Оглавление

Предисловие от издательства 9

Отзывы экспертов 10

Рекомендация. Все к лучшему 12

Рекомендация. Путеводитель по будущему управления данными 14

Глубокое изучение основ и теории 14
Систематическое объяснение практических приемов 14
Полное расширение в горизонтальном и вертикальном направлениях 15

Введение 16

Для чего написана эта книга 16
Для кого написана эта книга 17
Ключевые особенности 18
Как читать эту книгу 18
Благодарности 22

ЧАСТЬ 1. ЗНАКОМСТВО С ВЕКТОРНЫМИ БАЗАМИ ДАННЫХ 23

Глава 1. Основы векторных баз данных 24

1.1. Вектор 24
1.1.1. Что такое вектор 24
1.1.2. Все может быть вектором 26
1.1.3. Сходство между векторами 28
1.1.4. Примеры использования сходства 30
1.2. База данных 34
1.2.1. Что такое база данных 34
1.2.2. Реляционные базы данных 37
1.2.3. Нереляционные базы данных 38
1.2.4. Ограничения традиционных баз данных 39
1.3. Зачем нужны векторные базы данных 40
1.3.1. Различия между векторными и традиционными данными 40
1.3.2. Появление векторных баз данных 42
1.3.3. Интеллектуальная платформа хранения в эпоху больших моделей 43
1.4. Резюме 45

Глава 2. Краткая история векторных баз данных 46
2.1. Период зарождения (1980–2012) 46
2.1.1. Стремительное развитие глубоких нейронных сетей 47
2.1.2. Связь глубоких нейронных сетей и векторных баз данных 49
2.2. Период становления (2012–2017) 50
2.3. Период развития (2017 год – настоящее время) 51
2.3.1. Развитие отрасли 52
2.3.2. Сравнение возможностей ключевых продуктов 53
2.3.3. Технологическая архитектура ключевых продуктов 56
2.4. Резюме 60

Глава 3. Основные возможности векторных баз данных 62
3.1. Основные возможности 62
3.1.1. Логические уровни 63
3.1.2. Индексы 69
3.1.3. Ключевые характеристики 73
3.2. Расширенные возможности 74
3.2.1. Динамическая схема 75
3.2.2. Механизм псевдонимов 76
3.2.3. Векторизация 77
3.2.4. Гибридные запросы 78
3.3. Резюме 79

ЧАСТЬ 2. СОЗДАНИЕ ВЕКТОРНОЙ БАЗЫ ДАННЫХ 81

Глава 4. Реализация автономной векторной базы данных 82
4.1. Реализация индекса векторных данных 82
4.1.1. Основные функции библиотеки FAISS 83
4.1.2. Реализация плоского индекса 90
4.1.3. Основные функции библиотеки HNSWLib 98
4.1.4. Реализация HNSW-индекса 107
4.2. Реализация гибридного индекса данных 112
4.2.1. Реализация индекса скалярных данных 112
4.2.2. Единый вход управления 114
4.2.3. Реализация фильтрующего индекса 119
4.3. Реализация восстановления системы после сбоев 128
4.3.1. Персистентность данных на основе журнала 129
4.3.2. Персистентность данных на основе моментальных снимков 135
4.4. Резюме 142

Глава 5. Реализация распределенной векторной базы данных 144
5.1. Управление данными в кластере 146
5.1.1. Знакомство с NuRaft 148

5.1.2. Установление отношения главный–подчиненный 152
5.1.3. Реализация репликации данных 158
5.2. Управление трафиком кластера 162
5.2.1. Управление метаданными кластера 163
5.2.2. Единый входной поток 167
5.2.3. Разделение чтения и записи 171
5.2.4. Обеспечение согласованности чтения и записи 173
5.3 Управление исключениями в кластере 174
5.3.1. Обнаружение нового главного узла 174
5.3.2. Обнаружение отказа подчиненного узла 176
5.3.3. Реализация переключения при отказе 178
5.4. Сегментирование кластера 180
5.4.1. Настройка стратегии сегментирования кластера 182
5.4.2. Перенаправление запросов в соответствии со стратегией сегментирования 184
5.5. Резюме 194

Глава 6. Оптимизация векторной базы данных 197

6.1. Оптимизация производительности 198
6.1.1. Оптимизация векторных вычислений с использованием набора инструкций 198
6.1.2. Оптимизация алгоритмов запросов 200
6.1.3. Оптимизация протокола связи 203
6.1.4. Настройка инструмента для эталонного тестирования 206
6.2. Оптимизация затрат 213
6.2.1. Гибридное развертывание нескольких модулей 214
6.2.2. Развертывание на одном узле 217
6.3. Оптимизация удобства использования 218
6.3.1. Пакет SDK 218
6.3.2. Аутентификация доступа 222
6.3.3. Резервное копирование данных 232
6.4. Резюме 235

ЧАСТЬ 3. ПРАКТИКА И ПЕРСПЕКТИВЫ РАЗВИТИЯ ВЕКТОРНЫХ БАЗ ДАННЫХ 237

Глава 7. Практические примеры использования векторных баз данных 238

7.1. Создание системы поиска по изображениям 238
7.1.1. Векторизация изображений 239
7.1.2. Загрузка и запрос изображений 241
7.1.3. Обзор системы 244
7.2. Создание персональной базы знаний 245
7.2.1. Предобработка знаний 246
7.2.2. Векторизация знаний 247
7.2.3. Управление базой знаний 248

7.2.4. Система вопросов и ответов 250
7.2.5. Обзор системы 252
7.3. Резюме 253

Глава 8. Перспективы развития 255
8.1. С точки зрения эволюции отрасли 256
8.1.1. Новая парадигма управления данными человеком 256
8.1.2. Векторные данные устраняют различия в форматах данных 259
8.1.3. Ключевые аспекты платформизации векторной базы данных 260
8.2. С точки зрения применения в отрасли 262
8.2.2. Снижение порога использования RAG 264
8.3. Резюме 266

Предисловие от издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в основном тексте или программном коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу dmkpress@gmail.com, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Отзывы экспертов

В эпоху стремительного развития искусственного интеллекта информационно-коммуникационная сфера глубоко интегрируется с интеллектуальными технологиями. Можно с уверенностью сказать, что векторные базы данных будут играть ключевую роль в обработке данных в этой области. Книга «Векторные базы данных» является плодом творчества Ло Юня и его команды специалистов в области ИИ. В ней не только подробно раскрываются принципы работы векторных баз данных, но и предлагаются богатые примеры, практические приемы и ценные идеи, которые будут полезны исследователям и новаторам в области интеллектуальных технологий.

– Ван Цзяндань (Wang Jiangdan),
член Китайской академии инженерных наук

Ло Юнь – один из первых специалистов в области облачных вычислений и опытный эксперт в области баз данных, сетей и распределенных систем. Эта книга является практическим руководством, в котором доступным языком объясняется, как создать высокопроизводительную векторную базу данных, и я настоятельно рекомендую ее к прочтению.

– Лю Ин (Liu Ying),
вице-президент Tencent Cloud

Развитие ИИ-технологий требует унифицированного представления мультимодальных данных и управления ими. Решение этой задачи привело к появлению новой технологии – векторных баз данных. Книга в ясной и краткой форме знакомит с их основными понятиями, начиная с самых основ и постепенно углубляясь в детали, с акцентом на практическое применение. Это отличное справочное пособие для изучения векторных баз данных!

– Ду Сяоюн (Du Xiaoyong),
профессор Института информатики Китайского народного университета,
директор Ведущей лаборатории по инженерии данных и знаний
Министерства образования КНР

Книга «Векторные базы данных» содержит многолетний опыт Ло Юня и команды Tencent Cloud Database, накопленный при обслуживании группы компаний Tencent и ее внешних клиентов. Содержание книги легко усваивается и идеально подходит для технических специалистов, интересующихся технологиями векторных баз данных.

– Ли Голян (Li Guoliang),
профессор Университета Цинхуа,
IEEE Fellow

В условиях «демократизации ИИ-технологий» векторные базы данных быстро становятся основой интеллектуальных приложений. Эта книга призвана устранить недостаток подробной документации и содержит доступные объяснения и практические руководства. Рекомендуется к прочтению специалистам в области баз данных и искусственного интеллекта.

– Лю Чжиюань (Liu Zhiyuan),
доцент Университета Цинхуа

Автор, обладая глубокими академическими знаниями и богатым практическим опытом, подробно излагает сложные концепции технологии векторных баз данных и ведет читателей от самых базовых сведений к созданию полноценной векторной базы данных. Книга содержит как теорию векторных технологий, так и практический опыт в распределенных базах данных, а также описывает соответствующие области применения, что делает ее полезной как для профессионалов в области баз данных, так и для тех, кто интересуется ИИ-технологиями.

– Ян Чэнху (Yang Chenghu),
соучредитель и технический директор Beijing Fengqing Technology

Рекомендация. Все к лучшему

Гай Гоцян (Gai Guoqiang),
основатель Enmotech,
обладатель титула MVP (самый ценный эксперт) в Kunpeng

Ло Юнь в предисловии выразил особую благодарность своему редактору: «Именно с вашего письма установились наши взаимоотношения». Эта фраза вернула меня в 2004 год, когда я только что прибыл в Пекин. Именно благодаря одному письму я установил взаимоотношения с моим редактором и издательством связи и телекоммуникаций – с тех пор нас связывает двадцатилетняя дружба.

Наши отношения с технологиями часто начинаются случайно, но развиваются только благодаря упорству.

Автор и его команда многие годы исследовали задачи извлечения векторов, движимые потребностями в вычислениях над неструктурированными данными. Они создавали и постоянно совершенствовали инновационные продукты, обслуживая огромное количество пользователей и накапливая ценный опыт первопроходцев. Впоследствии с появлением больших языковых моделей роль векторных баз данных стала очевидна всему миру.

Я полагаю, что достижения в любом технологическом направлении происходят именно так. Благодаря упорной работе и удовлетворению определенных потребностей векторные базы данных достигли момента, когда они смогли проявить себя и произвести большое впечатление.

В июле 2023 года на презентации VectorDB компании Tencent Cloud я стал свидетелем глубокого понимания и проницательности Ло Юня в отношении векторных баз данных. В контексте обслуживания широкого набора индивидуальных сценариев у Tencent появилась возможность удовлетворять самые разнообразные потребности в векторных вычислениях, и именно поэтому их векторная база данных демонстрирует выдающуюся техническую конкурентоспособность.

Теория без практики остается неполной, истинное понимание приходит только через личный опыт. Автор утверждает, что это книга ориентирована на практику и требует активного участия. Только через практическое применение и личный опыт можно понять суть технологии. Самое важное при чтении книги – это понять суть мыслей и опыта автора. Я считаю, что «практика и участие» – это основное знание, которое автор стремится передать.

Конечно, первые две и последняя главы книги взаимосвязаны по содержанию, и их можно рассматривать как научно-популярный материал, который можно прочитать без практических упражнений и получить базовое представление о векторных базах данных. Эти три главы охватывают прошлое, настоящее и будущее векторных баз данных, и в них повсюду виден личный опыт и глубокие размышления автора.

При написании своей новой книги я обращался к автору за советом по разделу о векторных базах данных и получил много полезной информации. С публикацией его собственной работы заполнился пробел в литературе о векторных базах данных, что является важным вкладом в отрасль. Я уверен, что все читатели смогут извлечь пользу из этой книги.

Чтение этой книги имеет и другую практическую значимость: мы можем узнать, как различные продукты Tencent, которые мы используем ежедневно, способны быстро и точно отвечать, эффективно и интеллектуально искать изображения, рекомендовать персонализированный контент и многое другое. За всем этим стоят высокоскоростные векторные базы данных.

Все можно представить в виде векторов, и чтение книги обязательно принесет вам необходимые для этого знания.

Рекомендация. Путеводитель по будущему управления данными

Ван Хаофэн (Wang Haofen),
приглашенный исследователь Университета Тунцзи,
инициатор создания OpenKG (Китайское сообщество открытых знаний)

В эпоху стремительного развития искусственного интеллекта и больших языковых моделей векторные базы данных стали основой новой инфраструктуры. Их важность сопоставима с реляционными базами данных эпохи сетей персональных компьютеров и большими данными эпохи мобильного интернета. Поэтому существует потребность в систематичном учебном пособии, подходящем как для специалистов смежных областей, которые хотят понять суть векторных баз данных, так и для профессионалов отрасли, которые стремятся овладеть этой технологией. В условиях стремительного развития технологий и сильного желания обучающихся получить соответствующие знания я рад стать свидетелем выхода в свет книги «Векторные базы данных».

Эта книга написана Ло Юнем, руководителем направления векторных баз данных в Tencent Cloud, обладающим богатым практическим опытом. Она всесторонне и подробно освещает построение векторных баз данных и их практическое применение в различных сценариях. Целью книги является помощь читателям в овладении теоретическими знаниями и в создании распределенной векторной базы данных с нуля. Кроме того, книга демонстрирует широкие перспективы и мощный потенциал векторных баз данных. При внимательном рассмотрении можно выделить три основные особенности этой книги.

Глубокое изучение основ и теории

Книга начинается с базовых знаний с подробным изложением концепции, принципов и архитектуры векторных баз данных. Ло Юнь простым и понятным языком с использованием доступных аналогий объясняет ключевые идеи и технические аспекты векторных баз данных. Также он углубленно рассматривает такие важные технологии, как векторизация, запросы на сходство и векторные индексы, помогая читателям сформировать полную систему знаний и заложить прочную теоретическую основу для последующей практической работы.

Систематическое объяснение практических приемов

Будучи специалистом в области искусственного интеллекта, я глубоко осознаю важность выбора подходящей векторной базы данных для успеха проекта, особенно в процессе построения генерации, дополненной поиском. Благодаря

систематическому изложению читатели могут полностью освоить процесс построения векторной базы данных, включая импорт данных, построение индексов, оптимизацию запросов и настройку производительности. Ло Юнь делится множеством базовых деталей и практических советов из собственного опыта, что позволяет читателям не только узнать, как что-то делается, но и почему это делается именно так. Это бесценно для любого специалиста, желающего применять векторные базы данных в реальных проектах.

Полное расширение в горизонтальном и вертикальном направлениях

В эпоху искусственного интеллекта векторные базы данных не ограничиваются одним приложением. Их можно расширять горизонтально, добавляя поддержку различных модальностей, включая текст, изображения, аудио и видео, что позволяет гибко переносить их в разные сценарии. Книга объясняет, как построить расширяемую систему векторных баз данных, способную гибко адаптироваться к различным типам данных и приложениям. Также можно расширять в вертикальном направлении: в книге рассматривается, как интегрировать различные универсальные или специализированные большие языковые модели и бесшовно соединять их с производственными системами и системами больших данных, реализуя интеграцию исследований и эксплуатации.

«Векторные базы данных» – это техническая книга, сосредоточенная в равной степени на теории и практике, а также важное руководство, раскрывающее направление будущего управления данными. Искренне рекомендую эту книгу всем, кто интересуется векторными базами данных, и надеюсь, что она принесет вдохновение и реальную помощь в вашем обучении и работе.

Введение

2023 год называют «годом демократизации ИИ». За последние год-два мы вступили в новую эпоху стремительного роста технологий искусственного интеллекта. Непрерывно появляются новые термины, концепции и технологии, такие как трансформер, GPT, большие языковые модели, ИИ-агент, цепочка мышления, векторные базы данных и др. Мы потратили немало времени на изучение и освоение этих новых технологий. В этот период мы заметили частое упоминание некоторых традиционных математических концепций, таких как векторы, матрицы и тензоры. Особенно часто обсуждаются векторы.

Следует упомянуть, что вектор – это не только традиционное математическое понятие. С помощью моделей глубокого обучения можно преобразовывать различные данные, созданные человеческим обществом, в данные, представленные векторами. Это можно выразить фразой «все можно сделать вектором». Данные, представленные векторами, называются «векторными данными». На основе информации, содержащейся в векторных данных, ИИ-технологии могут легко понимать различные данные человеческого общества. Таким образом, векторные данные постепенно становятся основными данными при внедрении интеллектуальных технологий.

Поскольку векторные данные приобретают все большее значение, эффективное управление ими становится все более важным. За десятилетия своего развития технологии баз данных стали важной ветвью компьютерных наук, специально предназначенной для эффективного управления данными. Однако в силу различий подходов традиционные технологии баз данных не способны справиться с современными вызовами, характеризующимися высокой размерностью, высокой точностью и большими масштабами. Векторные базы данных были специально разработаны с учетом особенностей векторных данных, что позволяет им лучше хранить, индексировать и выполнять запросы таких данных. Эта технология появилась относительно недавно, но в ней уже появилось много деталей, которые стоит изучить и обсудить.

Для чего написана эта книга

Будучи профессионалом с более чем десятилетним опытом работы в компьютерной индустрии, я глубоко понимаю силу «обучения через практику». В период обучения в аспирантуре я получил «инсайты» во время работы с исходным кодом из классических книг, и эти моменты до сих пор остаются в моей памяти. К тому времени я уже мог решать множество задач по алгоритмам и программированию, но мое понимание механизмов работы компьютерных систем оставалось поверхностным. В частности, я понимал лишь на теоретическом уровне, как операционная система управляет процессами, распределяет ресурсы памяти и выполняет операции ввода-вывода. В итоге благодаря углубленному изучению и практической работе с богатым набором исходного

кода из книг «Ядро Linux» авторов Д. Бовет и М. Чезати и «Understanding Linux Network Internals: Guided Tour to Networking on Linux» автора Ch. Benvenuti я осознал, что в компьютерной области нет ничего мистического – это, по сути, код, который управляет нулями и единицами.

С тех пор я выработал привычку глубже разбирать новые технологии через практику. Практическая работа с кодом, стоящим за технологией, позволяет изучить ее более основательно, чем просто понять основные концепции. Для многих начинающих специалистов базы данных кажутся сложными, и они считают, что за ними скрываются «космические технологии». На самом деле базы данных – это технология, основанная на операционных системах и теории распределенных систем и реализующая сложное управление данными через программное обеспечение. За годы развития появилось множество теоретических и практических книг, которые могут служить справочником для новичков по этой теме.

Что касается новой технологии векторных баз данных, на рынке пока нет книги, которая бы направляла читателя в создании векторной базы данных с нуля. К счастью, я и моя команда работаем в этой области уже много лет, обслуживая многочисленных клиентов и накапливая ценный опыт первопроходцев этого направления. Основываясь на этом опыте, я надеюсь начать с самого простого кода программы «Hello World!» и провести вас через процесс написания собственной векторной базы данных на уровне исходного кода. Я верю, что благодаря этому процессу «высокотехнологичная» векторная база данных перестанет казаться такой загадочной.

Я понимаю, что из-за ограничений моего личного понимания и навыков писателя эта книга не может охватить все темы и аспекты и не гарантирует абсолютной безошибочности. Но если эта книга поможет вам лучше понять эту технологию и создать собственную упрощенную версию векторной базы данных, то моя цель будет достигнута.

Для кого написана эта книга

Эта книга является практическим руководством, в котором также объясняются базовые принципы. Технические аспекты в книге будут поняты начинающими программистами. Тем не менее для эффективного чтения необходимо знать основы программирования, так как в книге представлено немало исходного кода, с которым вам предстоит работать.

- Если вас интересуют векторные базы данных и вы хотите глубже понять процесс их создания на уровне исходного кода, эта книга научит вас создавать распределенные векторные базы данных с нуля. В книге рассматриваются такие темы, как создание системы индексов в автономной подсистеме базы данных, повышение способности системы к восстановлению после сбоя, а также реализация распределенной и кластерной работы, включая репликацию данных, управление трафиком и метаданными.
- Если вы интересуетесь базами данных и хотите глубже понять процесс их создания на уровне исходного кода, эта книга также будет полезна для вас – полный процесс создания распределенной векторной базы данных охватывает ключевые знания в этой области.

- Если вас интересует разработка ИИ-приложений и вы хотите понять, как создавать и управлять векторными данными, лежащими в их основе, книга расскажет о связи векторных данных с большими языковыми моделями и проведет вас через весь процесс выполнения запросов к векторным базам данных. Это поможет вам более эффективно интегрировать векторные базы данных для оптимизации ИИ-приложений, обновления знаний и более эффективного решения задач, возникающих при внедрении таких приложений.
- Если вы являетесь экспертом в разработке ИИ-приложений или в области баз данных и хотели бы внести вклад в улучшение этой книги и развитие отрасли в целом, эта книга также заслуживает вашего внимания. Чтение книги может вдохновить вас на более ценные размышления. Векторные базы данных – это относительно новая область, и обмен информацией, безусловно, будет способствовать ее прогрессу.

Ключевые особенности

Эта книга исповедует «практический подход». Как упоминалось ранее, этот принцип принес мне много пользы, поэтому я надеюсь через эту книгу поделиться этим методом с вами. Книга не нацелена на глубокое изучение принципов ИИ-технологий, а больше сосредоточена на векторных базах данных, лежащих в их основе. Направляя вас шаг за шагом к созданию собственной векторной базы данных, книга постепенно приводит к пониманию, как работает векторная база данных. Конечно, перед практическими занятиями будут представлены некоторые максимально краткие базовые знания с той технической глубиной, которая необходима для специалистов по базам данных.

Основная часть книги (главы 4–7) состоит преимущественно из фрагментов кода (с подробными комментариями) и анализа принципов его реализации. Весь код написан мной с нуля и охватывает темы от самых простых плоских индексов в векторной базе данных до несколько более сложных HNSW-индексов. В этом процессе я буду использовать существующий открытый исходный код для создания собственного сервиса. Это эффективная практика разработки программного обеспечения, ведь нам не нужно каждый раз изобретать велосипед.

Для еще более эффективного применения этого метода обучения я от начала до конца проведу вас через процесс создания ИИ-приложения на базе созданной ранее векторной базы данных, реализуя принцип «догфудинг». Представьте, что все – от веб-страницы до базового уровня хранения векторов, индексации и выполнения запросов – создано вашими руками. Вы, безусловно, испытаете огромное чувство удовлетворения, и с помощью этой книги вы достигнете такого уровня мастерства.

Как читать эту книгу

Структура

Книга разделена на три основные части, содержание которых организовано от простого к сложному, от основ к практике. Можно читать книгу по порядку

глав, но, если вы уже знакомы с материалом какой-то части, можно пропустить ее. Если вы хотите с нуля создать векторную базу данных, я настоятельно рекомендую читать по порядку.

Первая часть. Знакомство с векторными базами данных (главы 1–3)

Эта вводная часть книги призвана помочь заложить необходимые теоретические основы.

- Глава 1. Введение в базовые знания о векторах и векторных базах данных, определение основных понятий: вектор, база данных, векторная база данных.
- Глава 2. Сведения о развитии векторных баз данных как новой технологии, включая ключевые достижения, рост связанных компаний и стартапов, а также наиболее распространенные архитектурные модели в индустрии. Эта глава поможет понять эволюцию технологий на макроуровне.
- Глава 3. На примере Tencent Cloud VectorDB мы исследуем реализацию основных функций векторных баз данных. Если вы планируете самостоятельно разрабатывать или углубленно использовать векторные базы данных, эта глава познакомит вас с базовыми знаниями.

Вторая часть. Создание векторной базы данных (главы 4–6)

Эта часть является основной и подробно описывает, как создать с нуля и оптимизировать векторную базу данных. Здесь обсуждаются различные технические детали и приводятся примеры.

- Глава 4. Создание автономной векторной базы данных с нуля. Глава основана на концепции модульной архитектуры. Здесь используются некоторые зрелые открытые компоненты, постепенно добавляются общие программные и базовые технологии, применяются технологии персистентности и восстановления после сбоя из индустрии баз данных. В итоге демонстрируется процесс создания автономной векторной базы данных, которая пока мала, но уже полноценна и сопоставима с прототипами некоторых открытых автономных векторных баз данных.
- Глава 5. На основе автономной векторной базы данных продолжается разработка распределенной векторной базы данных, включая эффективное управление метаданными, управление трафиком и обработку исключений. Начиная с этой главы, наша векторная база данных будет обладать прототипным уровнем поддержки управления векторными данными определенного масштаба. Если вы хотите понять механизмы распределенных систем на низком уровне и узнать, как расширить автономную систему до распределенной, не пропускайте эту главу.
- Глава 6. На основе распределенной векторной базы данных, построенной в предыдущих двух главах, демонстрируются приемы оптимизации, включая производительность, издержки и удобство использования. В части оптимизации производительности будет рассмотрено, как использо-

вать вычислительные возможности CPU/GPU для повышения производительности. В части оптимизации издержек будет уделено внимание упрощению модели развертывания бизнеса для снижения операционных затрат. В заключение обсуждается, как сделать продукт векторной базы данных более удобным для разработчиков, например, предоставляя хороший пакет SDK и механизмы резервного копирования данных, что крайне важно для продвижения новой технологии.

Третья часть. Практика и перспективы векторных баз данных (главы 7–8)

Эта часть является заключительной и сосредоточена на практике применения и перспективах. Изучив эту часть, вы сможете не только соединить теоретические знания из книги с практическими навыками, но и заглянуть в будущее технологий, что поможет вам найти направление для дальнейшего обучения и планирования карьеры.

- Глава 7. В этой главе на практике проверяются методы использования разработанной нами векторной базы данных в ИИ-приложениях. Мы реализуем упомянутые в главе 1 сценарии поиска по изображению и базы знаний, которые в последние годы широко используются в ИИ-приложениях.
- Глава 8. Эта глава вернет нас к истории развития компьютерной индустрии, чтобы исследовать будущее векторных баз данных. С одной стороны, здесь обсуждается, почему векторные базы данных могут стать платформенной технологией. С другой стороны, на основе текущего применения векторных баз данных в сценариях генерации, дополненной поиском, развивается обсуждение будущего развития векторных баз данных. В этой главе я бы хотел вместе с вами поразмышлять и спрогнозировать долгосрочные тенденции векторных баз данных.

Ресурсы и соглашения

Весь исходный код можно скачать на домашней странице книги в сообществе Turing (www.ituring.com.cn/book/3305). Код включает десятки тысяч строк кода и позволяет сочетать чтение с практическими упражнениями.

1. Иерархические ментальные карты

В некоторых главах представлены иерархические ментальные карты, которые призваны помочь вам быстро освоить основные функции и состав модулей векторной базы данных. Перед чтением материала я настоятельно рекомендую ознакомиться с соответствующей ментальной картой, чтобы получить общее представление.

2. Итерации и обновления версий

В этой книге мы вместе создадим векторную базу данных, которая пройдет через несколько итераций и обновлений версий от v0.0.1 до v0.6. Переход от версии v0.0.1 к версии v0.0.2 сопровождается незначительными изменениями и в книге называется итерацией версии. А переход от v0.0.2 к v0.1, который вклю-

Конец ознакомительного фрагмента.
Приобрести книгу можно
в интернет-магазине
«Электронный универс»
e-Univers.ru