Посвящается нашим студентам

Содержание

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, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

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 электронной почты dmkpress@gmail.com.

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

Предисловие

В настоящей книге рассматриваются и обсуждаются модели, системы и фреймворки программирования, специально сконструированные для обработки и анализа крупных наборов данных. В частности, в ней дается подробное описание свойств и механизмов главных парадигм программирования, используемых в анализе больших данных, таких как модели на основе MapReduce, рабочих потоков, массового синхронного параллелизма, передачи сообщений и SQL-подобные. Более того, в главах книги на примерах программирования описаны наиболее часто используемые фреймворки, такие как Hadoop, Spark, Storm и MPI, специально сконструированные для анализа крупных коллекций данных.

Мировая путина, интернет вещей и платформы социальных сетей обеспечивают условия для порождения и сбора огромных объемов цифровых данных, поступающих из самых разных источников, включая блоги, датчики, мобильные устройства, носимые трекеры, спутники и камеры наблюдения. Эти данные, общепринято именуемые «большими данными», бросают вызов существующим системам и способностями по хранению, обработке и анализу. По этой причине в настоящее время изучаются, конструируются, разрабатываются и внедряются новые модели, языки, инструменты, системы и алгоритмы, способные эффективно собирать, хранить и анализировать большие данные, а также усваивать из них полезную информацию.

В этой книге описывается и приводится обзор параллельных и распределенных парадигм, языков и систем, используемых сегодня для анализа больших данных и усвоения из них действенной информации на масштабируемых компьютерах. В частности, в ней дается подробное описание свойств и механизмов главных парадигм параллельного программирования, а на примерах программирования иллюстрируются наиболее широко используемые фреймворки, предназначенные для анализа больших данных. Более того, в книге обсуждаются и сравниваются разные фреймворки, выделяются главные характеристики каждого из них, их распространенность (сообщество разработчиков и пользователей), а также главные преимущества и недостатки их использования в реализации приложений по анализу больших данных. Конечная цель данного тома – помочь конструкторам и разработчикам приложений по обработке больших данных, определив и выбрав наилучший или наиболее подходящий инструмент(ы) программирования в зависимости от их навыков, наличия оборудования, областей применения и целей, а также учитывая поддержку, предоставляемую сообществом разработчиков. По каждому языку программирования/фреймворку представлены реально-практические примеры программирования, демонстрирующие способы разработки и реализации приложений по обработке больших данных.

Об авторах

Доменико Талия – профессор кафедры Компьютерной инженерии в Университете Калабрии, Италия, и почетный профессор Университета Амити, Индия. Является старшим ассоциированным редактором журнала ACM Computing Surveys, ассоциированным редактором журнала Computer, а также членом Редакционного совета журналов Future Generation Computer Systems, IEEE Transactions on Parallel and Distributed Systems, International Journal of Web and Grid Services, Journal of Cloud Computing, Big Data and Cognitive Computing и International Journal of Next-Generation Computing. В сферу его научных интересов входят высокопроизводительные вычисления, большие данные, машинное обучение, параллельный и распределенный анализ данных, облачные вычисления, анализ социальных сетей, распределенное обнаружение знаний, одноранговые системы и конкурентные модели программирования. Является автором нескольких книг и более 400 научных статей.

Паоло Трунфио – доцент кафедры Компьютерной инженерии в Университете Калабрии, Италия. В 2007 году работал приглашенным исследователем в шведском Институте вычислительных наук (SICS) в Стокгольме, Швеция. В настоящее время является ассоциированным редактором журналов Journal of Big Data, IEEE Transactions on Cloud Computing и ACM Computing Surveys и членом редколлегий нескольких научных журналов, включая Future Generation Computer Systems, Big Data and Cognitive Computing, International Journal of Web Information Systems и International Journal of Parallel, Emergent and Distributed Systems. В сферу его научных интересов входят облачные вычисления, большие данные, анализ социальных сетей, параллельное и распределенное обнаружение знаний и одноранговые системы.

Фабрицио Мароззо – доцент кафедры Компьютерной инженерии в Университете Калабрии, Италия. Получил степень доктора философии в области системной и компьютерной инженерии в Университете Калабрии. В 2011–2012 годах проходил стажировку в Барселонском суперкомпьютерном центре в Исследовательской группе по решеточным вычислениям на факультете Вычислительных наук. Входит в состав редакционных советов нескольких журналов, включая IEEE Access, IEEE Transactions on Big Data, Journal of Big Data, Big Data and Cognitive Computing, Algorithms, Frontiers in Big Data, Heliyon и SN Computer Science. В сферу его научных интересов входят анализ больших данных, анализ социальных сетей, высокопроизводительные вычисления, облачные и периферийные вычисления, а также машинное обучение.

Лорис Белькастро – исследователь в области компьютерной инженерии в Университете Калабрии, Италия. Получил степень доктора философии в области информационно-коммуникационной инженерии в Университете Калабрии. В 2012 году стал стипендиатом в Институте высокопроизводительных вычислений и сетей Национального исследовательского совета Италии (ICAR-CNR). Выступал в качестве приглашенного редактора в многочисленных журналах, включая Future Generation Computer Systems, Journal of Big Data, Sensors, Algorithms, Applied Sciences и Frontiers in Big Data. В сферу его научных интересов входят облачные и периферийные вычисления, большие данные, анализ социальных сетей, параллельный и распределенный анализ данных.

Риккардо Кантини – исследователь в области компьютерной инженерии в Университете Калабрии, Италия. В том же университете получил степень доктора философии в области информационных и коммуникационных технологий. В 2021–2022 годах состоял приглашенным исследователем в Барселонском суперкомпьютерном центре, работая вместе с Группой по рабочим потокам и распределенным вычислениям на факультете Вычислительных наук. В сферу его научных интересов входят анализ социальных сетей и больших данных, машинное и глубокое обучение, обработка естественного языка, анализ мнений, выявление тем, периферийные вычисления и высокопроизводительная аналитика данных.

Алессио Орсино – в настоящее время Алессио Орсино получает степень доктора философии в области информационных и коммуникационных технологий в Университете Калабрии, Италия. В 2023 году работал приглашенным исследователем на факультете Вычислительных наук и технологий Кембриджского университета, сотрудничая с Лабораторией исследования мобильных систем. В сферу его научных интересов входят анализ больших данных, параллельные и распределенные вычисления, высокопроизводительная аналитика данных, облачные и периферийные вычисления, а также машинное обучение.

Признательности

Хотели бы поблагодарить Рози Уильямсон и Логеша Арумугама за их полезную поддержку и комментарии к ранним черновикам этой книги и во время всех редакционных процессов. Выражаем свою признательность за частичную финансовую поддержку проекта eFlows4HPC (Европейская комиссия по программе исследований и инноваций Horizon 2020 и EuroHPC JU по контракту 955558) и проекта «PNRR MUR PE0000013-FAIR» – CUP H23C22000860006.

Глава 1

Введение

Архитектуры параллельных вычислений и языки и инструменты масштабируемого программирования играют ключевую роль в конструировании и реализации сложных программных приложений, ориентированных на управление крупными наборами данных, хранящихся в файловых системах, базах данных, архивах и озерах данных, и их анализ. Эта книга написана для тех, кто интересуется программированием приложений по обработке больших данных на мультиядерных компьютерах, на облачных платформах, в системах распределенных вычислений и на массивно-параллельных машинах. Студенты, разработчики и ученые, желающие узнать о наиболее эффективных фреймворках программирования приложений, интенсивных по привлечению данных, найдут в этой книге руководство, которое знакомит с моделями, языками и инструментами эффективного управления большими данными и их анализа, а также обсуждение вопросов выбора среды, наиболее подходящей для достижения целевых задач приложения. Эта глава знакомит с целями книги, иллюстрирует темы и описывает ее организацию.

1.1. Мотивация и цели

Современный мир генерирует беспрецедентное количество данных, и способность добывать из них ценные сведения имеет решающее значение для успеха во многих областях, включая предпринимательство, науку и управление, обеспечивая инновации и принятие обоснованных решений. Самое лучшее, что можно сделать, чтобы задействовать ценность огромного количества имеющихся данных, состоит в реализации масштабируемых приложений по управлению данными и проведению их анализа, которые эффективно добывают из них полезные закономерности, модели и тренды. Задача программирования приложений по обработке больших данных сложна и многогранна и требует технических знаний и глубокого понимания самых разных понятий и концепций, включая аналитику данных, распределенные вычис-

ления, параллельную обработку и машинное обучение. Несмотря на значительные трудности, сфера больших данных постоянно растет, а вместе с ней растет и спрос на квалифицированных специалистов, способных конструировать и строить эффективные и масштабируемые приложения по обработке крупных объемов данных. Способность работать с большими данными стала важнейшим навыком на современном рынке труда, и овладение им может открывать массу карьерных возможностей.

Эта книга призвана стать незаменимым руководством для разработчиков, стремящихся создавать надежные и масштабируемые приложения по обработке больших данных. Благодаря всеобъемлющему охвату главных инструментов и фреймворков она предлагает глубокое понимание принципов и практик, необходимых для реализации эффективных приложений по проведению анализа больших данных, охватывая широкий спектр тем, включая системы распределенного хранения и вычисления, масштабируемую обработку данных, управление данными и машинное обучение, с использованием популярных инструментов и фреймворков, таких как Hadoop, Spark, Hive, MPI и Storm. Книга предлагает практический подход к конструированию приложений по обработке больших данных, тем самым превращая ее в руководство, которое подойдет для разработчиков с разным уровнем опыта.

Одним из ключевых преимуществ этой книги является ее акцент на масштабируемости. По сути дела, обсуждаемые здесь инструменты и фреймворки специально сконструированы для работы с крупными наборами данных и выполнения сложных заданий по обработке за счет привлечения параллелизма. Их освоение позволяет разработчикам создавать приложения, способные обрабатывать огромные объемы данных.

Еще одним существенным преимуществом книги является ее практический подход. В книге приводятся реально-практические примеры, которые показывают читателям, как применять полученные знания, и помогают приобретать опыт работы с различными практическими вариантами использования. В дополнение к этому в книгу включено сравнение инструментов обработки больших данных в реально-практических приложениях, которое показывает порядок использования больших данных в различных сценариях и областях, обеспечивая читателям глубокое понимание потенциальных приложений по обработке больших данных и предоставляя руководство по выбору правильных инструментов по каждому конкретному варианту использования.

Мы также освещаем некоторые последние тренды, такие как вычисления экзафлопсного масштаба, параллельное и распределенное машинное обучение, и обсуждаем порядок их возможного привлечения для анализа и обработки крупных наборов данных.

К концу этой книги читатель получит глубокое понимание принципов и методов, используемых для разработки масштабируемых и надежных приложений по обработке больших данных, а также практический опыт работы с некоторыми наиболее широко используемыми инструментами в этой области.

1.2. Главные темы

Эта книга представляет собой всеобъемлющее руководство, в котором обследуются главные парадигмы и фреймворки, используемые для обработки и анализа больших данных, и помогает программистам и разработчикам в выборе самых лучших инструментов программирования в зависимости от их навыков, наличия оборудования, областей применения и целей. Книга охватывает широкий спектр тем, связанных с обработкой, управлением и анализом больших данных, включая:

- главные системы распределенного хранения данных, являющиеся важнейшим элементом противостояния текущему экспоненциальному росту требований к хранению данных, обеспечивающие масштабируемость, эффективность, отказоустойчивость, доступность и состыкуемость;
- главные принципы, положенные в основу процессов анализа данных и науки о данных, а также их развитие на масштабируемых вычислительных системах;
- преимущества таких технологий, как высокопроизводительные, облачные и распределенные вычисления, которые способствуют обработке крупных объемов данных в реально-практических контекстах;
- главные модели программирования для больших данных, такие как MapReduce, рабочие потоки и передача сообщений, являющиеся ключевыми парадигмами, помогающими пользователям выражать параллельные алгоритмы и приложения, предоставляя абстракции для архитектуры параллельных вычислений;
- новейшие предложения в области вычислений экзафлопсного масштаба, ориентированные на масштабируемые технологические решения и инструменты в широком спектре научных областей, включая физику, биологию и симуляцию природных явлений;
- наиболее часто используемые инструменты программирования для обработки больших данных, предлагающие как универсальные, так и специализированные решения по работе с разными видами данных, от структурированных данных до графов и потоков, и областями, включая пакетные, потоковые, графовые и запросные приложения;
- ключевые характеристики, плюсы и минусы разных фреймворков относительно конкретных классов приложений, с целью оказания помощи программистам в выборе наиболее подходящего фреймворка, а также другие важные факторы, которые могут влиять на этот выбор, такие как тип данных, масштаб инфраструктуры, навыки разработчиков и размер сообщества.

1.3. Аудитория и организация

Эта книга предназначена для студентов и инженеров-изыскателей, изучающих обработку и аналитику больших данных, разработчиков программного обеспечения и профессионалов бизнеса, заинтересованных в использовании больших данных в своих организациях. Читатели, как правило, должны хорошо понимать языки программирования, такие как Java, Python или Scala, и иметь базовые знания о главных понятиях и концепциях параллельного и распределенного программирования. С целью удовлетворения потребностей такого широкого круга читателей в книге содержится исчерпывающий обзор фреймворков программирования масштабируемых распределенных и параллельных приложений, интенсивных по привлечению данных. Она представляет собой ценный ресурс для студентов, стремящихся глубже понять эти концепции и методы, а также для профессионалов, работающих на фабриках по производству программного обеспечения и в компаниях, занимающихся наукой о данных, которые могут извлечь пользу из практических идей и реальных приложений, представленных в главах книги.

Читатели могут свободно приспосабливать чтение книги под свои потребности, основываясь на собственных навыках и знакомстве с темой. Книгу можно читать целиком либо сосредотачиваться на отдельных интересующих разделах, не чувствуя себя обязанными внимательно прочитывать каждое слово, перед тем как продолжать дальше.

Книга состоит из пяти глав, которые вкратце описаны ниже.

Глава 2 «*Концепции больших данных*» знакомит с областью больших данных путем введения главных принципов и особенностей управления большими данными и их анализа. В частности, обсуждаются техники анализа данных и подходы на основе науки о данных, а также исследуется их развитие на масштабируемых вычислительных системах. Также рассматриваются такие технологии, как высокопроизводительные, облачные и распределенные вычисления, объясняя их полезность в обработке больших данных.

Глава 3 «*Модели программирования для больших данных*» посвящена главным моделям программирования, разработанным и используемым для реализации крупномасштабных приложений по обработке больших данных, включая модели на основе MapReduce, рабочих потоков, массового синхронного параллелизма, передачи сообщений, разделенного глобального адресного пространства и SQL-подобные модели. В книге также представлены последние предложения в области вычислений экзафлопсного масштаба. В данной главе рассматриваются ключевые характеристики и механизмы каждой модели программирования, которые можно использовать для обработки и анализа больших данных.

Глава 4 «*Инструменты для приложений по обработке больших данных*» описывает языки программирования, библиотеки и инструменты, используемые для конструирования масштабируемых приложений по обработке больших данных. Фреймворки, включая Hadoop, Spark, Storm и MPI, представлены

с описанием их характеристик и механизмов программирования. По каждому инструменту программирования приводится несколько реально-практических примеров приложений по обработке больших данных.

Глава 5 «*Сравнение инструментов программирования*» посвящена сравнению инструментов программирования, представленных в предыдущей главе, путем выделения главных характеристик, преимуществ и недостатков их использования в разных типах приложений, таких как пакетные, потоковые, графовые и запросные приложения. В ней также обсуждается сообщество разработчиков, проводится сравнение этих фреймворков по их распространенности и популярности с точки зрения конечных пользователей и разработчиков.

Глава 6 «*Выбор правильного фреймворка для приручения больших данных*» завершает книгу обсуждением главных факторов, которые могут влиять на выбор фреймворка, наиболее подходящего для обработки и анализа больших данных. Основное внимание уделяется характеристикам входных данных, классу приложений и масштабу инфраструктуры, а также приводятся многие другие факторы, которые в той или иной степени могут влиять на решения разработчиков, включая квалификацию конструктора или разработчика, размер сообщества, конфиденциальность данных, требования к обеспечению безопасности, доступный бюджет, интегрируемость и совместимость.

1.4. Онлайновые ресурсы

Онлайновый репозиторий, включающий все исходные коды и наборы данных, использованные в примерах, приводимых в главах книги, доступен читателю по адресу https://bigdataprogramming.github.io. Репозиторий предоставляет Docker-контейнеры для бесшовного исполнения предложенных примеров. Также имеется краткое руководство по установке, компиляции и запуску программ.

Копия слайдов, основанных на содержимом книги, предназначенных для использования в образовательных целях, находится на веб-сайте издательства. Инструкции по доступу к слайдам приведены на стр. 253.

Глава 2

Концепции больших данных

Эта глава знакомит с областью больших данных путем введения главных принципов и описания особенностей управления большими данными и их анализа. В частности, обсуждаются техники анализа данных и подходы на основе науки о данных, а также исследуется их развитие на масштабируемых вычислительных системах. Помимо этого, рассматриваются такие технологии, как высокопроизводительные, облачные и распределенные вычисления, объясняя их полезность в обработке больших данных.

2.1. Принципы и характеристики больших данных

За последние несколько лет способность генерирования и сбора данных выросла в геометрической прогрессии. В эпоху интернета вещей (IoT) из нескольких источников, таких как датчики, мобильные устройства, веб-приложения и услуги, генерируются и собираются огромные объемы цифровых данных. Более того, с широким принятием мобильных устройств миллионы людей ежедневно пользуются социальными сетями и производят огромные объемы цифровых данных, которые можно эффективно задействовать для добычи ценной информации о динамике и поведении людей. Такие данные, которые принято называть «большими данными», содержат ценную информацию о действиях, интересах и поведении пользователей, что делает их идеально пригодными для очень широкого круга приложений.

В настоящее время термин «большие данные» часто используется неправильно, но он очень важен в вычислительной науке для понимания предпринимательской и человеческой деятельности. В технической литературе

было предложено несколько определений данного термина, однако достичь глобального консенсуса относительно его содержимого было нелегко. Несмотря на то что указанный термин в явном виде не упоминается, первое его определение было предложено Дугом Лэйни (Doug Laney, аналитиком META Group, ныне компания Gartner) в отчете 2001 года (Laney и соавт., 2001), который предположил, что тремя аспектами трудностей в управлении данными являются объем, разнообразие и скорость. Впоследствии компания Gartner предложила более формальное определение (Gartner, Inc., дата публикации отсутствует): «*Большие данные – это крупные объемы, высокая быстрота и/или большое разнообразие информационных активов, которые нуждаются в экономически эффективных, инновационных формах обработки, позволяющих улучшать понимание, принятие решений и автоматизацию процессов*».

В таком определении для описания больших данных применена трехмерная модель, также именуемая моделью «3V» (то есть *volume*, *velocity* и *variety* – объем, быстрота и разнообразие). В частности, объем относится к количеству генерируемых данных, быстрота – к скорости, с которой эти данные генерируются, а разнообразие – к разнородности структуры и формата данных, поступающих из разных источников. Давайте обсудим эти три свойства больших данных более подробно.

- *Объем* – это, пожалуй, самое первое свойство, которое приходит на ум при мысли о больших данных. Поскольку каждый день создаются экзабайты данных, то уже не редкость, когда на устройствах хранения в крупных компаниях размещаются даже петабайты данных. Такой объем данных нередко бросает серьезный вызов способности управлять, так как требует нестандартных решений в области хранения и управления.
- *Быстрота*, по сути, служит мерой скорости поступления данных. Некоторые данные поступают в режиме реального времени, а другие – с задержкой, спорадически, отправляемые партиями или пакетами. Таким образом, может случиться так, что сбор данных, происходящих из разных источников, поступающих в одном и том же темпе, может поставить систему сбора в затруднительное положение, так как традиционные вычислительные системы не смогут работать на данных, поступающих быстрее, чем они способны их осмысливать. В качестве примера достаточно взять систему, которая собирает данные из сети датчиков, состоящей из тысяч устройств, посылающих данные с интервалом порядка секунд.
- *Разнообразие* означает, что данные могут собираться из разных источников и предъявляться в самых разных форматах, таких как видео, текст, аудио, CSV и PDF. Слияние или перевод этих данных в общий формат может нуждаться в больших усилиях и продвинутых аналитических навыках, чтобы понимать эти входные данные и делать их управляемыми и пригодными для анализа.

Согласно определению, данному компанией Gartner, большие данные характеризуются не только крупным размером наборов данных, но и их разно-