

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

Aleksandar Bradic, Igor Bogicevic

December 25, 2009

Contents

1	Introduction	1
1.1	Enter Real Time	1
1.2	Problems, Pitfalls and Challenges	1
2	The nature of large-scale data	3
2.1	Data Archives	3
2.2	Data Streams	3
3	The challenges of real-time information processing	5
3.1	Problem description	5
4	The nature of real-time data	7
4.1	Stochastic processes	7
4.2	Discrete-time	7
4.3	Continuous-time	7
5	Fundamental Algorithms in Data Analytics and IR	9
5.1	Statistical analysis framework	9
5.1.1	Regression analysis	9
5.1.2	Forecasting	9
5.1.3	Parameter estimation	9
5.1.4	Non-parametric methods	9
6	Advanced Algorithms	11
6.1	Online learning algorithms	11
6.2	Kernel Methods	11
7	Software toolkits for large-scale data analysis	13
7.1	Hadoop	13
7.2	Mahout	13
7.3	voidbase	13
8	Large-scale IR Cookbook	15
8.1	Building AVMs on vertical data	15
8.2	Model selection in the real world	15
9	Moving from batch to real-time	17
9.1	Paradigm shift	17

10 Concurrency : a new frontier	19
10.1 Problem Description	19
10.2 Data Structures	19
10.3 Algorithms	19
11 Real-world real-time applications	21
11.1 Web Analytics	21
11.2 Media analysis	21
11.3 Econometrics	21
11.4 Quantitive Finance	22
11.5 Online collaboration	23
12 Algorithms and Data Structure in support of large-scale real-time framework	25
12.1 Convolutional procedures	25
12.1.1 Example : Viterbi algorithm	25
12.2 Convolutional representation of fundamental algebraic operations	25
12.2.1 Average,Mean,Median,Variance	25
12.2.2 Matrix operations	25
12.3 Randomized Algorithms	25
12.3.1 Fast vs. Convolutional	25
12.4 Queue-based structures	25
13 voidbase : queue-based computing framework	27
13.1 Overview	27
13.2 Paradigms	27
14 voidbase cookbook	29
14.1 Simple Markov process tracking	29
14.2 Monte Carlo simulation	29
14.3 Zero-development dynamic resource monitoring framework	29
14.4 Automatic trend detection toolkit	29
14.5 Building automated news-based algorithmic trading app	29
15 Future challenges in Real-Time Large-Scale analytical processing	31
15.1 Representation problem	31
15.2 Fundamental limits	31

Chapter 1

Introduction

1.1 Enter Real Time

1.2 Problems, Pitfalls and Challenges

Chapter 2

The nature of large-scale data

2.1 Data Archives

2.2 Data Streams

Chapter 3

The challenges of real-time information processing

3.1 Problem description

Chapter 4

The nature of real-time data

4.1 Stochastic processes

4.2 Discrete-time

4.3 Continuous-time

Chapter 5

Fundamental Algorithms in Data Analytics and IR

5.1 Statistical analysis framework

5.1.1 Regression analysis

5.1.2 Forecasting

5.1.3 Parameter estimation

5.1.4 Non-parametric methods

Chapter 6

Advanced Algorithms

6.1 Online learning algorithms

6.2 Kernel Methods

Chapter 7

Software toolkits for large-scale data analysis

7.1 Hadoop

7.2 Mahout

7.3 voidbase

Chapter 8

Large-scale IR Cookbook

8.1 Building AVMs on vertical data

8.2 Model selection in the real world

Chapter 9

Moving from batch to real-time

9.1 Paradigm shift

Chapter 10

Concurrency : a new frontier

10.1 Problem Description

10.2 Data Structures

10.3 Algorithms

Chapter 11

Real-world real-time applications

11.1 Web Analytics

11.2 Media analysis

11.3 Econometrics

11.4 Quantitative Finance

In this chapter we describe computational aspects related to Quantitative Finance applications :

- security pricing
- stochastic process as a central concept in quant finance, as well as the central object in real-time analytics
- drawing analogies
- equivalents of financial concepts in fields such as web analytics
- continuous vs discrete variables
- real-time discrete-time
- real-time continuous-time
- notes :
 - Markov process - variances of the changes in successive time periods are additive
 - analogies
 - Twitter topic process
 - Trend detection and keyword bidding
 - Economy of online auctions
 - Towards efficient online marketplaces
 - Prediction markets
 - Financial software deals with real-time, but not large-scale data

11.5 Online collaboration

Chapter 12

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

12.1 Convolutional procedures

12.1.1 Example : Viterbi algorithm

12.2 Convolutional representation of fundamental algebraic operations

12.2.1 Average,Mean,Median,Variance

12.2.2 Matrix operations

12.3 Randomized Algorithms

12.3.1 Fast vs. Convolutional

12.4 Queue-based structures

Chapter 13

voidbase : queue-based computing framework

13.1 Overview

13.2 Paradigms

Chapter 14

voidbase cookbook

- 14.1 Simple Markov process tracking
- 14.2 Monte Carlo simulation
- 14.3 Zero-development dynamic resource monitoring framework
- 14.4 Automatic trend detection toolkit
- 14.5 Building automated news-based algorithmic trading app

Chapter 15

Future challenges in Real-Time Large-Scale analytical processing

15.1 Representation problem

15.2 Fundamental limits

