Big Data Frameworks CSE6701

Prof. Ramesh Ragala

December 10, 2018

Introduction

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

UNIT - I : Introduction to Big Data

- 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

UNIT - II: Hadoop 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

UNIT - III: Hadoop Ecosystem

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

UNIT - IV: Spark Framework

- 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 wi Spark Shell

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

UNIT - VI: Spark SQL and GraphX

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

UNIT - VII: Spark Streaming

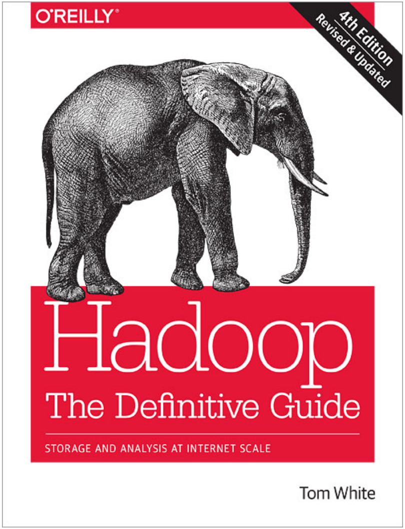
- Spark Streaming Overview
- Errors and Recovery
- Streaming Source
- Streaming Live Data with Spark

UNIT - VIII: Recent Trends

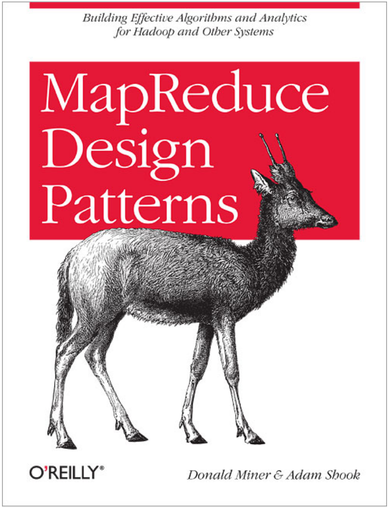
- **Guest Lecture from Industry experts**

Introduction

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

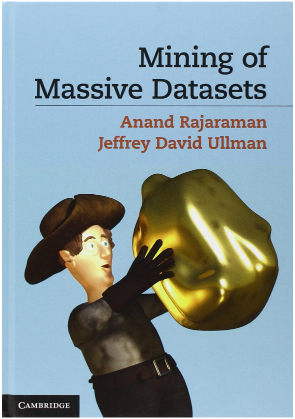


- **MapReduce Design Patterns, Ist Edition** by Donald Miner, Adam Shook


Building Effective Algorithms and Analytics for Hadoop and Other Systems
MapReduce Design Patterns
O'REILLY®
Donald Miner & Adam Shook

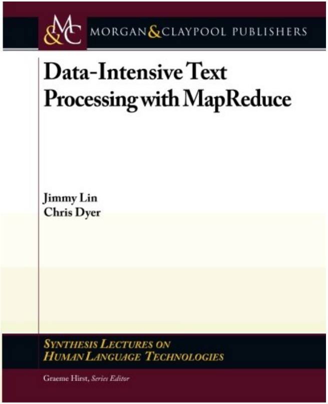
Introduction

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



Introduction

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



MORGAN&CLAYPOOL PUBLISHERS

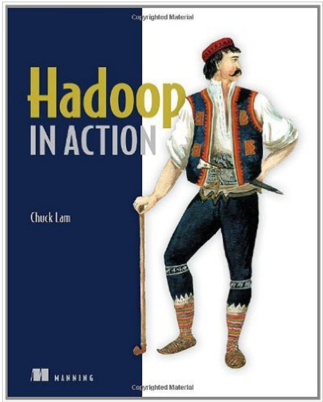
Data-Intensive Text Processing with MapReduce

Jimmy Lin
Chris Dyer

Synthesis Lectures on Human Language Technologies

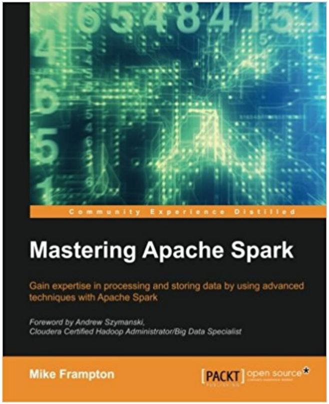
Graeme Hirst, *Series Editor*

- **Hadoop in Action** by Chuck Lam



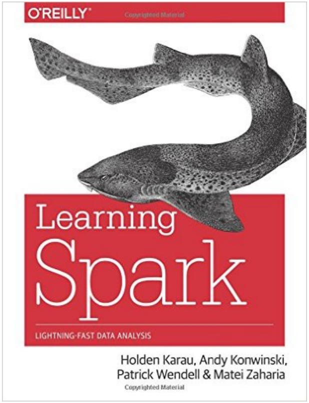
Introduction

- **Mastering Apache Spark** by Mike Frampton


Community Experience Distilled
Mastering Apache Spark
Gain expertise in processing and storing data by using advanced techniques with Apache Spark
Foreword by Andrew Szymanski,
Cloudera Certified Hadoop Administrator/Big Data Specialist
Mike Frampton
[PACKT] open source

Introduction

- **Learning Spark: Lightning-Fast Big Data Analysis** by Holdern karau, Andy Konwinski, Patrick wendell and Metai Zaharia



O'REILLY

Learning Spark

LIGHTNING-FAST DATA ANALYSIS

Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia

Communication

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

Lab Experiments

- 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