# DA 331 - Big Data Analytics : Tools & Techniques

Fall 2023

| | | | |
|---|---|---|---|
| **Instructor:** | Chiranjib Sur | **Time:** | Tuesday - 9:00-9:55 (5103)<br>Wednesday - 10:00-10:55 (5103)<br>Thursday - 11:00-12:55 (MDSAI Lab) |
| **Email:** | chiranjib@iitg.ac.in | **Place:** | 5103. |

**Course Pages:**

1. `chiranjibsuruf.github.io/courses/fall23da331.html`

**Office Hours:** After class (Tuesday 10:00-11:00, Wednesday 11:00-12:00), or by appointment, or post your questions in Microsoft Teams Group.

**Text Book:** Various interesting and useful topics that will be touched during the course are discussed in the following textbooks.

- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman, Cambridge University Press, 2nd Edition, 2014. [Link (http://mmds.org/)]
- Materials and Chapters will be referred when required.

**References:** No Need to Buy.

- Materials and Chapters related to Tools and Technology will be referred when required.

**Objectives:** Big Data Analytics is one of the most highly sought-after skills in the industry. In this course, you will learn the foundations of Big Data Tools, understand how to build a scalable system, and learn how to lead successful deployment projects and solve critical problems.

**Prerequisites:** An undergraduate-level understanding of Data Structure, Database Systems, Operating Systems, and Competitive Programming Skills is required to be successful. Being Comfortable with Probability theory and with Linear algebra is assumed.

**Tentative Course Outline:**

Fundamentals of Big Data: Understanding big data, datasets, data analysis, data analytics, big data characteristics, types of data, case studies; **(M1)**

Big data adoption and planning considerations: data procurement, big data analytics lifecycle, case study examples; **(M2)**

Big data storage concepts: cluster computing, file system, distributed file systems, Relational & non-relational databases, scaling up & scaling out storage; **(M3)**

No-SQL: Data types, Creating, Updating & Deleting documents, Querying, An example No-SQL database; **(M4)**

Distributed computing framework: Introduction, file system, MapReduce programming model, examples of distributed computing environment framework; **(M5)**

Stream data processing: tools such as Apache Spark, Apache Storm; Analytics with distributed computing framework: supervised learning examples, unsupervised learning examples. **(M6)**

| Lecture | Lab |
|---|---|
| Big Data – Introduction **(M1)**<br>Distributed Operating System, Grid Computing, Cloud Computing | No Lab |
| Distributed System<br>Chord, Gossip, Tapestry, Pastry, Blockchain | Implementation in Scala |
| Big Data Data-Structures<br>KD Trees, Bloom Filter | No Lab |
| Big Data Technologies (Retrieval Techs) **(M6)**<br>Apache Spark, ML in Apache Spark, Apache Airflow, Kafka, Hive | Implementation in Python |
| Big Data Databases **(M3, M4)**<br>Mongo dB, Apache Cassandra | Assignments in JavaScripts |
| Big Data Processing **(M5)**<br>Map Reduce and Hadoop, Apache Pig | Assignments in Python<br>for Hadoop |
| Case Studies **(M1, M2)**<br>more to be added | Parallel with Final Project |
| Fundamental Big Data Problems<br>K-Means, XBoost, Connected Component Problem<br>more to be added | No Lab |

Please note: The syllabus is subject to change.

**Grading Policy:** Refer Website.

**Important Dates:** Will be announced later.

**Class Policy:**

- Regular attendance is not essential but expected.

**Academic Honesty:** We encourage students to form groups to discuss different topics. Students may discuss and work on programming assignments and quizzes in groups. However, each student must write down the solutions independently and without referring to written notes from the joint session. In other words, each student must understand the solution well enough in order to reconstruct it by him/herself. In addition, each student should submit his/her own code and mention anyone he/she collaborated with.

Refer this CODE OF CONDUCT PLEDGE for IIT Guwahati