VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Large Scale Data Processing

Prof. Ramesh Ragala

VIT Chennai

Introduction

- **Course Objective:**
  - Understand different characteristics of big data.
  - Understand the requirement of big data frameworks
  - Learn the concepts of distributed file system
  - Provide MapReduce programming environment
  - Understand need of inverted indexing and graph data analytics

Introduction

- **Expected Course Outcome**:
  - Define the characteristics of big data and explain the data science life cycle.
  - Differentiate between conventional and contemporary distributed framework.
  - Characterize storage and processing of large data.
  - Implement and demonstrate the use of the Hadoop eco- system
  - Compare scalable frameworks for large data.

Introduction

- **Expected Course Outcome**:
  - Identify independent tasks in a program that may be parallelized.
  - Decompose a problem into map and reduce operations for implementation.
  - Recognize different input output formats for map reduce programs.
  - Design programs to analyze large scale text data.
  - Identify problems suitable for use of graph mining in large data processing.

UNIT – I: Introduction to Big Data and Analytics

- Big Data Overview
- Characteristics of Big Data
- Business Intelligence vs Data Analytics

Module – II: Need of Data Analytics

- Data Analytics Life Cycle
- Data Analytics in Industries
- Exploring Big Data
- Challenges in handling Big Data

Module – III: Big Data Tools

- Need of Big Data Tools
- Understanding Distributed System
- Overview of Hadoop
- Comparing SQL databases and Hadoop
- Hadoop Eco System
- HDFS: Distributed File System
- Design of HDFS
- Writing Files to HDFS
- Reading Files from HDFS

Module – IV: Hadoop Architecture

- Hadoop Daemons
- Hadoop Cluster Architecture
- YARN Yet Another Resource Negotiator
- Advantages of YARN

Module – V: Introduction to MapReduce

- Developing MapReduce Program
- Anatomy of MapReduce Code
- Simple MapReduce Code
  - Counting Things
- Map Phase
- Shuffle and Sorting Phase
- Reduce Phase
- Master Slave Architecture
- Job Processing in Hadoop
- MapReduce Pipelining

Module – VI: MapReduce Programming Concepts

- Use of Combiner
- Block Vs Split Size
- Working with Input and Output Formats
  - Key
  - Text
  - Sequence
  - Nline File format
  - XML file Format

Module – VII: Inverted Indexing and Graph Analytics

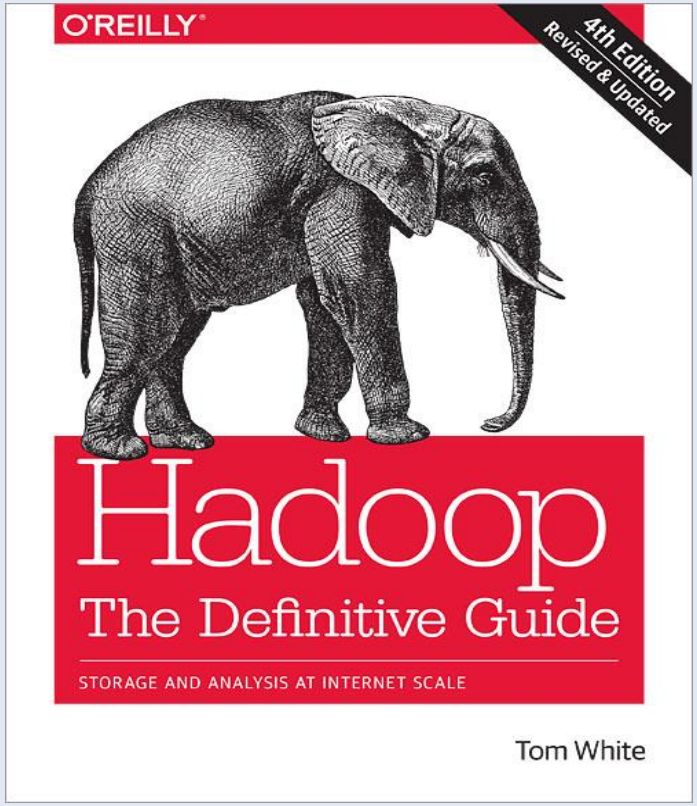
- Web Crawling
- Inverted Index
- Baseline and revised Implementation
- Graph Representation
- Parallel Breath First Search
- Page Rank
- Issues with graph Processing

Module – VIII: Recent Trends

Guest Lecture from Industry Expert

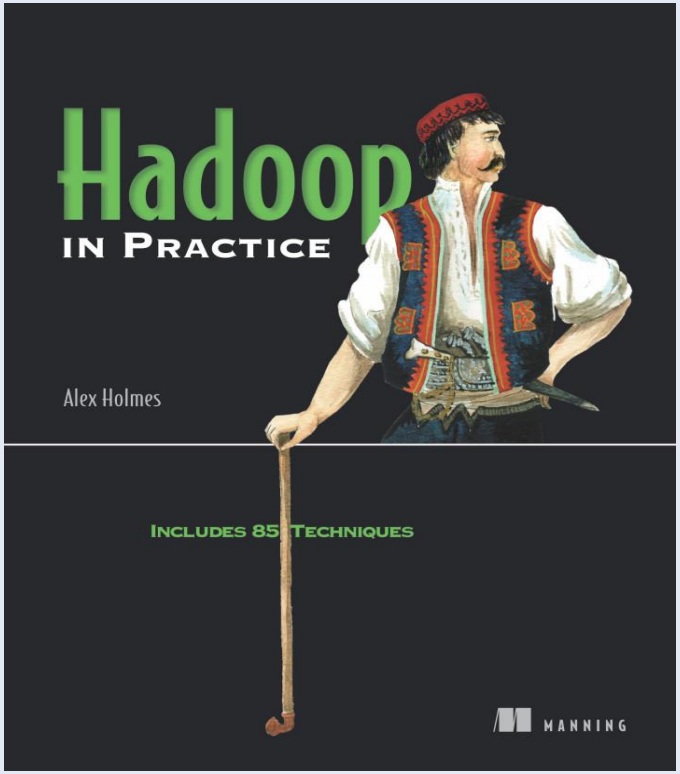
Text Books

- **Tom White, Hadoop The Definitive Guide, O'Reilly, 4th Edition, 2015**



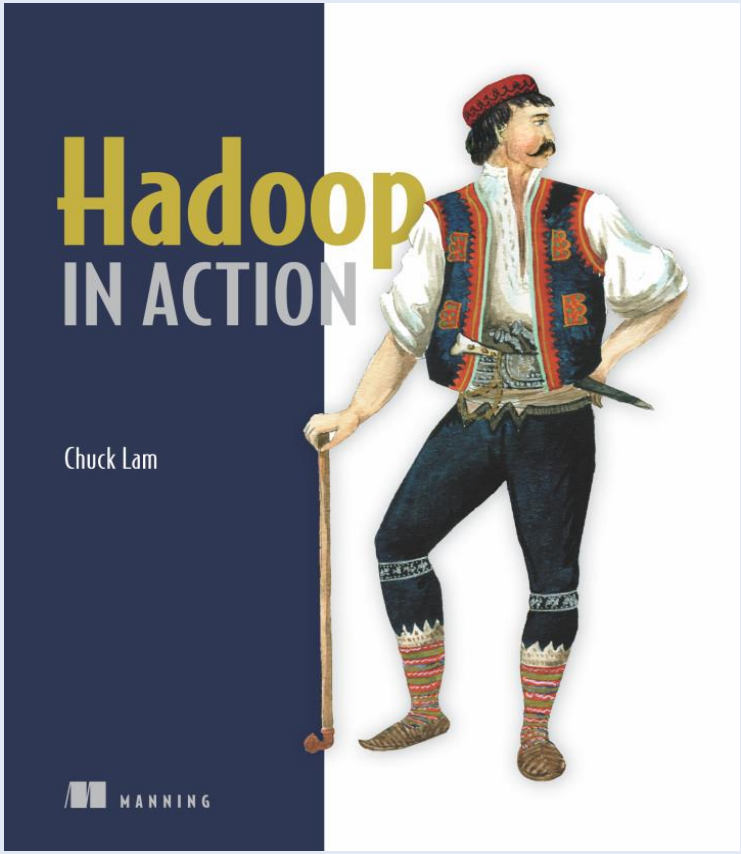
Reference Books

- **Alex Holmes, Hadoop in Practice, Manning Shelter Island, 2012**



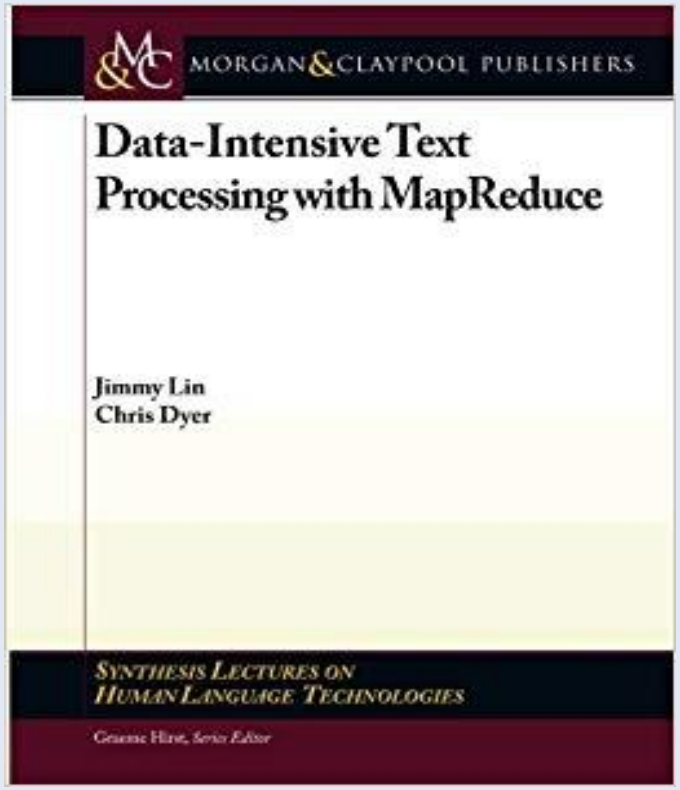
Reference Books

- **Chuck Lam, Hadoop in Action. Manning Shelter Island, 2011**



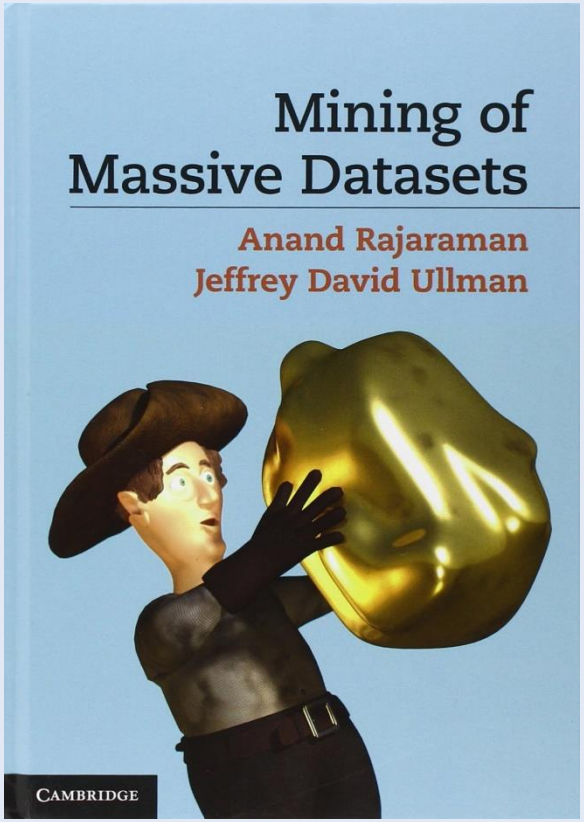
Reference Books

- **Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with Map Reduce, 2010**



Reference Books

- **Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets, 2011**



Lab Experiments

- Setting up Hadoop in Single node and Multinode environment
- Command line interface with HDFS
- Counting things using MapReduce
- Map Reduce Program to show the need of Combiner
- Map Reduce I/O Formats – key- value, Text
- Map Reduce I/O Formats – N line
- Multiline I/O
- Parallel Breadth First Search
- Sequence file Input / Output Formats

Lab Experiments

- Baseline Inverted Indexing using Map Reduce
- Revised Inverted Indexing using Map Reduce
- Matrix Factorization using Map Reduce
- Video Processing using Map Reduce
- BioInformatics (Protein/Gene Sequence etc) processing with MapReduce

Thank you