Big Data Analytics Course Outline

Session1: Introduction to Big Data, Hadoop and MapReduce

- Which data is called as Big Data, Applications of Big Data
- Introduction to AWS Academy, connection & basic operations
- Traditional Data warehousing & Big Data
- Introduction to Distributed Computing Environment, Hadoop & Its Ecosystem
- HDFS Architectures, HDFS Commands
- MapReduce

Session 2: Data Ingestion with Hive Sqoop

- Intro to Hive
- Hive Architecture
- HQL
- Bucketing and partitioning in Hive
- Importing/Exporting data to/from Hive using Sqoop
- File formats -ORC, Parquet, Avro
- Flume* (Optional)

Session 3: Introduction to NoSQL, MongoDB & PyMongo

- Introduction to NoSQL
- Cap Theorem & NoSQL Database types
- MongoDB & Its Features
- MongoDB & Its Features
- MongoDB collections, documents and operations
- Intro to PyMongo, Install PyMongo & Python Driver, connect to MongoDB
- Perform basic Create, Retrieve, Update and Delete (CRUD) operations using PyMongo
- Sharding* (optional)

*Optional/not part of end semester exam

Session 4: Apache Spark & Spark SQL

- Introduction to Spark
- Spark Architecture
- PySpark and Data Bricks
- Introduction to Spark SQL
- Spark SQL as an ETL tool
- Spark SQL Performance Tuning

Session 5: PySpark ML

- Intro to Spark ML
- Spark ML Pipeline - Transformers, Estimators
- Spark ML Component Flow
- Spark ML Data Types
- Spark ML Algorithms
- Building Pipeline
- Model Persistence

Session 6: Stream Processing: Spark-Streaming and Kafka*

- What is Streaming Data
- Intro to Spark Streaming & Its working
- Spark Streaming + Kafka Example
- Intro to Kafka & Its working
- Kafka Commands

Session 7: Case Study*

*Optional/not part of end semester exam