

Содержание

Предисловие	19
Введение	21
Замечания	22
О чем рассказано в книге?	23
Что нового во втором издании?	23
Что нового в третьем издании?	23
Использование примеров кода	24
Благодарности	25
От издательства	26
Глава 1. Знакомство с Nadoop	27
Данные!	27
Хранение и анализ данных	29
Сравнение с другими системами	30

Hadoop и РСУБД	31
Распределенные вычисления	34
Добровольные вычисления	35
Краткая история Hadoop	36
Apache Hadoop и экосистема Hadoop	41
Выпуски Hadoop	42
О чем рассказано в книге	44
Имена конфигураций	44
MapReduce API	45
Совместимость	45
Глава 2. MapReduce	47
Набор метеорологических данных	47
Формат данных	48
Анализ данных средствами Unix	49
Анализ данных в Hadoop	51
Отображение и свертка	51
Программа MapReduce на языке Java	53
Тестовый запуск	57
MapReduce в перспективе	63
Поток данных	64
Комбинирующие функции	68
Определение комбинирующей функции	69
Выполнение распределенного задания MapReduce	70
Hadoop Streaming	71
Ruby	71
Python	74
Hadoop Pipes	75
Компилирование и запуск	77
Глава 3. HDFS	79
Строение HDFS	80
Основные концепции HDFS	81
Блоки	81
Узлы имен и узлы данных	83

HDFS Federation	84
Высокая доступность HDFS	85
Преодоление сбоев и изоляция	86
Интерфейс командной строки	87
Основные операции файловой системы	88
Файловые системы Hadoop	90
Интерфейсы	93
Интерфейс Java	95
Чтение данных Hadoop по URL-адресу	95
Чтение данных с использованием FileSystem API	97
Запись данных	100
Получение информации от файловой системы	103
Удаление данных	109
Поток данных	109
Чтение файла	109
Запись в файлы	113
Модель целостности	116
Перемещение данных: Flume и Sqoop	118
Параллельное копирование с использованием distcp	118
Сбалансированность кластеров HDFS	121
HAR	121
Использование HAR	121
Ограничения	123
Глава 4. Ввод/вывод в Hadoop	125
Целостность данных	125
Целостность данных в HDFS	126
LocalFileSystem	127
ChecksumFileSystem	128
Сжатие	128
Кодеки	130
Сжатие и разбиение входных данных	135
Использование сжатия в MapReduce	137
Сериализация	140
Интерфейс Writable	141

Классы Writable	144
Пользовательские реализации Writable	153
Программные среды сериализации	158
Avro	161
Типы данных и схемы Avro	162
Сериализация и десериализация в памяти	166
Файлы данных Avro	170
Совместимость	172
Преобразование схемы	175
Порядок сортировки	177
Avro и MapReduce	179
Сортировка с использованием Avro MapReduce	183
Avro MapReduce в других языках	186
Файловые структуры данных	186
SequenceFile	186
Запись SequenceFile	187
Чтение из SequenceFile	189
MapFile	195
Глава 5. Разработка приложений MapReduce	202
API конфигурации	203
Объединение ресурсов	204
Расширение переменных	205
Настройка среды разработки	206
Управление конфигурацией	208
GenericOptionsParser, Tool и ToolRunner	211
Написание модульных тестов с MRUnit	215
Функция отображения	215
Функция свертки	218
Локальное выполнение с тестовыми данными	219
Локальный запуск задания	219
Тестирование управляющей программы	223
Запуск в кластере	225
Упаковка задания	225
Запуск задания	227

Веб-интерфейс MapReduce	229
Получение результатов	232
Отладка задания	235
Журналы Hadoop	240
Удаленная отладка	242
Оптимизация задания	243
Профилирование	244
Модель MapReduce	247
Разложение задачи на задания MapReduce	248
JobControl	249
Apache Oozie	250
Определение потока операций Oozie	251
Глава 6. Как работает MapReduce	256
Выполнение задания MapReduce	256
Классическая реализация MapReduce (MapReduce 1)	257
Отправка заданий	258
YARN (MapReduce 2)	265
Сбои	271
Сбои в классической модели MapReduce	272
Сбои в YARN	274
Планирование заданий	277
Fair Scheduler	277
Capacity Scheduler	278
Тасовка и сортировка	279
На стороне отображения	279
На стороне свертки	281
Настройка конфигурации	283
Выполнение задач	287
Среда выполнения задач	287
Спекулятивное выполнение	288
OutputCommitter	290
Файлы побочных эффектов	292
Повторное использование JVM задач	292
Пропуск некорректных записей	293

Глава 7. Типы и форматы MapReduce	296
Типы MapReduce	296
Задание MapReduce по умолчанию	299
Форматы входных данных	309
Входные сплиты и записи	309
FileInputFormat	311
Входные пути FileInputFormat	311
Текстовые входные данные	322
Двоичные входные данные	326
Множественные источники входных данных	328
Операции ввода (и вывода) с базами данных	329
Форматы выходных данных	329
Текстовые выходные данные	330
Двоичные выходные данные	330
Множественный вывод	331
Отложенный вывод	336
Вывод в базы данных	336
Глава 8. Дополнительные возможности MapReduce	337
Счетчики	337
Встроенные счетчики	338
Счетчики Java, определяемые пользователем	344
Пользовательские счетчики в Streaming	349
Сортировка	350
Подготовка	350
Частичная сортировка	351
Полная сортировка	357
Вторичная сортировка	361
Соединения	368
Соединения на стороне отображения	369
Соединения на стороне свертки	370
Распространение побочных данных	374
Использование конфигурации задания	375
Распределенный кэш	375
Библиотечные классы MapReduce	383

Глава 9. Создание кластера Hadoop	384
Оборудование кластера	384
Сетевая топология.	387
Настройка и установка кластера	389
Установка Java	389
Создание пользователя Hadoop	390
Установка Hadoop	390
Тестирование установки	391
Конфигурация SSH	391
Конфигурация Hadoop	392
Управление конфигурацией	393
Настройки окружения	396
Важные свойства демонов Hadoop	401
Адреса и порты демонов Hadoop	407
Другие свойства Hadoop	408
Создание учетных записей пользователей	412
Конфигурация YARN	412
Важные свойства демонов YARN	413
Адреса и порты демонов YARN	417
Безопасность	419
Kerberos и Hadoop	420
Маркеры делегирования	423
Другие улучшения в области безопасности	424
Тестирование кластера Hadoop	426
Пользовательские задания	429
Hadoop в облаке	429
Apache Whirr	430
Глава 10. Администрирование Hadoop	435
HDFS	435
Дисковые структуры данных	435
Безопасный режим	441
Журналы аудита	443
Инструменты	444

Мониторинг	450
Ведение журналов	450
Метрики	451
Сопровождение	458
Стандартные административные процедуры	458
Включение и исключение узлов	459
Обновления	463
Глава 11. Pig	467
Установка и запуск Pig	469
Режимы исполнения	469
Запуск программ Pig	471
Grunt	471
Редакторы Pig Latin	472
Пример	472
Генерирование примеров	475
Сравнение с базами данных	477
Pig Latin	478
Структура	478
Инструкции	479
Выражение	485
Типы	487
Схемы	489
Функции	494
Макросы	496
Пользовательские функции	498
Фильтрующая пользовательская функция	498
Вычисляющая пользовательская функция	502
Пользовательская функция загрузки	504
Операторы обработки данных	508
Загрузка и сохранение	508
Фильтрация данных	508
Группировка и соединение данных	512
Сортировка данных	518
Комбинирование и разбиение данных	518
Практическое использование Pig	519

Параллелизм	519
Подстановка параметров	520
Глава 12. Hive	523
Установка Hive	524
Оболочка Hive	525
Пример	526
Администрирование Hive	528
Настройка конфигурации Hive	528
Сервисные функции Hive	530
Метахранилище	533
Сравнение с традиционными базами данных	535
Проверка схемы при чтении и записи	536
Обновления, транзакции и индексы	536
HiveQL	537
Типы данных	539
Операторы и функции	542
Таблицы	543
Управляемые и внешние таблицы	543
Разделы и гнезда	545
Форматы хранения данных	550
Импортирование данных	557
Модификация таблиц	559
Удаление таблиц	560
Запросы к данным	560
Сортировка и агрегирование	560
Сценарии MapReduce	561
Подзапросы	566
Пользовательские функции	568
Написание пользовательской функции	570
Написание UDAF	572
Глава 13. HBase	577
Знакомство с HBase	577
История	578

Концепции	578
Краткий обзор модели данных	578
Реализация.	580
Установка	583
Пробный запуск	584
Клиенты	586
Java	586
Avro, REST и Thrift	591
Пример	592
Схемы	592
Загрузка данных	593
Веб-запросы	597
HBase и РСУБД	600
Масштабирование успешного сервиса	601
HBase	603
Пример из практики: HBase в Streamy.com	603
Переход на HBase	605
Глава 14. ZooKeeper	607
Установка и запуск ZooKeeper	609
Пример	611
Реализация списка принадлежности в ZooKeeper	611
Создание группы	612
Присоединение к группе	614
Вывод списка участников группы.	616
Удаление группы.	618
Сервис ZooKeeper	619
Модель данных	619
Операции.	622
Реализация.	627
Согласованность данных	628
Сеансы.	630
Состояния	632
Построение приложений с использованием ZooKeeper	633
Конфигурация	633
Отказоустойчивое приложение ZooKeeper	637

Блокировка.	641
Другие распределенные структуры данных и протоколы	643
Практическое использование ZooKeeper.	644
Надежность и производительность	644
Конфигурация.	645
Глава 15. Sqoop.	647
Установка и запуск Sqoop	647
Коннекторы Sqoop	649
Пример импортирования.	649
Текстовые и двоичные форматы	652
Сгенерированный код	653
Другие системы сериализации	653
Подробнее об импортировании	654
Управление импортированием	656
Импортирование и согласованность данных	656
Прямое импортирование.	657
Работа с импортированными данными	657
Импортирование данных и Hive.	658
Импортирование больших объектов.	661
Экспортирование	663
Подробнее об экспортировании	665
Экспортирование и транзакционность.	666
Экспортирование и SequenceFile.	667