Big Data Engineering In details

From Beginner to Professional

Mostafa Alaa Mohamed
Senior Big Data Engineer
MoustafaAlaa Moustafa Alaa @Moustafa_alaa22
✉ mustafa.alaa.mohamed@gmail.com

[1]Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

Table of Contents I

1. Course Introduction
   - Learning Objectives
   - Getting max benefit from this course
   - Assignments and Labs
   - Course Textbook
2. Introduction To Hadoop and Map-Reduce
   - From DWH to Big Data
   - Distributed Systems Concepts
   - Hadoop Architecture
     - Storage
     - YARN
     - Hadoop I/O
     - Processing
   - Map-Reduce
     - Map-Reduce Components
     - Word-Count Example
   - Hive

Table of Contents II

3 Spark Framework

- Spark Basics
- Spark Programming using RDDs
  - Spark RDD
  - Spark Working With Key/Value Pairs
- Spark Datasets/Dataframe
  - Spark SQL
  - Dataframes/Datasets vs. RDDs
- Spark on Production
- Spark For Batch Processing
- 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

Table of Contents III

4 Real World Applications

- Big Data Development Life Cycle
- Template for ETL Application
- Template for QA
- Template for Streaming Applications
- Template for Machine Learning Applications

5 Appendix

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

Course Introduction

Learning Objectives

- Understand the data management life-cycle.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL processing for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL processing for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL processing for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL processing for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.
- Building real-life examples.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL processing for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.
- Building real-life examples.
- Applying machine learning over Big Data.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL processing for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.
- Building real-life examples.
- Applying machine learning over Big Data.
- Understanding of the DevOps tools and function in data life-cycle.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.

Getting max benefit from this course

Take the course advantage

- Follow the videos order 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.

Getting max benefit from this course

Take the course advantage

- Follow the videos order 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.

Getting max benefit from this course

Take the course advantage

- Follow the videos order 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 the online meeting or discussions.

Getting max benefit from this course

Take the course advantage

- Follow the videos order 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 the online meeting or discussions.

Assignments and Labs

Remark

- Full project code.

Assignments and Labs

Remark

- Full project code.
- Notebooks (Jupyter or Zeppelin).

Assignments and Labs

Remark

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

Textbooks

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.

Textbooks

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau

Textbooks

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau
- High Performance Spark Best Practices for Scaling and Optimizing Apache Spark By Holden Karau, Rachel Warren.

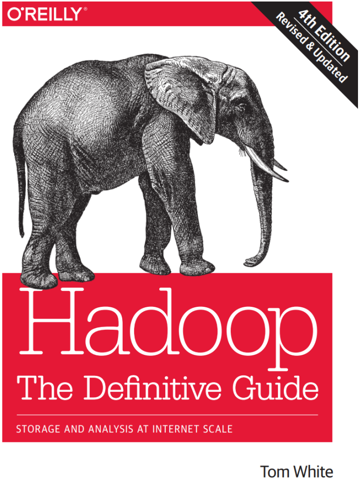
Textbooks

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau
- High Performance Spark Best Practices for Scaling and Optimizing Apache Spark By Holden Karau, Rachel Warren.
- Kafka: The Definitive Guide by Todd Palino, Gwen Shapira, Neha Narkhede.

Textbooks

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau
- High Performance Spark Best Practices for Scaling and Optimizing Apache Spark By Holden Karau, Rachel Warren.
- Kafka: The Definitive Guide by Todd Palino, Gwen Shapira, Neha Narkhede.
- Guide to High Performance Distributed Computing: Case Studies with Hadoop, Scalding and Spark (Computer Communications and Networks) 2015th Edition

Textbooks Cont.



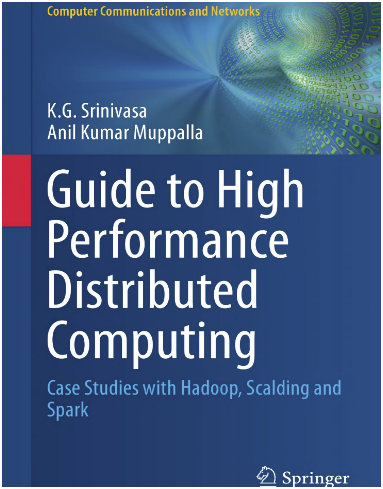
O'REILLY

4th Edition
Revised & Updated

Hadoop
The Definitive Guide

STORAGE AND ANALYSIS AT INTERNET SCALE

Tom White



Computer Communications and Networks

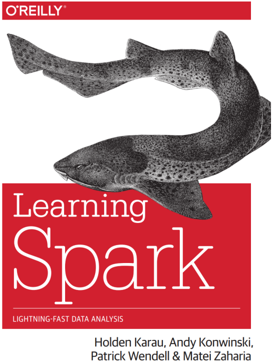
K.G. Srinivasa
Anil Kumar Muppalla

Guide to High Performance Distributed Computing

Case Studies with Hadoop, Scalding and Spark

Springer

Textbooks Cont.



O'REILLY

Learning Spark

LIGHTNING-FAST DATA ANALYSIS

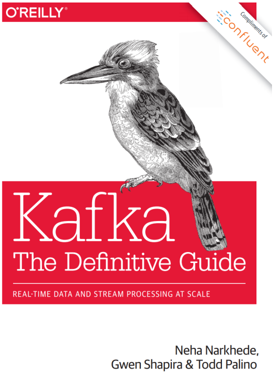
Holden Karau, Andy Konwinski,
Patrick Wendell & Matei Zaharia

O'REILLY

High Performance Spark

BEST PRACTICES FOR SCALING
& OPTIMIZING APACHE SPARK

Holden Karau &
Rachel Warren



O'REILLY

Compliments of confluent

Kafka The Definitive Guide

REAL-TIME DATA AND STREAM PROCESSING AT SCALE

Neha Narkhede,
Gwen Shapira & Todd Palino

Introduction To Hadoop and Map-Reduce

Spark Framework

Spark Framework: Spark Basics

- Any Big Data solution working based distributed systems.

Spark Framework: Spark Basics

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Basics

- Any Big Data solution working based distributed systems.

Spark Basics

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark on Production

- Any Big Data solution working based distributed systems.

Spark on Production

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark on Production

- Any Big Data solution working based distributed systems.

Spark on Production

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Batch Processing

- Any Big Data solution working based distributed systems.

Spark For Batch Processing

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Data Scientist

- Any Big Data solution working based distributed systems.

Spark For Data Scientist

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Data Scientist

- Any Big Data solution working based distributed systems.

Spark For Data Scientist

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Data Scientist

- Any Big Data solution working based distributed systems.

Spark For Data Scientist

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Real World Applications

Appendix

Appendix A- Shell Programming

- Any Big Data solution working based distributed systems.

Appendix A- Shell Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix B- Java Programming

- Any Big Data solution working based distributed systems.

Appendix B- Java Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix C- Scala Programming

- Any Big Data solution working based distributed systems.

Appendix C- Scala Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix D- SQL Programming

- Any Big Data solution working based distributed systems.

Appendix D- SQL Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix E- Oozie Programming

- Any Big Data solution working based distributed systems.

Appendix E- Oozie Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix F- DWH Concepts

- Any Big Data solution working based distributed systems.

Appendix F- DWH Concepts

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix G- Machine Learning Concepts Data Engineers

- Any Big Data solution working based distributed systems.

Appendix G- Machine Learning Concepts Data Engineers

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix H- Docker for Data Engineers

- Any Big Data solution working based distributed systems.

Appendix H- Docker for Data Engineers

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?