

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Big Data Problem	1
1.2	A Star Is Born: The MapReduce/Hadoop Framework	4
1.3	From Big Data to Big Graphs	9
1.4	Use Cases for Big Graphs	11
1.4.1	Social Networks	11
1.4.2	Web Graph	11
1.4.3	Knowledge Bases and Linked Data	14
1.4.4	Road Networks and Location-Based Services	16
1.4.5	Chemical and Biological Networks	17
1.5	Graph Databases	17
1.6	Does Hadoop Work Well for Big Graph Processing?	19
1.7	BSP Programming Model and Google Pregel	22
1.8	Pregel Extensions	24
1.9	Giraph: BSP + Hadoop for Graph Processing	27
1.10	Book Roadmap	31
1.11	How to Use the Code Examples of This Book?	32
<b>2</b>	<b>Installing and Getting Giraph Ready to Use</b>	<b>35</b>
2.1	Installing Hadoop	35
2.1.1	Single-Node Local Mode Installation	36
2.1.2	Single-Node Pseudo-Distributed Installation	38
2.1.3	Multi-node Cluster Installation	43
2.2	Monitoring Hadoop	46
2.2.1	Hadoop Web User Interfaces	46
2.3	Installing Giraph	50
2.4	Installing Giraph from Source Code	51
2.5	Running an Example Giraph Job	53
2.5.1	Hadoop Local Mode	53
2.5.2	Pseudo-Distributed and Clustered Hadoop Mode	55
2.6	Monitoring Giraph Application Life Cycle	56
2.6.1	MasterObserver	57
2.6.2	WorkerObserver	57

2.7	Monitoring Giraph Jobs . . . . .	57
2.7.1	Using Hadoop Commands to Monitor Giraph Jobs. . . . .	58
2.7.2	Using Hadoop UI to Monitor Giraph Jobs . . . . .	62
2.8	Configuring Giraph. . . . .	62
2.8.1	Giraph-Specific Options . . . . .	62
2.8.2	Job-Specific Options . . . . .	63
2.8.3	Algorithm-Specific Options. . . . .	65
2.8.4	Giraph Configuration Summary. . . . .	66
2.9	Creating a Giraph IDE Project . . . . .	66
2.9.1	Eclipse . . . . .	68
2.9.2	Eclipse with Maven. . . . .	73
2.9.3	IntelliJ IDE . . . . .	76
2.9.4	IntelliJ IDE with Maven . . . . .	79
2.10	Debugging Local Zookeeper. . . . .	81
<b>3</b>	<b>Getting Started with Giraph Programming</b> . . . . .	<b>87</b>
3.1	Giraph Graph Model. . . . .	87
3.1.1	The Edge Interface. . . . .	87
3.1.2	The Vertex Interface . . . . .	88
3.1.3	The Computation Interface. . . . .	91
3.2	Vertex Similarity Algorithm . . . . .	93
3.2.1	Implementation in Giraph . . . . .	95
3.3	Writing a Basic Giraph Job . . . . .	98
3.4	Writing the Driver Program . . . . .	99
3.4.1	Using main Method. . . . .	101
3.4.2	Using Tool Interface . . . . .	101
3.4.3	Using GiraphRunner Class . . . . .	103
3.5	Preparing Graph Data for Giraph Input . . . . .	105
3.5.1	Hadoop Distributed File System . . . . .	105
3.5.2	Hadoop Input Formats. . . . .	105
3.5.3	Giraph Input Formats . . . . .	106
3.5.4	Giraph Vertex Input Formats. . . . .	107
3.5.5	Edge Input Formats. . . . .	113
3.6	Preparing Graph Data for Giraph Output . . . . .	113
3.6.1	Vertex Output Formats . . . . .	113
3.6.2	Edge Output Formats . . . . .	116
<b>4</b>	<b>Popular Graph Algorithms on Giraph</b> . . . . .	<b>119</b>
4.1	PageRank . . . . .	119
4.1.1	Example . . . . .	120
4.1.2	Implementation Details . . . . .	122
4.2	Connected Components . . . . .	123
4.2.1	Example . . . . .	123
4.2.2	Implementation Details . . . . .	124

4.3	Shortest Path. . . . .	127
4.3.1	Example . . . . .	127
4.3.2	Implementation Details . . . . .	129
4.4	Triangle Closing . . . . .	131
4.4.1	Example . . . . .	131
4.4.2	Implementation Details . . . . .	133
4.5	Maximal Bipartite Graph Matching . . . . .	135
<b>5</b>	<b>Advanced Giraph Programming . . . . .</b>	<b>141</b>
5.1	Algorithm Optimization . . . . .	141
5.1.1	MasterCompute . . . . .	141
5.1.2	Data Sharing Across Nodes. . . . .	143
5.1.3	Combiners. . . . .	148
5.1.4	Coarse-Grained Processing . . . . .	149
5.2	Dealing with More Complex Graph Algorithms . . . . .	150
5.2.1	Graph Mutations . . . . .	150
5.2.2	Undirected Graphs . . . . .	151
5.2.3	Synchronizing the States of Neighboring Vertices. . . . .	154
5.2.4	Implementing Graph Coloring. . . . .	156
5.3	Performance Optimizations . . . . .	162
5.3.1	Multithreading. . . . .	162
5.3.2	Message Exchange Tuning . . . . .	163
5.3.3	Controlling the OutEdges Class . . . . .	163
5.3.4	Out-of-Core Processing . . . . .	164
5.4	Giraph Custom Partitioner . . . . .	166
5.5	Advanced Giraph I/O . . . . .	166
5.5.1	Writing a Custom Input Format . . . . .	167
5.5.2	Writing a Custom Output Format . . . . .	168
5.6	Analyzing Giraph Errors. . . . .	169
5.6.1	CountersExceededException . . . . .	170
5.6.2	ClassNotFoundException . . . . .	170
5.6.3	FileAlreadyExistsException. . . . .	170
5.6.4	OutOfMemoryError. . . . .	171
5.7	Failure Recovery. . . . .	171
5.7.1	Checkpointing. . . . .	171
5.7.2	Retrying and Recovering Failed Jobs . . . . .	172
<b>6</b>	<b>Related Large-Scale Graph Processing Systems . . . . .</b>	<b>175</b>
6.1	GraphX. . . . .	175
6.1.1	Spark and RDD. . . . .	175
6.1.2	GraphX RGD . . . . .	179
6.1.3	Examples. . . . .	183
6.2	GraphLab . . . . .	189
6.2.1	Programming Model . . . . .	189
6.2.2	Consistency Model . . . . .	194
	<b>References . . . . .</b>	<b>195</b>

Large-Scale Graph Processing Using Apache Giraph

Sakr, S.; Orakzai, F.M.; Abdelaziz, I.; Khayyat, Z.

2016, XXV, 197 p. 102 illus., 87 illus. in color.,

Hardcover

ISBN: 978-3-319-47430-4