## Contents

Acknowledgements \hspace{1em} page x

Dedication  \hspace{1em} xi

Preface  \hspace{1em} xii

### Part I  Fundamentals  \hspace{1em} 1

1 What Brought Us Here?  \hspace{1em} 3

1.1 Overview  \hspace{1em} 3

1.2 Towards continuous data processing: the requirements  \hspace{1em} 3

1.3 Stream processing foundations  \hspace{1em} 7

1.3.1 Data management technologies  \hspace{1em} 8

1.3.2 Parallel and distributed systems  \hspace{1em} 15

1.3.3 Signal processing, statistics and data mining  \hspace{1em} 17

1.3.4 Optimization theory  \hspace{1em} 20

1.4 Stream processing – tying it all together  \hspace{1em} 24

2 Introduction to Stream Processing  \hspace{1em} 37

2.1 Overview  \hspace{1em} 37

2.2 Stream processing applications  \hspace{1em} 38

2.2.1 Network monitoring for cybersecurity  \hspace{1em} 39

2.2.2 Transportation grid monitoring and optimization  \hspace{1em} 41

2.2.3 Healthcare and patient monitoring  \hspace{1em} 43

2.2.4 Discussion  \hspace{1em} 45

2.3 Information flow processing technologies  \hspace{1em} 46

2.3.1 Active databases  \hspace{1em} 46

2.3.2 Continuous queries  \hspace{1em} 47

2.3.3 Publish/subscribe systems  \hspace{1em} 48

2.3.4 Complex event processing systems  \hspace{1em} 49

2.3.5 ETL and SCADA systems  \hspace{1em} 50

2.4 Stream processing systems  \hspace{1em} 50

2.4.1 Data  \hspace{1em} 51

2.4.2 Processing  \hspace{1em} 55

2.4.3 System architecture  \hspace{1em} 60

2.4.4 Implementations  \hspace{1em} 64
2.4.4.1 Academic systems 64
2.4.4.2 Commercial systems 68
2.4.4.3 Open-source systems 72
2.4.5 Discussion 73
2.5 Concluding remarks 76
2.6 Exercises 77

Part II Application Development 85

3 Application Development - The Basics 87
3.1 Overview 87
3.2 Characteristics of stream processing applications 87
3.3 Stream processing languages 91
  3.3.1 Features of stream processing languages 91
  3.3.2 Approaches to stream processing language design 94
3.4 Introduction to SPL 98
  3.4.1 Language origins 98
  3.4.2 A “Hello World” application in SPL 99
3.5 Common stream processing operators 104
  3.5.1 Stream relational operators 104
  3.5.2 Utility operators 108
  3.5.3 Adapter operators 109
3.6 Concluding remarks 113
3.7 Programming exercises 114

4 Application Development - Data Flow Programming 119
4.1 Overview 119
4.2 Flow composition 119
  4.2.1 Static composition 121
  4.2.2 Dynamic composition 126
  4.2.3 Nested composition 137
4.3 Flow manipulation 143
  4.3.1 Operator state 144
  4.3.2 Selectivity and arity 147
  4.3.3 Using parameters 148
  4.3.4 Output assignments and output functions 150
  4.3.5 Punctuations 152
  4.3.6 Windowing 154
4.4 Concluding remarks 161
4.5 Programming exercises 162

5 Large-Scale Development - Modularity, Extensibility, and Distribution 167
5.1 Overview 167
5.2 Modularity and extensibility 167
9.3 Non-functional principles and design patterns 344
  9.3.1 Application design and composition 344
    9.3.1.1 Principles of application design and composition 345
  9.3.2 Parallelization 349
    9.3.2.1 Principles of parallelization 351
    9.3.2.2 Patterns of parallelization 356
  9.3.3 Performance optimization 361
    9.3.3.1 Principles of optimization 361
    9.3.3.2 Patterns of optimization 366
  9.3.4 Fault-tolerance 370
    9.3.4.1 Principles of fault tolerance 372

9.4 Concluding remarks 376

10 Stream Processing and Mining Algorithms: Data Pre-processing and Transformation 381
  10.1 Overview 381
  10.2 The mining process 382
  10.3 Notation 383
  10.4 Descriptive statistics 384
    10.4.1 Illustrative technique: BasicCounting 387
    10.4.2 Advanced reading 392
  10.5 Sampling 393
    10.5.1 Illustrative technique: reservoir sampling 395
    10.5.2 Advanced reading 397
  10.6 Sketches 398
    10.6.1 Illustrative technique: Count-Min sketch 401
    10.6.2 Advanced reading 403
  10.7 Quantization 404
    10.7.1 Illustrative techniques: binary clipping and moment preserving quantization 407
    10.7.2 Advanced reading 410
  10.8 Dimensionality reduction 410
    10.8.1 Illustrative technique: SPIRIT 413
    10.8.2 Advanced reading 416
  10.9 Transforms 416
    10.9.1 Illustrative technique: the Haar transform 421
    10.9.2 Advanced reading 424
  10.10 Concluding remarks 425

11 Stream Processing and Mining Algorithms: Modeling and Evaluation 431
  11.1 Overview 431
  11.2 Offline modeling and online evaluation 432
  11.3 Data stream classification 438
    11.3.1 Illustrative technique: VFDT 442
11.3.2 Advanced reading 447
11.4 Data stream clustering 448
   11.4.1 Illustrative technique: CluStream microclustering 454
   11.4.2 Advanced reading 458
11.5 Data stream regression 459
   11.5.1 Illustrative technique: linear regression with SGD 463
   11.5.2 Advanced Reading 465
11.6 Data stream frequent pattern mining 466
   11.6.1 Illustrative technique: lossy counting 471
   11.6.2 Advanced reading 474
11.7 Anomaly detection 474
   11.7.1 Illustrative technique: micro-clustering-based anomaly detection 479
   11.7.2 Advanced reading 480
11.8 Concluding remarks 480

Part V Case Studies 485

12 Applications 487
   12.1 Overview 487
   12.2 The Operations Monitoring application 488
      12.2.1 Motivation 488
      12.2.2 Requirements 489
      12.2.3 Design 491
      12.2.4 Analytics 499
      12.2.5 Fault tolerance 500
   12.3 The Patient Monitoring application 500
      12.3.1 Motivation 501
      12.3.2 Requirements 501
      12.3.3 Design 503
      12.3.4 Evaluation 510
   12.4 The Semiconductor Process Control application 515
      12.4.1 Motivation 515
      12.4.2 Requirements 517
      12.4.3 Design 520
      12.4.4 Evaluation 527
      12.4.5 User interface 529
   12.5 Concluding remarks 531

Part VI Closing Notes 535

13 Conclusion 537
   13.1 Book summary 537
   13.2 Challenges and open problems 538
13.2.1 Software engineering 539
13.2.2 Scaling up and distributed computing 541
13.2.3 Integration 543
13.2.4 Analytics 545
13.3 Where do we go from here? 547

_Acronyms_ 551
_General Index_ 557
_Keywords and Identifiers_ 585