

Большие данные

Современные фреймворки
и разработка приложений

Д. Талия, П. Трунфио, Ф. Мароццо,
А. Белькастро, Р. Кантини, А. Орсино



Содержание

От издательства	9
Предисловие	10
Об авторах	11
Признательности	13
 Глава 1. Введение	 14
1.1. Мотивация и цели	14
1.2. Главные темы	16
1.3. Аудитория и организация	17
1.4. Онлайн-ресурсы	18
 Глава 2. Концепции больших данных	 19
2.1. Принципы и характеристики больших данных	19
2.2. Концепции науки о данных	22
2.2.1. Процессы науки о данных	24
2.2.2. Навыки, специфичные для науки о данных	26
2.3. Хранение больших данных	28
2.3.1. MongoDB	30
2.3.2. Google Bigtable	32
2.3.3. HBase	33
2.3.4. Redis	34
2.3.5. DynamoDB	35
2.3.6. Apache Cassandra	36
2.3.7. Графовая система управления базами данных Neo4j	37
2.3.8. Сводные соображения о хранилищах NoSQL	38
2.4. Масштабируемый анализ данных	41
2.5. Параллельные вычисления	45
2.5.1. Базовые понятия и определения	45
2.5.2. Параллельные архитектуры	46
2.5.3. Аппаратные платформы	48
2.5.4. Метрики производительности	50
2.6. Облачные вычисления	51
2.6.1. Модели распределения и развертывания облачных услуг	52
2.6.2. Облачные услуги для больших данных	54
2.7. На пути к эксафлопсным вычислениям	59

2.7.1. Главные трудности систем экзафлопсного масштаба	61
2.8. Параллельное и распределенное машинное обучение	63
2.8.1. Стратегии параллельного обучения	64
2.8.2. Стратегии распределенного обучения	67

Глава 3. Модели программирования для больших данных

3.1. Параллельное программирование для приложений по обработке больших данных	72
3.1.1. Необходимость в моделях параллельного программирования	73
3.1.2. Характеристики моделей программирования	73
3.2. Модель на основе MapReduce	74
3.2.1. Ключевые идеи, положенные в основу MapReduce	75
3.2.2. Модель программирования	75
3.2.3. Программы MapReduce	76
3.2.4. Применения и соображения о производительности	79
3.3. Модель на основе рабочего потока	81
3.3.1. Шаблоны рабочих потоков	82
3.3.2. Ориентированные ациклические графы	84
3.4. Модель на основе передачи сообщений	85
3.4.1. От коллективной памяти к передаче сообщений	86
3.4.2. Примитивы передачи сообщений	87
3.4.3. Групповая коммуникация	89
3.5. Модель на основе массового синхронного параллелизма	90
3.5.1. Супершаг	91
3.5.2. Стоимость алгоритма массового синхронного параллелизма	93
3.5.3. Модель на основе массового синхронного параллелизма с коллективной памятью	94
3.6. SQL-подобная модель	96
3.6.1. От модели NoSQL к SQL-подобной модели	96
3.6.2. Зачем использовать язык SQL на больших данных?	97
3.6.3. Разбиение данных на разделы	99
3.7. Модель на основе разделенного глобального адресного пространства	99
3.7.1. Параллелизм в модели на основе разделенного глобального адресного пространства	101
3.7.2. Память и функция стоимости	101
3.7.3. Распределение данных по местам	102
3.8. Модели для систем экзафлопсного масштаба	102
3.8.1. Роль моделей программирования в системах экзафлопсного масштаба	103
3.8.2. Требования к моделям экзафлопсного масштаба	104
3.8.3. Ограничения существующих моделей программирования	104
3.8.4. Модели программирования для систем экзафлопсного масштаба	106

Глава 4. Инструменты обработки больших данных

4.1. Главные характеристики	109
4.2. Инструменты программирования на основе модели MapReduce	110

4.2.1. Apache Hadoop	111
4.3. Инструменты программирования на основе рабочих потоков	120
4.3.1. Apache Spark	121
4.3.2. Apache Storm	134
4.3.3. Apache Airflow	145
4.4. Инструменты программирования на основе передачи сообщений	154
4.4.1. Интерфейс передачи сообщений	154
4.5. Инструменты программирования на основе массового синхронного параллелизма	161
4.5.1. Spark GraphX	161
4.6. Инструменты SQL-подобного программирования	169
4.6.1. Apache Hive	170
4.6.2. Apache Pig	176
4.7. Инструменты программирования на основе разделенного глобального адресного пространства	182
4.7.1. UPC++	183

Глава 5. Сравнение инструментов программирования 190

5.1. Перед проведением анализа инструментов	190
5.2. Сравнительный анализ характеристик систем	191
5.2.1. Характеристики систем	191
5.2.2. Распространенность систем	195
5.2.3. Преимущества и недостатки	197
5.3. Сравнительный анализ на примерах приложений	199
5.3.1. Пакетное приложение: Apache Spark в сопоставлении с Apache Hadoop	199
5.3.2. Потокное приложение: Apache Storm в сопоставлении с Apache Spark Streaming	212
5.3.3. Приложение на основе SQL: Apache Hive в сопоставлении с Apache Spark SQL	221
5.3.4. Графовое приложение: MPI в сопоставлении с Apache Spark GraphX	229

Глава 6. Выбор правильного фреймворка для приручения

больших данных	243
6.1. Входные данные	244
6.2. Класс приложения	246
6.3. Инфраструктура	248
6.4. Другие факторы	251

Сопутствующие материалы 253

Библиография 254

Предметный указатель 262