

# Understanding Big Data

Analytics for Enterprise Class  
Hadoop and Streaming Data

**Paul C. Zikopoulos**  
**Chris Eaton**  
**Dirk deRoos**  
**Thomas Deutsch**  
**George Lapis**



New York Chicago San Francisco  
Lisbon London Madrid Mexico City  
Milan New Delhi San Juan  
Seoul Singapore Sydney Toronto

# CONTENTS AT A GLANCE

## PART I

### Big Data: From the Business Perspective

<b>1</b>	What Is Big Data? Hint: You're a Part of It Every Day . . .	3
<b>2</b>	Why Is Big Data Important? . . . . .	15
<b>3</b>	Why IBM for Big Data? . . . . .	35

## PART II

### Big Data: From the Technology Perspective

<b>4</b>	All About Hadoop: The Big Data Lingo Chapter . . . . .	51
<b>5</b>	InfoSphere BigInsights: Analytics for Big Data At Rest . . . . .	81
<b>6</b>	IBM InfoSphere Streams: Analytics for Big Data in Motion . . . . .	123

# CONTENTS

Foreword .....	xv
Acknowledgments .....	xxi
About this Book .....	xxiii

## PART I

### Big Data: From the Business Perspective

<b>1</b>	What Is Big Data? Hint: You're a Part of It Every Day . . .	3
	Characteristics of Big Data .....	5
	Can There Be Enough? The Volume of Data .....	5
	Variety Is the Spice of Life .....	7
	How Fast Is Fast? The Velocity of Data .....	8
	Data in the Warehouse and Data in Hadoop (It's Not a Versus Thing) .....	9
	Wrapping It Up .....	12
<b>2</b>	Why Is Big Data Important? .....	15
	When to Consider a Big Data Solution .....	15
	Big Data Use Cases: Patterns for Big Data Deployment . . . .	17
	IT for IT Log Analytics .....	18
	The Fraud Detection Pattern .....	20
	They Said What? The Social Media Pattern .....	24
	The Call Center Mantra: "This Call May Be Recorded for Quality Assurance Purposes" .....	26
	Risk: Patterns for Modeling and Management .....	29
	Big Data and the Energy Sector .....	31
<b>3</b>	Why IBM for Big Data? .....	35
	Big Data Has No Big Brother: It's Ready, but Still Young .....	37
	What Can Your Big Data Partner Do for You? .....	39
	The IBM \$100 Million Big Data Investment .....	40
	A History of Big Data Innovation .....	40
	Domain Expertise Matters .....	49

**PART II**  
**Big Data: From the Technology Perspective**

<b>4</b>	<b>All About Hadoop:</b>	
	The Big Data Lingo Chapter .....	53
	Just the Facts:	
	The History of Hadoop .....	54
	Components of Hadoop .....	55
	The Hadoop Distributed File System .....	56
	The Basics of MapReduce .....	60
	Hadoop Common Components .....	63
	Application Development in Hadoop .....	64
	Pig and PigLatin .....	65
	Hive .....	67
	Jaql .....	68
	Getting Your Data into Hadoop .....	73
	Basic Copy Data .....	73
	Flume .....	74
	Other Hadoop Components .....	76
	ZooKeeper .....	76
	HBase .....	77
	Oozie .....	78
	Lucene .....	78
	Avro .....	80
	Wrapping It Up .....	80
<b>5</b>	<b>InfoSphere BigInsights: Analytics for Big</b>	
	<b>Data at Rest .....</b>	<b>81</b>
	Ease of Use: A Simple Installation Process .....	82
	Hadoop Components Included in BigInsights 1.2 .....	84
	A Hadoop-Ready Enterprise-Quality	
	File System: GPFS-SNC .....	85
	Extending GPFS for Hadoop:	
	GPFS Shared Nothing Cluster .....	86
	What Does a GPFS-SNC Cluster Look Like? .....	88
	GPFS-SNC Failover Scenarios .....	91
	GPFS-SNC POSIX-Compliance .....	92
	GPFS-SNC Performance .....	94
	GPFS-SNC Hadoop Gives Enterprise Qualities .....	95

Compression .....	95
Splittable Compression .....	96
Compression and Decompression .....	97
Administrative Tooling .....	99
Security .....	102
Enterprise Integration .....	103
Netezza .....	103
DB2 for Linux, UNIX, and Windows .....	104
JDBC Module .....	104
InfoSphere Streams .....	105
InfoSphere DataStage .....	105
R Statistical Analysis Applications .....	106
Improved Workload Scheduling: Intelligent Scheduler .....	106
Adaptive MapReduce .....	107
Data Discovery and Visualization: BigSheets .....	109
Advanced Text Analytics Toolkit .....	112
Machine Learning Analytics .....	118
Large-Scale Indexing .....	118
BigInsights Summed Up .....	121
<b>6 IBM InfoSphere Streams: Analytics for Big Data</b>	
in Motion .....	123
InfoSphere Streams Basics .....	124
Industry Use Cases for InfoSphere Streams .....	125
How InfoSphere Streams Works .....	129
What's a Stream? .....	130
The Streams Processing Language .....	131
Source and Sink Adapters .....	133
Operators .....	134
Streams Toolkits .....	137
Enterprise Class .....	138
High Availability .....	139
Consumability: Making the Platform Easy to Use .....	140
Integration is the Apex of Enterprise Class Analysis .....	141