

Contents

Website	xi
Acknowledgments	xiii
Preface	xv
1 Orienting in the Cloud Universe	1
1.1 Cloud: Computer, Assistant, and Platform	1
1.2 The Cloud Landscape	3
1.3 A Guide to This Book	7
1.4 Accessing the Cloud: Web, APIs, and SDKs	8
1.5 Tools Used in This Book	12
1.6 Summary	15
1.7 Resources	16
 Part I. Managing Data in the Cloud	 17
2 Storage as a Service	21
2.1 Three Motivating Examples	22
2.2 Storage Models	23
2.3 The Cloud Storage Landscape	29
2.4 Summary	35
2.5 Resources	36
 3 Using Cloud Storage Services	 37
3.1 Two Access Methods: Portals and APIs	37
3.2 Using Amazon Cloud Storage Services	38
3.3 Using Microsoft Azure Storage Services	42

3.4	Using Google Cloud Storage Services	46
3.5	Using OpenStack Cloud Storage Services	50
3.6	Transferring and Sharing Data with Globus	51
3.7	Summary	56
3.8	Resources	57

Part II. Computing in the Cloud 59

Computing as a Service 63

4.1	Virtual Machines and Containers	64
4.2	Advanced Computing Services	66
4.3	Serverless Computing	67
4.4	Pros and Cons of Public Cloud Computing	68
4.5	Summary	70
4.6	Resources	71

Using and Managing Virtual Machines 73

5.1	Historical Roots	74
5.2	Amazon's Elastic Compute Cloud	75
5.3	Azure VMs	80
5.4	Google Cloud VM Services	82
5.5	Jetstream VM Services	82
5.6	Summary	83
5.7	Resources	84

Using and Managing Containers 85

6.1	Container Basics	85
6.2	Docker and the Hub	87
6.3	Containers for Science	90
6.4	Creating Your Own Container	91
6.5	Summary	92
6.6	Resources	93

Scaling Deployments 95

7.1	Paradigms of Parallel Computing in the Cloud	96
7.2	SPMD and HPC-style Parallelism	97
7.3	Many Task Parallelism	107
7.4	MapReduce and Bulk Synchronous Parallelism	108
7.5	Graph Dataflow Execution and Spark	109

7.6	Agents and Microservices	110
7.7	HTCondor	128
7.8	Summary	128
7.9	Resources	129

Part III. The Cloud as Platform 131

8 Data Analytics in the Cloud 135

8.1	Hadoop and YARN	136
8.2	Spark	137
8.3	Amazon Elastic MapReduce	143
8.4	Azure HDInsight and Data Lake	147
8.5	Amazon Athena Analytics	149
8.6	Google Cloud Datalab	150
8.7	Summary	157
8.8	Resources	158

9 Streaming Data to the Cloud 161

9.1	Scientific Stream Examples	162
9.2	Basic Design Challenges of Streaming Systems	166
9.3	Amazon Kinesis and Firehose	167
9.4	Kinesis, Spark, and the Array of Things	170
9.5	Streaming Data with Azure	175
9.6	Kafka, Storm, and Heron Streams	180
9.7	Google Dataflow and Apache Beam	184
9.8	Apache Flink	187
9.9	Summary	189
9.10	Resources	190

10 Machine Learning in the Cloud 191

10.1	Spark Machine Learning Library (MLlib)	192
10.2	Azure Machine Learning Workspace	197
10.3	Amazon Machine Learning Platform	202
10.4	Deep Learning: A Shallow Introduction	204
10.5	Amazon MXNet Virtual Machine Image	212
10.6	Google TensorFlow in the Cloud	215
10.7	Microsoft Cognitive Toolkit	218
10.8	Summary	220
10.9	Resources	222

The Globus Research Data Management Platform	225
11.1 Challenges and Opportunities of Distributed Data	226
11.2 The Globus Platform	226
11.3 Identity and Credential Management	230
11.4 Building a Remotely Accessible Service	240
11.5 The Research Data Portal Design Pattern	243
11.6 The Portal Design Pattern Revisited	250
11.7 Closing the Loop: From Portal to Graph Service	252
11.8 Summary	255
11.9 Resources	255

Part IV. Building Your Own Cloud 257

Building Your Own Cloud with Eucalyptus (with Rich Wolski)	261
12.1 Implementing Cloud Infrastructure Abstractions	262
12.2 Deployment Planning	263
12.3 Single-cluster Eucalyptus Cloud	267
12.4 Summary	281
12.5 Resources	281

Building Your Own Cloud with OpenStack (with Stig Telfer)	283
13.1 OpenStack Core Services	284
13.2 HPC in an OpenStack Environment	284
13.3 Considerations for Scientific Workloads	285
13.4 OpenStack Deployment	288
13.5 Example Deployment	289
13.6 Summary	295
13.7 Resources	296

Building Your Own SaaS	297
14.1 The Meaning of SaaS	298
14.2 SaaS Architecture	299
14.3 SaaS and Science	301
14.4 The Globus Genomics Bioinformatics System	303
14.5 The Globus Research Data Management Service	307
14.6 Summary	310
14.7 Resources	310

Part V. Security and Other Topics 311

15 Security and Privacy	315
15.1 Thinking about Security in the Cloud	315
15.2 Role-based Access Control	319
15.3 Secure Data in the Cloud	320
15.4 Secure Your VMs and Containers	324
15.5 Secure Access to Cloud Software Services	327
15.6 Summary	327
15.7 Resources	328

16 History, Critiques, Futures	329
16.1 Historical Perspectives	329
16.2 Critiques	332
16.3 Futures	335
16.4 Resources	339

17 Jupyter Notebooks	341
17.1 Environment	342
17.2 The Notebooks	342
17.3 Resources	344

18 Afterword: A Discovery Cloud 345

Bibliography 347

Index 367