

Big Data Engineering In details

From Beginner to Professional

Mostafa Alaa Mohamed

Senior Big Data Engineer

Email: mustafa.alaa.mohamed@gmail.com

Linkedin: Mostafa Alaa¹

¹Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

Table of Contents I

1 Spark Basics

- Introduction to Spark
- Map-Reduce using Spark

2 Spark Programming using RDDs

- Spark RDD
- Spark Working With Key/Value Pairs
- Why we need RDD in Real World Applications

3 Spark Datasets/Dataframe

- Introduction to Datasets/Dataframe
- Spark SQL
- Dataframes/Datasets vs. RDDs

4 Spark on Production

- Go Production Part 1
- Go Production Part 2

5 Spark For Data Warehouse

- ETL pipeline End-to-End into Spark

Table of Contents II

6 Spark Streaming

- Introduction to Spark Streaming
- Spark Streaming with Kafka
- Spark Structure Streaming
- Spark Streaming Applications

7 Spark using other Programming Languages

- PySpsark for Python Geeks
- PySpsark Utilizing Python Features
- RSpark for R Geeks

8 Spark For Data Scientist

- Data Analysis using Spark
- Machine Learning using Spark
- Deep Learning using Spark

9 Spark Graph Dataframe/Graphx

- Introduction to Spark Graph Dataframe/GraphX

Table of Contents III

- Graph Applications using Spark

10 Tuning your Spark Jobs

- Tuning Spark Jobs Part 1
- Tuning Spark Jobs Part 2

11 Appendix

Spark Basics

Spark Basics: Introduction to Spark

- Any Big Data solution working based distributed systems.

Spark Basics: Introduction to Spark

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

Map-Reduce using Spark

- Any Big Data solution working based distributed systems.

Map-Reduce using Spark

- 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.

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

Spark Working With Key/Value Pairs

- Any Big Data solution working based distributed systems.

Spark Working With Key/Value Pairs

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

Why we need RDD in Real World Applications

- Any Big Data solution working based distributed systems.

Why we need RDD in Real World Applications

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

Spark Datasets/Dataframe

Introduction to Datasets/Dataframe

- Any Big Data solution working based distributed systems.

Introduction to Datasets/Dataframe

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

- Any Big Data solution working based distributed systems.

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

Dataframes/Datasets vs. RDDs

- Any Big Data solution working based distributed systems.

Dataframes/Datasets vs. RDDs

- 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.

Go Production Part 1

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

- Any Big Data solution working based distributed systems.

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

Spark For Data Warehouse

ETL pipeline End-to-End into Spark

- Any Big Data solution working based distributed systems.

ETL pipeline End-to-End into Spark

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

Spark Streaming

Introduction to Spark Streaming

- Any Big Data solution working based distributed systems.

Introduction to Spark Streaming

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

- Any Big Data solution working based distributed systems.

Spark Streaming with Kafka

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

- Any Big Data solution working based distributed systems.

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

Spark Streaming Applications

- Any Big Data solution working based distributed systems.

Spark Streaming Applications

- 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.

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

- Any Big Data solution working based distributed systems.

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

- Any Big Data solution working based distributed systems.

- 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.

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

Machine Learning using Spark

- Any Big Data solution working based distributed systems.

Machine Learning using Spark

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

- Any Big Data solution working based distributed systems.

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

Spark Graph Dataframe/Graphx

Introduction to Spark Graph Dataframe/GraphX

- Any Big Data solution working based distributed systems.

Introduction to Spark Graph Dataframe/GraphX

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

Graph Applications using Spark

- Any Big Data solution working based distributed systems.

Graph Applications using Spark

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

Tuning your Spark Jobs

Tuning Spark Jobs Part 1

- Any Big Data solution working based distributed systems.

Tuning Spark Jobs Part 1

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

Tuning Spark Jobs Part 2

- Any Big Data solution working based distributed systems.

Tuning Spark Jobs Part 2

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

Appendix

Appendix A: Hadoop Map-Reduce

- Any Big Data solution working based distributed systems.

Appendix A: Hadoop Map-Reduce

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

Appendix B: Introduction to Scala

- Any Big Data solution working based distributed systems.

Appendix B: Introduction to Scala

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

- Any Big Data solution working based distributed systems.

Appendix C: SQL Syntax and Tips for Data Science

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

Appendix D: DWH Concepts and Modeling

- Any Big Data solution working based distributed systems.

Appendix D: DWH Concepts and Modeling

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

Appendix E: Maven for Automation Cycle

- Any Big Data solution working based distributed systems.

Appendix E: Maven for Automation Cycle

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

Appendix F: How to use Shell Scripts to Automate your tasks

- Any Big Data solution working based distributed systems.

Appendix F: How to use Shell Scripts to Automate your tasks

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

Appendix G: Introduction to Oozie for Big Data Development

- Any Big Data solution working based distributed systems.

Appendix G: Introduction to Oozie for Big Data Development

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

Appendix H: Introduction to Deep Learning

- Any Big Data solution working based distributed systems.

Appendix H: Introduction to Deep Learning

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