**Design of Algorithms for Big Data Analytics**

**August 1-12, 2016**

School of Computer and Information Sciences, University of Hyderabad, Hyderabad 500046, INDIA.

**Overview**

Very large volumes of data are being collected in every domain of human endeavour. The term Big Data has become pervasive recently and denotes the problems related to the storage, management, and analysis of these large amounts of data. Business, sciences, and engineering are all becoming dependent on their ability to harness useful knowledge from very large datasets. For example, the research in medical and biological sciences used to focus primarily on wet-lab experiments that generated small sets of numbers, and these were then analysed by statisticians. Availability of technologies for high speed sequencing of genomic entities has the potential to provide very deep insights into the biological processes but it requires analysis of very large sized datasets. This is a trend in almost all the fields where one considers large amounts of sensor and recording data at very fine resolutions and need intelligent analysis. Businesses are also becoming increasingly dependent on analysis of marketing and financial data. Security analysis for every type of forensics – from cyber-crime to money laundering to law enforcement - is becoming dependent on analysis of mining of very large amounts of monitored data.

Computer hardware technologies have enabled very fast processors but the needs of processing Big Data surpass the capabilities of individual processors. Cloud computing technologies that enable hundreds of processors to work in parallel to process huge datasets have emerged in last few years and are being adopted successfully by various users. However, there is a need to develop dedicated algorithms for various types of data processing needs on these scaled up server systems. The topic of efficient algorithm design for such Big Data environments has been extensive research for the last few years. This course seeks to teach students the various aspects of analysing the huge data sets using Map-Reduce programming paradigm in the cloud infrastructure.

**Design of Algorithms for Big Data Analytics: August 1 - August 12**

Number of participants for the course will be limited to Fifty only.

<table>
<thead>
<tr>
<th>You Should Attend If...</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. You are a student pursuing MSc / MCA / Mtech / PhD in Computer and Information Sciences.</td>
<td></td>
</tr>
<tr>
<td>2. You are researchers in Biological and Medical Sciences and Engineering disciplines who need to process very large datasets.</td>
<td></td>
</tr>
<tr>
<td>3. You are Business Analysts and Business Intelligence specialists who are interested in designing systems for processing large data sets.</td>
<td></td>
</tr>
</tbody>
</table>

**Course Pre-Requisites**

- A basic understanding of higher level programming language such as JAVA or Python. At least one year’s experience in writing non-trivial computer programs.
- A basic understanding of the issues in designing efficient algorithms for problem solving by computers.
- Undergraduate level familiarity with basic concepts of statistics.

**The participation fee for taking the course is as follows:**

- Industry/Research Organizations: Full Course: Rs. 20,000/-
- Academic Institutions: Full Course: Rs. 5,000/-
- Fee by DD is payable to the “Finance Officer”, University of Hyderabad

(Note: Students of UOH are exempted from Registration Fee. Conditions apply.)

The above fee include all instructional materials, computer use for tutorials, 24 hr. free internet facility.

The participants will be provided with single bedded accommodation on payment basis, depending on availability.

**August 1, 2016:** Registration II Overview of the Course

**August 2, 2016**

**Lecture 1:** Vertical Scale-up systems such as GPU processors. Horizontal scale-up systems such as Apache Hadoop. Hardware and software aspects of data analysis on each type of system.

**August 3, 2016**

**Tutorial 1:** Become familiar with logging to in the Hadoop system, loading data into HDFS, and running of simple programs.

**August 4, 2016**

**Lecture 2:** Map-Reduce programming paradigm, its theoretical foundations, and its adaptation in Hadoop environments. Relationship between Functional programming paradigm and Simple counting based programs. Difference between the Hadoop and Spark environments and algorithm design aspects using both environments.

**Tutorial 2:** Writing of simple programs using Map-Reduce formalism and executing them in Hadoop and Spark environments.

**August 5, 2016**

**Lecture 3:** Design of Map-Reduce algorithms for computing first order statistical operations on large datasets. Design issues involved in computation of second order statistical quantities and Bayesian reasoning using Map-Reduce formalism and some possible algorithms for some of the operations.

**Tutorial 3:** Write programs for cloud environments to compute statistical summaries of large datasets.

**August 8, 2016**

**Lecture 4:** Design of clustering algorithms in Hadoop and Spark environments. Sequential, K-means, and some other algorithms. Difficulties in the design of some other clustering algorithms and some approximation approaches for them.

**Tutorial 4:** Implement clustering algorithms in Cloud environments.

**August 9, 2016**

**Lecture 5:** Design of classification and association algorithms in cloud computing environments. Decision tree and perceptron training algorithms to be analyzed in details.

**Tutorial 5:** Implement some classification algorithms in Map-Reduce environments.

**August 10, 2016**

**Lecture 6:** Overview of graph mining problems, network analysis algorithms, and bi-clustering algorithms in the cloud environments. Status of current research and algorithm design issues.

**August 12, 2016**

Examination (9:00 am to 11:00 am)

Valedictory (4pm to 5pm) followed by high tea.

**Course Co-ordinator**

Prof. Arun Agarwal

Phone: 040-23010780, E-mail: aruncs@uohyd.ernet.in, http://scis.uohyd.ac.in/

**The Faculty**

Prof. Raj Bhatnagar, is a Professor of Computer Science at University of Cincinnati, Ohio USA. He obtained his Ph.D in Computer Science from University of Maryland, College Park, in 1989 and B.Tech in Electrical Engineering from IIT Delhi in 1979. His area of research is data mining and pattern recognition. He has been working in this research area for more than twenty five years. His research projects have been funded by NSF, US Air Force, US DARPA, and a number of industrial sponsors. He has supervised graduate students for completion of twelve Ph.D dissertations and seventy M.S theses. His recent research projects include design of mining and analysis algorithms for Big Data situations in Biomedical, Manufacturing, GIS and Security applications. These problems have involved various structured and unstructured data. He has published more than eighty peer-reviewed publications. He has designed and taught graduate level classes on the topics of Data Mining, Big Data analysis, and Artificial Intelligence.