

УДК 004.6:004.42Apache Hadoop
ББК 32.973.26-018.2
Л92

Л92 Чак Лэм

Hadoop в действии. – М.: ДМК Пресс, 2019. – 424 с.: ил.

ISBN 978-5-97060-723-7

Обработка больших массивов данных с помощью традиционных СУБД может оказаться трудным делом. Apache Hadoop — это каркас для разработки приложений, предназначенных для выполнения в распределенном кластере, без применения SQL. Такие приложения прекрасно масштабируются и могут обрабатывать гигантские массивы данных. Если вам требуется произвести анализ данных, то Hadoop — как раз то, что надо.

Прочитав эту книгу, вы познакомитесь с предметом и научитесь писать программы в стиле MapReduce. После нескольких простых примеров автор быстро переходит к вопросу об использовании Hadoop для решения более сложных задач анализа данных. Описываются рекомендованные приемы и паттерны проектирования, полезные при программировании для MapReduce.

Для чтения книги требуется знание основ языка Java. Некоторое знакомство с математической статистикой поможет разобраться в более сложных примерах.

УДК 004.6:004.42Apache Hadoop
ББК 32.973.26-018.2

Original English language edition published by Manning Publications 178 South Hill Drive, Westampton NJ 08060 USA, USA. Copyright (c) 2011 by Manning Publications. Russian-language edition copyright (c) 2012 by ДМК Пресс. 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.

Материал, изложенный в данной книге, многократно проверен. Но, поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

ISBN 978-1-93518-219-1 (англ.)

© by Manning Publications Co. All rights reserved.

ISBN 978-5-97060-723-7 (рус.)

© Оформление, перевод на русский язык
ДМК Пресс



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

Оглавление



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