Ecole d'Eté CEA-EDF-Inria

16 au 20 Juin 2014 – CEA Cadarache

Parallel and Distributed Data Analytics (PDDA'2014)

Organizers:

- CEA: Michael Aupetit (LIST, Saclay)
- EDF: Georges Hébrail (R&D, Clamart)
- Inria: Patrick Valduriez (Zenith Project, Montpellier)

Abstract:

The term 'Big Data' refers to the problem of collecting, storing and analyzing the huge amount of digital data produced by the human activity which is supported by computers. The goal of such analyses is either to understand and synthesize underlying phenomena/behaviors or to predict future phenomena/behaviors. According to Gartner, main challenges in Big Data are related to the so-called 3V's: *Volume* (ability to process huge volumes of data), *Velocity* (ability to process data on the fly even if it arrives at high speed in a continuous stream), *Variety* (ability to process data from a large number of sources and/or heterogeneous type – text, image, audio, video …).
The processing steps in a 'Big Data' activity are multiple from data acquisition, communication, integration, cleaning, querying, visualization, to more complex algorithms performing data mining. The objective of this summer school is to focus on the data mining part of this process, by presenting new advances in the field of *large scale data mining* in relation to 2 of the 3 V's: *Volume* and *Velocity*.

A call for proposal will be sent to participants of the summer school for presenting applications related to the subjects covered by the school. Selected applications will be presented during the 'Application sessions'.

Preliminary schedule

	Course	Duration	Teacher	Contents
Monday morning	Upgrade and prerequisites on distributed and parallel data processing	3h	P.Valduriez (INRIA)	2h Lecture 1h Exercices
Monday afternoon	Upgrade and prerequisites on data analytics	3h	G.Hébrail (EDF)	1h30 Lecture 1h30 Practical work on computers
	Applications	45’	TBD from Call of proposal	
Tuesday morning	Large scale data visualization	3h	M.Aupetit (CEA)	Lecture
Tuesday afternoon	Large scale data visualization	2h	M.Aupetit (CEA)	Practical work on computers with assistant Jacques-Henri Sublemontier
	Applications	45’	TBD from Call of proposal	
Wednesday morning	Multi-site (distributed) large scale data analytics: P2P, Multi-Agent, privacy-preserving	3h	H.Kargupta (President, Agnik, Professor, University of Maryland Baltimore County)	Lecture
Wednesday afternoon	Multi-site (distributed) large scale data analytics: P2P, Multi-Agent, privacy-preserving	3h	H.Kargupta (President, Agnik, Professor, University of Maryland Baltimore County)	Practical work on computers with assistant: T.Allard (INRIA)
	Applications	45’	TBD from Call of proposal	
Thursday morning	Mono-site (centralized) large scale data analytics: Hadoop, HPC, GPU	3h	Pr David Bader (Georgia Tech)	Lecture
Thursday afternoon	Mono-site (centralized) large scale data analytics: Hadoop, HPC, GPU	3h	Pr David Bader (Georgia Tech)	Practical work on computers with assistant: F.Masseglia (INRIA)
	Applications	45’	TBD from Call of proposal	
Friday morning	Large scale data mining on Hadoop MapReduce: frequent itemsets and clustering	3h	F.Masseglia (INRIA)	1h30 Lecture 1h30 Practical work on computers
Friday afternoon	Data stream processing and analytics	3h	A.Bondu (EDF)	1h30 Lecture 1h30 Practical work on computers
	Synthesis	1h	Organizers and teachers	

Possible applications:

- Social network analysis
- Cybersecurity
- Computational Biology and Genomics
- Streaming Graph Analytics
- Distributed Connected Car and M2M environments
- Smart metering data

Professor	Course synopsis
David A. Bader is a Full Professor in the School of Computational Science and Engineering, College of Computing, at Georgia Institute of Technology, and Executive Director for High Performance Computing. Dr. Bader serves as a Board member of the Computing Research Association (CRA), and on the Steering Committees of the IPDPS and HiPC conferences. He is Program Chair for IPDPS 2014, and has served as the General Chair of IPDPS 2010 and Chair of SIAM PP12. He is an associate editor-in-chief of the Journal of Parallel and Distributed Computing (JPDC), and editor-in-chief of the IEEE Transactions on Parallel and Distributed Systems (TPDS). Dr. Bader's interests are at the intersection of high-performance computing and real-world applications, including computational biology and genomics and massive-scale data analytics. He is also a leading expert on multicore, manycore, and multithreaded computing for data-intensive applications such as those in massive-scale graph analytics. He has co-authored over 130 articles in peer-reviewed journals and conferences, and his main areas of research are in parallel algorithms, combinatorial optimization, massive-scale social networks, and computational biology and genomics. Prof. Bader is a Fellow of the IEEE and AAAS, a National Science Foundation CAREER Award recipient, and has received numerous industrial awards from IBM, NVIDIA, Intel, Cray, Oracle/Sun Microsystems, and Microsoft Research. Bader is a co-founder of the Graph500 List for benchmarking "Big Data" computing platforms. Bader is recognized as a "RockStar" of High Performance Computing by InsideHPC and as HPCwire's People to Watch in 2012. http://www.cc.gatech.edu/~bader	**Mono-site (centralized) large scale data analytics: Hadoop, HPC, GPU** Emerging real-world graph problems include detecting community structure in large social networks, improving the resilience of the electric power grid, and detecting and preventing disease in human populations. Unlike traditional applications in computational science and engineering, solving these problems at scale often raises new challenges because of sparsity and the lack of locality in the data, the need for additional research on scalable algorithms and development of frameworks for solving these problems on high performance computers, and the need for improved models that also capture the noise and bias inherent in the torrential data streams. In this talk, the speaker will discuss the opportunities and challenges in massive data-intensive computing for applications in computational biology, genomics, and security. This talk will highlight the importance of a portfolio of high performance computing and big data platforms for massive data analytics.

Hillol Kargupta, President, Agnik and Professor, University of Maryland Baltimore County	
Florent Masseglia is a scientific researcher in computer science at Inria since 2002. He did his PhD in computer science at Montpellier in co-supervision with the University of Saint-Quentin en Yvelines (graduated in 2001), in the area of knowledge discovery from databases. Until 2010 he worked in Sophia Antipolis on the development and application of these techniques for knowledge discovery in areas related to usage (such as navigation on Web sites) and large amounts of data. Since 2010, he works in Montpellier, in the Zenith team of Inria on the analysis of very large amounts of data related to life sciences (agronomy, biology, medicine). These data, derived from observations, experiments and simulation are indeed complex, often very large, and are at the heart of important issues to better understand the studied domains.	**Large scale data mining on Hadoop MapReduce: frequent itemsets and clustering** Frequent itemset mining and clustering are two major domains of data mining. They are very well studied, for centralized or distributed environments. However, with the MapReduce paradigm, implementing and running the famous algorithms for frequent itemset mining and clustering is not straightforward. The Apriori algorithm, for instance, requires as much MapReduce jobs as the length of the longest expected pattern, and communications between mappers and reducers depend on the number of candidates. For clustering, it is not a good solution to compare objects that are not located on the same mapper. This calls for specific solutions, not only because of the programming language, but also because of the mechanisms that manage a MapReduce job. In this practical work, we will have a survey on some famous data mining algorithms for itemset mining clustering. Then, we will implement some of these algorithms and discuss choices of implementation for MapReduce.
Dr. Hab. **Michael Aupetit** and Dr. **Jacques -Henri Sublemontier** CEA LIST, Data Analysis and Intelligent Systems Laboratory **Michael Aupetit** is engineer and research scientist at CEA LIST located in the Paris–Saclay campus. After earning a Ph.D in industrial engineering from the Institut National Polytechnique de Grenoble in 2001, he obtained the Habilitation for Research Supervision in Computer Science from the University Paris Sud 11 in July 2012. He is a CEA senior expert in data mining and visual analytics at the Data Analysis and Intelligent Systems Laboratory (LADIS). He focuses his research on data analysis and the design of user-interpretable models, and he is currently interested in massive data processing by these techniques. http://michael.aupetit.free.fr/ **Jacques-Henri Sublemontier** is a Research Engineer at CEA LIST Paris-Saclay. He received a Ph.D. degree in Computer Science from the university of Orléans, France, in 2012. Trained as general Computer Scientist, he has been concerned with HPC (High Performance Computing) and AI (Artificial Intelligence) challenges and techniques. He joined the CEA as a specialist in Data Mining and Distributed Computing at the LADIS lab. His main research interest is the development of unsupervised collaborative approaches (clustering and dimensionality reduction) for data analysis with a focus on scalability using distributed systems and technologies relying on algorithmic skeletons. http://jh.sublemontier.free.fr/	**Large scale data visualization** This course will focus on visual analytics of very large data. Visual analytics rely on the natural ability of the human visual system to process very quickly a lot of information through their graphical representation. These representations are a key to the visualization pipeline. This pipeline transforms the raw signals measured by sensors or obtained by calculation, into digital values and then as geometric representations. These representations are made graphically on a screen before being perceived by the visual system of the user and interpreted by its cognitive system. We will study the main elements of the visualization pipeline. We will begin with the physiological laws inducing graphics encodings to be preferred, and different types of information that we intend to visualize (histograms, scatter plots and parallel coordinates for the table data types, trees and matrices for biological data, graphs for social networks...). Making visualization interactive is crucial to the user to retrieve information from the data representation. We will see for each type of representation the main interaction techniques. Finally we will see how the scaling can be done at each stage of the visualization pipeline: distributing preprocessing calculations upstream, using techniques to synthetize instances or variables at the geometric encoding step, and using suitable graphic rendering techniques for display on the screen. A practice session will allow participants to experience the use of the cloud for very large data visualization.

Alexis Bondu is currently a researcher at EDF R&D in the team “statistics & decision support tools”. He is particularly interested in supervised learning, data streams processing, and time series analysis.	**Data stream processing and analytics** Data-streams processing is a recent domain of research which is complementary to the Big Data. This kind of algorithms analyze data on the fly, and could be qualified as designed to treat ``Fast Data''. This talk aims at providing an overview of data-streams processing approaches and consist of 3 parts : i) querying, ii) unsupervised learning, iii) supervised learning. At last, a "hands on" tutorial will provide you the opportunity of implementing online algorithms within the STORM platform. * Introduction : - Big Data vs. Fast Data - Application areas of data-stream processing - Specific contrains - Brief review of Complex Event Processing tools (CEP) * Part 1 - Querying : a comparison between BD and CEP - BD schema vs. streams connections - Punctual queries on DB vs. continuous queries on data-streams (notion of windowing) - Practical examples of queries on data-streams * Part 2 - Unsupervised Learning - Batch mode vs. online clustering - Micro-clustering as an evolving summary of a data-stream - ClusStream Algorithm - DenStream Algorithm (Feng Cao) - Trac-Streams and Rino-Streams Algorithms (O. Nasraoui) - Unresolved Issues : - user parameters to be adjusted - forgetting the ``old'' pieces of information, and not the none-informative ones - curse of dimensionality * Part 3 - Supervised Learning - From batch mode toward online learning - Anytime Algorithms - Incremental Algorithms - Online Algorithms - The two streams paradigm - Stability / Plasticity dilemma - Evaluation of online classifiers - Examples of online classifiers - Concept drift & Model Management * Part 4 - Conclusion : - How to handle very fast data-streams ? - Summarize and Analyze - Distributed Algorithms - Adaptive Algorithms (the quality of the analysis is the adjusting variable)

Tristan Allard is a post-doctoral researcher in the ZENITH team, a joint team between the LIRMM laboratory and INRIA - the French national research institute in Computer Science. He conducted his Ph.D. thesis in Computer Science in the SMIS team and received it from the University of Versailles in December 2011. Dr Allard's research interests are in the area of privacy-preserving personal data management, which key ingredients encompass distributed algorithms, cryptographic schemes, and sanitization models and mechanisms. He also frequently serves as an external reviewer for major database conferences and journals. For more information, please visit: https://sites.google.com/site/tristanallardhome/home	