eksploatacji, prostoty i łatwości modyfikowania.

## Łatwość eksploatacji — ułatwianie życia pracownikom operacyjnym

Pojawiają się sugestie, że „dobry zespół operacyjny często potrafi poradzić sobie z ograniczeniami kiepskiego (lub niekompletnego) oprogramowania, jednak dobre oprogramowanie nie będzie działać stabilnie przy kiepskim zespole operacyjnym” [12]. Choć niektóre aspekty eksploatacji można i należy automatyzować, to ludzie odpowiadają za konfigurowanie automatyzacji i zapewnianie jej poprawnego działania.

Zespół operacyjny jest niezbędny do zapewniania płynnego działania systemów informatycznych. Dobry zespół operacyjny zwykle odpowiada za czynniki wymienione poniżej i inne [29]:

- Monitorowanie stanu systemu i szybkie przywracanie usługi, jeśli znajdzie się w nieodpowiednim stanie.
- Wykrywanie przyczyn problemów takich jak awarie systemu lub spadek wydajności.
- Zapewnianie aktualności oprogramowania i platform (w tym: instalowanie poprawek bezpieczeństwa).
- Badanie tego, jak różne systemy wpływają na siebie, co pozwala uniknąć wprowadzania mogących rodzić problemy zmian, zanim spowodują szkody.
- Przewidywanie przyszłych problemów i rozwiązywanie ich przed wystąpieniem (planowanie wydajności).
- Opracowywanie dobrych praktyk i narzędzi wdrażania oprogramowania, zarządzania konfiguracją itd.
- Wykonywanie złożonych zadań z zakresu eksploatacji (np. przenoszenie aplikacji z jednej platformy na inną).
- Utrzymywanie bezpieczeństwa systemu w trakcie wprowadzania zmian w konfiguracji.
- Definiowanie procesów, dzięki którym eksploatacja jest przewidywalna, i pomaganie w utrzymywaniu stabilności środowiska produkcyjnego.
- Zachowywanie w organizacji wiedzy na temat systemu, nawet gdy poszczególne osoby przychodzą do firmy i z niej odchodzą.

Łatwość eksploatacji oznacza, że wykonywanie rutynowych zadań jest proste, dzięki czemu zespół operacyjny może się skoncentrować na działaniach o wysokiej wartości. Systemy danych mogą na różne sposoby ułatwiać wykonywanie rutynowych prac. Oto niektóre z tych sposobów:

- Zapewnianie wglądu w pracę środowiska uruchomieniowego i wewnętrzne mechanizmy systemu za pomocą dobrego systemu monitorowania.
- Zapewnianie dobrej obsługi automatyzacji i integracji ze standardowymi narzędziami.
- Unikanie zależności od pojedynczych maszyn (pozwala to na wyłączenie maszyny na potrzeby konserwacji, a cały system działa wtedy bez zakłóceń).
- Zapewnianie dobrej dokumentacji i łatwego do zrozumienia modelu operacyjnego („Jeśli zrobię X, nastąpi Y”).
- Zapewnianie skutecznego działania domyślnego, ale też dawanie administratorom swobody do zmiany w razie potrzeby wprowadzenia ustawień domyślnych.
- Umożliwianie automatycznej naprawy tam, gdzie to wskazane, i zapewnianie administratorom ręcznej kontroli nad stanem systemu, jeśli to konieczne.
- Zapewnianie przewidywalnej pracy i minimalizowanie niespodzianek.

## Prostota — zarządzanie złożonością

W niewielkich projektach informatycznych kod może być cudownie prosty i zrozumiały. Jednak gdy projekt się rozrasta, kod często staje się bardzo skomplikowany i trudny do analizy. Ta złożoność spowalnia wszystkich, którzy muszą pracować nad systemem, co dodatkowo zwiększa koszty konserwacji. Projekt informatyczny pogrążony w złożoności można nazwać *wielką bryłą błota* [30].

Istnieją różne objawy złożoności: znaczny wzrost ilości miejsca potrzebnego na stan, ściśle powiązanie między modułami, skomplikowane zależności, niespójne nazewnictwo i terminologia, sztuczki służące rozwiązaniu problemów z wydajnością, specjalne przypadki rozwiązujące problemy z innych miejsc itd. Wiele na ten temat już napisano [31, 32, 33].

Gdy złożoność utrudnia konserwację, często następuje przekroczenie budżetów i harmonogramów. W skomplikowanym oprogramowaniu rośnie też ryzyko pojawienia się błędów w trakcie wprowadzania zmian. Gdy system dla programistów jest trudny do zrozumienia i analizy, łatwiej może być przeoczyć ukryte założenia, niezamierzone konsekwencje i nieoczekiwane interakcje. Natomiast ograniczenie złożoności znacznie ułatwia konserwację oprogramowania. Dlatego prostota powinna być ważnym celem w budowanych systemach.

Upraszczenie systemu nie zawsze oznacza ograniczenia jego funkcji. Może też polegać na eliminowaniu *przypadkowej* złożoności. Moseley i Marks [32] definiują złożoność jako przypadkową, jeśli nie jest nieodłączna od problemu rozwiązywanego przez oprogramowanie (z perspektywy użytkownika), a powstaje tylko z powodu implementacji.

Jedno z najlepszych narzędzi do eliminowania przypadkowej złożoności stanowi *abstrakcja*. Dobra abstrakcja pozwala ukryć wiele szczegółów implementacji za przejrzystą i łatwą do zrozumienia fasadą. Dobrą abstrakcją można też wykorzystać w wielu różnych aplikacjach. Takie ponowne wykorzystanie kodu jest nie tylko wydajniejsze od wielokrotnego pisania podobnego rozwiązania, ale też prowadzi do powstawania oprogramowania wyższej jakości, ponieważ wzrost poziomu wyodrębnionego komponentu przynosi korzyść we wszystkich aplikacjach, którego go używają.

Na przykład wysokopoziomowe języki programowania są abstrakcjami, które ukrywają kod maszynowy, rejestry procesora i wywołania systemowe. SQL to abstrakcja ukrywająca złożone struktury danych używane na dysku i w pamięci, równoległe żądania od innych klientów i niespójności występujące po awariach. Oczywiście w trakcie programowania w języku wysokopoziomowym nadal posługujesz się kodem maszynowym, ale nie robisz tego *bezpośrednio*, ponieważ abstrakcja w postaci języka programowania sprawia, że nie musisz myśleć o takim kodzie.

Jednak znajdowanie wartościowych abstrakcji jest bardzo trudne. Choć w obszarze systemów rozproszonych istnieje wiele dobrych algorytmów, nie ma jasności, jak łączyć je w abstrakcje pomagające utrzymać złożoność systemu na akceptowalnym poziomie.

W tej książce zwracam uwagę na przydatne abstrakcje, które umożliwiają wyodrębnienie fragmentów dużych systemów i przekształcenie ich w dobrze zdefiniowane komponenty wielokrotnego użytku.



## Łatwość modyfikowania — umożliwianie prostego wprowadzania zmian

Jest bardzo mało prawdopodobne, że wymagania stawiane Twojemu systemowi pozostaną na zawsze niezmiennie. Z dużo większym prawdopodobieństwem będą się ciągle zmieniać. Poznasz nowe fakty, pojawią się nieprzewidziane wcześniej przypadki użycia, zmieniają się priorytety biznesowe, użytkownicy będą żądać nowych funkcji, starsze platformy zostaną zastąpione nowymi, zmieniają się wymogi prawne lub ustawowe, rozrastanie się systemu wymusi zmiany architektoniczne itd.

W obszarze procesów organizacyjnych model dostosowywania się do zmian opisano w postaci wzorców *zwinnych* (ang. *agile*). Społeczność skupiona wokół metod zwinnych opracowała też narzędzia techniczne i wzorce pomocne w rozwijaniu oprogramowania w często zmieniającym się środowisku. Te narzędzia to np. programowanie sterowane testami (ang. *test-driven development* — **TDD**) i refaktoryzacja.

Większość technik zwinnych działa w stosunkowo niewielkiej, lokalnej skali (np. kilku plików z kodem źródłowym jednej aplikacji). W tej książce szukam sposobów zwiększenia zwinności na poziomie większego systemu danych, który może obejmować kilka różnych aplikacji lub usług o rozmaitych cechach. Jak np. przeprowadzić „refaktoryzację” architektury tworzenia osi czasu na Twitterze (zob. punkt „Opisywanie obciążenia”), aby przejść od podejścia nr 1 do podejścia nr 2?

Łatwość modyfikowania systemu danych i dostosowywania go do zmieniających się wymagań jest ściśle powiązana z prostotą i abstrakcjami. Proste i łatwe do zrozumienia systemy są zwykle łatwiejsze do modyfikowania niż złożone. Jednak ponieważ jest to tak ważne zagadnienie, do opisu zwinności na poziomie systemu danych używa się innego słowa — *łatwość modyfikowania* (ang. *evolvability*) [34].

## Podsumowanie

W tym rozdziale opisano podstawowe sposoby myślenia o aplikacjach intensywnie przetwarzających dane. Te zasady posłużą za wskazówki w dalszych rozdziałach w trakcie omawiania szczegółów technicznych.

Aplikacja musi spełniać różne wymagania, aby była przydatna. Istnieją *wymagania funkcjonalne* (określające, co aplikacja powinna robić — np. umożliwiać przechowywanie, pobieranie, wyszukiwanie i przetwarzanie danych na różne sposoby), a także *wymagania нефunkcjonalne* (są to ogólne cechy takie jak bezpieczeństwo, niezawodność, zgodność z prawem, skalowalność, kompatybilność i łatwość konserwacji). W tym rozdziale szczegółowo omówiono niezawodność, skalowalność i łatwość konserwacji.

*Niezawodność* oznacza, że system działa prawidłowo nawet po wystąpieniu błędów. Źródłem błędów może być sprzęt (te kłopoty są przeważnie losowe i nieskorelowane), oprogramowanie (wtedy problemy są zwykle powtarzalne i trudno sobie z nimi poradzić) oraz ludzie (którzy, co nieuniknione, od czasu do czasu popełniają pomyłki). Techniki zapewniania odporności na błędy pozwalają ukryć niektóre rodzaje usterek przed użytkownikami końcowymi.

*Skalowalność* oznacza stosowanie strategii utrzymywania wysokiej wydajności nawet w obliczu wzrostu obciążenia. Aby omówić skalowalność, najpierw potrzebne są metody ilościowego opisu obciążenia i wydajności. W ramach przykładu opisywania obciążenia pokrótce przedstawiano os



czasu z Twittera. Jako przykład sposobu pomiaru wydajności posłużyły percentyle czasów odpowiedzi. W skalowalnym systemie można dodać moc obliczeniową, aby system pozostał niezawodny przy wysokim obciążeniu.

*Łatwość konserwacji* ma wiele aspektów, jednak istotę tego zagadnienia stanowi ułatwianie pracy inżynierom i zespołom operacyjnym pracującym z danym systemem. Wartościowe abstrakcje pomagają ograniczyć złożoność i sprawiają, że system łatwiej jest modyfikować i dostosowywać do nowych przypadków użycia. Wysoka jakość operacyjna wymaga dobrego wglądu w stan systemu i skutecznych sposobów zarządzania tym stanem.

Niestety, nie istnieją łatwe metody zapewniania niezawodności, skalowalności lub łatwości konserwacji. Są jednak wzorce i techniki, które często się pojawiają w aplikacjach różnego rodzaju. W kilku następnym rozdziałach przyjrzymy się przykładowym systemom danych i analizie tego, jak osiągać opisane tu cele.

Dalej, w części III, poznasz wzorce tworzenia systemów składających się z kilku współpracujących ze sobą komponentów. Są to systemy podobne do tego z rysunku 1.1.

## **Literatura cytowana**

- [1] Michael Stonebraker i Uğur Çetintemel, „*One Size Fits All*”: *An Idea Whose Time Has Come and Gone*, z: *21st International Conference on Data Engineering (ICDE)*, kwiecień 2005 (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.9136&rep=rep1&type=pdf>).
- [2] Walter L. Heimerdinger i Charles B. Weinstock, *A Conceptual Framework for System Fault Tolerance*, Technical Report CMU/SEI-92-TR-033, Software Engineering Institute, Carnegie Mellon University, październik 1992 (<https://www.sei.cmu.edu/reports/92tr033.pdf>).
- [3] Ding Yuan, Yu Luo, Xin Zhuang i in., *Simple Testing Can Prevent Most Critical Failures: An Analysis of Production Failures in Distributed Data-Intensive Systems*, z: *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, październik 2014 (<https://www.usenix.org/system/files/conference/osdi14/osdi14-paper-yuan.pdf>).
- [4] Yury Izrailevsky i Ariel Tseitlin, *The Netflix Simian Army*, techblog.netflix.com, 19 lipca 2011 (<https://medium.com/netflix-techblog/the-netflix-simian-army-16e57fbab116>).
- [5] Daniel Ford, François Labelle, Florentina I. Popovici i in., *Availability in Globally Distributed Storage Systems*, z: *9th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, październik 2010 (<https://static.googleusercontent.com/media/research.google.com/pl//pubs/archive/36737.pdf>).
- [6] Brian Beach, *Hard Drive Reliability Update — Sep 2014*, backblaze.com, 23 września 2014 (<https://www.backblaze.com/blog/hard-drive-reliability-update-september-2014/>).
- [7] Laurie Voss, *AWS: The Good, the Bad and the Ugly*, blog.awe.sm, 18 grudnia 2012 (<https://web.archive.org/web/20160429075023/http://blog.awe.sm/2012/12/18/aws-the-good-the-bad-and-the-ugly/>).

- [8] Haryadi S. Gunawi, Mingzhe Hao, Tanakorn Leesatapornwongsa i in., *What Bugs Live in the Cloud?*, z: *5th ACM Symposium on Cloud Computing (SoCC)*, listopad 2014 (<http://ucare.cs.uchicago.edu/pdf/socc14-cbs.pdf>; <https://dl.acm.org/citation.cfm?doid=2670979.2670986>).
- [9] Nelson Minar, *Leap Second Crashes Half the Internet*, *somebits.com*, 3 lipca 2012 (<http://www.somebits.com/weblog/tech/bad/leap-second-2012.html>).
- [10] Amazon Web Services, *Summary of the Amazon EC2 and Amazon RDS Service Disruption in the US East Region*, *aws.amazon.com*, 29 kwietnia 2011 (<https://aws.amazon.com/message/65648/>).
- [11] Richard I. Cook, *How Complex Systems Fail*, Cognitive Technologies Laboratory, kwiecień 2000 (<http://web.mit.edu/2.75/resources/random/How%20Complex%20Systems%20Fail.pdf>).
- [12] Jay Kreps, *Getting Real About Distributed System Reliability*, *blog.empathy-box.com*, 19 marca 2012 (<http://blog.empathybox.com/post/19574936361/getting-real-about-distributed-system-reliability>).
- [13] David Oppenheimer, Archana Ganapathi i David A. Patterson, *Why Do Internet Services Fail, and What Can Be Done About It?*, z: *4th USENIX Symposium on Internet Technologies and Systems (USITS)*, marzec 2003 ([http://static.usenix.org/legacy/events/usits03/tech/full\\_papers/oppenheimer/oppenheimer.pdf](http://static.usenix.org/legacy/events/usits03/tech/full_papers/oppenheimer/oppenheimer.pdf)).
- [14] Nathan Marz, *Principles of Software Engineering, Part 1*, *nathanmarz.com*, 2 kwietnia 2013 (<http://nathanmarz.com/blog/principles-of-software-engineering-part-1.html>).
- [15] Michael Jurewitz, *The Human Impact of Bugs*, *jury.me*, 15 marca 2013 (<http://jury.me/blog/2013/3/14/the-human-impact-of-bugs>).
- [16] Raffi Krikorian, *Timelines at Scale*, z: *QCon San Francisco*, listopad 2012.
- [17] Martin Fowler, *Patterns of Enterprise Application Architecture*, Addison Wesley, 2002, ISBN: 978-0-321-12742-6.
- [18] Kelly Sommers, *After all that run around, what caused 500ms disk latency even when we replaced physical server?*, *twitter.com*, 13 listopada 2014 (<https://twitter.com/kellabyte/status/532930540777635840>).
- [19] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani i in., *Dynamo: Amazon's Highly Available Key-Value Store*, z: *21st ACM Symposium on Operating Systems Principles (SOSP)*, październik 2007 (<http://www.allthingsdistributed.com/files/amazon-dynamo-sosp2007.pdf>).
- [20] Greg Linden, *Make Data Useful*, slajdy z prezentacji z kursu eksploracji danych na Uniwersytecie Stanforda (CS345), grudzień 2006 (<http://glinden.blogspot.co.uk/2006/12/slides-from-my-talk-at-stanford.html>).
- [21] Tammy Everts, *The Real Cost of Slow Time vs Downtime*, *webperformancetoday.com*, 12 listopada 2014 (<https://blog.radware.com/applicationdelivery/wpo/2014/11/real-cost-slow-time-vs-downtime-slides/>).

- [22] Jake Brutlag, *Speed Matters for Google Web Search*, [googleresearch.blogspot.co.uk](http://googleresearch.blogspot.co.uk), 22 czerwca 2009 (<https://research.googleblog.com/2009/06/speed-matters.html>).
- [23] Tyler Treat, *Everything You Know About Latency Is Wrong*, [bravenewgeek.com](http://bravenewgeek.com), 12 grudnia 2015 (<http://bravenewgeek.com/everything-you-know-about-latency-is-wrong/>).
- [24] Jeffrey Dean i Luiz André Barroso, *The Tail at Scale*, „Communications of the ACM”, rocznik 56, nr 2, s. 74 – 80, luty 2013 (<https://cacm.acm.org/magazines/2013/2/160173-the-tail-at-scale/abstract>; <https://dl.acm.org/citation.cfm?doid=2408776.2408794>).
- [25] Graham Cormode, Vladislav Shkapenyuk, Divesh Srivastava i Bojian Xu, *Forward Decay: A Practical Time Decay Model for Streaming Systems*, z: *25th IEEE International Conference on Data Engineering (ICDE)*, marzec 2009 (<http://dimacs.rutgers.edu/~graham/pubs/papers/fwdddecay.pdf>).
- [26] Ted Dunning i Otmar Ertl, *Computing Extremely Accurate Quantiles Using t-Digests*, [github.com](https://github.com), marzec 2014 (<https://github.com/tdunning/t-digest>).
- [27] Gil Tene, *HdrHistogram*, [hdrhistogram.org](http://www.hdrhistogram.org) (<http://www.hdrhistogram.org/>).
- [28] Baron Schwartz, *Why Percentiles Don't Work the Way You Think*, [vividcortex.com](http://www.vividcortex.com), 7 grudnia 2015 (<https://www.vividcortex.com/blog/why-percentiles-dont-work-the-way-you-think>).
- [29] James Hamilton, *On Designing and Deploying Internet-Scale Services*, z: *21st Large Installation System Administration Conference (LISA)*, listopad 2007 ([https://www.usenix.org/legacy/events/lisa07/tech/full\\_papers/hamilton/hamilton.pdf](https://www.usenix.org/legacy/events/lisa07/tech/full_papers/hamilton/hamilton.pdf)).
- [30] Brian Foote i Joseph Yoder, *Big Ball of Mud*, z: *4th Conference on Pattern Languages of Programs (PLoP)*, wrzesień 1997 (<http://www.laputan.org/pub/foote/mud.pdf>).
- [31] Frederick P. Brooks, *No Silver Bullet — Essence and Accident in Software Engineering*, z: „The Mythical Man-Month”, wydanie rocznicowe, Addison-Wesley, 1995, ISBN: 978-0-201-83595-3.
- [32] Ben Moseley i Peter Marks, *Out of the Tar Pit*, z: „BCS Software Practice Advancement” (SPA), 2006 (<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.93.8928>).
- [33] Rich Hickey, *Simple Made Easy*, z: „Strange Loop”, wrzesień 2011 (<https://www.infoq.com/presentations/Simple-Made-Easy>).
- [34] Hongyu Pei Breivold, Ivica Crnkovic i Peter J. Eriksson, *Analyzing Software Evolvability*, z: *32nd Annual IEEE International Computer Software and Applications Conference (COMPSAC)*, lipiec 2008 (<http://www.mrtc.mdh.se/publications/1478.pdf>; <http://ieeexplore.ieee.org/document/4591576/>).



## A

ACID, 225  
agregacje zmaterializowane, 108  
agregowanie  
  danych, 108  
  w pamięci, 383  
aktor, 143  
aktualizowanie, 241, 243  
  danych, 136  
  stanu pochodnego, 479  
  widoków zmaterializowanych, 450  
aktualność, 505  
aktualny stan systemu, 442  
aktywność, 303  
  użytkownika, 392  
algorytmy  
  osiągania konsensusu, 355  
  ograniczenia, 357  
  rozwiązywania konfliktów, 478  
  SSI, 258  
analitka, 98  
  prognostyczna, 513  
  schematy, 101  
analiza  
  dziennika, 381  
  przepływów danych, 475  
  strumieni, 449  
  zdarzeń, 392  
Apache Avro, 128  
Apache Thrift, 124  
API, 413, 440  
architektura  
  bez zasobów współdzielonych, 152  
  lambda, 480  
  SIMD, 106

asymetria, 395  
asymetryczne obciążenie robocze, 206  
asynchroniczne wykrywanie konfliktów, 175  
atomowe  
  operacje zapisu, 242  
  zatwierdzanie transakcji, 343, 362  
atomowość, 225, 228  
audyty, 511, 512  
automatyczne  
  rozwiązywanie konfliktów, 178  
  wykrywanie utraconych aktualizacji, 243  
Avro, 128  
awaria, 84, 302, 359, 461  
  koordynatora, 348, 352  
  lidera, 162  
  obserwatora, 162  
awarie częściowe, 272

## B

baza danych DynamoDB, 180  
bazy danych  
  federacyjne, 484  
  kolumnowe, 103  
    agregowanie, 108  
    kompresja kolumn, 105  
    sortowanie, 107  
    zapis, 108  
  MPP, 402, 404  
  OLTP, 101  
  oparte na dokumentach, 49, 51  
  oparte na grafach, 69  
  relacyjne, 51  
  z podziałem na komponenty, 484  
BBS, bulletin board system, 384

- b-drzewo, 89
  - optymalizacja, 91
  - zapewnianie niezawodności, 91
- bezpieczeństwo, 303
- biblioteka
  - Protocol Buffers, 124
  - Thrift, 124
- blokady, 243, 296, 323, 362
  - dwuetapowe, 254, 264
    - implementowanie, 255
    - wydajność, 256
  - oparte na predykcjach, 256
  - zakresów indeksu, 257
- bloki, 89
- błędy, 272, 404, 412, 459
  - bizantyjskie, 298, 302
  - ludzkie, 24
  - programowe, 23, 509
  - sieci, 276
  - sprzętowe, 23
  - typu awaria
    - przerwanie pracy, 302
    - przywrócenie stanu, 302
- brokery komunikatów, 142, 429
  - oparty na dziennikach, 462
  - w modelu AMQP, 462
- brudne
  - odczyty, 234, 263
  - zapisy, 235, 263
- BSP, bulk synchronous parallel, 411

## C

- CEP, complex event processing, 449
- chmura, 273
- CQRS, command query responsibility segregation, 445
- Cypher, 63
- czas
  - wystąpienia zdarzenia, 452
  - zwracania odpowiedzi, 293
- członkostwo, 358

## D

- dane
  - archiwalne, 136
  - jako
    - zasoby, 520
    - źródło władzy, 520
  - pochodne, 375, 475
  - rozproszone, 151
  - typu klucz-wartość, 203
- Datalog, 69
- DDD, domain-driven design, 441
- drzewa zrównoważone, 87
- drzewo LSM, 85, 87, 92
- duplikaty, 499
- dynamiczne określanie typów, 133
- Dynamo, 180
- dysk SSD, 84, 97
- dziennik, 80, 82
  - do przechowywania komunikatów, 432
  - logiczny, 165
  - par klucz-wartość, 83
  - replikacji, 163
  - WAL, 91, 164
  - zapisu z wyprzedzeniem, 91
  - zdarzeń, 442, 445
  - z obsługą komunikatów, 433
- dzierżawy, 362

## E

- eksploatacja, 33, 213
- elastyczność schematu, 52
- eliminowanie entropii, 181
- entropia, 181
- ETL, Extract-Transform-Load, 100
- event sourcing, 441
- ewolucja schematów, 127
  - reguły, 131

## F

- fantomy, 245, 248, 264
- federacyjne bazy danych, 484

filozofia Uniksa, 383  
eksperymenty, 386  
jednolity interfejs, 384  
przezroczystość, 386

format

Avro, 128  
BinaryProtocol, 125  
JSOM, 121  
SSTable, 85  
XML, 121

formaty

przechowywania danych, 80  
kodowania danych, 120

## G

generowanie

danych, 474  
kodu, 133  
numerów porządkowych, 336  
widoków, 445

GFS, Google File System, 387

GiST, Generalized Search Tree, 96

graf, 60, 410

właściwości, 61

zależności przyczynowych, 191

grupowanie, 391

gwarancje

czasu zwracania odpowiedzi, 293  
spójności, 316  
unikatowości, 323  
uporządkowania, 331, 339

## H

Hadoop, 402

haszowanie, 81

spójne, 205

hurtownie danych, 99, 101, 110

## I

idempotencja, 460

operacji, 498

identyfikatory operacji, 500

implementowanie

bazy liniowej, 340

blokad dwuetapowych, 255

dzienników replikacji, 163

izolacji snapshotów, 238

odczytu, 236

przechwytywania zmian w danych, 438

rozgłaszania z uporządkowaniem całkowitym, 342

systemów liniowych, 325

indeksy, 81, 240

dzielone na podstawie pojęć, 217

dzielone według dokumentów, 217

GiST, 96

klastrowe, 95

pokrywające, 95

pomocnicze, 94, 207, 488

rozmyte, 96

wielokolumnowe, 95

wyszukiwania, 399

wyszukiwania pełnotekstowego, 488

z dołączonymi kolumnami, 95

integralność, 505, 509

integrowanie danych, 474

interfejs API, 413, 440

inwigilacja, 517

izolacja, 226, 229

snapshotów, 236, 238, 240

transakcji, 233

## J

jednoczesność, 189

język

Avro IDL, 128

Cypher, 63

Datalog, 69

SPARQL, 66, 68

WSDL, 138

języki

deklaratywne zapytań, 413

z dynamicznym określeniem typów, 133

zapytań o dane, 54

JSOM, 121

## K

klauzula GROUP BY, 394  
klienty stanowe, 493

- przekazywanie zmian stanu, 494
- tryb offline, 493

klucz główny, 94kod aplikacji

- do rozwiązywania konfliktów, 177
- jako funkcja generująca, 488
- oddzielony od stanu, 488
- zmiany stanu, 489

kodowanie danych, 120, 141

- binarne, 122, 124
- CompactProtocol, 126

kolejki, 279kolumny, 105komponenty, 482kompresja

- dziennika, 83, 440
- kolumn, 105
- segmentów, 83

komunikaty, 142, 427

- starsze, 436

konflikty, 177

- automatyczne rozwiązywanie, 178
- niestandardowy kod, 177
- unikanie, 175
- wykrywanie, 175

konsensus, 343, 354–356konserwacja, 33

- aplikacji, 19

konsumenci, 433, 435kontrola współbieżności, 84, 446konwergencja baz

- opartych na dokumentach, 54
- relacyjnych, 54

koordynacja, 358, 508kopie zapasowe, 237korekcja błędów, 275kostka

- danych, 109
- OLAP, 109

koszty liniowości, 327kwora, 182, 326, 356

- niepełne, 185

## L

lambda, 480lider, 158, 296limity czasu, 278liniowe operacje atomowe, 359liniowość, 317, 319, 322, 326

- koszty, 327
- opóźnienia sieciowe, 330

logiczne znaczniki czasu, 478lokalność danych, 53LSM, Log-Structured Merge-Tree, 88, 92

## Ł

łańcuch wywołań, 382łatwość

- eksploatacji, 33
- konserwacji, 33
- modyfikowania, 36

łączenie powiązanych danych, 394

## M

MapReduce, 57, 386, 388, 406

- dane wyjściowe, 399
- przepływ pracy, 390, 398
- rozproszone wykonywanie operacji, 389
- wykonywanie zadań, 388

materializowanie, 410

- konfliktów, 249
- stanu pośredniego, 406

mechanizm hinted-handoff, 185memtable, 87metabaza, 484migracja schematów, 480mikroporcje, 459model

- aktora, 143
- asynchroniczny, 301
- częściowo synchroniczny, 301
- danych, 41
  - RDF, 67
  - typu graf, 60
- MapReduce, 57



- NoSQL, 43
- obiektowy, 43
- oparty na dokumentach, 42
- przetwarzania, 403
- relacyjny, 42, 50
- sieciowy, 49, 69
- synchroniczny, 301
- triplestore, 66
- modyfikowanie, 36
  - aplikacji, 479
- monitorowanie nieaktualnych danych, 185
- MPP, massively parallel processing, 402
- MVCC, 260

## N

- narzędzia unixowe, 380
- niemodyfikowalność, 443
  - ograniczenia, 446
- niezawodność, 22
  - b-drzew, 91
- numerowanie epok, 356

## O

- obciążenie, 26, 30
  - asymetryczne, 206
- obserwator, 158
  - tworzenie, 161
- obserwowanie stanu pochodnego, 491
- obsługa
  - błędów, 232
  - danych wyjściowych, 401
  - komunikatów, 427, 433
  - prze stojów węzłów, 181
  - strumieni zmian, 440
- odciążanie hot spotów, 206
- odczyt, 495
  - brudny, 234
  - fantomów, 264
  - monotoniczny, 169
  - powtarzalny, 236, 241
  - własnych zapisów, 167
  - zatwierdzonych danych, 234
  - ze spójnym przedrostkiem, 170

- zegara, 289
- zniekształcony, 237
- odporność na błędy, 354, 409, 412, 416, 459
- odtworzenie stanu po awarii, 461
- odzworowywanie modeli systemu, 303
- odzyskiwanie
  - pamięci, 294
  - stanu po awarii koordynatora, 352
  - stanu z nadrabianiem zaległości, 162
- ograniczenia
  - niemodyfikowalności, 446
  - transakcji rozproszonych, 353
  - uporządkowania całkowitego, 476
- okno
  - przesuwne, 455
  - rozłączne, 455
  - sesyjne, 455
  - skokowe, 455
- określanie spójności, 184
- OLAP, online analytic processing, 99
- OLTP, online transaction processing, 98
- operacje na obiektach, 228
- opis
  - obciążenia, 26
  - wydajności, 28
- opóźnienia
  - nieograniczone, 278
  - sieciowe, 281, 330
  - zmiennie, 283
  - replikacji, 166, 171
- optymalizacje
  - b-drzew, 91
  - wydajności, 88
- ORM, object-relational mapping, 232
- osiąganie konsensusu, 354, 355

## P

- pamięć
  - podręczna, 488, 492
  - RAM, 97
- para klucz-wartość, 82
- partycje, 201
  - przeciążone, 203
  - strategie równoważenia, 210

perspektywy punktów końcowych, 501  
 pętle sprzężenia zwrotnego, 515  
 platformy z rozproszonymi aktorami, 143  
 pliki  
   samoopisowe, 133  
   SSTable, 85, 86, 87  
 podział, 153  
   asymetryczny, 203  
   baz danych, 482  
   dzienników na partycje, 432  
   indeksów pomocniczych  
     na podstawie dokumentów, 207  
     według pojęć, 208  
   na komponenty, 485, 486, 487  
   na partycje, 201, 253  
     dane typu klucz-wartość, 203  
     dynamiczny, 212  
     indeksy pomocnicze, 207  
     proporcjonalnie do liczby węzłów, 213  
     przedziały kluczy, 203  
     skrótowy kluczy, 204  
     stała liczba, 211  
     według przedziałów kluczy, 217  
     z użyciem skrótów, 217  
   sieci na partycje, 277  
 ponowne  
   dostarczanie, 431  
   przetwarzanie, 479, 498  
 poprawność systemów przepływu danych, 506  
 porządkowanie zdarzeń, 477  
 potwierdzenia, 431  
 powiadomienia o zmianach, 359  
 poziomy izolacji, 233  
 predykat, 256  
 procedury składowane, 251, 252  
 proces ETL, 100  
 procesy wsadowe, 399, 401  
 producenci, 433, 435  
 projektowanie aplikacji, 487  
 Protocol Buffers, 124  
 protokół  
   IP, 275  
   SOAP, 138  
   TCP, 275, 279  
   UDP, 280  
 prywatność, 516, 518  
 przechowywanie  
   dziennika, 82  
   komunikatów, 432  
   wartości w indeksie, 95  
 przechwytywanie zmian w danych, 438  
   implementowanie, 438  
 przeciążenie sieci, 279  
 przekazywanie komunikatów, 142, 451  
 przełączanie awaryjne, 162  
 przenoszenie danych  
   do hurtowni, 100  
   do pamięci, 106  
 przepływ  
   danych, 134, 487, 489  
     przekazywanie komunikatów, 142  
     z użyciem usług, 137  
     z wykorzystaniem baz, 135  
   pracy, 390  
 przestoje węzłów, 161, 181  
 przestrzeganie więzów, 502  
 przesyłanie  
   dziennika WAL, 164  
   komunikatów, 503  
   strumieni zdarzeń, 426  
 przeszukiwanie strumieni, 450  
 przetwarzanie  
   analityczne w czasie rzeczywistym, 99  
   danych  
     ponowne, 479  
     z wielu partycji, 496  
   iteracyjne, 410  
   komunikatów, 350  
   strumieniowe, 425, 447, 478  
     zastosowania, 448  
   transakcji, 98  
     w czasie rzeczywistym, 98  
   w chmurze, 273  
   wektorowe, 106  
   wsadowe, 379, 380, 478  
   złożonych zdarzeń, 449  
   żądań, 504  
 przyczynowość, 331, 477  
 przydział zadań do węzłów, 360  
 przywracanie stanu po awarii, 84

punkty  
kontrolne, 459  
końcowe, 495, 498

## R

RDD, resilient distributed dataset, 409  
referencje do dokumentów, 391  
reguły  
ewolucji schematów, 131  
widoczności, 240  
rekordy częściowo zapisane, 84  
relacja  
„zdarzyło się wcześniej”, 189, 190  
wiele do jednego, 46  
wiele do wielu, 46  
relacyjne bazy typu MPP, 216  
replika, 158  
nadrzędna, 158  
podrzędna, 158  
replikacja, 153, 157  
asynchroniczna, 159, 167  
bez lidera, 180, 194  
oparta na instrukcjach, 163  
oparta na liderze, 158  
oparta na wyzwalaczach, 165  
opóźnienie, 166, 171  
synchroniczna, 159  
z jednym liderem, 194, 356  
z użyciem dziennika logicznego, 165  
z wieloma liderami, 171, 194  
konflikt przy zapisie, 174  
przypadki użycia, 172  
topologie, 178  
REST, 137  
rezerwa dynamiczna, 158  
rodzaje okien, 454  
rozgałęzianie, 430  
rozgłaszanie z uporządkowaniem całkowitym, 339,  
355, 362, 503  
implementowanie, 342  
implementowanie bazy liniowej, 340  
rozproszone  
bazy danych, 402  
systemy plików, 386, 402  
wykonywanie operacji, 389

równoległe wykonywanie zapytań, 216  
równoległość, 189  
równoważenie  
automatyczne, 213  
obciążenia, 430  
partycji, 210  
ręczne, 213  
RPC, 137  
kodowanie danych, 141  
modyfikacje, 141  
wywołania, 139, 140

## S

samoregulacja, 521  
scalanie segmentów, 83–86  
schemat, 133  
generowany dynamicznie, 132  
gwiazdy, 101, 102  
odczytu, 130  
płatka śniegu, 103  
zapisu, 130, 131  
segmenty, 89  
sekwencyjne wykonywanie transakcji, 250, 254  
sekwencyjność, 322  
sieci synchroniczne i asynchroniczne, 281  
sieć semantyczna, 67  
SIMD, single-instruction-multi-data, 106  
skalowalność, 26, 151  
skalowanie odczytów, 166  
składowanie danych  
archiwalnych, 136  
łączenie technologii, 483  
skrótów kluczy, 204  
słabe formy kłamania, 300  
słownik, 81  
snapshot, 236, 240, 439  
SOA, service-oriented architecture, 137  
sortowanie, 107, 383  
przez scalanie, 85, 393  
w bazach kolumnowych, 107  
według klucza, 85  
SPARQL, 66, 68

- spójność, 184, 226, 315
  - odczytu po zapisie, 167
  - odczytu własnych zapisów, 167
  - ostateczna, 166
  - przyczynowa, 334
- SQL
  - zapytania o grafy, 64
- SSI, serializable snapshot isolation, 258, 264
- SSTable, Sorted String Table, 85, 87
- stan, 443
  - pochodny, 491
- strony, 89
- struktura
  - danych, 79
    - b-drzewo, 89
    - drzewo LSM, 92
    - dziennik, 82
    - indeks, 81
    - memtable, 87
    - tablica z haszowaniem, 81
  - grafu, 61
- struktury drzewiaste, 87
- strumienie, 426, 436, 443
  - zdarzeń, 426, 495
- superkomputery, 273
- synchroniczne wykrywanie konfliktów, 175
- synchronizacja
  - systemów, 437
  - zegarów, 285
- system
  - analityczny, 99
  - BBS, 384
  - bez zasobów współużytkowanych, 274
  - Bitcask, 82
  - CEP, 449
  - danych, 20
    - pochodnych, 375
    - unikający koordynacji, 508
  - liniowy, 319
    - implementacja, 325
  - obietnic, 347
  - obliczeniowy o wysokiej wydajności, 273
  - obsługi komunikatów, 427
  - OLAP, 99
  - OLTP, 99, 110
  - plików, 386
    - GFS, 387
    - HDFS, 387
    - QFS, 387
  - Pregel, 411
  - przepływu danych, 407, 506
  - przetwarzania
    - strumieniowego, 380, 451, 490
    - wsadowego, 379
  - RDBMS, 42
  - rozproszony, 271
  - składowania danych
    - oparty na stronach, 89
    - struktura memtable, 87
    - ze strukturą dziennika, 80, 111
  - z podziałem na komponenty, 486
  - zapisu, 375
  - zintegrowany, 486

## Ś

śledzenie, 516

## T

- tabela faktów, 101
- tablica z haszowaniem, 81, 84
- technika MVCC, 260
- technologie składowania danych, 483
- Thrift, 124
- tokeny odgradzające, 297
- topologie replikacji z wieloma liderami, 178
- transakcje, 223
  - anulowania, 232
  - blokady dwuetapowe, 254
  - dla wielu obiektów, 231
  - gwarancje ACID, 225
  - izolacja, 233
  - komercyjne, 98
  - OLTP, 251
  - rozproszone, 343, 349, 475
    - ograniczenia, 353
  - sekwencje rzeczywiste, 251
  - sekwencyjność, 250

- w czasie rzeczywistym, 98
- w procedurach składowanych, 251
- XA, 351
- trasowanie żądań, 214
- trwałość, 227
- twierdzenie CAP, 328
- tworzenie
  - drzew LSM, 87
  - globalnych snapshotów, 290
  - indeksów wyszukiwania, 399
  - indeksu, 483
  - obserwatorów, 161
  - plików SSTable, 87
  - systemów danych, 512
- typy danych, 128

## U

- uczenie maszynowe, 488
- ujednolicanie przetwarzania wsadowego i strumieniowego, 481
- unikatowość, 323, 503
- uporządkowanie, 331
  - całkowite, 333, 339
    - ograniczenia, 476
    - operacji, 359
  - przyczynowe, 333
  - według numerów porządkowych, 335
  - znaczniki czasu, 338
- usługa, 379, 490
  - REST, 137
  - RPC, 137
- usługi
  - internetowe, 138
  - zarządzania członkostwem i koordynacją, 362
  - związane z członkostwem, 361
- uspójnianie przy odczycie, 181
- ustalanie zakończenia okna, 453
- ustawodawstwo, 521
- usuwanie rekordów, 84
- utrata aktualizacji, 241, 264
- utrzymywanie
  - blokad, 352
  - synchronizacji systemów, 437
- uzyskiwanie aktualnego stanu, 442

## W

- WAL, write-ahead log, 91
- wektory wersji, 192
- weryfikowanie, 510
- węzeł, 181
- widoki zmaterializowane, 109, 450, 492
- więzy, 323, 507
  - integralności, 238
  - unikatowości, 362, 502
- wnioskowanie na temat czasu, 451
- wsadowe przepływy pracy, 398
- współbieżność, 84, 446
- wstrzymywanie procesów, 291
- wyбір lidera, 323, 343
- wydajność, 28, 88, 98, 256
  - techniki SSI, 262
- wykonywanie równoległe, 412
- wykorzystanie danych, 518
- wykrywanie
  - awarii, 359
  - błędów, 277
  - jednoczesnych zapisów, 187
  - konfliktów, 175
  - odczytu nieaktualnych wersji, 260
  - usług, 360
  - utraconych aktualizacji, 243
  - zapisów, 261
- wymuszanie przestrzegania więzów, 502
- wyszukiwanie pełnotekstowe, 96
- wywołania RPC, 139, 140, 451
- wyzwalacz, 165
- względność, 189

## X

- XML, 121

## Z

- zadania w modelu MapReduce, 388
- zależności czasowe
  - między kanałami, 324
  - w złączeniach, 458

zapis  
brudny, 235  
jednoczesny, 187  
scalanie, 192  
pojedynczego obiektu, 230  
sekwencyjny, 84  
zniękształcający, 245–248, 264

zapobieganie  
duplikatom, 499  
utracie aktualizacji, 241, 243  
zduplikowanym żądaniom, 500

zapytania  
analityczne, 238  
deklaratywne, 56  
o grafy, 64  
w modelu MapReduce, 57

zarządzanie  
członkostwem i koordynacją, 358  
współbieżnością  
optymistyczne, 259  
pesymistyczne, 259  
złożonością, 35

zasada punktów końcowych, 498, 500, 511

zasoby współdzielone, 152

zatwierdzanie  
atomowe, 460  
w jednym węźle, 344  
w systemie rozproszonym, 344  
dwuetapowe, 344  
trzyetapowe, 349

zdarzenia, 443  
niemodyfikowalne, 444

zegary, 283, 545  
czasu rzeczywistego, 284  
monotoniczne, 284  
synchronizowane, 287, 290  
wektorowe, 193

złączanie  
po stronie mappera, 396  
przez scalanie, 398  
strumieni, 455  
z partycjami i haszowaniem, 397  
z podziałem i haszowaniem, 416  
z rozsyłaniem i haszowaniem, 396, 416  
z sortowaniem przez scalanie, 393, 416

złączenia, 391  
strumień-strumień, 456, 463  
strumień-tabela, 456, 463  
tabela-tabela, 457, 463

złożoność, 35

znaczniki  
czasu, 287  
czasu Lamporta, 337  
pól, 127

## Ż

żądania  
dotyczące wielu partycji, 504  
trasowanie, 214

# PROGRAM PARTNERSKI

GRUPY WYDAWNICZEJ HELION



1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW  
w działający bankomat!

**Dowiedz się więcej i dołącz już dzisiaj!**

<http://program-partnerski.helion.pl>

GRUPA WYDAWNICZA

 **Helion SA**



## Poznaj systemy, w których liczą się dane!

Przetwarzanie i bezpieczne przechowywanie danych absorbuje uwagę inżynierów oprogramowania w coraz większym stopniu. W ostatnich latach pojawiło się wiele bardzo różnych rozwiązań w dziedzinie baz danych, systemów rozproszonych i metodyce budowania aplikacji. Sprzyjają temu zarówno rozwój technologii, rosnące potrzeby dostępu do danych, jak i malejąca tolerancja na przestoje spowodowane awarią czy konserwacją systemu. To wszystko sprawia, że zespoły projektujące aplikacje muszą cały czas aktualizować swoją wiedzę i znakomicie orientować się w zakresie słabych i silnych stron poszczególnych rozwiązań oraz możliwości ich stosowania.

I właśnie ta książka Ci to ułatwi. Dzięki niej będziesz się orientować w świecie szybko zmieniających się technologii przetwarzania i przechowywania danych. Znajdziesz tu przykłady skutecznych systemów spełniających wymogi skalowalności, wydajności i niezawodności. Zapoznasz się z wewnętrznymi mechanizmami tych systemów, analizami najważniejszych algorytmów, omówieniem zasad działania i koniecznymi kompromisami. Przy okazji przyswoisz umiejętność elastycznego myślenia o systemach danych. To zaowocuje intuicyjnością w postrzeganiu tego, jak i dlaczego te systemy działają – a w efekcie pozwoli analizować ich pracę, podejmować trafne decyzje projektowe i wyszukiwać źródła pojawiających się problemów.

**Martin Kleppmann** – bada systemy rozproszone. Pracuje na Uniwersytecie Cambridge w Wielkiej Brytanii. Wcześniej był inżynierem oprogramowania w takich firmach jak LinkedIn czy Rapportive, gdzie pracował nad działającą w dużej skali infrastrukturą do obsługi danych. Jest blogerem, często występuje na konferencjach i rozwija oprogramowanie open source. Wierzy, że ważne idee nauki i techniki powinny być przystępne dla każdego, a pełniejsze ich zrozumienie umożliwi tworzenie lepszego oprogramowania.

W tej książce między innymi:

- co to właściwie znaczy: niezawodność, skalowalność i łatwość konserwacji
- różne modele danych i obsługa zapytań
- replikacja, dzielenie danych, transakcje
- dane pochodne i ich przetwarzanie
- przetwarzanie strumieniowe

<b>Helion</b> ul. Kościuszki 1c 44-100 Gliwice tel.: 32 230 98 63 helion@helion.pl	<i>Sprawdź nasze szkolenia!</i> <b>SZKOLENIA</b>  AKADEMIA IT & BUSINESS <a href="http://WWW.SZKOLENIA.HELION.PL">WWW.SZKOLENIA.HELION.PL</a>	<b>KOD KORZYŚCI</b> Śledź po więcej!  ISBN 978-83-283-4065-7  9 788328 340657
<b>INFORMATYKA W NAJLEPSZYM WYDANIU</b>		Cena: 89,00 zł