

(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

¹Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

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

1 Introduction To Distributed Systems

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

1 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

1 Introduction to Functional Programming

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

1 Spark Framework

- Spark Philosophy towards the Engine and the Programming languages
- Spark Basics



Table of Contents V

- 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
- 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



Table of Contents VI

- Tuning your Spark Jobs
- Further Readings and Assignment

1 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



Table of Contents VII

- Kafka Architecture
- Kafka Topics
- 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)



Table of Contents VIII

- Enterprise source tools
- How to choose the right tool?
- 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



Table of Contents IX

12 Data Architecture Design

- Further Readings and Assignment

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



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