HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Big Data Frameworks Course

Prof. Sasu Tarkoma

10.3.2015

Contents

Course Overview

- This course examines current and emerging Big Data frameworks with focus on Data Science applications.
- The course starts with an introduction to MapReduce-based systems and then focuses on Spark and the Berkeley Data Analytics (BDAS) architecture.
- The course covers traditional MapReduce processes, streaming operation, machine learning and SQL integration.
- The course consists of the lectures and the assignments.

- Course is focused on assignments/exercises
  - Running distributed code!

Data Science Education

- “Data Science study profile”: An MSc level programme that combines elements from different subfields of computer science
- Trains new generations of data scientists for the industry, academia, and administration
- Organized together by two sub-programmes of the Department of Computer Science
  - the Algorithms, Data Analytics and Machine Learning sub-programme
  - the Networking and Services sub-programme
- Language of education is English
- Data Science Study Profile:
  - http://www.cs.helsinki.fi/en/studies/datascience/datascience

Degree Requirements (1/3)

- Compulsory courses of the profile (15 cr):
  - Introduction to machine learning (5 cr)
  - Distributed Systems (5 cr)

- Either one of the following:
  - Design and analysis of algorithms (5 cr)
  - Programming in C (5 op) (bachelor-level course) and Distributed Systems Project (5 cr)
- Seminars (6 cr)
- Master’s thesis (40 cr)

Degree Requirements (2/3)

- Elective courses of the profile, at least 2 of the following (9 cr):
    - Supervised machine learning (4 cr)
    - Unsupervised machine learning (4 cr) and project (3 cr)
    - Data mining (4 cr) and project (2 cr)
    - Probabilistic models (4 cr) and project (2 cr)
    - Distributed Systems Project (5 cr) (unless already included compulsory courses)
    - **Big Data Frameworks (5 cr)**

Degree requirements (3/3)

- Optional courses (10 cr), examples:
  - String processing algorithms, Data Compression Techniques, Randomized Algorithms, Approximation Algorithms, Information Theoretic Modelling, Introduction to Computational Creativity, Natural Language Processing
  - Internet-protocols, Overlay and P2P Networks, Service Ecosystems, Interactive Systems, Human Computer Interaction, Interface Technologies

  - Any project work directly related to the courses above

General Info

Advanced course, 5 credits

Part of the Data Science profile

Course components

- Lectures
- Assignments/exercises and tutorial
- Reading list
- Exam

Team:

- Professor Sasu Tarkoma
- Dr. Eemil Lagerpetz, Dr. Mohammad Hoque, Ella Peltonen

Lectures

- Lectures
  - Tuesdays 12-14 in D122 10.3 – 28.4.

- Assignments/exercises
  - Friday 10-12 D122 13.3 – 30.4.

- Course exam
  - Friday 8.5. 9:00 at B123

- Optional Separate exam (assignments are mandatory)
  - Tuesday 16.6. 16:99 in B123.

Schedule

Tuesday 10.3. Introduction and the Big Data Challenge

Friday 13.3. Intro to Scala and Spark. Problem sheet #1 available.

Tuesday 17.3. MapReduce and Spark: Overview

Friday 20.3. Exercises for Problem Sheet #1. Problem Sheet #2 available.

Tuesday 24.3. Advanced and Professional Spark

Friday 27.3. Exercises for Problem Sheet #2. Problem Sheet #3 available.

Tuesday 31.3. Distributed algorithms for Big Data

Easter break

Friday 10.4. Exercises for Problem Sheet #3. Problem Sheet #4 available.

Tuesday 14.4. MLBase and Streaming Spark

Friday 17.4 Exercises for Problem Sheet #4. Problem Sheet #5 available.

Tuesday 21.4. Big Data and Spark Use Cases

Tuesday 28.4. Summary

Wednesday 29.4 10-12 Final exercises for Problem Sheet #5.

Assignments/exercises

Environment: Spark 1.2, we use Scala 2.10, no support for Python! Scala IDE Eclipse recommended, check Spark version (there are large differences)

Weekly exercise Problem Sheet

Detailed instructions provided in the problem sheet

Completed questions contribute to the grade

Total points determine 40% of the grade

The last Problem Sheet is more involving and contributes more to the points

Moodle is used to return the answers

IRCnet channel #tkt-bdf

Exam material (in addition to slides and exercises)

Articles (part of the exam material):

1. Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Originally OSDI 2004. CACM Volume 51 Issue 1, January 2008. http://dl.acm.org/citation.cfm?id=1327492
2. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. Matei Zaharia et al. NSDI (2012) usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf
3. HaLoop: Efficient Iterative Data Processing on Large Clusters by Yingyi Bu et al. In VLDB'10: The 36th International Conference on Very Large Data Bases, Singapore, 24-30 September, 2010.
4. MLbase: A Distributed Machine-learning System. Tim Kraska et al. CIDR 2013. http://www.cs.ucla.edu/~ameet/mlbase.pdf

Additional material (not part of the exam):

http://spark.apache.org

http://spark.apache.org/docs/latest/programming-guide.html

www.databricks.com

Grading

Course grading will be based on the final exam and the assignments/exercises.

Exam 60% and exercises 40% of the grade.

- Exam
  - Friday 8.5. 9:00 at B123

Main theme	**Prerequisites**	**Approaches learning goals**	**Meets learning goals**	**Deepens learning goals**
Big Data Frameworks: definitions and systems	Basics of data communications and distributed systems (Introduction to Data Communications, Distributed Systems)	Knowledge of how to define the concepts of MapReduce and variants and state their central features Ability to describe at least one system in detail	Ability of being able to compare different Big Data frameworks in a qualitative manner Ability to assess the suitability of different systems to different use cases	Ability to give one's own definition of the central concepts and discuss the key design and deployment issues
Internal operation and implementation of a Big Data framework	Basics of data communications and distributed systems (Introduction to Data Communications, Distributed Systems) Big-O-notation and basics of algorithmic complexity Basics of reliability in distributed systems	Knowledge of the design and implementation level concepts of Big Data frameworks, specifically Hadoop and Spark. Knowledge of how distributed state is maintained and synchronized. Understanding of the communication and computational costs in Big Data processing. Ability to describe at least one algorithm in detail	Ability of being able to compare different Big Data frameworks based on their design and implementation. Ability of designing distributed Big Data systems building on existing frameworks for batch and streaming processing. Knowledge of key performance issues and the ability to analyze these systems Knowledge of the most important factors pertaining to reliability	The knowledge of designing a Big Data platform for a given problem Familiarity with the state of the art
Distributed algorithms for Big Data frameworks	Basics of algorithm design and machine learning	Knowledge of the basic design of a distributed algorithm for MapReduce and Spark. Ability to use graph processing and machine learning in a distributed cluster environment	Ability to design and implement a solution that uses distributed algorithms for a large dataset Ability to create both batch and streaming solutions	Design and implementation of a new machine learning algorithm for Big Data Familiarity with the state of the art
Data Science applications	-	Knowledge of the basic Data Science use cases based on Big Data frameworks	Knowledge of at least two Data Science use cases and how they use the Big Data framework Knowledge of Data Science pipelines	Familiarity with the state of the art Automation of Data Science pipelines

Contact information

Lecturer prof. Sasu Tarkoma (contact info on homepage)

Assignments: Ella Peltonen, Eemil Lagerspetz, Mohammad Hoque (@cs.helsinki.fi)

Course homepage can be found: www.cs.helsinki.fi/courses

Questions?