

# **DA512 Big Data Processing using Hadoop**

## **Storing Big Data : Apache Hadoop**

- Introduction to Apache Hadoop and the Hadoop Ecosystem
- Apache Hadoop Overview
- Apache Hadoop Cluster Components
- HDFS Architecture
- YARN Architecture

## **Processing Big Data: Apache Spark**

- What is Apache Spark?
- Getting Started with RDD and DataFrames
- RDD and DataFrame Operations
- Spark SQL

## **Machine Learning on Big Data : Apache Spark ML**

- Spark ML
- Regression, Classification and Clustering
- Feature Selection and Vectorizing
- Parameter Tuning, Cross Validation

## **Processing Data Streams : Apache Spark Streaming & Kafka**

- Apache Spark Streaming Overview
- Sliding Window Operations
- Structured Streaming
- Apache Kafka Data Sources
- Kafka Topics
- Kafka Producers and Consumers

## **Grading Policy**

- Homework-1 ----- %20
- Homework-2 ----- %20
- Homework-3 ----- %20
- Final Exam ----- %40