UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Dottorato di Ricerca in Ingegneria dell'Informazione

Doctoral Course

"Cloud Computing for Big Data Analysis"

Claudio Lucchese, Franco Maria Nardini, Nicola Tonellotto

High Performance Computing Lab

ISTI-CNR, Pisa, Italy

Short Abstract: In this course, we will discuss the characteristics and benefits of cloud computing as the current technological trend to deliver on-demand computing resources over the Internet on a pay-for-use basis, and the Map Reduce programming paradigm, daily used by large IT companies to process huge amounts of data on large-scale distributed platforms, together with the Apache Hadoop framework, its open source de-facto standard implementation. Furthermore, we will present and discuss some problems and solutions for cloud data management systems. To this end, we will introduce the consensus problem in asynchronous distributed platforms, presenting impossibility results of distributed systems theory, and we will discuss algorithms and solutions for data consistency, availability and fault tolerance. Eventually we will present big data analysis techniques, such as clustering, regression and graph analysis, as fundamental tools to model and extract knowledge from data, with a focus on information retrieval problems.

Course Contents in brief:

- Introduction
- Concepts and techniques for Cloud computing (**2 hours**)
  - Cloud characteristics and benefits
  - Designing applications for the Cloud
  - Virtualization mechanisms
- Cloud Data Management problems and solutions (**6 hours**)
  - Availability, consistency and fault tolerance: impossibility results
  - Strong consistency: classical solutions
  - Weak consistency: Amazon solutions
- Programming for Big Data problems (**6 hours**)
  - MapReduce programming and design patterns
  - Apache Hadoop and PIG frameworks
  - Streaming Data Analysis
- Big Data Analysis Techniques (**6 hours**)
  - Clustering and regression
  - Graph analysis
  - Machine learning techniques for information retrieval

Total # of hours: 20

CV of the Teachers

Claudio Lucchese.

Claudio Lucchese (http://hpc.isti.cnr.it/□claudio/) received summa cum laude the MSc in Computer Science from the Ca' Foscari University of Venice in October 2003. He received Ph.D. in Computer Science from the same university, and since 2007 he is researcher at the ISTI-CNR in Pisa, an institute of the Italian National Research Council. His research interests are mainly in the field of data mining and data mining techniques for information retrieval. He published about 50 papers in peer reviewed international conferences and journals. He has teaching experience at the University of Venice (Dept. Computer Science - 2010), Pisa (Dept. of Computer Science in the Humanities - 2011) and Firenze (Dept. of Compute Science 2011-2012).

Franco Maria Nardini.

Franco Maria Nardini (http://hpc.isti.cnr.it/□nardini/) is currently a Researcher at ISTI–CNR in Pisa. He received his Ph.D. in Information Engineering from the University of Pisa in 2012. His research interests are focused on Web Information Retrieval, Data Mining, and Machine Learning. Franco Maria Nardini is member of the program committee of important conferences in IR and DM like ACM CIKM and SIGKDD. He authored more than 25 papers spanning from Web Information Retrieval to Data mining in peer reviewed international journal and conferences.

Nicola Tonellotto.

Nicola Tonellotto (http://hpc.isti.cnr.it/□khast/) received his Ph.D. in Information Engineering from the University of Pisa and in Computer Engineering from the Technical University of Dortmund in 2008. He is researcher at ISTI-CNR since 2006, and he is contract Professor at the Computer Science Department of the University of Pisa since 2009, where he teaches courses on high performance computing and distributed enabling platforms. His main research interests include Cloud computing and Web information retrieval. He co-authored more than 40 papers in highly relevant venues spanning from distributed and parallel computing to IR and Grid/Cloud related conferences in peer reviewed international journal and conferences.

Room and Schedule

Room: *Aula Riunioni del Dipartimento di Ingegneria dell'Informazione, Via G. Caruso 16, Pisa – Ground Floor*

Schedule:

June, 08, 15:00-17:00

June, 09: 15:00-18:00

June, 10: 15:00-18:00

June, 11: 15:00-18:00

June, 12: 15:00-18:00

June, 18: 15:00-18:00

June, 22: 15:00-18:00