GIAN Course on

Big Data Stream Analytics

26 October – 01 November 2016

Overview

Over the last decade, we have witnessed the emergence of data-intensive applications that need to handle very large flows of data. Examples of such applications include network monitoring, financial applications, security log, and sensors applications. For all these applications, there is a growing need for (new) techniques capable of monitoring or analyzing these streams to detect outliers, intrusions, unusual or anomalous activities, complex correlations, extreme events, or the emergence of patterns. These techniques are grouped together under the term "Big data stream analytics". These techniques must be capable of processing the input data quick enough to keep pace with the rate of the stream. The solution adopted to process such data streams is to trade off accuracy for space. Actually, one of the specificities of data intensive applications is that they do not require accurate responses to queries, only (high quality) approximate responses computed by summarizing the data are acceptable. Streaming algorithms, through synopses, handle such features. They read their input data sequentially and there is no requirement regarding the order in which data items are received. They use working memory whose size is much smaller than both the input size, and the domain from which items are drawn. There is a large panel of streaming algorithms that vary according to the number of passes they need to process their input process, the size of the memory they use, the time needed to process each read item, or whether they are randomized or deterministic.

Objectives:

The objective of the course is to bring closer the vast applications of data streaming analysis techniques. The course aims at proposing a comprehensive survey of these techniques, across several practical usages (Statistical metrics over distribution of stream, distributed system safety and network security, cloud monitoring and so on). The course will be presented with some famous models of this research area and advanced algorithms, basics of which will be covered as well. The course is aimed at the more general audience including mathematicians, statisticians, computer scientists and electrical engineers. Applications will be illustrated by practical examples. The participants' knowledge about the course content will be raised to the level such that they will be able to use the methods for their own applications and research.

Tentative Schedule of the Course

Day 1	Lecture 1: Panorama of Big Data Lecture 2: Data stream analysis: Model and Basic tools 1. *The data stream model* 2. *Background algorithms* Tutorial 1: Performance analysis of basic algorithms
Day 2	Lecture 3 and 4: Similarity metrics over distributed streams: *AnKLE*, *Codeviation metric and Sketch-* metric* Tutorial 2: Performance analysis of basic algorithms – Part 2
Day 3	Lecture 5: Distributed monitoring: the need of sampling: 1. *System size estimation by sampling* 2. *Byzantine tolerant uniform node sampling service* Lecture 6: Identifying Global Icebergs in Distributed Streams using probability distribution learning Tutorial 3: Performance analysis of basic algorithms – Mini-defense of results

Day 4	Lecture 7 and 8: Usage of stream analysis methods for Stream processing: 1. *Overview of Apache Storm* 2. *Key-grouping: Load-balancing using sketching techniques* 3. *Shuffle grouping: Proactive online scheduling* Tutorial 4: Large scale data stream processing – Application with Apache Storm
Day 5	Lecture 9 and 10: Enhancement of a core sketching algorithm: 1. *CASE: Estimating the Frequency of Data Items in Massive* 2. *Efficiently Summarizing Data Streams over Sliding Windows* Tutorial 5: Large scale data stream processing – Part 2
You Should Attend If…	▪ You are students at all levels (BTech/MSc/MTech/PhD) or Faculty from reputed academic institutions and technical institutions. ▪ You are executives, engineers and researchers from manufacturing, service and government organizations including R&D laboratories.
Max. No. of participants	50
Credit Points	The course carries ONE credit. All the participants will be provided a certificate after completion of the course
Fees	Participants from abroad: US $300 MSc/M.Phil/B.Tech/M. Tech. Students: Rs. 1000 Research Scholars: Rs. 2000 Faculty from Academia: Rs. 3000 Government Research Organization: Rs. 5000 Industry Participants: Rs. 8000 The above fee includes all instructional materials, computer use for tutorials and assignments, laboratory equipment usage charges, 24 hr free internet facility. The participants will be provided with accommodation on payment basis.

The Expert Faculty

Dr Yann Busnel is currently an Associate Professor at the Ensai, the national school for Statistics and Information Analysis since September 2014. He is head of the Computer Science Department and co-head of the MSc in Big Data. He is member of CREST (Research Center in Economics et Statistics), Laboratory of Statistics et Models. Since April 2015, he is also associated member of Inria Research Center Rennes - Bretagne Atlantique, in the Dionysos team. Finally, he is also associated member of LINA (Computer Science Laboratory of Nantes Atlantic). Previously, he spent 5 years as Assistant Professor at the University of Nantes, in the Computer Science Department. He obtained his PhD in Computer Science at the University of Rennes (France) in November 2008. Then, he spent one year in Italy, at the University "La Sapienza" of Rome, between 2008 and 2009. In 2016, he has been granted as Invited Professor at La Sapienza for 3 month, and holds a national grant of Scientific Excellence since 2011.

His research topics are (but not limited to): large-scale distributed data streams (for Big Data or Safety context for instance) and distributed system models. He has published more than 40 international papers in these fields.

The Host Faculty:

Dr. M. Ashok Kumar is an Assistant Professor in the Discipline of Mathematics, IIT Indore. His research interest broadly lies in Information Theory, Statistics and Probability. He is particularly interested in Measures of Information, Statistical Inference Based on Distance Functions, and Information Geometry.

Dr. Sk. Safique Ahmad is an Assistant Professor and Head of the Discipline of Mathematics, IIT Indore. His research interests lie in Numerical Linear Algebra, Stability Stochastic Differential Equations (SDEs) and Quaternion Linear Algebra.

Duration:

26 October – 01 November, 2016

Course Coordinators

Dr. M. Ashok Kumar
Email: ashokm[at]iiti.ac.in
Phone: +91 731 2438 968
Home page: http://iiti.ac.in/people/~ashokm/

&

Dr. Sk. Safique Ahmad
Email: safique[at]iiti.ac.in
Phone: +91 731 2438 947
Home page: http://iiti.ac.in/people/~safique/

Discipline of Mathematics
IIT Indore, Khandwa Road, Simrol, Indore
452020
Madhya Pradesh

……………………………………………………………………………………

Course website:
http://iiti.ac.in/people/~ashokm/GianBigData.html