ELEN E6889

Large-Scale Stream Processing

Deepak S. Turaga
Oden Technologies
deepak.turaga@gmail.com or dt2261@columbia.edu

Objectives

- Introduction to “Big Data” problems
  - Focus on streaming data and Internet of Things
  - Interaction of Machine Learning with streaming data
- Hands on exposure to stream processing
  - Systems and Programming
    - Spark Streaming and Apache Beam
    - On the cloud
  - Algorithms
- Fun “research-like” projects
  - Financial, Healthcare, Social Media, Natural Systems

Logistics

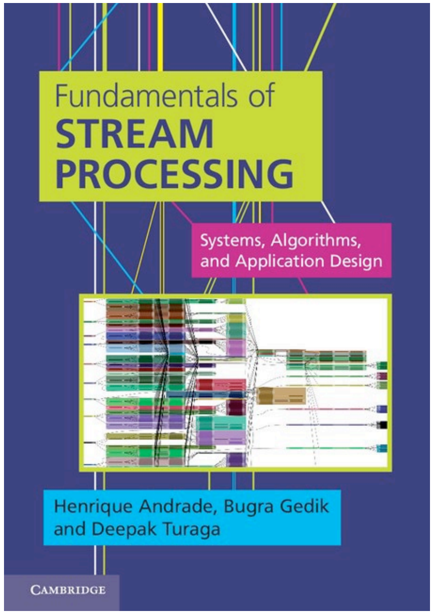
- Location
  - Lecture: Thursday 7:00 PM – 9:30 PM
  - 633 Mudd
- Office Hours
  - Thursday 6:00-6:45 PM – before class
  - Email welcome
    - dt2261@columbia.edu

- Homework
  - Two programming exercises that emphasize application development Spark/Beam programming
  - Typically Python/Java

- Seminar/Projects (group based)
  - Explore emerging areas
  - Written report and presentation
  - Demo

Logistics

- Course Evaluation
  - Homework – 30%
  - Seminar – 30%
  - Project – 40%
    - Graduate course ⇒ projects should tackle state of the art


Fundamentals of
STREAM
PROCESSING
Systems, Algorithms, and Application Design
Henrique Andrade, Bugra Gedik and Deepak Turaga
CAMBRIDGE

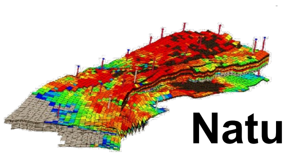
- Reference Material
  - Text Book
  - Tutorial and exploratory papers on current topics

- Course Prerequisites
  - Basics of Data Management and Relational Databases - Preferred
  - Basic Signal Processing and Time Series Analysis - Preferred
  - Basic Statistics and Data Analysis Techniques - Preferred
  - Basics of Distributed Systems - Preferred
  - Basics of Optimization Theory - Preferred
  - Programming skills in Python or Java - Mandatory

Logistics: Stream Processing Systems

- Spark and Spark Streaming
  - https://spark.apache.org/streaming/
- Apache Beam
  - https://beam.apache.org/

“Big Data” and Stream Processing



Natural Systems

- Seismic monitoring
- Wildfire management
- Water management

Stock market

- Impact of weather on securities prices
- Analyze market data at ultra-low latencies

Law Enforcement

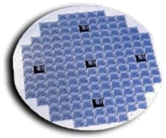
- Real-time multimodal surveillance
- Fraud detection and prevention

Transportation

- Intelligent traffic management

Social Media

- Sentiment analysis
- Customer profiling



Manufacturing

- Process control for microchip fabrication

Radio Astronomy

- Detection of transient events

Health & Life Sciences

- Neonatal ICU monitoring
- Epidemic early warning system
- Remote healthcare monitoring

Telecom

- Processing of Call Detail records
- Real-time services, billing, advertizing
- Business intelligence
- Churn Analysis, Fraud Detection

ELEN E6889 Course Outline

- Motivation for Large-Scale Stream Processing
  - Applications and Target Domains
- Fundamentals of Stream Processing
  - Systems: Distributed Systems, Transport, Control and Management
  - Development: Programming Stream Processing Applications
  - Algorithms: Stream Mining, Approximation Algorithms
- Advanced Research Topics
- Hands-on Exposure to Stream Processing
  - Programming Exercises on Stream Processing System
  - Student Projects – Concepts in practice
- Emerging Areas of Interest

Course Schedule

Date	Title	Description
Jan 19	Introduction	Introduction, History of Stream Processing, and Systems Concepts
Jan 26	Stream Processing Systems I	Spark and Spark Streaming
Feb 02	Stream Processing Systems II	Apache Beam (HW1)
Feb 09	Stream Relational Processing	Fundamental Concepts, Operations, Patterns and Optimizations
Feb 16	Stream Data Preprocessing	Descriptive Statistics, Sampling, Sketches (HW2)
Feb 23	Stream Data Preprocessing	Transforms, Dimensionality Reduction, Quantization (Seminar Topic Selection)
Mar 02	Seminar Presentations	
Mar 09	Seminar Presentations	
Mar 23	Stream Data Preprocessing	Transforms, Dimensionality Reduction
Mar 30	Project Proposal Reviews	
Apr 6	Stream Data Mining	Classification and Regression
Apr 13	Interim Project Help	
Apr 20	Stream Data Mining	Clustering, Frequent Patterns and Anomaly Detection
Apr 27	End-to-End Applications	Streaming in Practice
May 4	Advanced Topics	Mining Topologies, Distributed and Online Learning
May 11	Project Presentations	

Seminars and Projects

- Seminars
  - Research papers from the area
- Seminars from 2021 and 2022
  - Systems
  - Optimization and Algorithms
  - Stream Mining and Analysis
  - https://sites.google.com/site/fundamentalsofstreamprocessing/home

Projects

- Applications
  - Social Media, Financial, Healthcare, Natural Systems, Multimedia
- Systems
  - Performance optimization, fault tolerance
- Algorithms
  - Analytics, preprocessing and mining, data generation
- Others
  - Anything of your choice based on our discussion

Previous Project Topics

- Healthcare
  - Covid 19 Predictions with Various Methods
  - Heartbeat data analysis
  - States Coronavirus Sentiment Analysis based on Tweets
- Spatio-temporal Analysis
  - Spatio-temporal localized analysis of twitter content
  - Route optimization and visualization for NYC data
  - A real-time people counting application based on streaming data
  - Citi Bike Demand Prediction with Weather Information
- Social Media Analysis
  - Youtube video recommendation based on twitter
  - Sentiment analysis based on youtube
  - Dynamic Word Cloud Generator on Twitter
  - Real-time hotspot issue detection
  - Real-time twitter hot topic mining
  - Wikistats
- Financial Data Analysis
  - Stock/Bitcoin price prediction
  - Predicting stock prices with multi-dimensional data
  - Sentiment based stock price prediction
  - Evaluating the Correlation of Streamed Sentiments with Multiple Cryptocurrencies
- Deep Learning and Stream Processing
  - Real-time facial recognition
  - Deep learning model based speech to text

Comments/Questions (?)