



ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ	10
БЛАГОДАРНОСТИ	12
ОБ ЭТОЙ КНИГЕ.....	14
Структура книги.....	15
Графическое выделение и загрузка исходного кода	15
АВТОР В СЕТИ	16
ОБ АВТОРЕ	17
ОБ ИЛЛЮСТРАЦИИ НА ОБЛОЖКЕ	18
 ЧАСТЬ 1.	
Нadooр – каркас распределенного программирования.....	19
ГЛАВА 1. Введение в Hadooр	21
1.1. Зачем написана книга «Hadooр в действии»?	22
1.2. Что такое Hadooр?	23
1.3. Сравнение Hadooр с другими распределенными системами.....	24
1.4. Сравнение СУБД на основе SQL с Hadooр	26
1.5. Знакомство с MapReduce.....	29
1.5.1. Масштабирование простой программы вручную	30
1.5.2. Масштабирование той же программы с помощью MapReduce	33
1.6. Подсчет слов с помощью Hadooр – ваша первая программа.....	36
1.7. История Hadooр.....	43
1.8. Резюме	44
1.9. Ресурсы	45

ГЛАВА 2. Запуск Hadoop 46

2.1. Структурные элементы Hadoop	46
2.1.1. NameNode	47
2.1.2. DataNode	47
2.1.3. Secondary NameNode	49
2.1.4. JobTracker	49
2.1.5. TaskTracker	50
2.2. Настройка SSH для кластера Hadoop	52
2.2.1. Определение общей учетной записи	52
2.2.2. Проверка правильности установки SSH	52
2.2.3. Генерация пары ключей	53
2.2.4. Распространение открытого ключа и проверка возможности входа в систему	53
2.3. Запуск Hadoop	54
2.3.1. Локальный (автономный) режим	55
2.3.2. Псевдораспределенный режим	56
2.3.3. Полностью распределенный режим	58
2.4. Веб-интерфейс для мониторинга кластера	62
2.5. Резюме	63

ГЛАВА 3. Компоненты Hadoop 65

3.1. Работа с файлами в системе HDFS	65
3.1.1. Основные команды для работы с файлами	66
3.1.2. Чтение и запись в HDFS из программы	71
3.2. Анатомия MapReduce-программы	74
3.2.1. Типы данных в Hadoop	76
3.2.2. Распределитель	78
3.2.3. Редуктор	79
3.2.4. Разбивка — направление выхода распределителя	80
3.2.5. Комбинатор — локальная редукция	81
3.2.6. Подсчет слов с помощью готовых классов распределителя и редуктора	81
3.3. Чтение и запись	83
3.3.1. Интерфейс InputFormat	85
3.3.2. Интерфейс OutputFormat	91

3.4. Резюме	93
-------------------	----

ЧАСТЬ 2.

Hadoop в действии95

Глава 4. Создание простых MapReduce-программ . 97

4.1. Получение набора данных о патентах	98
4.1.1. Данные о цитировании патентов	99
4.1.2. Данные об описаниях патентов	101
4.2. Определение шаблона MapReduce-программы	102
4.3. Подсчет всякой всячины	108
4.4. Адаптация к изменениям в API Hadoop	114
4.5. Интерфейс Hadoop Streaming	118
4.5.1. Интерфейс Streaming и команды Unix	119
4.5.2. Streaming и скрипты	120
4.5.3. Интерфейс Streaming и пары ключ/значение	126
4.5.4. Интерфейс Streaming и пакет Aggregate	131
4.6. Повышение производительности с помощью комбинаторов.....	137
4.7. Упражнения	142
4.8. Резюме	144
4.9. Дополнительные ресурсы	145

ГЛАВА 5. Углубленное изучение MapReduce 147

5.1. Сцепление задач MapReduce	148
5.1.1. Последовательное сцепление задач MapReduce	148
5.1.2. Сцепление задач MapReduce со сложными зависимостями	148
5.1.3. Включение в цепочку шагов пред- и постобработки	149
5.2. Соединение данных из разных источников.....	154
5.2.1. Соединение на стороне редуктора	155
5.2.2. Построение реплицированных соединений с помощью класса DistributedCache	166
5.2.3. Полусоединение: соединение на стороне редуктора с фильтрацией на стороне распределителя	171
5.3. Создание фильтра Блума	173

5.3.1. Что делает фильтр Блума?	173
5.3.2. Реализация фильтра Блума	176
5.3.3. Фильтр Блума в Hadoop версии 0.20+	184
5.4. Упражнения	184
5.5. Резюме	187
5.6. Дополнительные ресурсы	187
ГЛАВА 6. Практическое программирование	189
6.1. Разработка MapReduce-программ	190
6.1.1. Локальный режим	191
6.1.2. Псевдораспределенный режим	197
6.2. Мониторинг и отладка в производственном кластере	203
6.2.1. Счетчики	203
6.2.2. Пропуск плохих записей	205
6.2.3. Перезапуск сбойных заданий с помощью IsolationRunner	210
6.3. Оптимизация производительности	211
6.3.1. Уменьшение сетевого трафика с помощью комбинатора	212
6.3.2. Уменьшение объема выходных данных	212
6.3.3. Использование сжатия	213
6.3.4. Повторное использование JVM	216
6.3.5. Наблюдаемое исполнение	217
6.3.6. Переработка кода и модификация алгоритмов	219
6.4. Резюме	221
ГЛАВА 7. Сборник рецептов	222
7.1. Передача нестандартных параметров задаче	222
7.2. Получение информации о конкретном задании	226
7.3. Разбиение на несколько выходных файлов	227
7.4. Ввод и вывод в базу данных	234
7.5. Сортировка выходных данных	236
7.6. Резюме	238
ГЛАВА 8. Администрирование Hadoop	239
8.1. Практическая настройка параметров	240

8.2. Проверка состояния системы	243
8.3. Установка прав доступа	245
8.4. Управление квотами	246
8.5. Включение корзины	247
8.6. Удаление узлов DataNode.....	247
8.7. Добавление узлов DataNode	249
8.8. Управление узлами NameNode и Secondary NameNode...	250
8.9. Восстановление после сбоя узла NameNode.....	252
8.10. Проектирование топологии сети и осведомленность о стойках	254
8.11. Планирование задач, поступающих от нескольких пользователей.....	257
8.11.1. Организация нескольких узлов JobTracker.....	257
8.11.2. Справедливый планировщик	258
8.12. Резюме	261

ЧАСТЬ 3.

Hadoop в реальной жизни 263

ГЛАВА 9. Эксплуатация Hadoop в облаке 265

9.1. Введение в Amazon Web Services.....	266
9.2. Настройка AWS	267
9.2.1. Получение учетных данных для аутентификации в AWS.....	268
9.2.2. Получение командных утилит	271
9.2.3. Подготовка пары ключей для работы с SSH	273
9.3. Настройка Hadoop в EC2	275
9.3.1. Задание параметров защиты	275
9.3.2. Конфигурирование типа кластера	276
9.4. Запуск MapReduce-программ в среде EC2.....	278
9.4.1. Перенос своего кода в кластер Hadoop	279
9.4.2. Доступ к данным из кластера Hadoop	279
9.5. Очистка и останов экземпляров EC2	285
9.6. Amazon Elastic MapReduce и другие службы AWS.....	285
9.6.1. Amazon Elastic MapReduce.....	286

9.6.2. AWS Import/Export.....	287
9.7. Резюме	288
ГЛАВА 10. Программирование с помощью Pig	289
10.1. Научитесь думать по-свински.....	290
10.1.1. Язык описания потоков данных.....	290
10.1.2. Типы данных	291
10.1.3. Определенные пользователем функции	291
10.2. Установка Pig	291
10.3. Запуск Pig	293
10.3.1. Управление оболочкой Grunt	294
10.4. Изучение языка Pig Latin с помощью Grunt	295
10.5. Учимся говорить на Pig Latin	302
10.5.1. Типы данных и схемы.....	302
10.5.2. Выражения и функции	304
10.5.3. Реляционные операторы	307
10.5.4. Оптимизация исполнения.....	317
10.6. Определяемые пользователем функции	317
10.6.1. Использование UDF.....	318
10.6.2. Создание UDF	319
10.7. Работа со скриптами.....	322
10.7.1. Комментарии	322
10.7.2. Подстановка параметров	323
10.7.3. Режим многозапросного исполнения	324
10.8. Pig в действии: отыскание похожих патентов.....	326
10.9. Резюме	332
ГЛАВА 11. Hive и другие	333
11.1. Hive	334
11.1.1. Установка и настройка Hive.....	335
11.1.2. Примеры запросов	338
11.1.3. Детали языка HiveQL	342
11.1.4. Hive: подводя итоги	352
11.2. Другие проекты, связанные с Hadoop.....	353
11.2.1. HBase	353

11.2.2. ZooKeeper	353
11.2.3. Cascading	354
11.2.4. Cloudera	354
11.2.5. Katta	355
11.2.6. CloudBase.....	355
11.2.7. Aster Data и Greenplum.....	356
11.2.8. Hama и Mahout	356
11.2.9. search-hadoop.com.....	356
11.3. Резюме	357
ГЛАВА 12. Примеры применения	358
12.1. Преобразование 11 миллионов изображений из архива газеты New York Times	358
12.2. Добыча данных в компании China Mobile	360
12.3. Рекомендование лучших веб-сайтов на StumbleUpon....	367
12.3.1. Как мы пришли к распределенной обработке в StumbleUpon	368
12.3.2. HBase и StumbleUpon	369
12.3.3. Другие применения Hadoop на сайте StumbleUpon.....	379
12.4. Построение аналитической системы для внутрикорпоративного поиска – проект IBM ES2	381
12.4.1. Архитектура ES2	386
12.4.2. Робот ES2.....	387
12.4.3. Аналитические средства в ES2	390
12.4.4. Выводы	400
12.4.5. Библиография.....	401
ПРИЛОЖЕНИЕ. Команды HDFS.....	403
ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ	408



ПРЕДИСЛОВИЕ

Уже давно данные очаровывают меня. Еще на младших курсах, изучая электротехнику, я открыл для себя цифровую обработку сигналов и прикипел к ней всей душой. Я понял, что музыку, фотографии и множество других вещей можно рассматривать как данные. Оказалось, что создавать и усиливать эмоциональные переживания можно с помощью вычислений. Ничего интереснее я даже представить себе не мог.

Со временем передо мной раскрылись новые, не менее удивительные аспекты данных. Вот уже несколько лет, как я занимаюсь социальными сетями и большими массивами данных. Именно большие массивы бросают вызов моему интеллекту. Статистический анализ данных я освоил уже давно, а для работы с новыми типами данных были нужны «всего лишь» новые математические методы. Конечно, это не просто, но по крайней мере я этому учился, а уж в ресурсах для пополнения знаний недостатка и вовсе нет. С другой стороны, для работы с большими объемами данных необходим совсем другой системный подход и новые способы программирования. Этому меня не учили, но – и это даже важнее – не один я оказался в таком положении. Знания о практических методах обработки больших массивов данных сродни черной магии. Существует немало инструментов и приемов масштабируемой обработки данных, в частности, кэширование (например, с помощью программы `memcached`), репликация, секционирование и, конечно, MapReduce/Hadoop. Последние несколько лет я потратил на приобретение и совершенствование навыков в этой области.

Лично для меня наиболее сложной оказалась средняя часть кривой обучения. Поначалу не так уж трудно найти блоги с вводным материалом и презентации, объясняющие, как написать пример уровня «Здравствуй, мир». А приобретя начальные знания, вы уже можете задавать дополнительные вопросы в списках рассылки, общаться со специалистами на различных встречах и конференциях и даже самостоятельно читать исходный код. Но вот в середине зияет гигантский

провал —аппетит уже разыгрался, а как сформулировать последующие вопросы, неясно. Эта проблема стоит особенно остро для таких новейших технологий, как Nadoor. Необходимо некое упорядоченное изложение, которое начиналось бы с примера «Здравствуй, мир» и доводило читателя до точки, в которой он смог бы без особых усилий применить Nadoor на практике. Именно так я и представляю себе эту книгу. К счастью, я обнаружил, что серия «In Action», выходящая в издательстве Manning, отлично согласуется с моей целью, и к тому же там есть прекрасные редакторы, помогавшие мне на всем протяжении работы.

Я получал удовольствие, когда писал книгу, и надеюсь, что она станет для вас началом увлекательного путешествия в мир Nadoor.



БЛАГОДАРНОСТИ

В эту книгу внесло вклад много людей. Прежде всего, я хочу поблагодарить Джеймса Уоррена (James Warren). Он возглавлял аналитический отдел в компании RockYou и вместе с ним мы стремились популяризировать Hadoop в этой организации. Я многому научился у него, он даже помогал мне писать ранние варианты текста.

На мое счастье нашлось немало людей, познакомивших меня с интересными примерами за рамками индустрии Web 2.0. В частности, я хочу выразить благодарность Чжи Гуо Люо (Zhiguo Luo), Мень Ху (Meng Xu), Шаолинь Сун (Shaoling Sun), Кену Макиннису (Ken MacInnis), Райану Роусону (Ryan Rawson), Вуку Эрчеговачу (Vuk Ercegovic), Раджасекару Кришнамурти (Rajasekar Krishnamurthy), Срираму Рагхавану (Sriram Raghavan), Фредерику Рейсу (Frederick Reiss), Юджину Шекита (Eugene Shekita), Сандипу Тата (Sandeep Tata), Шивакумару Вайтианатхану (Shivakumar Vaithyanathan) и Хуай Ю Чжу (Huaiyu Zhu).

Я также благодарен рецензентам этой книги. Они прислали ценные отзывы на первые варианты. В частности, технический редактор Пол О'Рорк (Paul O'Rorke) вышел далеко за пределы формальных обязанностей и внес ряд крайне полезных предложений о том, как улучшить рукопись. Я надеюсь, что когда-нибудь он сам напишет книгу. Кроме того, я получал истинное удовольствие от долгих бесед с Джонатаном Као (Jonathan Cao). Его опыт работы с базами данных и крупномасштабными системами позволил мне шире взглянуть на возможности Hadoop и лучше понять их.

По ходу работы текст книги многократно читали и другие рецензенты, которых я хотел бы поблагодарить за их бесценный труд: Пол Стусяк (Paul Stusiak), Филипп К. Дженерт (Philipp K. Janert), Амин Мохаммед-Коулмен (Amin Mohammed-Coleman), Джон С. Гриффин (John S. Griffin), Марко Угетти (Marco Ughetti), Рик Вагнер (Rick Wagner), Кеннет Делонг (Kenneth DeLong), Джош Паттерсон (Josh Patterson), Срини Пенчикала (Srini Penchikala), Костантино Кербо (Costantino Cerbo), Стив Логран (Steve Loughran), Ара Абрахамян

(Ara Abrahamian), Бен Холл (Ben Hall), Эндрю Зимер (Andrew Siemer), Роберт Хансон (Robert Hanson), Кит Ким (Keith Kim), Сопан Шивейл (Sopan Shewale), Марион Стуртеван (Marion Sturtevant), Крис Чендлер (Chris Chandler), Эрик Реймонд (Eric Raymond) и Джероен Бенкхейсен (Jeroen Benckhuijsen).

Мне посчастливилось работать с коллективом чудесных людей в издательстве Manning. Отдельное спасибо Трою Мотту (Troy Mott), который побудил меня взяться за это дело и терпеливо дожидался, пока я его закончу. Спасибо также Таре Уолш (Tara Walsh), Карен Тегтмейер (Karen Tegtmeier), Марьян Бейс (Marjan Bace), Мэри Пирджис (Mary Piergies) Синтии Кэйн (Cynthia Kane), Стивену Хонгу (Steven Hong), Рашель Шредер (Rachel Schroeder), Кэти Теннант (Katie Tennant) и Морин Спенсер (Maureen Spencer). Поддержка с их стороны была просто феноменальной. Лучшей компании для работы я себе даже представить не могу.

Само собой, добрых слов заслуживают все, кто внес свой вклад в разработку каркаса Hadoop и помогает развиваться сформировавшейся вокруг него экосистеме. Начало положил Дуг Каттинг (Doug Cutting), а компания Yahoo оказалась достаточно прозорливой, чтобы поддержать его с самых первых шагов. Теперь компания Cloudera представляет Hadoop вниманию более широкой аудитории корпоративных пользователей. Сейчас самое подходящее время присоединиться к растущему сообществу Hadoop.

И наконец, я хочу сказать спасибо всем своим друзьям, семье и коллегам, которые поддерживали меня на протяжении работы над книгой.



ОБ ЭТОЙ КНИГЕ

Hadoop – это каркас с открытым исходным кодом, в котором реализован алгоритм распределения и редукции MapReduce, лежащий в основе подхода Google к организации запросов к распределенным наборам данных, которые и составляют Интернет. В связи с таким определением возникает очевидный вопрос: что такое *распределение* (map) и зачем нужна последующая *редукция* (reduce)? Зачастую массивные наборы данных с трудом поддаются анализу и опросу традиционными средствами, особенно когда сами запросы весьма сложны. По существу, алгоритм MapReduce разбивает запрос и набор данных на несколько частей – это этап распределения. Отдельные компоненты запроса можно обработать параллельно, а затем *свести* полученные результаты воедино (*редуцировать*).

Читатель этой книги узнает, как использовать Hadoop и писать MapReduce-программы. Книга предназначена для программистов, архитекторов и руководителей проектов, связанных с обработкой большого объема данных в офлайн-режиме. Будет описано, как получить копию Hadoop, как организовать кластер и как писать программы анализа. Мы начнем с применения Hadoop в конфигурации по умолчанию к решению нескольких простых задач, например, об анализе изменения частоты вхождения слов в корпус документов; это позволит уяснить основные идеи Hadoop и MapReduce. Далее мы перейдем к базовым концепциям MapReduce-приложений, разрабатываемых с помощью Hadoop, и по ходу дела изучим компоненты каркаса, применение Hadoop к широкому спектру задач анализа данных и многочисленным примерам Hadoop в действии.

Алгоритм MapReduce сложен как концептуально, так и в плане реализации, а от пользователей каркаса Hadoop требуется изучить все средства и хитрости, необходимые для работы с ним. В этой книге не просто рассказывается о том, как запустить Hadoop; прочитав ее, вы научитесь писать полезные программы с использованием MapReduce.

Предполагается, что читатель хотя бы немного владеет языком Java, поскольку именно на нем написано большинство примеров.

Знакомство с основами математической статистики (например, гистограммами и корреляцией) поможет разобраться в более сложных примерах обработки данных.

Структура книги

Эта книга состоит из 12 глав, разбитых на три части.

Часть 1 состоит из трех глав и представляет собой введение в каркас Hadoop. Здесь излагаются те базовые сведения, которые необходимо знать для понимания и использования каркаса. Описывается, из каких аппаратных компонентов состоит кластер Hadoop, рассказывается об установке и конфигурировании системы. Также в части 1 дается общее представление о каркасе MapReduce и приводится пример первой MapReduce-программы.

Часть 2 «Hadoop в действии» состоит из пяти глав, в которых описаны практические навыки, необходимые для составления и запуска программ обработки данных в среде Hadoop. Здесь мы рассмотрим многочисленные примеры применения Hadoop к анализу набора данных о патентах, в том числе и такие нетривиальные алгоритмы, как фильтр Блума. Мы также поговорим о приемах программирования и администрирования, чрезвычайно полезных при работе с Hadoop в производственной среде.

Часть 3 «Hadoop в реальной жизни», в которую входят последние четыре главы, посвящена обширной экосистеме, сложившейся вокруг Hadoop. Существуют облачные службы, которые позволяют обойтись без покупки собственного оборудования для создания кластера. Имеются также дополнительные пакеты, предлагающие высокоуровневые абстракции, надстроенные над MapReduce. Наконец, мы рассмотрим несколько примеров практического применения Hadoop к решению реальных задач бизнеса.

В приложении приведен список команд HDFS.

Графическое выделение и загрузка исходного кода

Исходный код в листингах и в основном тексте набран моноширинным шрифтом. Многие листинги сопровождаются аннотациями, в которых излагаются важные концепции. В некоторых случаях в листингах присутствуют нумерованные маркеры, с которыми соотносятся последующие пояснения.

Код всех приведенных в книге примеров можно скачать с сайта издательства по адресу www.manning.com/HadoopinAction.



АВТОР В СЕТИ

Приобретение книги «*Hadoop в действии*» открывает бесплатный доступ к закрытому форуму, организованному издательством Manning Publications, где вы можете оставить свои комментарии к книге, задать технические вопросы и получить помощь от автора и других пользователей. Получить доступ к форуму и подписаться на список рассылки можно на странице www.manning.com/HadoopinAction. Там же написано, как зайти на форум после регистрации, на какую помощь можно рассчитывать, и изложены правила поведения в форуме.

Издательство Manning обязуется предоставлять читателям площадку для общения с другими читателями и автором. Однако это не означает, что автор обязан как-то участвовать в обсуждениях; его присутствие на форуме остается чисто добровольным (и не оплачивается). Мы советуем задавать автору хитроумные вопросы, чтобы его интерес к форуму не угасал!

Форум автора в сети и архивы будут доступны на сайте издательства до тех пор, пока книга не перестанет печататься.



ОБ АВТОРЕ

В настоящее время Чак Лэм занят организацией социальной сети для пользователей мобильных устройств под названием *RollCall*. Это будет социальный секретарь для активных людей.

Раньше Чак работал техническим руководителем проекта в компании RockYou. В этой должности он разрабатывал социальные приложения и инфраструктуру обработки данных, рассчитанную на сотни миллионов пользователей. Он применял A/B тестирование и статистический анализ для оптимизации скорости распространения социального приложения среди пользователей (его *виральности*). Ему удавалось значительно – иногда на порядок – повысить частоту посещения страниц.

Обработкой больших объемов данных Чак заинтересовался во время работы над кандидатской диссертацией в Стэнфордском университете. Он узнал о том, какой серьезный эффект большие массивы данных могут оказать на машинное обучение, и начал изучать последствия. В его диссертации «Вычислительные методы сбора данных» впервые были исследованы новые подходы к сбору данных для машинного обучения – с применением идей, заимствованных из программ с открытым исходным кодом и онлайн-овых игр.



Об иллюстрации на обложке

Рисунок на обложке книги «Nadoor в действии» называется «Юноша из селения Кистане в Далмации». Репродукция взята из альбома традиционной хорватской одежды середины девятнадцатого века, составленного Николой Арсеновичем и опубликованного этнографическим музеем города Сплит, Хорватия, в 2003 году. Рисунок был получен при содействии библиотекаря этнографического музея Сплита, который расположен в той части средневекового центра города, которая восходит еще к временам римского владычества: рядом с руинами дворца императора Диоклетиана, датируемыми примерно 304 годом до н. э. В альбоме собраны тщательно раскрашенные изображения людей из разных районов Хорватии, сопровождаемые описаниями костюмов и деталей повседневного быта.

Сейчас Кистане – небольшой городок в Буковице, одном из географических регионов Хорватии. Он расположен в северной Далмации, области с богатой историей времен римлян и венецианцев. Хорватское слово «тапок» означает «холостяк», «кавалер», «ухажер» – неженатый молодой человек брачного возраста. Изображенный на обложке юноша выглядит очень щеголевато в выглаженной белой полотняной рубашке и расшитом жилете – вероятно, это его лучший наряд, в котором он ходит в церковь и на праздники, или оделся для свидания с девушкой.

Манера одеваться и общий уклад жизни за прошедшие 200 лет сильно изменились, и различия между областями, когда-то столь разительные, сгладились. Теперь трудно отличить друг от друга даже выходцев с разных континентов, что уж говорить о деревеньках и городках, разделенных несколькими километрами. Мы обменяли культурное разнообразие на иное устройство личной жизни – основанное на многостороннем и стремительном технологическом развитии.

Издательство Manning откликается на новации и инициативы в компьютерной отрасли обложками своих книг, на которых представлено широкое разнообразие местных укладов быта в позапрошлом веке. Мы возвращаем его в том виде, в каком оно запечатлено на иллюстрациях в старых книгах и собраниях, – примером может служить обложка этой книги.



Часть 1

Hadoop – каркас распределенного программирования

В части 1 закладываются основы для понимания и использования Hadoop. Мы расскажем, из каких аппаратных компонентов состоит кластер Hadoop, а также опишем процедуру установки и конфигурирования, необходимую для создания работоспособной системы. Мы дадим общее представление о каркасе MapReduce, после чего вы напишете и запустите свою первую MapReduce-программу.



ГЛАВА 1.

Введение в Hadoop

В этой главе:

- Основы составления масштабируемых, распределенных программ для обработки данных.
- Что представляют собой Hadoop и MapReduce.
- Создание и запуск простой MapReduce-программы.

В современном мире данные окружают нас со всех сторон. Мы загружаем на серверы видео, делаем фотографии с помощью мобильных телефонов, посылаем текстовые сообщения друзьям, изменяем их статус в системе Facebook, оставляем комментарии на разных сайтах, щелкаем по рекламным баннерам и т. д. Компьютеры со своей стороны также порождают и хранят все больше и больше данных. Возможно, даже эта книга представлена у вас в электронном виде, и читаете вы ее на экране своего компьютера. И уж точно факт покупки этой книги отражен в базе данных какого-нибудь розничного продавца¹.

Проблемы, вызванные экспоненциальным ростом данных, первыми ощутили на себе такие находящиеся на переднем крае технологий компании, как Google, Yahoo, Amazon и Microsoft. Им приходилось просеивать терабайты и петабайты данных, чтобы понять, какие веб-сайты популярны, какие книги пользуются спросом, какая реклама находит отклик у пользователей. Имеющиеся инструменты оказались не приспособлены к обработке столь больших объемов данных. Google стала первой компанией, начавшей пропагандировать систему *MapReduce*, которую применила для масштабирования обработки данных. Эта система вызвала большой интерес, так как многие другие компании уже столкнулись с аналогичными проблемами, а заново изобретать собственный инструментарий было нерентабельно.

¹ Вы ведь читаете легальную копию, правда?

Дуг Каттинг разглядел открывающиеся возможности и принялся за разработку версии MapReduce с открытым исходным кодом, которую назвал Hadoop. Вскоре Yahoo и другие компании объединили усилия в поддержку этого начинания. Ныне Hadoop составляет основную часть вычислительной инфраструктуры многих работающих в веб компаний, в частности, Yahoo, Facebook, LinkedIn и Twitter. Примериваются к ней и более традиционные организации, например, средства массовой информации и телекоммуникационные компании. В главе 12 описаны примеры использования Hadoop в таких компаниях, как New York Times, China Mobile и IBM.

Hadoop и, более общо, технологии распределенной обработки больших массивов данных, быстро становятся важным умением для широкого круга программистов. Сегодня программисту для эффективной работы необходимо знать о реляционных базах данных, сетях и безопасности, хотя еще пару десятков лет назад все это считалось факультативными навыками. Так и базовые знания о распределенной обработке данных скоро станут неременной частью багажа любого программиста. Ведущие университеты, например Стэнфордский и Карнеги-Меллона, уже ввели в программу факультетов информатики курс по изучению Hadoop. Эта книга поможет вам, программисту-практику, быстро освоить Hadoop и начать использовать его для обработки собственных наборов данных.

Эта глава содержит несколько более формальное введение в каркас Hadoop, рассматриваемый в контексте распределенных систем и систем обработки данных. Мы дадим обзор модели программирования MapReduce. Простой пример программы подсчета слов с помощью имеющихся инструментов поможет уяснить, какие проблемы возникают при обработке больших массивов данных. Затем вы напишете программу, основанную на использовании Hadoop, что позволит лучше оценить простоту этого каркаса. Мы также поговорим об истории Hadoop и некоторых перспективах парадигмы MapReduce. Но сначала позвольте мне вкратце объяснить, для чего я написал эту книгу и чем она может быть вам полезна.

1.1. Зачем написана книга «Hadoop в действии»?

На своем опыте я убедился, что поначалу Hadoop завораживает своими возможностями, но дело с трудом продвигается дальше написания элементарных примеров. Документация на официальном сайте

Конец ознакомительного фрагмента.

Приобрести книгу можно
в интернет-магазине «Электронный универс»
(e-Univers.ru)