**MHRD Scheme on Global Initiative on Academic Network(GIAN)**

**COURSE TITLE**

**Parallel and Distributed Data Stream Mining**

## Overview

Data is being continuously collected from a variety of sensor sources, such as Twitter feeds, news streams, and environmental sensors. It is a significant challenge to continuously monitor such data and derive insights in a timely manner. This course on data stream analysis focuses on methods and software for deriving patterns and aggregates from data streams in real-time. The course will focus on (1) Parallel and distributed methods for data stream mining, (2) Methods for mining from graphical data, where each stream item represents a relationship between entities.

## Objectives

The main objectives of the course are:

1. Introduce the student to use cases of stream processing, the data stream model and graph stream model
2. Present algorithmic techniques for graph stream processing, including random sampling, graph sketches, and merge-and-reduce.
3. Show their application to problems such as subgraph counting, graph connectivity, random sampling from graphs, graph matchings, etc
4. Present current techniques on monitoring parallel and distributed streams, including algorithms in the continuous distributed monitoring model, and the parallel streaming model
5. Provide practical perspective on building software for stream processing
6. Provide experience with an open source tool, Apache Flink

## Teaching Faculty with allotment of Lectures and Tutorials

1. Prof. Srikanta Tirthapura (ST) : 8 hrs lectures and 5 hrs tutorials
2. Prof. Sonali Agarwal (SA): 5 hrs lectures and 4 hrs tutorials

## Course details

**Duration:** Dec 19 – Dec 23, 2017: 13 hrs lectures and 10 hrs Tutorials

**Lecture Schedule:**

***Day 1: Basic Fundamentals of Stream Processing***

Lecture 1: 1:30 hrs. (10:00 am to 11:30 am): ST
Fundamentals of Stream Processing, Models, Algorithms, and Systems

Lecture 2: 1:30 hrs. (12:00 pm to 1:30 pm): SA
High Velocity Data Stream mining algorithms & techniques: I

Tutorial 1: 2 hrs. (3:30 pm to 5:30 pm): ST
Basics of Stream Processing Software using Apache Flink

***Day 2: Data Stream Sampling and Mining***

Lecture 3: 1:30 hrs. (10:00 am to 11:30 am): SA
High Velocity Data Stream mining algorithms and techniques: II

Lecture 4: 1 hrs. (12:00 pm to 1:00 pm): ST
Random Sampling from Data Streams

Tutorial 2: 2 hrs. (3:30 pm to 5:30): ST
Processing Streaming Data Using an Operator Pipeline

***Day 3: Graph Stream Processing***

Lecture 5: 1:30 hrs. (10:00 am to 11:30 am): ST
Graph Stream Processing: I

Lecture 6: 1 hrs. (12:00 pm to 1:00 pm): ST
Graph Stream Processing: II

Tutorial 3: 2 hrs. (3:30 pm to 5:30):SA
Distributed stream processing using Apache Flink

***Day 4: Distributed/Parallel Stream Processing and Challenges***

Lecture 7: 1:30 hrs. (10:00 am to 11:30 pm):ST
Parallel and Distributed Stream Processing

Lecture 8: 1 hrs. (12:00 pm to 1:00 pm): SA
Performance Measures and Challenges of Data Streams

Tutorial 4: 2 hrs. (3:30 pm to 5:30):SA
Use cases of real-time Data Stream Mining

***Day 5: Stream Learning and Complex Event Processing***

Lecture 9: 1:30 hrs. (10:00 am to 11:30 am): ST
Machine Learning on Data Streams

Lecture 10: 1 hrs. (12:00 pm to 1:00 pm):SA
Complex Event Processing on Data Streams

Tutorial 5: 2 hrs. (3:30 pm to 4:30):ST
Rump session for participants