(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

[1]Big Data & Analytics Department, Epam Systems

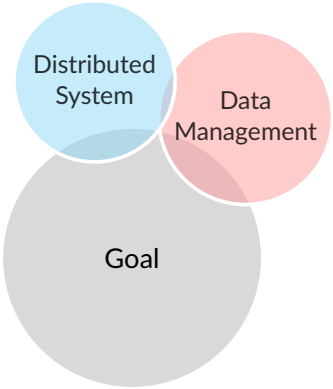
The Definitive Guide to Big Data Engineering Tasks

Course Introduction

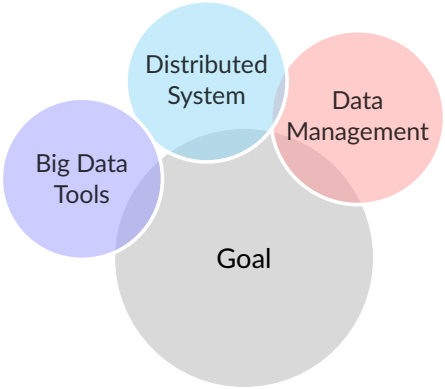
Course Target

Course Target

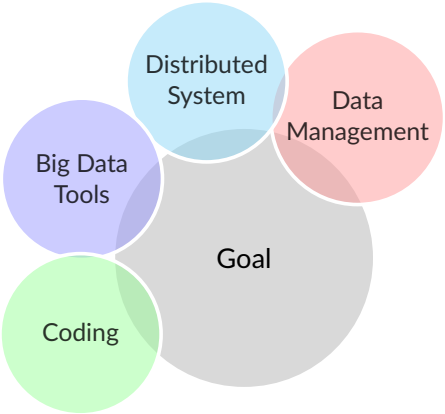
Course Target



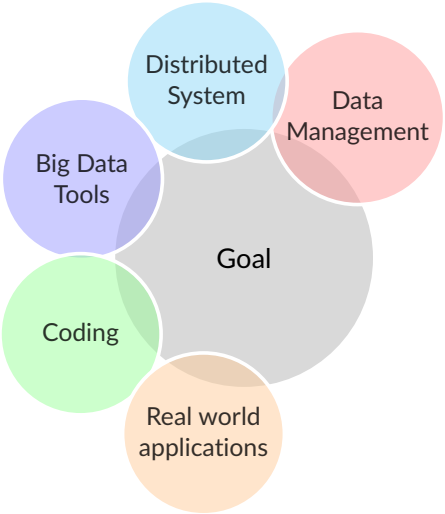
Course Target



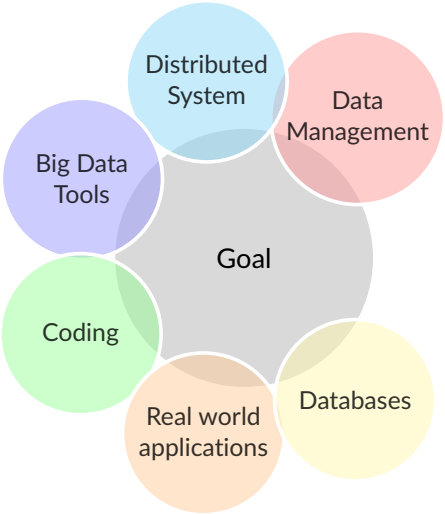
Course Target



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/steaming) 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/steaming) 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/steaming) 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/steaming) 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/steaming) 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/steaming) 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/steaming) 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.

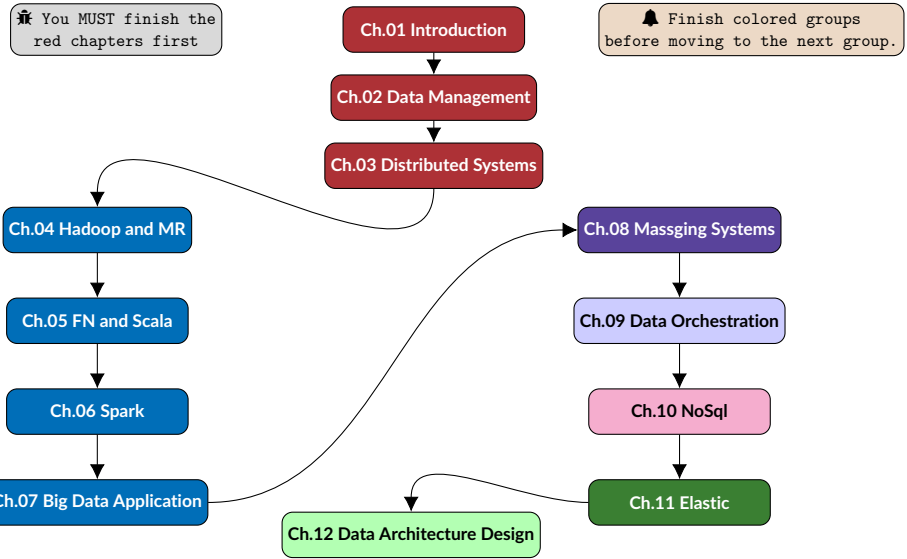
Getting max benefit from this course

Take the course advantage

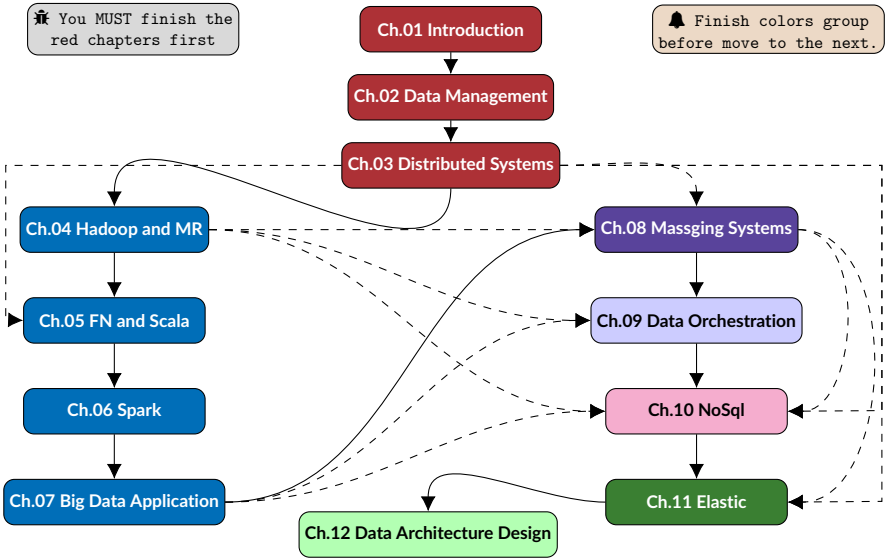
- Follow the order of the videos as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.
- Do the assignments.
- Ask your questions.
- Join online meetings or discussions.

Chapter Dependencies

Chapters Dependencies



Chapter Dependencies (Jump Out Path)



Assignments, Labs, and Text Books

Assignments and Labs

Remark

- Full project code.

Assignments and Labs

Remark

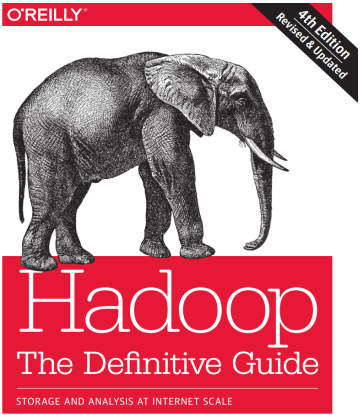
- Full project code.
- Notebooks (Jupyter or Zeppelin).

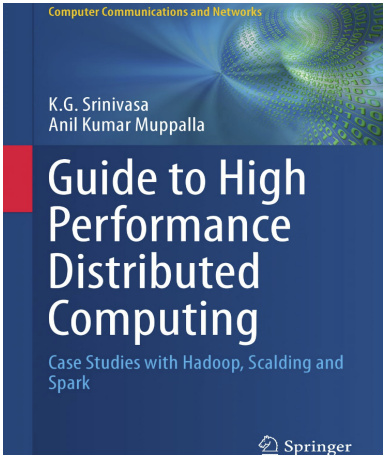
Assignments and Labs

Remark

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

Textbooks-1


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

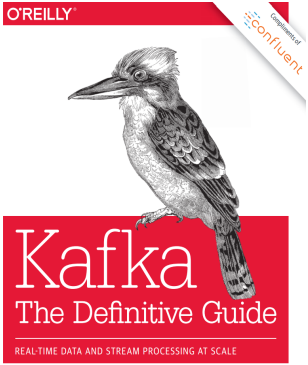
Textbooks-2


Functional Programming
scala
IN
Paul Chiusano
Rúnar Bjarnason
Foreword by Martin Odersky
MANNING

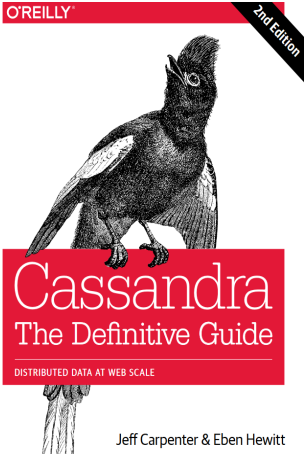

O'REILLY
High Performance
Spark
BEST PRACTICES FOR SCALING
& OPTIMIZING APACHE SPARK
Holden Karau &
Rachel Warren


O'REILLY
Learning
Spark
LIGHTNING-FAST DATA ANALYSIS
Holden Karau, Andy Konwinski,
Patrick Wendell & Matei Zaharia

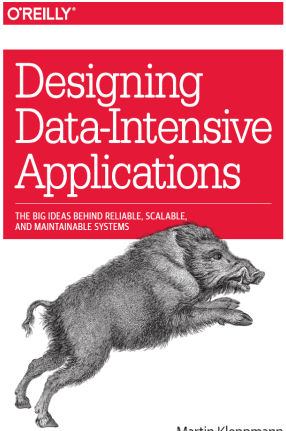
Textbooks-3



Neha Narkhede,
Gwen Shapira & Todd Palino



Jeff Carpenter & Eben Hewitt



Martin Kleppmann

Videos classification

- First 5 sec for every video contains a classification for this video as the following:

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.
    - Developer.

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.
    - Developer.
    - DevOps.

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.
    - Developer.
    - DevOps.
    - Business.

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.
    - Developer.
    - DevOps.
    - Business.
  - Watching Method:

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.
    - Developer.
    - DevOps.
    - Business.
  - Watching Method:
    - On Computer (Focus and rewrite coding).

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.
    - Developer.
    - DevOps.
    - Business.
  - Watching Method:
    - On Computer (Focus and rewrite coding).
    - On Mobile/Tablet (charts and points you need to watch).

Videos classification

- First 5 sec for every video contains a classification for this video as the following:
  - Length:
    - Short (2:5 min).
    - Medium (6:12 min).
    - Long (12:20 min).
  - Audience: Development, DevOps, Business.
    - Developer.
    - DevOps.
    - Business.
  - Watching Method:
    - On Computer (Focus and rewrite coding).
    - On Mobile/Tablet (charts and points you need to watch).
    - Just listening (You can listen anywhere walking, driving, etc).

Videos classification

Watching Method / Audience	Computer	Mobile/Tablet	Just listening
Developer	●		
DevOps			●
Business		●	

Figure: Video classification
The green circle ● means short video.
The blue circle ● means medium video.
The red circle ● means long video

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.