DS256 Jan 3:1

Scalable Systems for Data Science

Instructor

Yogesh Simmhan
Email: simmhan@iisc.ac.in

Teaching Assistant

Aakash Khochare
Email: aakhochare@IISc.ac.in

Department: Department of Computational and Data Sciences

Course Time: Tue/Thu 330-5PM

Lecture venue: CDS 202

Detailed Course Page: http://cds.iisc.ac.in/courses/ds256/

Announcements

The first lecture will be on Jan 9, 330PM at CDS 202.

Brief description of the course

This course will teach the fundamental Systems aspects of designing and using Big Data platforms, which are a specialization of scalable systems for data science applications. It will cover topics on: Why Big Data platforms are necessary? How they are designed? What are the programming abstractions (e.g. MapReduce) that are used to compose data science applications? How the programming models are translated to scalable runtime execution on clusters and Clouds (e.g. Hadoop)? How do you design algorithms for analyzing large datasets? How do you map them to Big Data platforms? and How can these be used to develop Big Data applications in an integrated manner?

Prerequisites

This is an introductory course on platforms and tools required to develop analytics over Big Data. However, it builds upon prior knowledge that students have on computing and software systems, programming, data structures and algorithms. Students must be familiar with Data Structures (e.g. Arrays, Queues, Trees, Hashmaps, Graphs) and Algorithms (e.g. Sorting, Searching, Graph traversal, String algorithms, etc.).

It is recommended that students have good programming skills (preferably in Java) which is necessary for the programming assignments and projects. Familiarity with one or more of the following courses will also be helpful (although not mandatory): DS 221 (ISS), DS 295 (Parallel Programming), E0 253 (Operating Systems), E0 264 (Distributed Computing Systems), SE252 (Introduction to Cloud Computing), E0 225 (Design and Analysis of Algorithms), E0 232 (Probability and Statistics), E0 259 (Data Analytics).

Syllabus

This course will teach the fundamental Systems aspects of designing and using Big Data platforms, which are a specialization of scalable systems for data science applications. This course will address three facets of these platforms.

The design of distributed program models and abstractions, such as MapReduce, Dataflow and Vertex-centric models, for processing volume, velocity and linked datasets, and for storing and querying over NoSQL datasets.

The approaches and design patterns to translate existing data-intensive algorithms and analytics into these distributed programming abstractions.

Distributed software architectures, runtime and storage strategies used by Big Data platforms such as Apache Hadoop, Spark, Storm, Giraph and Hive to execute applications developed using these models on commodity clusters and Clouds in a scalable manner.

It will cover topics on: Why Big Data platforms are necessary? How they are designed? What are the programming abstractions (e.g. MapReduce) that are used to compose data science applications? How the programming models are translated to scalable runtime execution on clusters and Clouds (e.g. Hadoop)? How do you design algorithms for analyzing large datasets? How do you map them to Big Data platforms? and

How can these be used to develop Big Data applications in an integrated manner?

As part of a hands-on Project in this course, students will work with real, large datasets and commodity clusters, and use scalable algorithms and platforms to develop a Big Data application. The emphasis will be on designing applications that show good â€œweak scalingâ€• as the size, speed or complexity of data increases, and using distributed systems such as commodity clusters and Clouds.

Besides class lectures, there will be several guest lectures by experts from the Industry who work on Big Data platforms, Cloud computing and data science.

Course outcomes

At the end of the course, students will have learned about the following concepts.

1) Types of Big Data, Design goals of Big Data platforms, and where in the systems landscape these platforms fall.

2) Distributed programming models for Big Data, including Map Reduce, Stream processing and Graph processing.

3) Runtime Systems for Big Data platforms and their optimizations on commodity clusters and Clouds.

4) Scaling data Science algorithms and analytics using Big Data platforms.

Grading policy

45% Homework Three programming assignments (5%+2*20% points)

30% Project One final project, to be done individually or in teams

20% Exams One Final exam

5% Participation Participation (i.e. not just â€œattendanceâ€•) in classroom discussions and online forum for the course

Assignments

Three programming assignments on big data platforms.

One project assignment as individual or teams.

Resources

Textbook:

Select chapters from Data-Intensive Text Processing with MapReduce, Jimmy Lin and Chris Dyer, 1st Edition, Morgan & Claypool Publishers, 2010

Select chapters from Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman and Jeff Ullman, 2nd Edition (v2.1), 2014.

Current literature and online documentation

Cluster Access: Students will validate their assignments and projects on the CDS turing cluster, and Cloud resources. Details for accessing the cluster and running programs on it will be covered in a lab session.

Hadoop on turing

Storm on turing

Giraph/GoFFish on turing