

BIG DATA FRAMEWORKS

CSE6701

Prof. Ramesh Ragala

December 3, 2018

COURSE OBJECTIVE

1. To impart an understanding of the challenges in storing and processing big data.
2. How to use different big data frameworks effectively to store and process big data.

EXPECTED OUTCOMES

On Completion of the course, the students will be able to

1. Discuss the challenges in Big Data.
2. Describe the need of different big data frameworks
3. Write Map Reduce programming in both Hadoop and Spark Framework
4. Write programs in Spark Streaming, SPARK SQL and GraphX

- Data Storage and Analysis
- Characteristics of Big Data
- Big Data Analytics
- Typical Analytical Architecture
- Requirement of New Analytical Architecture
- Challenges in Big Data Analytics
- Need of big data framework

- Requirement of Hadoop Framework
- Design Principle of Hadoop
- Comparison with other system
- Hadoop Components
- Hadoop - 1 vs Hadoop - 2
- Hadoop Daemons
- HDFS Commands
- Map Reduce Programming : Introduction
- I/O Formats
- Map side Join
- Reduce Side Join
- Secondary Storage sorting
- Pipelining Map Reduce jobs

- Introduction to Hadoop Ecosystem Technologies
- Serialization : AVRO
- Co-Ordination : Zookeeper
- Databases : HBase and Hive
- Scripting Language : Pig
- Streaming : Flink and Storm

- Overview of Spark
- Hadoop Vs Spark
- Cluster Design
- Cluster Management - Performance
- Application Programming Interface (API):
 - Spark Context
 - Resilient Distributed Datasets
 - Creating RDD
 - RDD Operations
 - Saving RDD
 - Lazy Operations
- Lazy Operations
- Spark Jobs

UNIT - V: INTERACTIVE DATA ANALYSIS WITH SPARK SHELL



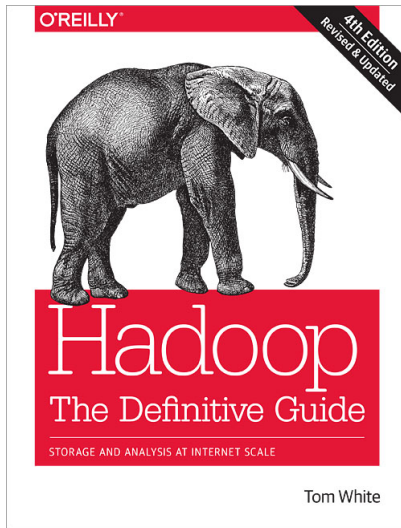
- Writing Spark Application
- Spark Programming in Scala
- Spark Programming in Java
- Spark integration with R
- Spark Programming with Python

- SQL Context
- Importing and Saving Data
- Data Frames
- Using SQL
- GraphX Overview
- Creating Graph
- Graph Algorithms

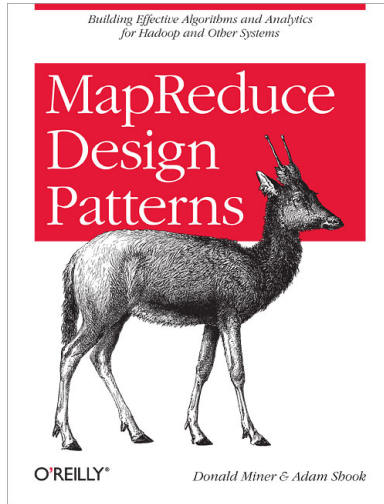
- Spark Streaming Overview
- Errors and Recovery
- Streaming Source
- Streaming Live Data with Spark

- **Guest Lecture from Industry experts**

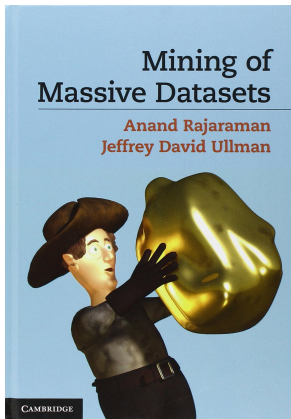
- **Hadoop: The Definitive Guide, 4th Edition** by Tom White



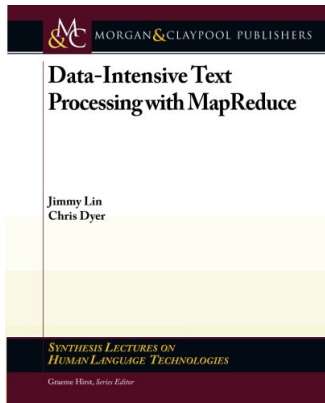
- **MapReduce Design Patterns, 1st Edition** by Donald Miner, Adam Shook



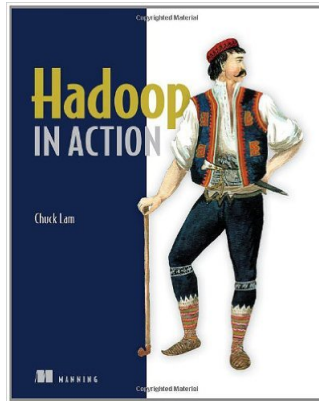
- **Mining of Massive Datasets** by Anand Rajaraman and Jeffrey David Ullman



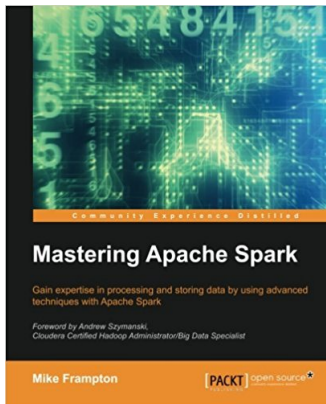
- **Data-Intensive Text Processing with MapReduce** by Jimmy Lin, Chris Dyer and Graeme Hirst



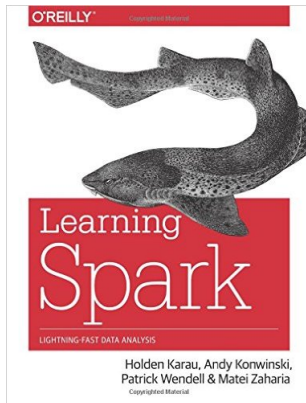
- **Hadoop in Action** by Chuck Lam



- **Mastering Apache Spark** by Mike Frampton



- **Learning Spark: Lightning-Fast Big Data Analysis** by Holden Karau, Andy Konwinski, Patrick Wendell and Matei Zaharia



- **email ID:** ramesh.ragala@vit.ac.in
- **Mobile No:** 9087277270
- **Room No:AB1-604, Cabin No: 8**

- HDFS Commands
- Map Reduce Program to show the need of Combiner
- Map Reduce I/O Format - Text, Key-Value
- Map Reduce I/O Format - NLine, Multiline
- Sequence file I/O Format
- Secondary Sorting
- Distributed Cache, Map Side Join and Reduce Side Join
- Building and Running Spark Application
- Wordcount in Hadoop and Spark
- Manipulating RDD
- Inverted Index using Spark
- Sequence Alignment problem in Spark
- Implementation of Matrix algorithm in Spark
- Spark Sql Programming
- Building Spark Streaming Application