

(Big) Data Engineering In Depth

From Beginner to Professional

Mostafa Alaa Mohamed

Senior Big Data Engineer

 MoustafaAlaa  Moustafa Alaa  @Moustafa_alaa22

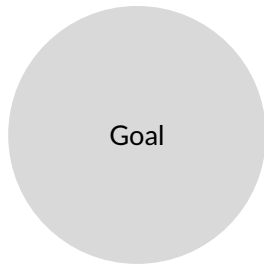
 mustafa.alaa.mohamed@gmail.com

¹Big Data & Analytics Department, Epam Systems

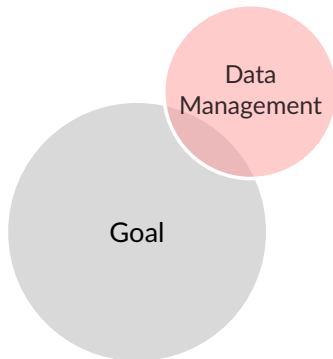
The Definitive Guide to Big Data Engineering Tasks

Course Introduction

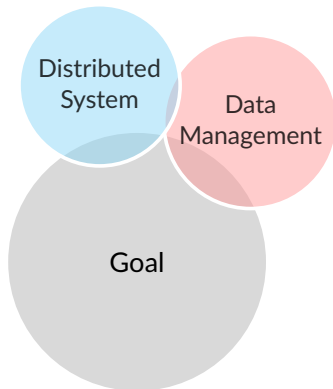
Course Target



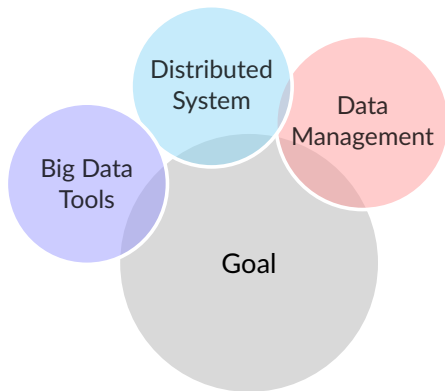
Course Target



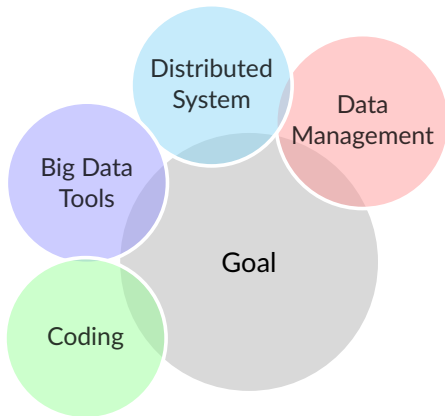
Course Target



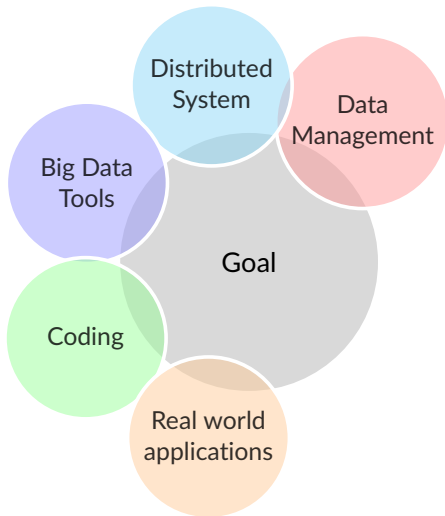
Course Target



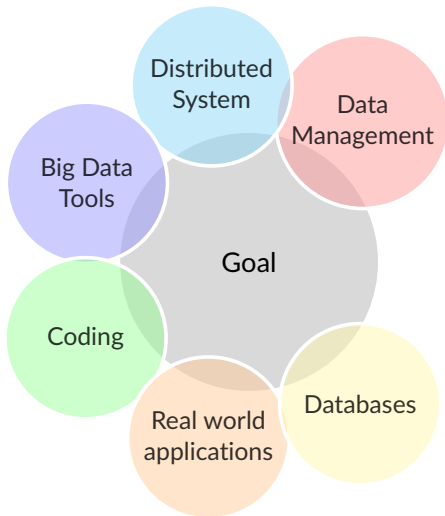
Course Target



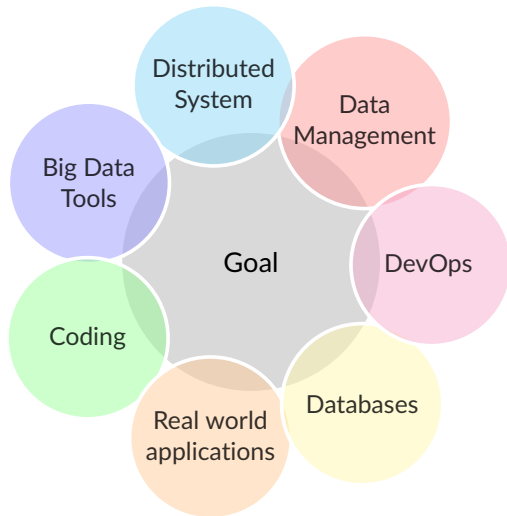
Course Target



Course Target



Course Target



Learning Objectives and Audience

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).
- Be familiar with ETL for (batch/streaming) data over distributed systems ex: Hadoop & Spark ([Ch.2,3,4,6](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).
- Be familiar with ETL for (batch/streaming) data over distributed systems ex: Hadoop & Spark ([Ch.2,3,4,6](#)).
- Apply QA and testing the data ([Ch.6,7](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).
- Be familiar with ETL for (batch/streaming) data over distributed systems ex: Hadoop & Spark ([Ch.2,3,4,6](#)).
- Apply QA and testing the data ([Ch.6,7](#)).
- Building real-life examples ([Ch.7](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).
- Be familiar with ETL for (batch/streaming) data over distributed systems ex: Hadoop & Spark ([Ch.2,3,4,6](#)).
- Apply QA and testing the data ([Ch.6,7](#)).
- Building real-life examples ([Ch.7](#)).
- Build and scale your data product ([Ch.7,8,9](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).
- Be familiar with ETL for (batch/streaming) data over distributed systems ex: Hadoop & Spark ([Ch.2,3,4,6](#)).
- Apply QA and testing the data ([Ch.6,7](#)).
- Building real-life examples ([Ch.7](#)).
- Build and scale your data product ([Ch.7,8,9](#)).
- Applying machine learning over big data ([Ch.6,7](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).
- Be familiar with ETL for (batch/streaming) data over distributed systems ex: Hadoop & Spark ([Ch.2,3,4,6](#)).
- Apply QA and testing the data ([Ch.6,7](#)).
- Building real-life examples ([Ch.7](#)).
- Build and scale your data product ([Ch.7,8,9](#)).
- Applying machine learning over big data ([Ch.6,7](#)).
- Automate the data life-cycle process end-to-end (e2e) ([Appx. H](#)).

Learning Objectives

- Simplify the concepts in data management ([Ch.2](#)).
- Understand the data management life-cycle ([Ch.2](#)).
- Illustrate the basics of distributed systems concepts ([Ch.3](#)).
- Be familiar with ETL for (batch/streaming) data over distributed systems ex: Hadoop & Spark ([Ch.2,3,4,6](#)).
- Apply QA and testing the data ([Ch.6,7](#)).
- Building real-life examples ([Ch.7](#)).
- Build and scale your data product ([Ch.7,8,9](#)).
- Applying machine learning over big data ([Ch.6,7](#)).
- Automate the data life-cycle process end-to-end (e2e) ([Appx. H](#)).
- Understanding of the DevOps tools and functions in data life-cycle ([Appx. H](#)).

Audience: Who Should Take This Course?

- Data Engineer who needs to get more knowledge in distributed systems and Big Data.

Audience: Who Should Take This Course?

- Data Engineer who needs to get more knowledge in distributed systems and Big Data.
- Data Warehouse Engineer who needs to know more about big data.

Audience: Who Should Take This Course?

- Data Engineer who needs to get more knowledge in distributed systems and Big Data.
- Data Warehouse Engineer who needs to know more about big data.
- A software developer who needs to change to the data engineering track.

Audience: Who Should Take This Course?

- Data Engineer who needs to get more knowledge in distributed systems and Big Data.
- Data Warehouse Engineer who needs to know more about big data.
- A software developer who needs to change to the data engineering track.
- DevOps engineer who needs to understand the concepts of big data.

Audience: Who Should Take This Course?

- Data Engineer who needs to get more knowledge in distributed systems and Big Data.
- Data Warehouse Engineer who needs to know more about big data.
- A software developer who needs to change to the data engineering track.
- DevOps engineer who needs to understand the concepts of big data.
- Business or entrepreneur who needs to get more information about how to build or manage a data product.

Getting max benefit from this course

Getting max benefit from this course

Take the course advantage

- Follow the order of the videos as described.

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).

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.

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.

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.

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

Chapters Dependencies

🔔 You MUST finish the red chapters first

🔔 Finish colored groups before moving to the next group.

Ch.01 Introduction

Ch.02 Data Management

Ch.03 Distributed Systems

Ch.04 Hadoop and MR

Ch.05 FN and Scala

Ch.06 Spark

Ch.07 Big Data Application

Ch.08 Massging Systems

Ch.09 Data Orchestration

Ch.10 NoSql

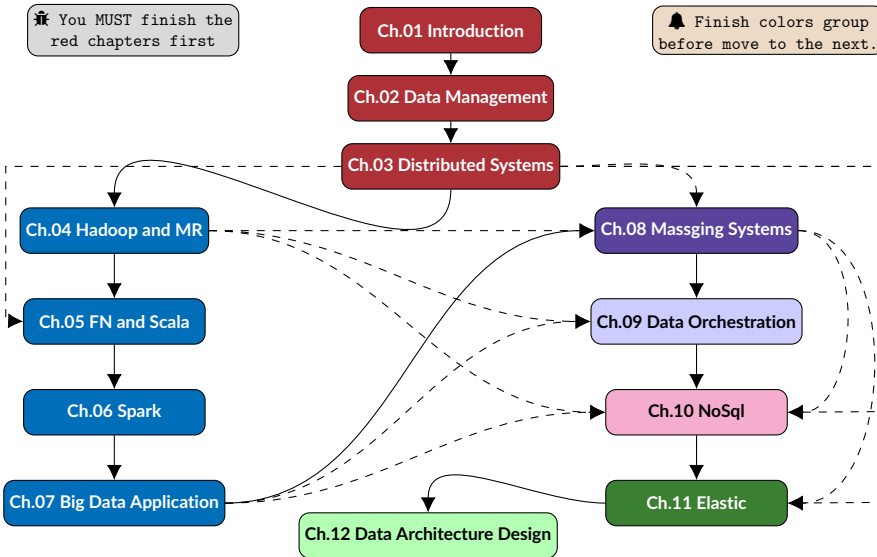
Ch.11 Elastic

Ch.12 Data Architecture Design

Chapter Dependencies (Jump Out Path)

🔔 You MUST finish the red chapters first

🔔 Finish colors group before move to the next.



Assignments, Labs, and Text Books

Remark

- Full project code.

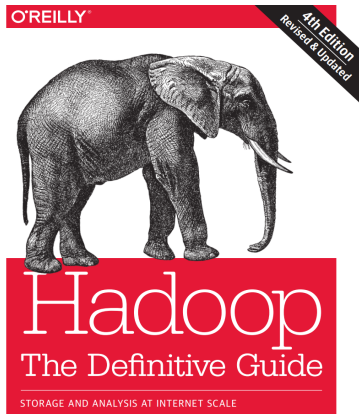
Remark

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

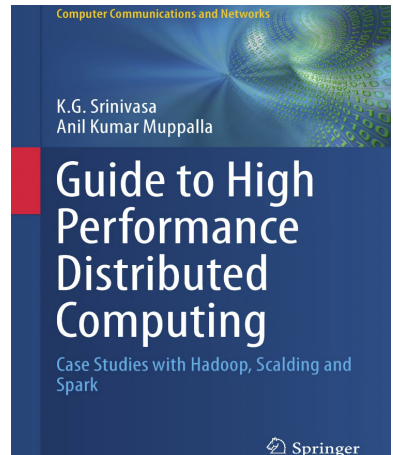
Assignments and Labs

Remark

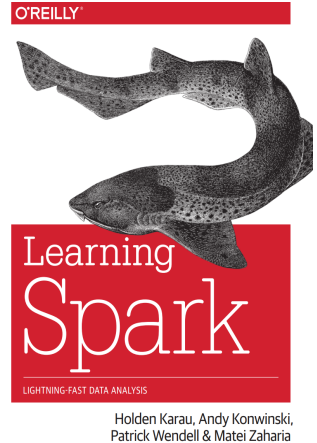
- Full project code.
- Notebooks (Jupyter or Zeppelin).
- Read the references.

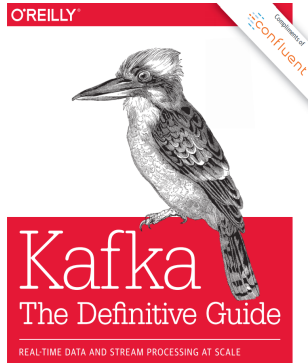


Tom White

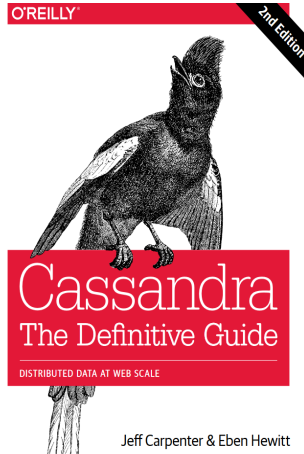


Textbooks-2

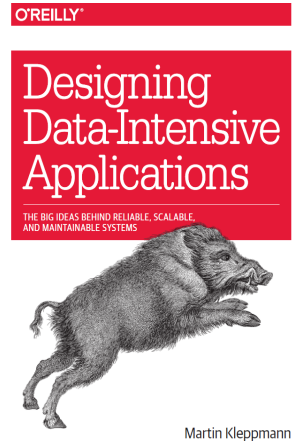




Neha Narkhede,
Gwen Shapira & Todd Palino



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.

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.