CS4830: Big Data Laboratory

Faculty: Prof. Balaraman Ravindran
Guest Faculty: Mr. Rangarajan Vasudevan
DCF Ext. (ravi@cse.iitm.ac.in)
Jan-Apr Semester 2020 ; 'T' Slot;
Slot is: Fri (2 - 4.40pm);
TA(s): Pranshu Malviya (cs19s031@smail.iitm.ac.in),
Bikash Kumar Behera (cs19m019@smail.iitm.ac.in),
Ramya Kasaraneni (cs19d003@smail.iitm.ac.in)

I. Course Objectives

This course will introduce students to practical aspects of analytics at a large scale, i.e. big data. The course will start with a basic introduction to big data and cloud concepts spanning hardware, systems and software, and then delve into the details of algorithm design and execution at large scale.

II. Learning Outcomes

- Introduction to Cloud Concepts: Cloud-Native architecture, serverless computing, message queues, PaaS, SaaS, IaaS
- Introduction to Big Data concepts: divide-and-conquer, parallel algorithms, distributed virtualized storage, distributed resource management, orchestration and scheduling, real-time processing.
- Technology deep-dive: Google Cloud Storage, GCP Dataflow, Google Pub/Sub, Cloud Functions
- Analytics at Large Scale: Google Colab, Data Studio, BigQuery, Integration with Tensorflow/Pytorch.

III. Classroom Mode

A Hands-on lab interspersed with theory sessions for the associated concepts.

IV. Textbook(s)/Reference(s)

1) Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. Mining of massive datasets. Cambridge university press, 2014.
2) Tom Laszewski, Kamal Arora, Erik Farr and Piyum Zonooz. Cloud Native Architectures: Design high-availability and cost-effective applications for the cloud. Packt Publishing Ltd, 2018.

V. Additional Resources

- Getting Started with GCP: (link)
- GCP VMs: (link)

VI. Course Requirements

You are *required* to attend all the lab sessions. If you miss any of them, it is your responsibility to find out what went on during the labs and submit the associated lab assignment on time. You are required to adhere to the submission deadlines of reports/assignments. Class participation is strongly encouraged to demonstrate an appropriate level of understanding of the material being discussed in the class.

VII. Planned Syllabus

The following topics will be covered:

- Introduction to cloud-native architecture, including microservices and service registry.
- Serverless computing using Google Cloud Functions, including the usage of message queues, notifications and triggers.
- Understanding the difference between PaaS, Saas and IaaS.
- Overview of spark and its divide-and-conquer technique.

- Batch processing using spark-like queries on GCP Dataflow and integration with Google Cloud Storage.
- Moving to real-time processing and contrasting it with batch processing, with a focus on Google Pub/Sub as an example.
- Introduction to Google Machine learning Engine and BigQuery.
- Deep-learning on the cloud using GCP. Distributed GPU-based processing on the cloud with Tensorflow.
- Stitching all the concepts together using an end-to-end application involving big data analysis utilizing the tools discussed above.

VIII. Tentative Grading Scheme

The following allocation of points is tentative. These may change during the semester.

- Every lab is for 10 marks (implementation + report)
- The final evaluation will be for 20 marks.

IX. Academic Honesty

Academic honesty is expected from each student participating in the course. NO sharing (willing, unwilling, knowing, unknowing) of reports/assignments between students, submission of downloaded material (from the Internet, Campus LAN, or anywhere else) is allowed.

The project work done as a part of this course cannot be used as-is, to meet any other degree requirements. The project must NOT be copied/downloaded material from the Internet or elsewhere.

Academic violations will be handled by IITM Senate Discipline and Welfare (DISCO) Committee. Typically, the first violation instance will result in ZERO marks for the corresponding component of the Course Grade and a drop of one- penalty in overall course grade. The second instance of code copying will result in a U Course Grade and/or other penalties. The DISCO Committee can also impose additional penalties.

Please protect your Moodle account password. Do not share it with ANYONE. Do not share your academic disk drive space on the Campus LAN.