


# (Big) Data Engineering In Depth

From Beginner to Professional

Mostafa Alaa Mohamed

Senior Big Data Engineer

 MoustafaAlaa **in** Moustafa Alaa  @Moustafa\_alaa22

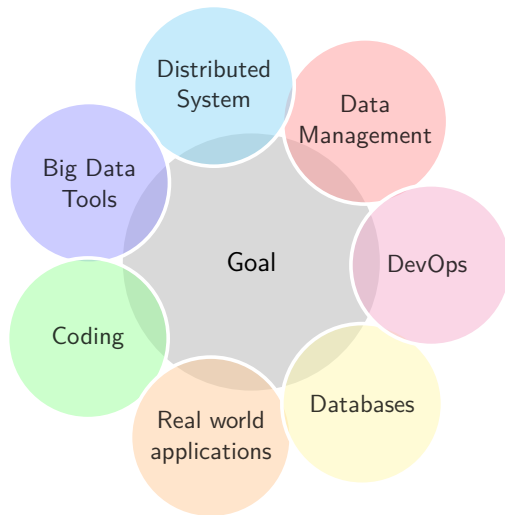
 mustafa.alaa.mohamed@gmail.com

<sup>1</sup>Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

# Course Introduction

# Course Target



## Learning Objectives and Audience

# Learning Objectives

Ch.2

Simplify the concepts in data management.  
Understand the data management life-cycle

Ch.3

Illustrate the basics of distributed systems concepts.

Ch.4/6

Be familiar with ETL for (batch/streaming) data  
over distributed systems ex: Hadoop & Spark.

Ch.6/7

Apply QA and testing the data.

Ch.7

Building real-life examples.

# Learning Objectives

Ch.6/7

Applying machine learning over big data.

Ch.9/12

Build and scale your data product.

Appx. H

Understanding of the DevOps tools and its functions in data life-cycle and development automation (e2e).

# Videos classification

| Watching Method<br>/ Audience | Computer | Mobile/Tablet | Just listening |
|-------------------------------|----------|---------------|----------------|
| Developer                     | ●        |               |                |
| DevOps                        |          | ●             |                |
| Business                      |          |               | ●              |

Table: Video classification

The green circle ● means short video.

The blue circle ● means medium video.

The red circle ● means long video

# Videos classification Example

| Watching Method<br>/ Audience | Computer | Mobile/Tablet | Just listening |
|-------------------------------|----------|---------------|----------------|
| Developer                     |          |               | ●              |
| DevOps                        |          |               | ●              |
| Business                      |          |               | ●              |

Table: Video classification

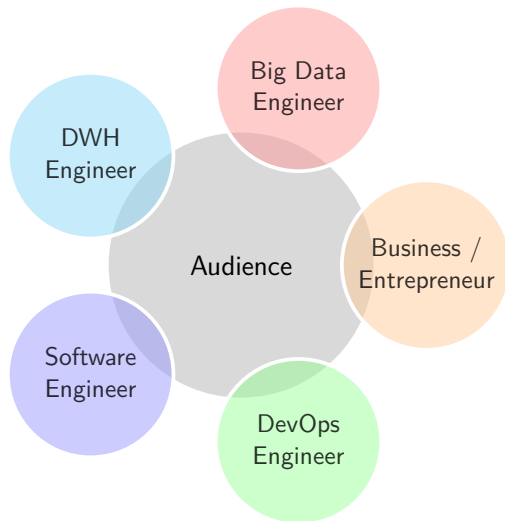
The green circle ● means short video.

The blue circle ● means medium video.

The red circle ● means long video



# Who Should Take This Course?



Getting max benefit from this course

# Getting max benefit from this course

## Take the course advantage

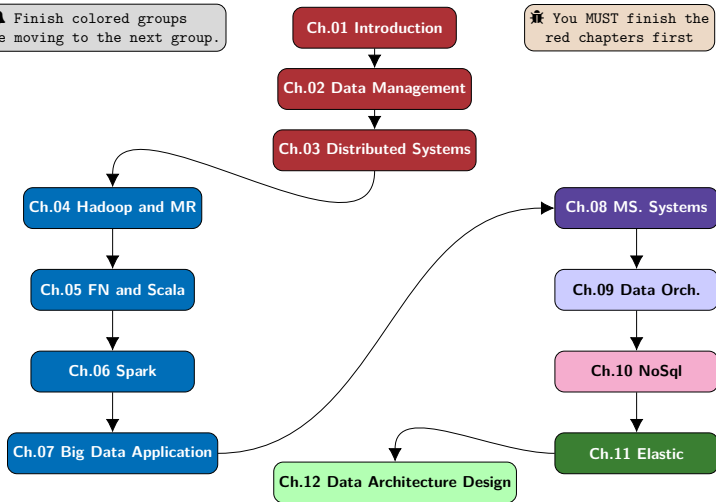
- Follow the order of the videos as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.
- Do the assignments.
- Ask your questions.
- Join online meetings or discussions.

## Chapter Dependencies

# Chapter Dependencies

🔔 Finish colored groups before moving to the next group.

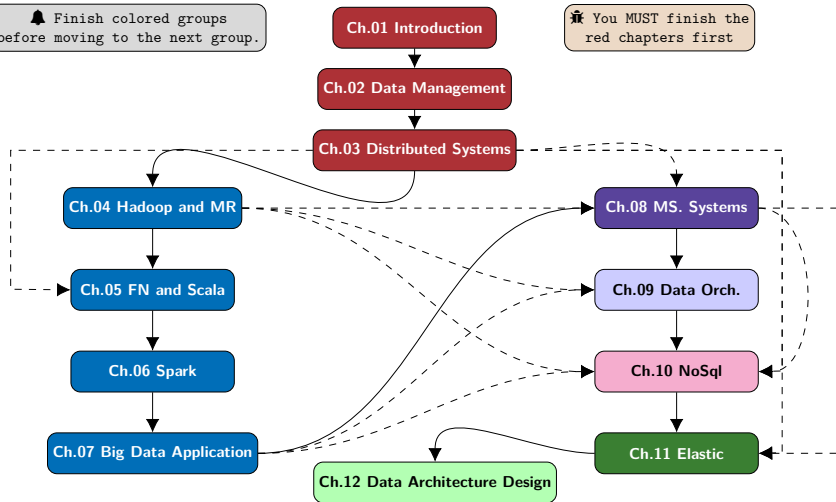
🔔 You MUST finish the red chapters first



# Chapter Dependencies (Jump Out Path)

🔔 Finish colored groups before moving to the next group.

🔔 You MUST finish the red chapters first



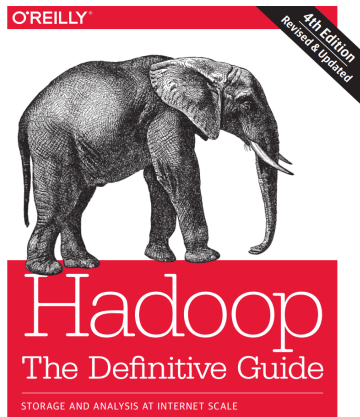
## Assignments, Labs, and Text Books

# Assignments and Labs

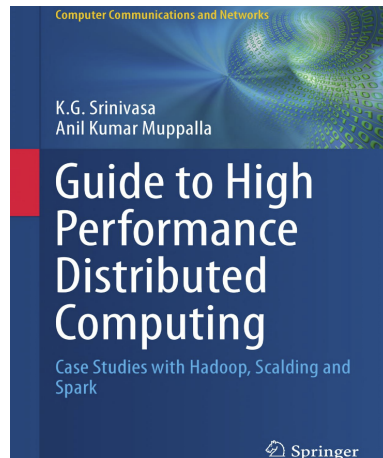
## Remark

- Full project code.
- Notebooks (Jupyter or Zeppelin).
- Read the references.



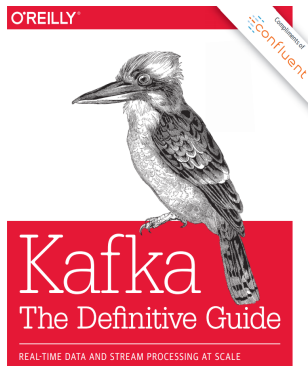


Tom White

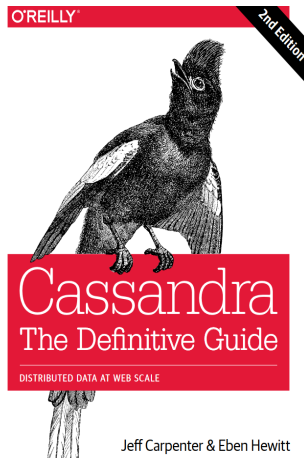


# Textbooks-2

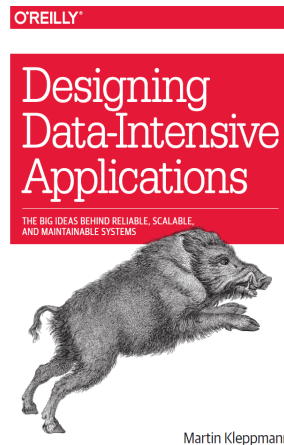




Neha Narkhede,  
Gwen Shapira & Todd Palino



Jeff Carpenter & Eben Hewitt



Martin Kleppmann

## Ugly but important

- User stories or technical discussions are not related to any of my current work or my previous companies.
- I am working at EPAM Systems. My company approved me for doing this online course public but the materials are not reviewed or assessed by my company. It is on my responsibilities.

# Table of Contents I

## 1 Course Introduction

- Learning Objectives and Audience
- Getting max benefit from this course
- Chapter Dependencies
- Assignments, Labs, and Text Books

## 1 Introduction To Data Management and Data Warehouse

- Data Management
- Data Abstraction
- Introduction to DWH
  - Motivation to Data Warehouse (DWH)
  - Differences Between DWH and Operational DB
  - Types of DWH
  - Use Cases of Operational DB vs DWH
- DWH Characteristics
- Hot vs Cold Storage

# Table of Contents II

- DWH Architecture
  - Source System Integration Process
  - Extraction Layer
  - Staging Layer
  - Data Modeling
  - ETL Process
  - Storage layer
  - Logical layer
  - Reporting (UI) layer
  - Metadata layer
  - System operations layer
- File Formats
- Data Encoding and Formats
- Data Compression Technique
- Data Archiving and Retention
- DWH On Cloud

# Table of Contents III

- Further Readings and Assignment

## 3 Introduction To Distributed Systems

- Distributed Systems Concepts
- Distributed Systems Architecture
- Distributed Systems Challenges
- Design Simple Distributed System
- Further Readings and Assignment

## 4 Hadoop and Map-Reduce

- Hadoop Architecture
  - Storage
  - YARN
  - Hadoop I/O
  - Processing
- Map-Reduce
  - Map-Reduce Components

# Table of Contents IV

- Word-Count Example
- Pig
- Hive
- ZooKeeper
- Further Readings and Assignment

## 5 Introduction to Functional Programming

- Why functional programming commonly used in distributed systems?
- Introduction to Scala
- Further Readings and Assignment

## 6 Spark Framework

- Spark Philosophy towards the Engine and the Programming languages
- Spark Basics
- Spark Programming using RDDs
  - Spark RDD



# Table of Contents V

- Spark Working With Key/Value Pairs
- Spark Datasets/Dataframe
  - Spark SQL
  - Dataframes/Datasets vs. RDDs
- Spark on Production
- Spark For Batch Processing
- Building custom input and output connector using Spark
- Spark Streaming
- Spark using other Programming Languages
  - PySpsark for Python Geeks
  - RSpark for R Geeks
- Spark For Data Scientist
- Spark Graph Dataframe/Graphx
- Tuning your Spark Jobs
- Further Readings and Assignment

# Table of Contents VI

## 7 Real World Applications

- Big Data Development Life Cycle
- Template Concept for Data Engineering
  - Template for ETL Application
  - Template for QA
  - Template for Streaming Applications
  - Template for Machine Learning Applications
- Further Readings and Assignment

## 8 Messaging Systems

- Motivation
- Messaging Systems Architecture
- JMS as an example
- Introduction to Kafka
  - Kafka Architecture
  - Kafka Topics

# Table of Contents VII

- Partitions
- Kafka Producers
- Kafka Consumers
- Kafka Connector
- Kafka Custom Connectors
- Kafka Configuration
- Kafka Configuration Optimizations
- Kafka Operations
- Kafka Integration with Enterprise tools
- Further Readings and Assignment

## 9 Data Orchestration

- Motivation
- Enterprise vs Open source tools
  - Open source tools (Oozie as an Example)
  - Enterprise source tools
  - How to choose the right tool?

# Table of Contents VIII

- Further Readings and Assignment

## 10 NOSQL

- Introduction to NoSQL Databases.
- Cassandra
  - Why Cassandra?
  - Introducing Cassandra
  - The Cassandra Data Model
  - Architecture
  - Reading and Writing Data
  - Integrating Hadoop
- Further Readings and Assignment

## 11 Elastic

- Further Readings and Assignment

## 12 Data Architecture Design

- Further Readings and Assignment

# Table of Contents IX

## 13 Appendix

- Appendix A- Shell Programming
- Appendix B- Java Programming
- Appendix C- Scala Programming
- Appendix D- SQL Programming
- Appendix E- Oozie Orchestration
- Appendix F- DWH Concepts and Data Modeling Design
- Appendix G- Machine Learning Concepts Data Engineers
- Appendix H- Docker for Data Engineers