

Contents

- 1 Introduction 1**
 - 1.1 The Big Data Phenomenon 1
 - 1.2 Big Data and Cloud Computing 3
 - 1.3 Big Data Storage Systems 5
 - 1.4 Big Data Processing and Analytics Systems 8
 - 1.5 Book Roadmap 11
- 2 General-Purpose Big Data Processing Systems 15**
 - 2.1 The Big Data Star: The Hadoop Framework 15
 - 2.1.1 The Original Architecture 15
 - 2.1.2 Enhancements of the MapReduce Framework 19
 - 2.1.3 Hadoop’s Ecosystem 27
 - 2.2 Spark 28
 - 2.3 Flink 33
 - 2.4 Hyracks/ASTERIX 36
- 3 Large-Scale Processing Systems of Structured Data 41**
 - 3.1 Why SQL-On-Hadoop? 41
 - 3.2 Hive 42
 - 3.3 Impala 44
 - 3.4 IBM Big SQL 45
 - 3.5 SPARK SQL 46
 - 3.6 HadoopDB 47
 - 3.7 Presto 48
 - 3.8 Tajo 50
 - 3.9 Google Big Query 50
 - 3.10 Phoenix 51
 - 3.11 Polybase 51

4	Large-Scale Graph Processing Systems	53
4.1	The Challenges of Big Graphs	53
4.2	Does Hadoop Work Well for Big Graphs?	54
4.3	Pregel Family of Systems	58
4.3.1	The Original Architecture	58
4.3.2	Giraph: BSP + Hadoop for Graph Processing	61
4.3.3	Pregel Extensions	63
4.4	GraphLab Family of Systems	66
4.4.1	GraphLab	66
4.4.2	PowerGraph	66
4.4.3	GraphChi	68
4.5	Other Systems	68
4.6	Large-Scale RDF Processing Systems	71
5	Large-Scale Stream Processing Systems	75
5.1	The Big Data Streaming Problem	75
5.2	Hadoop for Big Streams?!	76
5.3	Storm	79
5.4	Infosphere Streams	81
5.5	Other Big Stream Processing Systems	82
5.6	Big Data Pipelining Frameworks	84
5.6.1	Pig Latin	84
5.6.2	Tez	86
5.6.3	Other Pipelining Systems	88
6	Conclusions and Outlook	91
	References	97