Real-Time Large-Scale Data Analytics and Information Retrieval in Practice

Aleksandar Bradic, Igor Bogicevic

2009

Contents

1 Introduction **1**
- 1.1 Enter the real time . . . 1
- 1.2 Fundamental issues . . . 1

2 The nature of large-scale data **3**

3 The challenges of real-time information processing **5**
- 3.1 Problem description . . . 5

4 Software toolkits for large-scale data analysis **7**
- 4.1 Hadoop . . . 7
- 4.2 Mahout . . . 7

5 Fundamental Algorithms in Data Analytics and IR **9**
- 5.1 Basic statistical framework . . . 9

6 Advanced Algorithms **11**

7 Online learning algorithms **13**

8 Large-scale IR Cookbook **15**

9 Moving from batch to real-time **17**

10 Real-world real-time applications **19**
- 10.1 Web Analytics . . . 19
- 10.2 Media analysis . . . 19
- 10.3 Finance . . . 19
- 10.4 Online collaboration . . . 19

11 Algorithms and Data Structure in support of large-scale real-time framework **21**
- 11.1 Randomized Algorithms . . . 21
- 11.2 Queue-based structures . . . 21

12 VoidBase : queue-based computing framework **23**
- 12.1 Overview . . . 23
- 12.2 Cookbook . . . 23

Chapter 1

Introduction

1.1 Enter the real time

1.2 Fundamental issues

Chapter 2

The nature of large-scale data

Chapter 3

The challenges of real-time information processing

3.1 Problem description

Chapter 4

Software toolkits for large-scale data analysis

4.1 Hadoop

4.2 Mahout

Chapter 5

Fundamental Algorithms in Data Analytics and IR

5.1 Basic statistical framework

Chapter 6

Advanced Algorithms

Chapter 7

Online learning algorithms

Chapter 8

Large-scale IR Cookbook

Chapter 9

Moving from batch to real-time

Chapter 10

Real-world real-time applications

10.1 Web Analytics

10.2 Media analysis

10.3 Finance

10.4 Online collaboration

Chapter 11

Algorithms and Data Structure in support of large-scale real-time framework

11.1 Randomized Algorithms

11.2 Queue-based structures

Chapter 12

VoidBase : queue-based computing framework

12.1 Overview

12.2 Cookbook