

UNDERSTANDING BIG DATA SCALABILITY

BIG DATA SCALABILITY SERIES, PART I

CORY ISAACSON



PRAISE FOR *UNDERSTANDING BIG DATA SCALABILITY*

“This book is useful to anyone who works with data and wants to learn more about scaling. Cory helps you understand what causes databases to slow down as data volumes grow over time. He then reviews a number of strategies that you have at your disposal to manage the growth, including software and database tuning, hardware upgrades, read-replication, and ultimately horizontal partitioning of data.”

—Dan Lynn, *cofounder of FullContact*

“*Understanding Big Data Scalability* presents the fundamentals of scaling databases from a single node to large clusters. It provides a practical explanation of what Big Data systems are, and the fundamental issues to consider when optimizing for performance and scalability. Cory draws on his many years of database experience to explain the issues involved in working with data sets that can no longer be handled with single monolithic relational databases.

“When transitioning from a traditional relational database deployment, it is tempting to ignore traditional database discipline regarding data modeling and data integrity. Much of this has been motivated by the proliferation of schema-less NoSQL databases. In spite of this trend, Cory shows why it is still important to carefully structure your data to maintain data integrity and allow sharding in such a way as to avoid costly distributed scan/shuffle operations. He discusses a practical approach to this called *relational sharding*. This is a commonsense method that avoids the pitfalls of black-box sharding. Cory’s approach is particularly relevant now that relational data models are making a comeback via SQL interfaces to popular NoSQL databases and Hadoop distributions.

“*Understanding Big Data Scalability* addresses practical problems in Big Data processing systems using real-life examples. This book should be especially useful to database practitioners new to the process of scaling a database beyond a traditional single-node deployment.”

—Brian O’Kafka, *software architect*

Understanding Big Data Scalability: Big Data Scalability Series, Part I

Table of Contents

Contents

Preface

About the Author

Chapter 1: Introduction

What You Will Learn

The Challenge of Big Data

Today's Big Data Explosion

Managing and Capitalizing on the Current Data Boom

Your Role as a Data Architect

The Acceleration of Big Data Innovation

Background for This Book

Why the Focus on Database Sharding?

Summary

Chapter 2: Why Databases Slow Down

The Database Slowdown Curve

A Hard-Won Lesson

The Root Cause

The Lesson Learned

The Enemies of Database Performance

Enemy #1: Table Scans

Enemy #2: Slow Writes

Table of Contents

Enemy #3: Concurrency Contention

Enemy #4: Missing Indexes

How to Identify Database Slowdown Issues

Summary

Chapter 3: What Is Big Data?

What Is Big Data Anyhow?

A Formal Definition for Big Data

Practical Big Data Definition

Sources of Big Data

The Advent of the Search Engine

The Rise of Social Networks

Introducing the Social Network Application

The Adoption of the Smartphone and Tablets

Traditional Big Data Sources

Future Big Data Generators

Summary

Chapter 4: Big Data in the Real World

Some Real-World Examples of Big Data

FullContact

The Big Data Challenge

Adopting a Big Data Streaming Approach

The Big Data Repository

Positioned for Rapid Growth

Social Point

Scaling from the Start

The Current Cluster

The Database Cluster Payoff

Summary

Table of Contents

Chapter 5: Scaling Your Application

The Goals of a Scalable Application Platform

The Excitement of a High-Growth Success

Application Scalability Fundamentals

Detecting Performance Bottlenecks

Scalability and Reliability

Scaling Up

Scaling Out

A Typical Online Application Architecture

The Load Balancer Tier

The Application Server Tier

The Database Tier

Analytics Application Architectures

Scaling an Analytics Application

How to Scale a Traditional Online Application

Load Balancer Tier

Application Server Tier

Database Tier

Summary

Chapter 6: When to Scale Your Database

The Last Mile of Application Scalability

How Do You Know When to Scale Your Database?

Options for Increasing Database Performance

Performance Optimization on Your Monolithic Database

Vertical Scaling

Read Scaling

Implementing a Full Big Data Scalability Environment

Table of Contents

Indications of the Need for Scale

Slow Read Queries

Slow Database Writes

Summary

Chapter 7: All Data Is Relational

Relational Data Overview

The Meaning of Data

Relationships Matter

Why Data Modelling Is Critical to Success

Summary

Chapter 8: Its All About Sharding

Sharding: The Ultimate Answer to Database Slowdown

The Laws of Databases

Sharding Defined

Black-Box Sharding

Relational Sharding

Summary

Chapter 9: Scaling Big Data: The Endgame

The Game of Big Data Scalability

The Ideal Big Data Infrastructure

Scaling Big Data Theory

The Good, the Bad

. . . And the Ugly

The Big Data Endgame

Data Locality

Summary

Table of Contents

Index