Designs of Algorithms and Programming for Massive Data

DS8001

Andriy Miranskyy

Sep. 12, 2016

Outline

- Introduction
- Learning objectives
- Quiz

Who am I?

- Andriy Miranskyy is an assistant professor at the Department of Computer Science, Ryerson University.
- Research interests: mitigating risks in software engineering
  - Focusing on software quality assurance, program comprehension, software requirements, project risk management, Big Data, and Green IT.
- Ph.D. in Applied Math. at the University of Western Ontario.
- 15 years of software engineering experience in information management and pharmaceutical industries.
- Prior to joining Ryerson in January of 2014 worked in the IBM Information Management division at the IBM Toronto Software Laboratory.
- http://www.scs.ryerson.ca/~avm/

How to get in touch?

- Send email to avm@ryerson.ca
  - Expect response within 24 hours
- Please
  - Include [DS8001] into the subject line
  - Send email from your @ryerson.ca account
    - University policy

Who are you?

- Program
- Background
  - Economist
  - Computer Scientist
  - Biologist
  - Finance
  - Etc.

Preliminary Course Description

Available on D2L

Course Syllabus

- To introduce students to the theory and design of algorithms to acquire and process large dimensional data.
- Advanced data structures, graph algorithms, and algebraic algorithms.
- Complexity analysis, complexity classes, and NP-completeness, approximation algorithms and parallel algorithms.
- Study of algorithmic techniques and modeling frameworks that facilitate the analysis of massively large amounts of data.
- Introduction to information retrieval, streaming algorithms and analysis of web searches and crawls.

Prerequisites

- Undergraduate-level data structures and algorithms course(s)

Objectives

- Goal
  - To develop skills and knowledge enabling to design algorithms for processing large volumes of data
- Outcome. At the end of the course you will be able
  - To analyze algorithms and suggest potential ways to parallelize them;
  - To identify existing or design new algorithms required to process massive data.

Books for the course

- Mandatory ones: none
  - Could not find anything suitable
- Books that may be helpful
  - Kuan-Ching Li, Hai Jiang, Laurence T. Yang, Alfredo Cuzzocrea, (eds.) “Big Data: Algorithms, Analytics, and Applications” Chapman & Hall/CRC Big Data Series, 2015.
  - Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein “Introduction to Algorithms”, 3rd Edition, MIT Press, 2009
  - Cleve B. Moler, Numerical Computing with MATLAB, Revised Reprint, 2nd edition, Society for Industrial and Applied Mathematics, 2008. http://www.mathworks.com/moler/chapters.html

Evaluation

- Labs/Assignments 30%
- Midterm test 20%
- Project 30%
- Final exam 20%

- Total 100%

Project

- Individual or Team work
- Topic: I am open to your ideas (but will need to vet them)
  - E.g., parallelize an algorithm that you use on a regular basis to process “massive data”
- Deliverables: Report and Presentation

What is Massive Data

- The course is called “Designs of Algorithms and Programming for Massive Data”

- We know what the algorithms are…
- But what is Massive Data?
- These days the buzzword for large volumes of Data is Big Data

Big Data

- Big Data: “data sets so large or complex that traditional data processing applications are inadequate.”[1]
- Big Data technology and services market growth[2]
  - Compound annual rate of 27%
  - From $9.8 billion in 2012 to $32.4 billion in 2017

1. http://en.wikipedia.org/wiki/Big_data
2. D. Vesset, “Worldwide Big Data Technology and Services 2013–2017 Forecast,” IDC Market Analysis, 244979, Dec. 2013.

Big Data Characteristics

- 4 Vs
  - Velocity : needs to be processed in real time
  - Volume : mountain-ranges of historical data
  - Variety : structured or unstructured
  - Veracity : have to be cleaned
- Give people a dictionary and some free time…
  - Validity / Variability / Volatility : inconsistency
  - Value : but it is useful?
  - Your V goes here ☺

Big Data is a moving target…

IBM 350

Big Data is a moving target…

IBM 350:

- Announced in 1956 (60 years ago)
- 3.75 megabytes of storage (5 million characters)
  - Capacity could have been bigger, but marketing department thought that there would be no demand for such a massive product
- Available for rent for $3,200 per month
- Design motivated by the need for real time accounting in business
- Weight: > 1000kg

- https://en.wikipedia.org/wiki/History_of_IBM_magnetic_disk_drives#IBM_350

What is considered big nowadays?

- Yours truly was part of the the team at IBM that built 3PB relational database and got a Guinness world record for it.
  - Which was overturned next year ☺. Remember the moving target?
  - The database engine used was IBM DB2 – which is a definition of a traditional enterprise tool
  - However it did require the skillset that a classic DBA might not have

Rule of thumb (very informal)

From the *volume* perspective, whatever you cannot fit on a single modern computer can be considered Big Data. These days it is somewhere around 25 Tb.

Note that I am ignoring here other Big Data V-characteristics (velocity, veracity, etc.)

Massive data ≠ Massive computations

- For example, think of project of finding prime numbers: http://www.mersenne.org
- "GIMPS celebrated its 20th anniversary with the discovery of the largest known prime number, $2^{74,207,281}-1$. It has [...] 22,338,618 digits -- almost 5 million digits longer than the previous record prime number." Jan. 7, 2016

Today's Numbers	
Teams	1,027
Users	162,132
CPUs	1,342,899
TFLOP/s	262.767
GHz-Days	131,384

Big Data Generators

- Examples
  - Genetics
  - Physics
  - Social nets
  - Finance
- Enterprise[1] and smaller systems can generate Big Data in the form of logs and traces

```
#How to generate 2PB of raw text with 2 lines ☺
For i from 1 to 1E14 :
    log.debug(“I am in the for-loop”)
```

1. A. Mockus, “Engineering Big Data Solutions,” in Proc. of the on Future of Softw. Eng., 2014, pp. 85–99.

Typical answers to V-challenges

- Parallelize
  - On CPU cores of a single computer
  - On a cluster of computers
- Approximate
  - Sample data

Names & Photo

Take photo now

Do introduction on Monday

Please check if you can access from your device

- https://goo.gl/forms/nPbs7akSDSzCoRPw1

Quiz

- https://goo.gl/forms/AelP00JlNasZR0qy2
  - 15 minutes