

УДК 004.655
ББК 32.973.26-018.2
Е81

Ёсу М. Т., Вальдуриес П.

Е81 Принципы организации распределенных баз данных / пер. с англ.
А. А. Слинкина. – М.: ДМК Пресс, 2021. – 672 с.: ил.

ISBN 978-5-97060-391-8

В книге представлено подробное описание распределенных и параллельных баз данных с учетом новейших технологий. Авторы затрагивают такие темы, как проектирование распределенных и параллельных БД, контроль распределенных данных, распределенная обработка запросов и транзакций, интеграция баз данных. Отдельная глава посвящена обработке больших данных (в частности, обсуждаются распределенные системы хранения, потоковая обработка данных, платформы MapReduce и Spark, анализ графов и озера данных). Обработка веб-данных рассматривается с акцентом на технологию RDF, получившую широкое распространение.

В конце глав 2–12 приводятся упражнения, позволяющие закрепить теоретический материал. На сопроводительном сайте читатели найдут информацию об основах реляционных баз данных, обработке запросов, управлении транзакциями и компьютерных сетях. Кроме того, на сайте выложены все рисунки к книге, слайды и решения упражнений (только для преподавателей).

Издание может использоваться в качестве учебника для студентов и магистрантов, изучающих информатику и смежные дисциплины, а также заинтересует всех, кто занимается компьютерными науками.

УДК 004.655
ББК 32.973.26-018.2

Original English language edition published by Springer New York Heidelberg Dordrecht London. Copyright © Springer Science+Business Media New York 2011. Russian language edition copyright © 2021 by DMK Press. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-3-030-26252-5 (англ.)
ISBN 978-5-97060-391-8 (рус.)

© Springer Nature Switzerland AG, 2020
© Оформление, издание, перевод,
ДМК Пресс, 2021

Содержание

Об авторах	15
Предисловие	16
От издательства	19

Глава 1. Введение	20
1.1. Что такое распределенная система баз данных?	21
1.2. История распределенных СУБД	22
1.3. Различные способы доставки данных	24
1.4. Обещания распределенных СУБД	26
1.4.1. Прозрачное управление распределенными и реплицированными данными	27
1.4.2. Обеспечение надежности с помощью распределенных транзакций	29
1.4.3. Повышенная производительность	30
1.4.4. Масштабируемость	32
1.5. Вопросы проектирования	33
1.5.1. Проектирование распределенной базы данных	33
1.5.2. Контроль распределенных данных	34
1.5.3. Распределенная обработка запросов	34
1.5.4. Распределенное управление конкурентностью	34
1.5.5. Надежность распределенной СУБД	35
1.5.6. Репликация	35
1.5.7. Параллельные СУБД	35
1.5.8. Интеграция баз данных	36
1.5.9. Альтернативные подходы к распределению	36
1.5.10. Обработка больших данных и NoSQL	36
1.6. Архитектуры распределенных СУБД	37
1.6.1. Архитектурные модели для распределенных СУБД	37
1.6.1.1. Автономность	37
1.6.1.2. Распределение	39
1.6.1.3. Гетерогенность	39
1.6.2. Клиент-серверные системы	40
1.6.3. Одноранговые системы	42
1.6.4. Системы управления мультибазами данных	45
1.6.5. Облачные вычисления	47
1.7. Библиографические замечания	52

Глава 2. Проектирование распределенных и параллельных баз данных	53
2.1. Фрагментация данных	56
2.1.1. Горизонтальная фрагментация	58

2.1.1.1. Требования к дополнительной информации.....	58
2.1.1.2. Главная горизонтальная фрагментация.....	61
2.1.1.3. Производная горизонтальная фрагментация.....	67
2.1.1.4. Проверка корректности.....	71
2.1.2. Вертикальная фрагментация.....	72
2.1.2.1. Требования к дополнительной информации.....	73
2.1.2.2. Алгоритм кластеризации.....	75
2.1.2.3. Алгоритм расщепления.....	80
2.1.2.4. Проверка корректности.....	83
2.1.3. Гибридная фрагментация.....	83
2.2. Размещение.....	84
2.2.1. Дополнительная информация.....	86
2.2.2. Модель размещения.....	87
2.2.2.1. Полная стоимость.....	87
2.2.2.2. Ограничения.....	89
2.2.3. Методы решения.....	90
2.3. Комбинированные подходы.....	90
2.3.1. Методы секционирования, безразличные к рабочей нагрузке.....	91
2.3.2. Методы секционирования, учитывающие рабочую нагрузку.....	92
2.4. Адаптивные подходы.....	96
2.4.1. Обнаружение изменений рабочей нагрузки.....	97
2.4.2. Обнаружение проблемных участков.....	98
2.4.3. Инкрементная реконфигурация.....	98
2.5. Каталог данных.....	101
2.6. Заключение.....	102
2.7. Библиографические замечания.....	103
Упражнения.....	105

Глава 3. Контроль распределенных данных.....	109
3.1. Управление представлениями.....	110
3.1.1. Представления в централизованных СУБД.....	110
3.1.2. Представления в распределенных СУБД.....	113
3.1.3. Обслуживание материализованных представлений.....	115
3.2. Контроль доступа.....	121
3.2.1. Избирательный контроль доступа.....	122
3.2.2. Мандатный контроль доступа.....	125
3.2.3. Распределенный контроль доступа.....	127
3.3. Контроль семантической целостности.....	129
3.3.1. Централизованный контроль семантической целостности.....	131
3.3.1.1. Спецификация ограничений целостности.....	131
3.3.1.2. Проверка целостности.....	133
3.3.2. Распределенный контроль семантической целостности.....	136
3.3.2.1. Определение распределенных ограничений целостности.....	136
3.3.2.2. Проверка распределенных ограничений целостности.....	139
3.3.2.3. Итоги обсуждения распределенного контроля целостности.....	142
3.4. Заключение.....	143
3.5. Библиографические замечания.....	143
Упражнения.....	145

Глава 4. Распределенная обработка запросов	148
4.1. Общий обзор	149
4.1.1. Задача обработки запроса	149
4.1.2. Оптимизация запроса.....	152
4.1.2.1. Пространство поиска	152
4.1.2.2. Модель стоимости	153
4.1.2.3. Стратегия поиска	154
4.1.3. Уровни обработки запросов	155
4.1.3.1. Декомпозиция запроса	156
4.1.3.2. Локализация данных	157
4.1.3.3. Распределенная оптимизация	158
4.1.3.4. Распределенное выполнение	159
4.2. Локализация данных	160
4.2.1. Редукция для главной горизонтальной фрагментации	160
4.2.1.1. Редукция с помощью выборки.....	161
4.2.2. Редукция с помощью соединения.....	162
4.2.3. Редукция для вертикальной фрагментации	163
4.2.4. Редукция для производной фрагментации.....	165
4.2.5. Редукция для гибридной фрагментации.....	166
4.3. Порядок соединений в распределенных запросах	168
4.3.1. Деревья соединений	169
4.3.2. Порядок соединений.....	170
4.3.3. Алгоритмы на основе полусоединений.....	172
4.3.4. Сравнение соединения и полусоединения	176
4.4. Распределенная модель стоимости	177
4.4.1. Функции стоимости	177
4.4.2. Статистика базы данных	179
4.5. Оптимизация распределенных запросов.....	181
4.5.1. Динамический подход.....	181
4.5.2. Статический подход.....	185
4.5.3. Гибридный подход	188
4.6. Адаптивная обработка запроса.....	193
4.6.1. Процесс адаптивной обработки запросов.....	194
4.6.1.1. Отслеживаемые параметры	194
4.6.1.2. Адаптивные реакции	195
4.6.2. Вихревой подход	196
4.7. Заключение.....	197
4.8. Библиографические замечания	198
Упражнения.....	200
 Глава 5. Распределенная обработка транзакций	 203
5.1. Основные понятия и терминология	205
5.2. Распределенное управление конкурентностью	208
5.2.1. Алгоритмы на основе блокировки.....	209
5.2.1.1. Централизованный алгоритм 2PL	210
5.2.1.2. Распределенный 2PL.....	213

5.2.1.3. Управление распределенными взаимоблокировками.....	214
5.2.2. Алгоритмы на основе временных меток.....	217
5.2.2.1. Базовый алгоритм упорядочения временных меток	218
5.2.2.2. Консервативный УВМ-алгоритм.....	221
5.2.3. Многоверсионное управление конкурентностью	223
5.2.4. Оптимистические алгоритмы	225
5.3. Распределенное управление конкурентностью с помощью изоляции моментальных снимков.....	227
5.4. Надежность распределенных СУБД	230
5.4.1. Протокол двухфазной фиксации.....	232
5.4.2. Варианты 2PC	237
5.4.2.1. Протокол 2PC с предполагаемой отменой	239
5.4.2.2. Протокол 2PC с предполагаемой фиксацией	240
5.4.3. Обработка отказов узлов	241
5.4.3.1. Протоколы завершения и восстановления для 2PC.....	241
5.4.3.2. Протокол трехфазной фиксации	247
5.4.4. Разделение сети	248
5.4.4.1. Централизованные протоколы	251
5.4.4.2. Протоколы на основе голосования	251
5.4.5. Протокол достижения консенсуса Paxos	252
5.4.6. Архитектурные соображения	255
5.5. Современные подходы к горизонтальному масштабированию управления транзакциями.....	257
5.5.1. Spanner	258
5.5.2. LeanXcale.....	259
5.6. Заключение.....	260
5.7. Библиографические замечания.....	263
Упражнения.....	266

Глава 6. Репликация данных..... 270

6.1. Согласованность реплицированных баз данных	272
6.1.1. Взаимная согласованность	272
6.1.2. Взаимная согласованность и согласованность транзакций	274
6.2. Стратегии управления обновлениями	276
6.2.1. Энергичное распространение обновлений	276
6.2.2. Ленивое распространение обновлений.....	277
6.2.3. Централизованные методы.....	278
6.2.4. Распределенные методы	278
6.3. Протоколы репликации	279
6.3.1. Энергичные централизованные протоколы	279
6.3.1.1. Единственный главный узел с ограниченной прозрачностью репликации	280
6.3.1.2. Единственный главный узел с полной прозрачностью репликации	282
6.3.1.3. Ведущая копия с полной прозрачностью репликации	285
6.3.2. Энергичные распределенные протоколы	286
6.3.3. Ленивые централизованные протоколы.....	287

6.3.3.1. Единственный главный узел с ограниченной прозрачностью репликации	287
6.3.3.2. Единственный главный или ведущий узел с полной прозрачностью репликации	289
6.3.4. Ленивые распределенные протоколы	292
6.4. Групповая коммуникация	294
6.5. Репликация и отказы	298
6.5.1. Отказы и ленивая репликация	298
6.5.2. Отказы и энергичная репликация	298
6.6. Заключение.....	302
6.7. Библиографические замечания.....	303
Упражнения.....	304

Глава 7. Интеграция баз данных – системы управления

мультибазами данных.....	307
7.1. Интеграция баз данных	308
7.1.1. Методология проектирования снизу вверх.....	309
7.1.2. Сопоставление схем	313
7.1.2.1. Гетерогенность схем.....	316
7.1.2.2. Подходы на основе лингвистического сопоставления	317
7.1.2.3. Сопоставление на основе ограничений.....	319
7.1.2.4. Сопоставление на основе обучения	321
7.1.2.5. Комбинированные подходы к сопоставлению.....	321
7.1.3. Интеграция схем.....	322
7.1.4. Отображение схем	324
7.1.4.1. Создание отображения	324
7.1.4.2. Обслуживание отображений.....	330
7.1.5. Очистка данных.....	332
7.2. Обработка мультибазовых запросов.....	333
7.2.1. Проблемы обработки мультибазовых запросов.....	334
7.2.2. Архитектура обработки мультибазового запроса	336
7.2.3. Переписывание запросов с помощью представлений	338
7.2.3.1. Терминология Datalog.....	338
7.2.3.2. Переписывание в случае ГКП	339
7.2.3.3. Переписывание в случае ЛКП.....	340
7.2.4. Оптимизация и выполнение запроса	343
7.2.4.1. Моделирование гетерогенной стоимости	343
7.2.4.2. Гетерогенная оптимизация запроса	350
7.2.5. Трансляция и выполнение запроса.....	355
7.3. Заключение.....	358
7.4. Библиографические замечания.....	360
Упражнения.....	363

Глава 8. Параллельные системы баз данных

8.1. Цели	375
8.2. Параллельные архитектуры	378

8.2.1. Общая архитектура	379
8.2.2. Архитектура с общей памятью.....	380
8.2.2.1. Равномерный доступ к памяти (UMA).....	380
8.2.2.2. Неравномерный доступ к памяти (NUMA).....	381
8.2.3. Архитектура с общим диском	383
8.2.4. Архитектура без разделения ресурсов	384
8.3. Размещение данных	385
8.4. Параллельная обработка запросов	388
8.4.1. Параллельные алгоритмы обработки данных	388
8.4.1.1. Параллельные алгоритмы сортировки	389
8.4.1.2. Параллельные алгоритмы соединения.....	390
8.4.2. Оптимизация параллельных запросов.....	396
8.4.2.1. Пространство поиска	396
8.4.2.2. Модель стоимости.....	399
8.4.2.3. Стратегия поиска	400
8.5. Балансировка запроса	400
8.5.1. Проблемы параллельного выполнения	401
8.5.2. Внутриоператорная балансировка нагрузки	403
8.5.3. Межоператорная балансировка нагрузки	405
8.5.4. Внутризапросная балансировка нагрузки	406
8.6. Отказоустойчивость.....	410
8.7. Кластеры баз данных	412
8.7.1. Архитектура кластера баз данных.....	412
8.7.2. Репликация	414
8.7.3. Балансировка нагрузки	414
8.7.4. Обработка запросов	415
8.8. Резюме	418
8.9. Библиографические замечания	419
Упражнения.....	421

Глава 9. Управление данными в одноранговых системах

9.1. Инфраструктура	428
9.1.1. Неструктурированные P2P-сети	429
9.1.2. Структурированные P2P-сети	432
9.1.3. Суперодноранговые P2P-сети	437
9.1.4. Сравнение P2P-сетей.....	438
9.2. Отображение схем в P2P-системах.....	439
9.2.1. Попарное отображение схем.....	439
9.2.2. Отображение на основе методов машинного обучения	440
9.2.3. Отображение на основе общего согласия	441
9.2.4. Отображение схем методами информационного поиска.....	442
9.3. Запросы в P2P-системах	442
9.3.1. Получение первых k результатов.....	442
9.3.1.1. Базовые методы	443
9.3.1.2. Запросы типа «первые k» в неструктурированных системах	450
9.3.1.3. Запросы типа «первые k» в DHT-системах	452
9.3.1.4. Запросы типа «первые k» в суперодноранговых системах	455

9.3.2. Запросы с соединением	455
9.3.3. Запросы по диапазону	457
9.4. Согласованность реплик	460
9.4.1. Базовая поддержка в DHT	460
9.4.2. Актуальность данных в DHT	462
9.4.3. Урегулирование реплик	464
9.4.3.1. OceanStore	464
9.4.3.2. P-Grid	466
9.4.3.3. APPA	466
9.5. Блокчейн	468
9.5.1. Определение блокчейна	469
9.5.2. Инфраструктура блокчейна	471
9.5.2.1. Создание транзакции	471
9.5.2.2. Группировка транзакций в блоки	471
9.5.2.3. Консенсусная проверка блока	473
9.5.3. Блокчейн 2.0	474
9.5.4. Проблемы	475
9.6. Заключение	477
9.7. Библиографические замечания	478
Упражнения	480

Глава 10. Обработка больших данных

10.1. Распределенные системы хранения	485
10.1.1. Google File System	486
10.1.2. Сочетание объектного и файлового хранения	488
10.2. Каркасы для обработки больших данных	489
10.2.1. Обработка данных в MapReduce	490
10.2.1.1. Архитектура MapReduce	492
10.2.1.2. Языки высокого уровня для MapReduce	494
10.2.1.3. Реализация операторов базы данных в MapReduce	495
10.2.2. Обработка данных с помощью Spark	500
10.3. Управление потоковыми данными	505
10.3.1. Потоковые модели, языки и операторы	507
10.3.1.1. Модели данных	507
10.3.1.2. Модели и языки потоковых запросов	509
10.3.1.3. Потоковые операторы и их реализация	509
10.3.2. Обработка запросов к потокам данных	511
10.3.2.1. Выполнение оконного запроса	512
10.3.2.2. Управление нагрузкой	513
10.3.2.3. Обработка не по порядку	514
10.3.2.4. Многозапросная оптимизация	515
10.3.2.5. Параллельная обработка потоков данных	516
10.3.3. Отказоустойчивость СПД	520
10.4. Платформы для анализа графов	521
10.4.1. Разбиение графа	525
10.4.2. MapReduce и анализ графов	530
10.4.3. Специализированные системы анализа графов	531

10.4.4. Ориентированная на вершины пошагово-синхронная модель	534
10.4.5. Ориентированная на вершины асинхронная модель	537
10.4.6. Ориентированная на вершины модель сбора–обработки–распространения.....	540
10.4.7. Ориентированная на разделы пошагово-синхронная модель	541
10.4.8. Ориентированная на разделы асинхронная модель	542
10.4.9. Ориентированная на разделы модель сбора–обработки–распространения.....	543
10.4.10. Ориентированная на ребра пошагово-синхронная модель.....	543
10.4.11. Ориентированная на ребра асинхронная модель.....	544
10.4.12. Ориентированная на ребра модель сбора–обработки–распространения.....	544
10.5. Озера данных	544
10.5.1. Озера данных и хранилища данных	545
10.5.2. Архитектура.....	546
10.5.3. Проблемы.....	548
10.6. Заключение.....	549
10.7. Библиографические замечания.....	550
Упражнения.....	553

Глава 11. NoSQL, NewSQL и полихранилища..... 557

11.1. Причины появления NoSQL	558
11.2. Хранилища ключей и значений	560
11.2.1. DynamoDB	560
11.2.2. Другие хранилища ключей и значений.....	563
11.3. Документные хранилища	563
11.3.1. MongoDB	564
11.3.2. Другие документные хранилища.....	567
11.4. Хранилища с широкими столбцами	568
11.4.1. Bigtable	568
11.4.2. Другие хранилища с широкими столбцами	570
11.5. Графовые СУБД.....	570
11.5.1. Neo4j.....	571
11.5.2. Другие графовые базы данных.....	575
11.6. Гибридные склады данных.....	575
11.6.1. Многомодельные NoSQL-системы.....	575
11.6.2. СУБД типа NewSQL.....	576
11.6.2.1. F1	577
11.6.2.2. LeanXcale.....	578
11.7. Полихранилища.....	580
11.7.1. Слабо связанные полихранилища.....	580
11.7.1.1. BigIntegrator	581
11.7.1.2. Forward	583
11.7.1.3. QoX	584
11.7.2. Сильно связанные полихранилища	585
11.7.2.1. Polybase	586
11.7.2.2. HadoopDB	588

11.7.2.3. Estocada	589
11.7.3. Гибридные системы	590
11.7.3.1. Spark SQL	590
11.7.3.2. CloudMdsQL	592
11.7.3.3. BigDAWG	594
11.7.4. Заключительные замечания	594
11.8. Заключение	595
11.9. Библиографические замечания	597
Упражнения	598

Глава 12. Управление веб-данными 600

12.1. Управление веб-графом	601
12.2. Поиск в вебе	603
12.2.1. Обход веба роботом	604
12.2.2. Индексирование	607
12.2.2.1. Структурный индекс	607
12.2.2.2. Текстовый индекс	607
12.2.3. Ранжирование и анализ ссылок	608
12.2.4. Поиск по ключевым словам	609
12.3. Запросы к вебу	610
12.3.1. Веб как слабо структурированные данные	611
12.3.2. Языки веб-запросов	616
12.4. Вопросно-ответные системы	620
12.5. Поиск и опрос скрытого веба	625
12.5.1. Обход скрытого веба	625
12.5.1.1. Запрос через поисковый интерфейс	625
12.5.1.2. Анализ страниц результатов	626
12.5.2. Метапоиск	627
12.5.2.1. Выделение резюме содержимого	627
12.5.2.2. Категоризация баз данных	628
12.6. Интеграция веб-данных	629
12.6.1. Веб-таблицы и фьюжн-таблицы	630
12.6.2. Семантический веб и проект Linked Open Data	630
12.6.2.1. XML	633
12.6.2.2. RDF	636
12.6.2.3. Навигация и опрос в проекте LOD	647
12.6.3. Вопросы качества данных при интеграции веб-данных	648
12.6.3.1. Очистка структурированных веб-данных	649
12.6.3.2. Слияние веб-данных	651
12.6.3.3. Качество источника веб-данных	652
12.7. Библиографические замечания	655
Упражнения	658

Предметный указатель 660