

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

Aleksandar Bradic, Igor Bogicevic

December 25, 2009



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Enter Real Time . . . . .	1
1.2	Problems, Pitfalls and Challenges . . . . .	1
<b>2</b>	<b>The nature of large-scale data</b>	<b>3</b>
2.1	Data Archives . . . . .	3
2.2	Data Streams . . . . .	3
<b>3</b>	<b>The challenges of real-time information processing</b>	<b>5</b>
3.1	Problem description . . . . .	5
<b>4</b>	<b>The nature of real-time data</b>	<b>7</b>
4.1	Stochastic processes . . . . .	7
4.2	Discrete-time . . . . .	7
4.3	Continuous-time . . . . .	7
<b>5</b>	<b>Fundamental Algorithms in Data Analytics and IR</b>	<b>9</b>
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
<b>6</b>	<b>Advanced Algorithms</b>	<b>11</b>
6.1	Online learning algorithms . . . . .	11
6.2	Kernel Methods . . . . .	11
<b>7</b>	<b>Software toolkits for large-scale data analysis</b>	<b>13</b>
7.1	Hadoop . . . . .	13
7.2	Mahout . . . . .	13
7.3	voidbase . . . . .	13
<b>8</b>	<b>Large-scale IR Cookbook</b>	<b>15</b>
8.1	Building AVMs on vertical data . . . . .	15
8.2	Model selection in the real world . . . . .	15
<b>9</b>	<b>Moving from batch to real-time</b>	<b>17</b>
9.1	Paradigm shift . . . . .	17
<b>10</b>	<b>Concurrency : a new frontier</b>	<b>19</b>
10.1	New challenges . . . . .	19

<b>11 Real-world real-time applications</b>	<b>21</b>
11.1 Web Analytics . . . . .	21
11.2 Media analysis . . . . .	21
11.3 Econometrics . . . . .	21
11.4 Quantitive Finance . . . . .	22
11.5 Online collaboration . . . . .	23
<b>12 Algorithms and Data Structure in support of large-scale real-time framework</b>	<b>25</b>
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
<b>13 voidbase : queue-based computing framework</b>	<b>27</b>
13.1 Overview . . . . .	27
13.2 Paradigms . . . . .	27
<b>14 voidbase cookbook</b>	<b>29</b>
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
<b>15 Future challenges in Real-Time Large-Scale analytical processing</b>	<b>31</b>
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 New challenges





## 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

