Contents

Introduction                          xv

Chapter 1   Introduction to Streaming Data     1
Sources of Streaming Data            2
  Operational Monitoring             3
Web Analytics                        3
Online Advertising                   4
Social Media                         5
Mobile Data and the Internet of Things  5
Why Streaming Data Is Different      7
  Always On, Always Flowing          7
  Loosely Structured                8
  High-Cardinality Storage          9
Infrastructures and Algorithms      10
Conclusion                           10

Part I   Streaming Analytics Architecture   13

Chapter 2   Designing Real-Time Streaming Architectures  15
Real-Time Architecture Components    16
  Collection                         16
  Data Flow                          17
  Processing                         19
  Storage                            20
  Delivery                           22
Features of a Real-Time Architecture 24
  High Availability                 24
  Low Latency                        25
  Horizontal Scalability             26
<table>
<thead>
<tr>
<th>Chapter 3</th>
<th>Service Configuration and Coordination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Motivation for Configuration and Coordination Systems</td>
</tr>
<tr>
<td></td>
<td>Maintaining Distributed State</td>
</tr>
<tr>
<td></td>
<td>Unreliable Network Connections</td>
</tr>
<tr>
<td></td>
<td>Clock Synchronization</td>
</tr>
<tr>
<td></td>
<td>Consensus in an Unreliable World</td>
</tr>
<tr>
<td></td>
<td>Apache ZooKeeper</td>
</tr>
<tr>
<td></td>
<td>The znode</td>
</tr>
<tr>
<td></td>
<td>Watches and Notifications</td>
</tr>
<tr>
<td></td>
<td>Maintaining Consistency</td>
</tr>
<tr>
<td></td>
<td>Creating a ZooKeeper Cluster</td>
</tr>
<tr>
<td></td>
<td>ZooKeeper’s Native Java Client</td>
</tr>
<tr>
<td></td>
<td>The Curator Client</td>
</tr>
<tr>
<td></td>
<td>Curator Recipes</td>
</tr>
<tr>
<td></td>
<td>Conclusion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>Data-Flow Management in Streaming Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distributed Data Flows</td>
</tr>
<tr>
<td></td>
<td>At Least Once Delivery</td>
</tr>
<tr>
<td></td>
<td>The “n+1” Problem</td>
</tr>
<tr>
<td></td>
<td>Apache Kafka: High-Throughput Distributed Messaging</td>
</tr>
<tr>
<td></td>
<td>Design and Implementation</td>
</tr>
<tr>
<td></td>
<td>Configuring a Kafka Environment</td>
</tr>
<tr>
<td></td>
<td>Interacting with Kafka Brokers</td>
</tr>
<tr>
<td></td>
<td>Apache Flume: Distributed Log Collection</td>
</tr>
<tr>
<td></td>
<td>The Flume Agent</td>
</tr>
<tr>
<td></td>
<td>Configuring the Agent</td>
</tr>
<tr>
<td></td>
<td>The Flume Data Model</td>
</tr>
<tr>
<td></td>
<td>Channel Selectors</td>
</tr>
<tr>
<td></td>
<td>Flume Sources</td>
</tr>
<tr>
<td></td>
<td>Flume Sinks</td>
</tr>
<tr>
<td></td>
<td>Sink Processors</td>
</tr>
<tr>
<td></td>
<td>Flume Channels</td>
</tr>
</tbody>
</table>
## Part II Analysis and Visualization 225

### Chapter 7 Delivering Streaming Metrics 227

- Streaming Web Applications 228
  - Working with Node 229
  - Managing a Node Project with NPM 231
  - Developing Node Web Applications 235
  - A Basic Streaming Dashboard 238
  - Adding Streaming to Web Applications 242
- Visualizing Data 254
  - HTML5 Canvas and Inline SVG 254
  - Data-Driven Documents: D3.js 262
  - High-Level Tools 272
- Mobile Streaming Applications 277
- Conclusion 279

### Chapter 8 Exact Aggregation and Delivery 281

- Timed Counting and Summation 285
  - Counting in Bolts 286
  - Counting with Trident 288
  - Counting in Samza 289
- Multi-Resolution Time-Series Aggregation 290
  - Quantization Framework 290
- Stochastic Optimization 296
- Delivering Time-Series Data 297
  - Strip Charts with D3.js 298
  - High-Speed Canvas Charts 299
  - Horizon Charts 301
- Conclusion 303

### Chapter 9 Statistical Approximation of Streaming Data 305

- Numerical Libraries 306
- Probabilities and Distributions 307
  - Expectation and Variance 309
  - Statistical Distributions 310
  - Discrete Distributions 310
  - Continuous Distributions 312
  - Joint Distributions 315
- Working with Distributions 316
  - Inferring Parameters 316
  - The Delta Method 317
  - Distribution Inequalities 319
- Random Number Generation 319
  - Generating Specific Distributions 321
Contents

Sampling Procedures 324
  Sampling from a Fixed Population 325
  Sampling from a Streaming Population 326
  Biased Streaming Sampling 327
Conclusion 329

Chapter 10  Approximating Streaming Data with Sketching 331
  Registers and Hash Functions 332
    Registers 332
    Hash Functions 332
  Working with Sets 336
  The Bloom Filter 338
    The Algorithm 338
    Choosing a Filter Size 340
    Unions and Intersections 341
    Cardinality Estimation 342
    Interesting Variations 344
  Distinct Value Sketches 347
    The Min-Count Algorithm 348
    The HyperLogLog Algorithm 351
  The Count-Min Sketch 356
    Point Queries 356
    Count-Min Sketch Implementation 357
    Top-K and “Heavy Hitters” 358
    Range and Quantile Queries 360
  Other Applications 364
Conclusion 364

Chapter 11  Beyond Aggregation 367
  Models for Real-Time Data 368
    Simple Time-Series Models 369
    Linear Models 373
    Logistic Regression 378
    Neural Network Models 380
  Forecasting with Models 389
    Exponential Smoothing Methods 390
    Regression Methods 393
    Neural Network Methods 394
  Monitoring 396
    Outlier Detection 397
    Change Detection 399
    Real-Time Optimization 400
Conclusion 402

Index 403