

УДК 004.6
ББК 32.972.13
П86

П86 Эндрю Дж. Пселтис

Потоковая обработка данных. Конвейер реального времени / пер. с англ.
А. А. Слинкин – М.: ДМК Пресс, 2018. – 218 с.: ил.

ISBN 978-5-97060-606-3

Эта насыщенная идеями книга научит вас думать об эффективном взаимодействии с быстрыми потоками данных. В ней выдержан идеальный баланс между широкой картиной и деталями реализации. На содержательных примерах и практических задачах вы узнаете о проектировании приложений, которые читают, анализируют, разделяют и сохраняют потоковые данные. Попутно вы поймете, какую роль играют такие технологии, как Spark, Storm, Kafka, Flink, RabbitMQ и другие.

Издание ориентировано на разработчиков, знакомых с концепциями реляционных баз данных.

УДК 004.6
ББК 32.972.13

Original English language edition published by Manning Publications USA, USA.
Copyright © 2017 by Manning Publications. Russian language edition copyright © 2016 by DMK Press. 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 978-1-61729-228-6 (англ.)
ISBN 978-5-97060-606-3 (рус.)

Copyright © 2017 by Manning Publications Co.
© Оформление, перевод на русский язык,
издание, ДМК Пресс, 2018

Оглавление

Архитектурная диаграмма потоковой обработки данных	5
Предисловие	9
Благодарности	11
Об этой книге	12
Как работать с книгой?	12
Кому стоит прочитать эту книгу?	12
Структура книги	13
О коде	14
Об авторе	14
Автор в сети	15
Об иллюстрации на обложке	15
Часть I. Новый целостный подход	17
Глава 1. Введение в потоковую обработку данных	19
1.1. Что такое система реального времени?	20
1.2. Различия между системами реального времени и потоковыми системами	23
1.3. Архитектурная диаграмма	25
1.4. Безопасность в контексте потоковых систем	27
1.5. Как производится масштабирование?	27
1.6. Резюме	29
Глава 2. Получение данных от клиентов: внесение данных	31
2.1. Типичные паттерны взаимодействия	31
2.1.1. Запрос-ответ	32
2.1.2. Паттерн запрос-подтверждение	36
2.1.3. Паттерн издатель-подписчик	37
2.1.4. Паттерн одностороннего взаимодействия	39
2.1.5. Паттерн поток	40
2.2. Масштабирование паттернов взаимодействия	42
2.2.1. Паттерны запрос-ответ	43
2.2.2. Масштабирование паттерна поток	44
2.3. Отказоустойчивость	46
2.3.1. Протоколирование сообщений на стороне получателя	48
2.3.2. Протоколирование сообщений на стороне отправителя	51
2.3.3. Гибридное протоколирование сообщений	52
2.4. Опустимся на грешную землю	54
2.5. Резюме	55
Глава 3. Транспортировка данных из звена сбора данных: расчленение конвейера данных	56

3.1. Зачем нужно звено очереди сообщений	56
3.2. Основные концепции	58
3.2.1. Производитель, брокер и потребитель	59
3.2.2. Изоляция производителей от потребителей	61
3.2.3. Долговечные сообщения	62
3.2.4. Семантика доставки сообщений	65
3.3. Безопасность	69
3.4. Отказоустойчивость	70
3.5. Применение базовых концепций в конкретных задачах	73
3.6. Резюме	75
Глава 4. Анализ потоковых данных	77
4.1. Анализ данных в движении	77
4.2. Архитектуры распределенной обработки потоков	82
4.3. Ключевые функции систем потоковой обработки	88
4.3.1. Семантика доставки сообщений	89
4.4. Резюме	96
Глава 5. Алгоритмы анализа данных	97
5.1. Ограничения и их ослабление	98
5.2. К вопросу о времени	99
5.2.1. Скользящее окно	101
5.2.2. Прыгающие окна	103
5.3. Методы обобщения	106
5.3.1. Случайная выборка	106
5.3.2. Подсчет уникальных элементов	108
5.3.3. Частота	111
5.3.4. Вопрос о вхождении	113
5.4. Резюме	115
Глава 6. Сохранение результатов сбора или анализа данных	116
6.1. Когда нужно долговременное хранилище	118
6.2. Хранение данных в памяти	120
6.2.1. Встраиваемые хранилища в памяти с оптимизацией для флеш-памяти	121
6.2.2. Система кэширования	123
6.2.3. Базы данных и решетки данных в памяти	127
6.3. Примеры и упражнения	130
6.3.1. Сеансовая персонализация	130
6.3.2. Энергетическая компания следующего поколения	134
6.4. Резюме	135
Глава 7. Получение доступа к данным	136
7.1. Паттерны взаимодействия	137
7.1.1. Паттерн Data Sync	137
7.1.2. Удаленный вызов метода и удаленный вызов процедуры	139
7.1.3. Простой обмен сообщениями	140
7.1.4. Издатель-подписчик	141

7.2. Протоколы отправки данных клиентам	142
7.2.1. Веб-уведомления	143
7.2.2. Длинный HTTP-опрос	144
7.2.3. События, посылаемые сервером	146
7.2.4. Веб-сокеты	150
7.3. Фильтрация потока	154
7.3.1. Где производится фильтрация	154
7.3.2. Статическая и динамическая фильтрации	155
7.4. Пример: построение потокового API для сайта Meetup	156
7.5. Резюме	158
Глава 8. Возможности конечных устройств и ограничения	
доступа к данным	160
8.1. Основные концепции	162
8.1.1. Достаточная скорость чтения	163
8.1.2. Запоминание состояния	166
8.1.3. Смягчение последствий потери данных	168
8.1.4. Обработка ровно один раз	170
8.2. Все по-настоящему: компания SuperMediaMarkets	172
8.3. Введение в веб-клиент	176
8.3.1. Интеграция со службой потокового API	178
8.4. На пути к языку запросов	180
8.5. Резюме	181
Часть II. Потоки в реальном мире	182
Глава 9. Анализ приглашений Meetup.com в режиме реального	
времени	183
9.1. Звено сбора данных	185
9.1.1. Диаграмма последовательности службы сбора данных	185
9.2. Звено очереди сообщений	195
9.2.1. Установка и настройка Kafka	195
9.2.2. Интеграция службы сбора данных с Kafka	196
9.3. Звено анализа	198
9.3.1. Установка Storm и подготовка Kafka	199
9.3.2. Построение топологии Storm для нахождения <i>n</i> самых популярных тем	200
9.3.3. Интеграция звена анализа в конвейер	207
9.4. Хранилище данных в памяти	207
9.5. Звено доступа к данным	208
9.5.1. На пути к производственному режиму	213
9.6. Резюме	213
Предметный указатель	214