Table of Contents

Preface 1

Chapter 1: Defining IoT Analytics and Challenges 9

- **The situation** 9
- **Defining IoT analytics** 13
  - Defining analytics 13
  - Defining the Internet of Things 15
  - The concept of constrained 16
- **IoT analytics challenges** 16
  - The data volume 17
  - Problems with time 19
  - Problems with space 21
  - Data quality 22
  - Analytics challenges 23
- **Business value concerns** 24
- **Summary** 24

Chapter 2: IoT Devices and Networking Protocols 25

- **IoT devices** 26
  - The wild world of IoT devices 26
    - Healthcare 26
    - Manufacturing 26
    - Transportation and logistics 27
    - Retail 28
    - Oil and gas 28
    - Home automation or monitoring 28
    - Wearables 29
  - Sensor types 29
- **Networking basics** 30
- **IoT networking connectivity protocols** 32
  - Connectivity protocols (when the available power is limited) 32
    - Bluetooth Low Energy (also called Bluetooth Smart) 33
    - 6LoWPAN 34
    - ZigBee 37
      - Advantages of ZigBee 39
      - Disadvantages of ZigBee 39
      - Common use cases 39
    - NFC 40

Common use cases 41
Sigfox 41
Connectivity protocols (when power is not a problem) 42
Wi-Fi 42
Common use cases 43
Cellular (4G/LTE) 44
Common use cases 45
IoT networking data messaging protocols 45
Message Queue Telemetry Transport (MQTT) 46
Topics 47
Advantages to MQTT 48
Disadvantages to MQTT 50
QoS levels 50
QoS 0 51
QoS 1 51
QoS 2 52
Last Will and Testament (LWT) 53
Tips for analytics 54
Common use cases 54
Hyper-Text Transport Protocol (HTTP) 54
Representational State Transfer (REST) principles 55
HTTP and IoT 56
Advantages to HTTP 56
Disadvantages to HTTP 56
Constrained Application Protocol (CoAP) 56
Advantages to CoAP 58
Disadvantages to CoAP 59
Message reliability 59
Common use cases 59
Data Distribution Service (DDS) 60
Common use cases 61
Analyzing data to infer protocol and device characteristics 62
Summary 64
Chapter 3: IoT Analytics for the Cloud 65
Building elastic analytics 66
What is cloud infrastructure? 66
Elastic analytics concepts 68
Design with the endgame in mind 70
Designing for scale 70
Decouple key components 70
Encapsulate analytics 70
Decoupling with message queues 71
Distributed computing 74
Avoid containing analytics to one server 74

When to use distributed and when to use one server 74
Assuming that change is constant 75
Leverage managed services 75
Use Application Programming Interfaces (API) 77
Cloud security and analytics 78
Public/private keys 78
Public versus private subnets 78
Access restrictions 79
Securing customer data 79
The AWS overview 80
AWS key concepts 82
Regions 82
Availability Zones 82
Subnet 83
Security groups 83
AWS key core services 83
Virtual Private Cloud (VPC) 83
Identity and Access Management (IAM) 85
Elastic Compute (EC2) 85
Simple Storage Service (S3) 86
AWS key services for IoT analytics 86
Amazon Simple Queue Service (SQS) 87
Amazon Elastic Map Reduce (EMR) 87
AWS machine learning 88
Amazon Relational Database Service (RDS) 89
Amazon Redshift 89
Microsoft Azure overview 89
Azure Data Lake Store 89
Azure Analysis Services 90
HDInsight 91
The R server option 91
The ThingWorx overview 92
ThingWorx Core 93
ThingWorx Connection Services 93
ThingWorx Edge 94
ThingWorx concepts 95
Thing templates 95
Things 95
Properties 96
Services 96
Events 96
Thing shapes 97
Data shapes 97

Entities 97

Summary 97

Chapter 4: Creating an AWS Cloud Analytics Environment 99

The AWS CloudFormation overview 99

The AWS Virtual Private Cloud (VPC) setup walk-through 101

Creating a key pair for the NAT and bastion instances 103

Creating an S3 bucket to store data 105

Creating a VPC for IoT Analytics 107

What is a NAT gateway? 107

What is a bastion host? 108

Your VPC architecture 108

The VPC Creation walk-through 110

How to terminate and clean up the environment 119

Summary 122

Chapter 5: Collecting All That Data - Strategies and Techniques 123

Designing data processing for analytics 124

Amazon Kinesis 124

AWS Lambda 125

AWS Athena 125

The AWS IoT platform 126

Microsoft Azure IoT Hub 127

Applying big data technology to storage 128

Hadoop 128

Hadoop cluster architectures 131

What is a Node? 132

Node types 132

Hadoop Distributed File System 133

Parquet 135

Avro 138

Hive 139

Serialization/Deserialization (SerDe) 141

Hadoop MapReduce 142

Yet Another Resource Negotiator (YARN) 142

HBase 144

Amazon DynamoDB 144

Amazon S3 144

Apache Spark for data processing 145

What is Apache Spark? 145

Spark and big data analytics 146

Thinking about a single machine versus a cluster of machines 147

Using Spark for IoT data processing 148

To stream or not to stream 150
Lambda architectures 151
Handling change 152
Summary 153
Chapter 6: Getting to Know Your Data - Exploring IoT Data 155
Exploring and visualizing data 157
The Tableau overview 157
Techniques to understand data quality 159
Look at your data - au naturel 160
Data completeness 161
Data validity 167
Assessing Information Lag 169
Representativeness 170
Basic time series analysis 170
What is meant by time series? 171
Applying time series analysis 171
Get to know categories in the data 176
Bring in geography 176
Look for attributes that might have predictive value 178
R (the pirate's language...if he was a statistician) 178
Installing R and RStudio 178
Using R for statistical analysis 179
Summing it all up 183
Solving industry-specific analysis problems 184
Manufacturing 184
Healthcare 185
Retail 186
Summary 186
Chapter 7: Decorating Your Data - Adding External Datasets to Innovate 187
Adding internal datasets 188
Which ones and why? 189
Customer information 189
Production data 189
Field services 190
Financial 190
Adding external datasets 190
External datasets - geography 191
Elevation 191
SRTM elevation 191
National Elevation Dataset (NED) 192

Weather 193
Geographical features 194
Planet.osm 195
Google Maps API 196
USGS national transportation datasets 197
External datasets - demographic 198
The U.S. Census Bureau 198
CIA World Factbook 199
External datasets - economic 200
Organization for Economic Cooperation and Development (OECD) 200
Federal Reserve Economic Data (FRED) 202
Summary 203
Chapter 8: Communicating with Others - Visualization and Dashboarding 205
Common mistakes when designing visuals 207
The Hierarchy of Questions method 210
The Hierarchy of Questions method overview 211
Developing question trees 212
Pulling together the data 216
Aligning views with question flows 216
Designing visual analysis for IoT data 216
Using layout positioning to convey importance 217
Use color to highlight important data 217
The impact of using a single color to communicate importance 218
Be consistent across visuals 219
Make charts easy to interpret 219
Creating a dashboard with Tableau 220
The dashboard walk-through 220
Hierarchy of Questions example 220
Aligning visuals to the thought process 221
Creating individual views 222
Assembling views into a dashboard 226
Creating and visualizing alerts 228
Alert principles 229
Organizing alerts using a Tableau dashboard 229
Summary 233
Chapter 9: Applying Geospatial Analytics to IoT Data 235
Why do you need geospatial analytics for IoT? 237
The basics of geospatial analysis 239
Welcome to Null Island 239
Coordinate Reference Systems 240

    - The Earth is not a ball 240
- **Vector-based methods** 243
  - The bounding box 245
  - Contains 246
  - Buffer 247
    - Dilation and erosion 248
  - Simplify 249
  - Vector summary 250
- **Raster-based methods** 250
- **Storing geospatial data** 251
  - File formats 251
  - Spatial extensions for relational databases 253
  - Storing geospatial data in HDFS 253
  - Spatial indexing 254
    - R-tree 254
- **Processing geospatial data** 256
  - Geospatial analysis software 256
    - ArcGIS 256
    - QGIS 257
    - ogr2ogr 258
  - PostGIS spatial functions 259
  - Geospatial analysis in the big data world 260
- **Solving the pollution reporting problem** 261
- **Summary** 262

Chapter 10: Data Science for IoT Analytics 263

- **Machine learning (ML)** 264
  - What is machine learning? 265
    - Representation 267
    - Evaluation 267
    - Optimization 268
  - Generalization 268
  - Feature engineering with IoT data 269
    - Dealing with missing values 270
    - Centering and scaling 275
    - Time series handling 276
  - Validation methods 277
    - Cross-validation 277
    - Test set 279
    - Precision, recall, and specificity 279
  - Understanding the bias–variance tradeoff 281
    - Bias 282
    - Variance 283

Trade-off and complexity 283
Comparing different models to find the best fit using R 284
ROC curves 285
Area Under the Curve (AUC) 289
Random forest models using R 290
Random forest key concepts 291
Random forest R examples 292
Gradient Boosting Machines (GBM) using R 294
GBM key concepts 295
The Gradient Boosting Machines R example 296
Ensemble 297
Anomaly detection using R 298
Forecasting using ARIMA 299
Using R to forecast time series IoT data 300
Deep learning 301
Use cases for deep learning with IoT data 302
A Nickel Tour of deep learning 302
Setting up TensorFlow on AWS 304
Summary 304
Chapter 11: Strategies to Organize Data for Analytics 305
Linked Analytical Datasets 306
Analytical datasets 306
Building analytic datasets 307
Linking together datasets 309
Managing data lakes 312
When data lakes turn into data swamps 312
Data refineries 312
Developing a progression process 313
The data retention strategy 315
Goals 315
Retention strategies for IoT data 315
Reducing accessibility 316
Reducing the number of fields 316
Reduce the number of records 317
The retention strategy example 318
Summary 318
Chapter 12: The Economics of IoT Analytics 319
The economics of cloud computing and open source 320
Variable versus fixed costs 320
The option to quit 321
Cloud costs can escalate quickly 322

Monitoring cloud billing closely 323
Open source economics 323
Intellectual property considerations 323
Scale 324
Support 324
Cost considerations for IoT analytics 325
Cloud services costs 325
Expected usage considerations 325
Thinking about revenue opportunities 325
The economics of predictive maintenance example 328
Situation 328
The value formula 329
An example of making a value decision 329
Summary 337
Chapter 13: Bringing It All Together 339
Review 340
The IoT data flow 341
IoT exploratory analytics 343
IoT data science 344
Building revenue from IoT analytics 346
A sample project 347
Summary 348
Index 349