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

Big Data Analytics
ITA6008

Prof. Ramesh Ragala

https://github.com/rameshragala

June 28, 2021

Course Objective

1. To understand the big data platform and its use cases.
2. To impart knowledge in applying skills and tools to manage and analyze the big data.
3. To apply analytics on structured and unstructured data.

Expected Course Outcomes

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

1. Demonstrate knowledge of the fundamental elements and concepts related to big data.
2. Analyze the core architectural concepts to meet the challenges in implementing big data systems.
3. Design and develop a Big Data Environment according to the benchmarks.
4. Setup a Big Data Environment and implement security techniques.
5. Evaluate the use of data through cleansing, warehousing, analytics, and visualization to the ultimate business decision.
6. Analyze the data using various statistical methods.
7. Develop applications using large scale analytics tools to solve open big data problems.

Student Learning Outcome

2. Having a clear understanding of the subject related concepts and of contemporary issues.

7. Having computational thinking (Ability to translate vast data in to abstract concepts and to understand database reasoning).

14. Having an ability to design and conduct experiments, as well as to analyze and interpret data.

- Big Data Overview
- Characteristics of Big Data
- State of practice in analytics
- Role of Data Scientists
- Examples of Big Data Analytics
- Data Analytics Lifecycle

- Hadoop Components
- Design Principle of Hadoop
- Analyzing Big data with Hadoop
- Design of HDFS
- Developing a Map reduce Application

- Distributed File System(DFS)
- Map Reduce
- Algorithms using Map Reduce
- Communication cost Model
- Graph Model for Map Reduce Problem

- Setting up a Hadoop Cluster
- Hadoop Configuration
- Security in Hadoop
- Administering Hadoop
- Hadoop Benchmarks
- Hadoop in the cloud

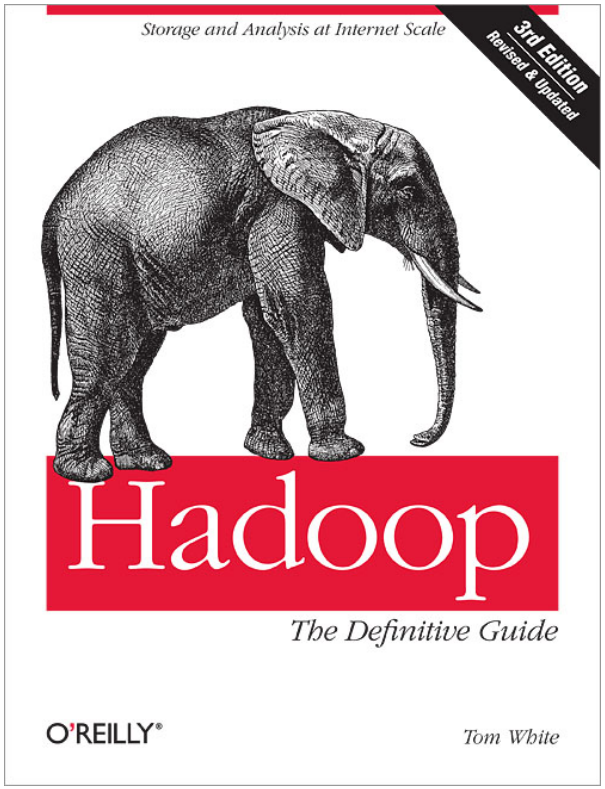
- Introduction to R
- R Graphical user interfaces
- Data import and export
- Attribute and Data Types
- Descriptive Statistics
- Exploratory Data Analysis

- Hypothesis Testing
- Difference of Means
- Wilcoxon Rank-Sum Test
- Type I and Type II errors
- Power and Sample size
- ANOVA

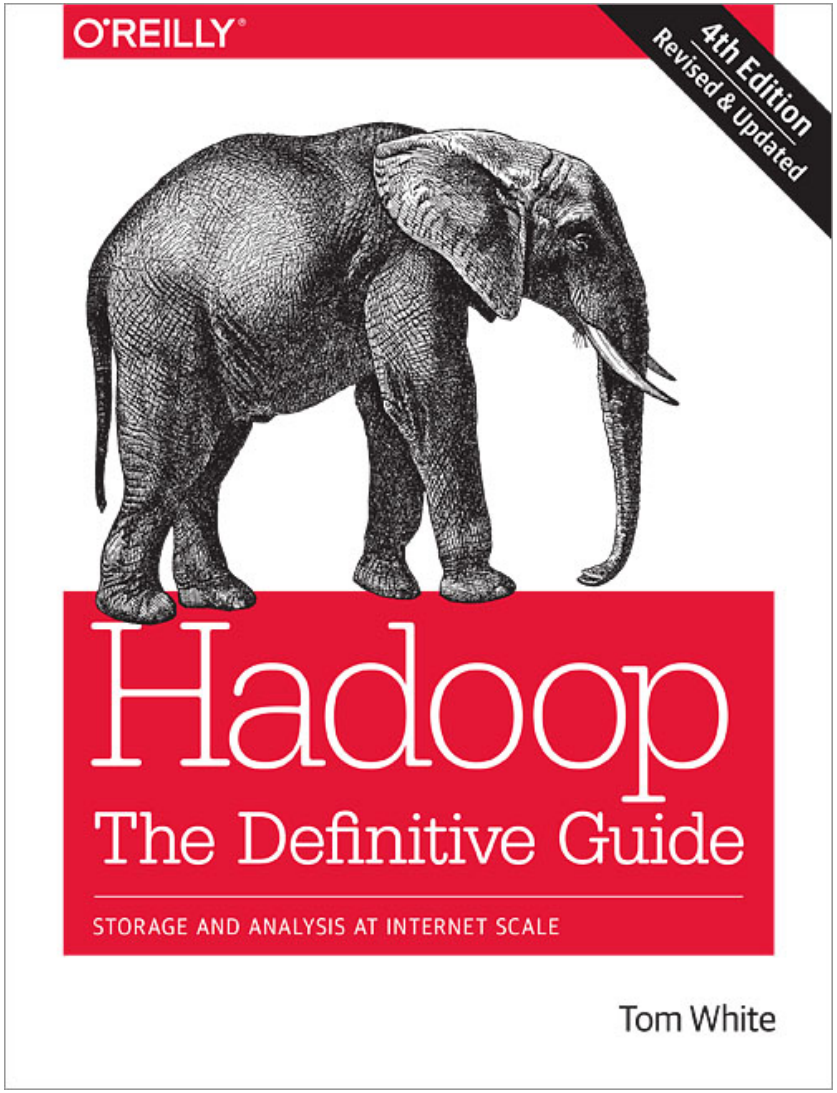
- Analytics for unstructured data
- The Hadoop ecosystem
  - pig
  - Hive
  - Hbase
  - Mahout
  - NoSQL

- **Guest Lecture from Industry experts**

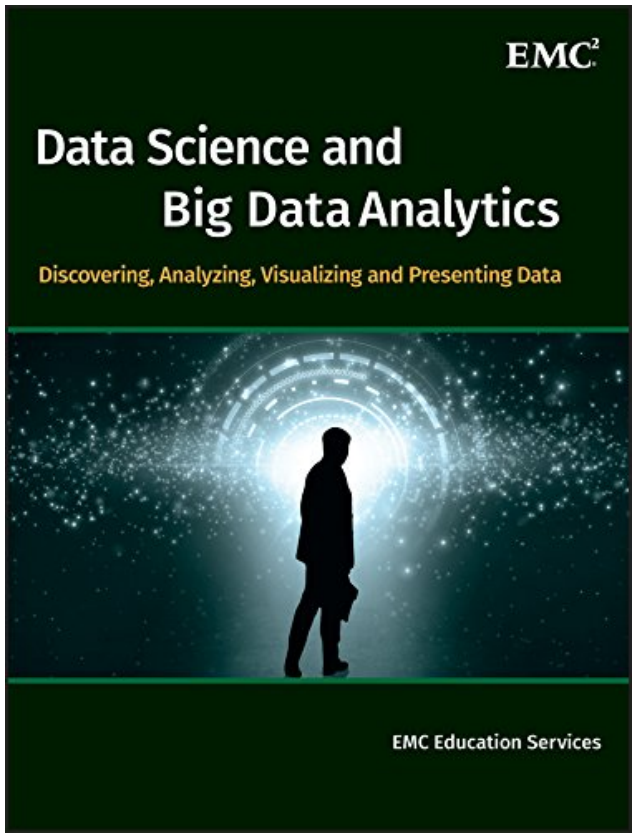
- **Hadoop: The Definitive Guide, 3rd Edition** by Tom White



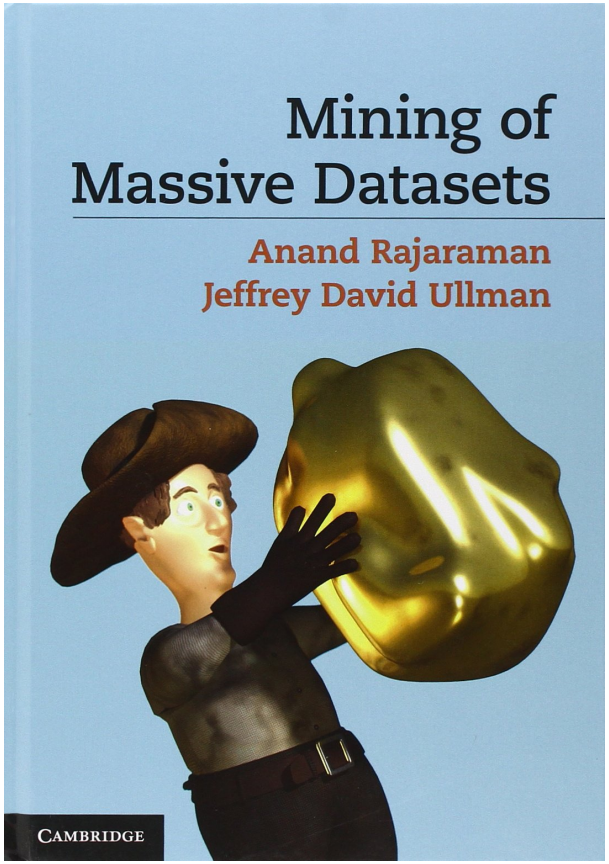
- **Hadoop: The Definitive Guide, 4^{th} Edition** by Tom White



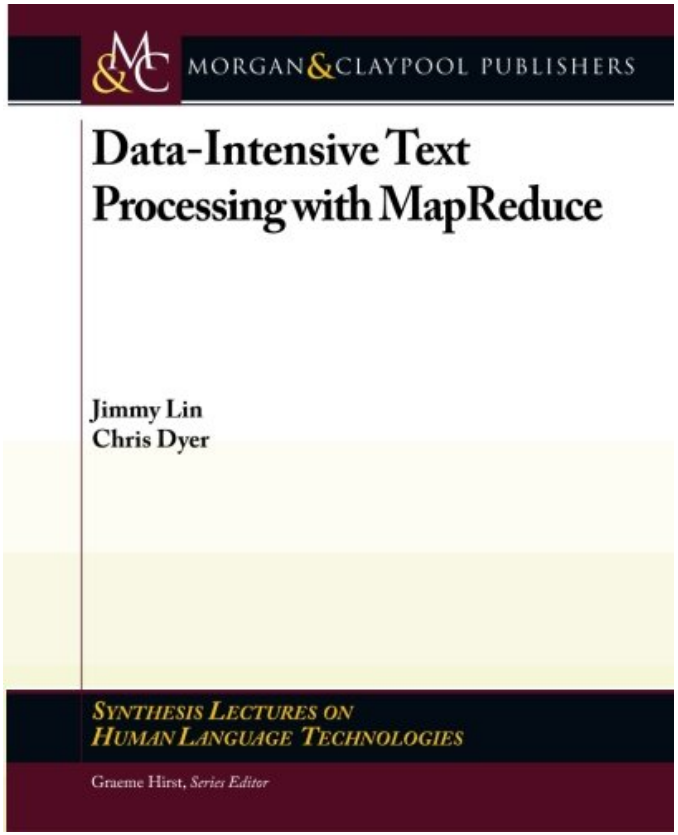
- **Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data 2015** by by EMC Education Services



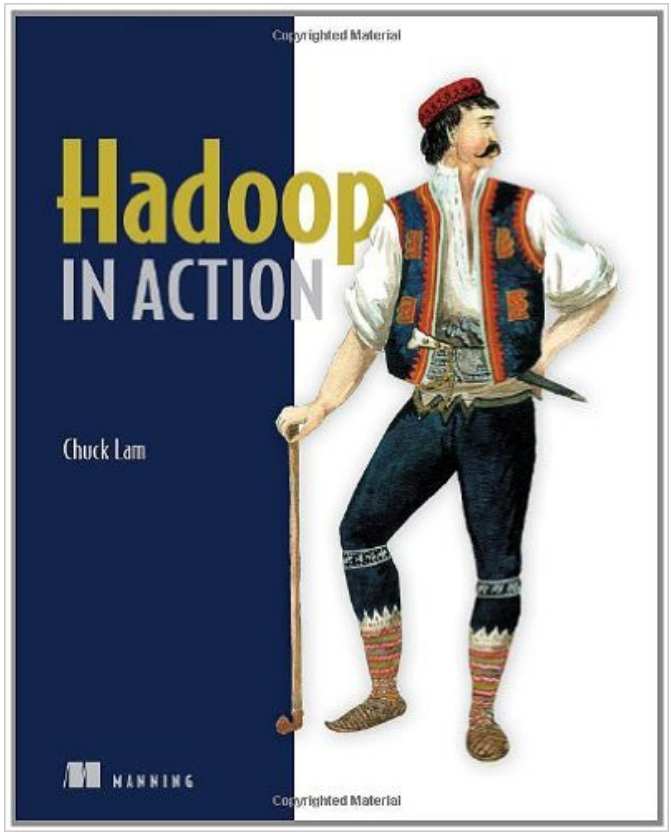
- **Mining of Massive Datasets** by Anand Rajaraman and Jeffrey David Ullman



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



- **Hadoop in Action** by Chuck Lam



Communication

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