

Przedmowa	13
CZĘŚĆ I. Mechanizmy pamięci masowej	19
1. Wprowadzenie i ogólny zarys	25
Architektura DBMS	26
Systemy DBMS oparte na pamięci kontra systemy oparte na dyskach	28
Trwałość w magazynach opartych na pamięci	29
Kolumnowe i wierszowe systemy DBMS	30
Wierszowy układ danych	31
Kolumnowy układ danych	31
Rozróżnienia i optymalizacje	32
Magazyny z szerokimi kolumnami	33
Pliki danych i pliki indeksowe	34
Pliki danych	35
Pliki indeksowe	36
Indeks główny jako pośrednik	37
Buforowanie, niezmiennosc i porządkowanie	38
Podsumowanie	40
2. Podstawy B-drzew	41
Drzewa wyszukiwania binarnego	41
Równoważenie drzewa	42
Drzewa dla pamięci masowych opartych na dyskach	44
Struktury oparte na dyskach	45
Dyski twarde	45
Dyski półprzewodnikowe	45
Struktury na dysku	47

Wszechobecne B-drzewa	48
Hierarchia B-drzewa	50
Klucze oddzielające	51
Złożoność przeszukiwania B-drzewa	52
Algorytm przeszukiwania B-drzewa	52
Liczenie kluczy	53
Dzielenie węzłów B-drzewa	53
Scalanie węzłów B-drzewa	56
Podsumowanie	57
3. Formaty plików	59
Motywacje	60
Kodowanie binarne	60
Typy podstawowe	61
Ciągi znaków i dane o zmiennym rozmiarze	62
Dane upakowane bitowo: wartości logiczne, wyliczenia i flagi	63
Zasady ogólne	64
Struktura strony	65
Strony podzielone na obszary	66
Układ komórek	67
Łączenie komórek w strony podzielone na obszary	69
Zarządzanie danymi o zmiennym rozmiarze	70
Wersjonowanie	71
Sumy kontrolne	72
Podsumowanie	73
4. Implementowanie B-drzew	74
Nagłówki strony	74
Magiczne liczby	74
Powiązania między rodzeństwem	75
Skrajne prawe wskaźniki	76
Najwyższe klucze węzłów	77
Strony przepełnienia	78
Wyszukiwanie binarne	79
Wyszukiwanie binarne ze wskaźnikami kierunku	80
Propagowanie podziałów i scaleń	80
Okrusзки	81
Przywracanie równowagi	82
Dołączanie tylko z prawej strony	83
Ładowanie masowe	84
Kompresja	84

Odkurczanie i konserwacja	86
Fragmentacja spowodowana aktualizacjami i usunięciami	87
Defragmentacja stron	87
Podsumowanie	88
5. Przetwarzanie transakcji i przywracanie poprzedniego stanu	90
Zarządzanie buforami	91
Semantyka buforowania	93
Zwalnianie pamięci podręcznej	94
Blokowanie stron w pamięci podręcznej	95
Zastępowanie stron	96
Przywracanie poprzedniego stanu	99
Semantyka dziennika	100
Działanie a dziennik danych	101
Zasady kradzieży i wymuszania	102
ARIES	103
Kontrola współbieżności	104
Serializowalność	105
Izolacja transakcji	106
Anomalie odczytu i zapisu	106
Poziomy izolacji	107
Optymistyczna kontrola współbieżności	108
Wielowersyjna kontrola współbieżności	109
Pesymistyczna kontrola współbieżności	110
Kontrola współbieżności oparta na blokadach	110
Podsumowanie	117
6. Odmiany B-drzewa	120
Kopiowanie przy zapisie	120
Implementowanie kopiowania przy zapisie: LMDB	121
Abstrakcja aktualizacji węzłów	122
Leniwe B-drzewa	123
WiredTiger	123
Drzewo z leniwą adaptacją	124
Drzewa FD	125
Kaskadowanie ułamkowe	126
Przebiegi logarytmiczne	127
Drzewa Bw	128
Łańcuchy aktualizacji	129
Ograniczanie współbieżności za pomocą porównywania i zamiany	129
Modyfikacje strukturalne	130
Konsolidacja i zbieranie śmieci	131

B-drzewa nieświadome pamięci podręcznej	132
Układ van Emde Boasa	133
Podsumowanie	134
7. Pamięć masowa o strukturze dziennika	136
Drzewa LSM	137
Struktura drzewa LSM	139
Aktualizacje i usuwanie	143
Wyszukiwanie w drzewie LSM	144
Iteracja przez scalanie	144
Uzgadnianie	146
Konserwacja w drzewach LSM	147
Odczyt, zapis i wzmocnienie przestrzenne	149
Hipoteza RUM	150
Szczegóły implementacji	151
Posortowane tabele ciągów	151
Filtry Blooma	152
Lista z przeskokami	154
Dostęp do dysku	156
Kompresja	157
Nieuporządkowana pamięć masowa LSM	158
Bitcask	158
WiscKey	159
Współbieżność w drzewach LSM	161
Układanie dzienników w stos	162
Warstwa translacji pamięci flash	162
Rejestrowanie systemu plików	164
LLAMA i uważne układanie na stosie	165
Dyski SSD z otwartym kanałem	166
Podsumowanie	167
Podsumowanie części I	169

CZĘŚĆ II. Systemy rozproszone 171

8. Wprowadzenie i przegląd	175
Współbieżne wykonywanie	175
Współdzielony stan w systemie rozproszonym	177
Błędy obliczeń rozproszonych	177
Przetwarzanie	179
Zegary i czas	180
Spójność stanu	180

Wykonywanie lokalne i zdalne	181
Potrzeba radzenia sobie z awariami	182
Partycje sieciowe i częściowe awarie	182
Awarie kaskadowe	183
Abstrakcje systemów rozproszonych	184
Łączy	185
Problem dwóch generałów	190
Niemożliwość FLP	191
Synchronizacja systemu	192
Modele awarii	193
Awaria systemu	193
Błędy pominięcia	194
Przypadkowe błędy	194
Radzenie sobie z awariami	195
Podsumowanie	195
9. Wykrywanie awarii	197
Puls i pingi	198
Detektor awarii bez limitu czasu	199
Zewnętrzne sprawdzanie pulsu	200
Detektor awarii Phi-Accural	201
Plotki i wykrywanie awarii	202
Odwracanie problemu wykrywania awarii	203
Podsumowanie	204
10. Wybór lidera	205
Algorytm tyrana	206
Przełączanie awaryjne na następny w kolejności proces	207
Zwykła optymalizacja kandydata	208
Algorytm zapraszania	209
Algorytm pierścieniowy	210
Podsumowanie	211
11. Replikacja i spójność	213
Osiąganie dostępności	214
Niesławny CAP	214
Ostrożne korzystanie z CAP	215
Zbiór i uzysk	216
Pamięć współdzielona	217
Porządkowanie	218

Modele spójności	219
Ścisła spójność	220
Linearyzowalność	220
Spójność sekwencyjna	225
Spójność przyczynowo-skutkowa	226
Modele sesji	230
Ostateczna spójność	231
Dostrajana spójność	231
Repliki świadków	233
Silna ostateczna spójność i typy CRDT	234
Podsumowanie	236
12. Antyentropia i rozpowszechnianie	239
Naprawa odczytu	240
Skrócone odczyty	241
Przekazanie ze wskazówką	242
Drzewa Merkle'a	243
Wektory wersji bitmapowej	244
Rozpowszechnianie plotek	245
Mechanika plotki	246
Sieci nakładkowe	246
Plotki hybrydowe	248
Widoki częściowe	249
Podsumowanie	250
13. Transakcje rozproszone	252
Sprawianie, aby działania wyglądały na niepodzielne	253
Zatwierdzanie dwufazowe	254
Awarie w grupach w 2PC	256
Awarie koordynatora w 2PC	257
Zatwierdzanie trójfazowe	258
Awarie koordynatora w 3PC	259
Transakcje rozproszone z użyciem Calvina	260
Transakcje rozproszone z użyciem Spannera	262
Podział bazy danych na partycje	264
Spójne obliczanie skrótów	265
Transakcje rozproszone z rozprzestrzenianiem	265
Unikanie koordynacji	268
Podsumowanie	270

14. Konsensus	272
Rozgłaszanie	273
Niepodzielne rozgłaszanie	274
Synchroniczność wirtualna	275
Niepodzielne rozgłoszenie Zookeeper (ZAB)	275
Paxos	277
Algorytm Paxos	278
Kworum w Paxosie	280
Scenariusze awarii	281
Multi-Paxos	283
Fast Paxos	284
Egalitarian Paxos	285
Flexible Paxos	287
Uogólnione rozwiązanie konsensusu	289
Raft	291
Rola lidera w algorytmie Raft	293
Scenariusze awarii	294
Konsensus bizantyński	296
Algorytm PBFT	296
Odzyskiwanie i punkty kontrolne	298
Podsumowanie	299
 Podsumowanie części II	 303
 Bibliografia	 307