

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

¹Big 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 Textbooks

2 Introduction To Distributed Systems (Hadoop as example)

- Data Management
- From DWH to Big Data
- Distributed Systems Concepts
- Hadoop Architecture
 - Storage
 - YARN
 - Hadoop I/O
 - Processing
- Map-Reduce
 - Map-Reduce Components
 - Word-Count Example

Table of Contents II

- Hive
- Assignment and Homework

3 Function Programming

- Why FN commonly used distributed systems?
- Introduction to Scala
- Assignment and Homework

4 Spark Framework

- Spark Philosophy towards the Engine and the Programming languages
- 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

Table of Contents III

- Spark For Batch Processing
- Building custom input and output connector using Spark
- Spark Streaming
- Spark using other Programming Languages
 - PySpark for Python Geeks
 - RSpark for R Geeks
- Spark For Data Scientist
- Spark Graph Dataframe/Graphx
- Tuning your Spark Jobs
- Assignment and Homework

5 Real World Applications

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

Table of Contents IV

- Assignment and Homework

6 Massaging Systems

- Motivation
- Massaging Systems Architecture
- JMS queue as an example
- Introduction to Kafka
 - 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
- Assignment and Homework

Table of Contents V

7 Elastic

- Assignment and Homework

8 NOSQL

- Introduction to NoSQL Databases.
- Cassandra
 - Why Cassandra?
 - Introducing Cassandra
 - The Cassandra Data Model
 - Architecture
 - Reading and Writing Data
 - Integrating Hadoop
- Assignment and Homework

9 Data Orchestration

- Motivation
- Enterprise vs Open source tools
 - Open source tools
 - Enterprise source tools

Table of Contents VI

- How to choose the right tool?
- Assignment and Homework

10

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

Course Introduction

Course Target



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 for (Batch/Streaming) 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 for (Batch/Streaming) 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 for (Batch/Streaming) 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 for (Batch/Streaming) 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 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 for (Batch/Streaming) 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 functions 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-1

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

Textbooks-1

- 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-1

- 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.

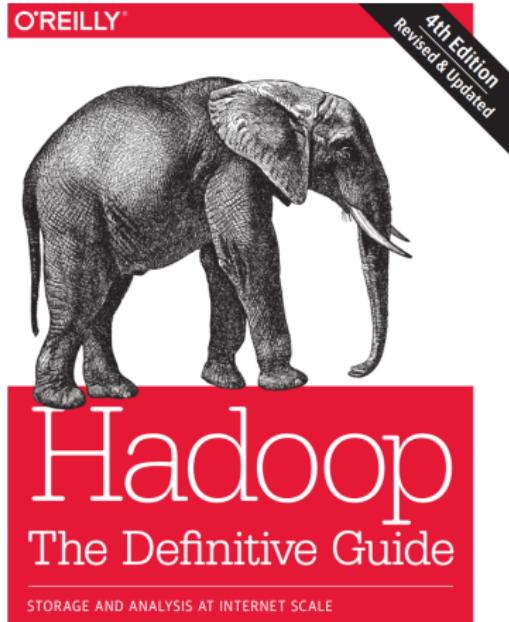
Textbooks-1

- 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.

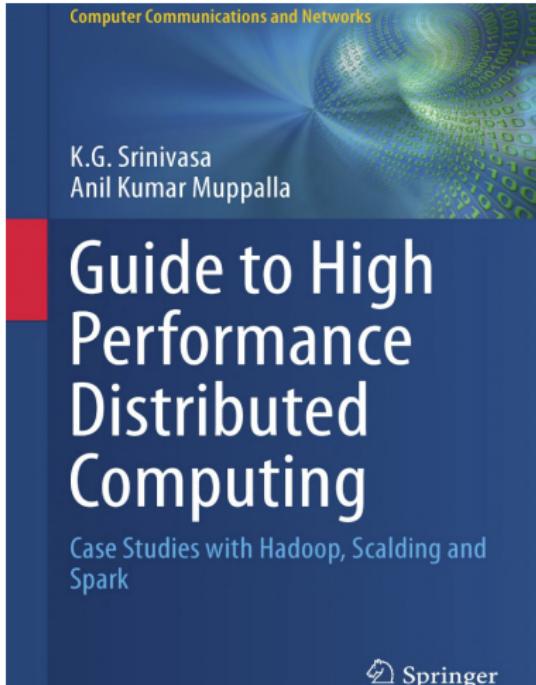
Textbooks-1

- 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-2



Tom White



Textbooks-3

O'REILLY®



Learning Spark

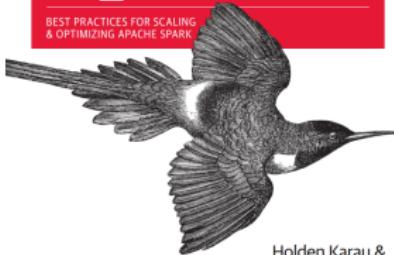
LIGHTNING FAST DATA ANALYSIS

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®



Kafka The Definitive Guide

REAL-TIME DATA AND STREAM PROCESSING AT SCALE

Neha Narkhede,
Gwen Shapira & Todd Palino

Companion to
confluent

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 own responsibilities.

Introduction To Distributed Systems (Hadoop as example)

Chapter Objectives

- What is data management?

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.
- Developing Map-Reduce applications.
- Using Hive QL over Map-Reduce.

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.
- Developing Map-Reduce applications.
- Using Hive QL over Map-Reduce.

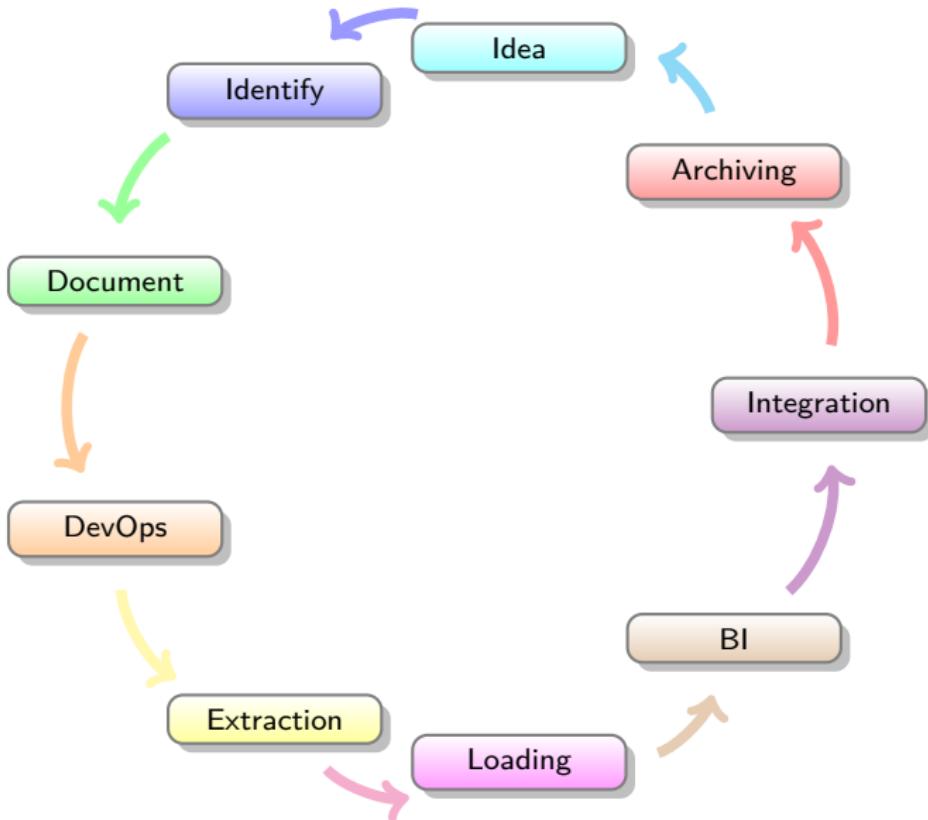
Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.
- Developing Map-Reduce applications.
- Using Hive QL over Map-Reduce.
- Hadoop advantages and disadvantages with use cases?

Data Management

- Data are a product.
- Data product has a life-cycle as following (simplified):
 - **Question**, Idea, or service.
 - **Identifying** the source of information and the data type ex: (text, images, videos, audio, or sensors).
 - **Document** all details regarding the data including quality, security, efficiency, and access (consideration during the cycle).
 - Delivery automation (Tools and Process) AKA **DevOps** cycle.
 - **Extraction** Process (collection).
 - **Transformation** ex: (cleansing, Apply business logic, Organize).
 - **Loading** or store the transformed data based on our usage or use case.
 - Business Intelligence (**BI**) or data discovery (continues process).
 - **Integration** and publishing.
 - Data retention or **archiving** process ex: (Hot or Cold storage).

Data Management Life-Cycle



Motivation to Data Warehouse

- Data could be a product for some companies.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
 - Performance.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
 - Performance.
 - Integration.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
 - Performance.
 - Integration.
 - Applying analytical functions.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
 - Performance.
 - Integration.
 - Applying analytical functions.
- Vendors who are working to solve the above challenges creating their own product of DWH and their ultimate work is to optimize the above points.

Motivation to Data Warehouse

Definition (What is Data Warehousing?)

A data warehousing is defined as a technique for collecting and managing data from varied sources to provide meaningful business insights. It is a blend of technologies and components which aids the strategic use of data.¹

- It is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.
- The data warehouse is not a product but an environment.
- It is an architectural construct of an information system which provides users with current and historical decision support information which is difficult to access or present in the traditional operational data store.
- The data warehouse is the core of the BI system which is built for data analysis and reporting.

¹The definition mentioned in this slides copied from guru99.com

Motivation to Data Warehouse

Data warehouse system is also known by the following name:

- Decision Support System (DSS)
- Executive Information System
- Management Information System
- Business Intelligence Solution
- Analytic Application
- Data Warehouse

The real concept was given by Inmon Bill. He was considered as a father of data warehouse. He had written about a variety of topics for building, usage, and maintenance of the warehouse & the Corporate Information Factory

Motivation to Data Warehouse

Types of Data Warehouse

- **Enterprise Data Warehouse:** It provides decision support service across the enterprise. It offers a unified approach for organizing and representing data. It also provide the ability to classify data according to the subject and give access according to those divisions.
- **Operational Data Store:** It also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.
- **Data Mart:** A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.
- Data warehouse designed to solve the **huge amount of data**.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.
- Data warehouse designed to solve the **huge amount of data**.
- Most of DWH can't solve the online transactions similar to the transaction DB.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.
- Data warehouse designed to solve the **huge amount of data**.
- Most of DWH can't solve the online transactions similar to the transaction DB.
- Transactions databases have a performance issue while handling a huge amount of data. So, analysis of a huge amount of data (including historical data) we used DWH for this purpose. On the other hand Transactions DB used for online or short historical data based on product type and requirements.

DWH vs Operational databases

Metric	Transactions DB	DWH
Volume	GB/TB	TB/PB
Historical rows	Short-term <1000M	Long-Term 1000M>
Orientation	Product	Subject or multi products
Business Units	Product team	Multi organizational units
Normalization	Normalized	Not required (De-normalized in many use cases)
Data Model	Relational	Star Schema or Multi-dim
Intelligence	Reporting	Advanced reporting and Machine Learning
Use cases	Online transactions & operations	Centralized storage (360°)

Transnational DB Use cases



Transnational DB Use cases



DWH Use cases



DWH Use cases



DWH Use cases



User stories Telecom company.

It has a CRM System backend database reporting the sales. vs Another backend database contains the CRM, Telecom signaling data, IN charging system, Billing

Decision is related to sales or CRM. Decision is related to company strategies.

Analytical model checking the fraud which require a CRM data with customer locations from signaling with Billing details from CAR table.

managing risk of the project in Transaction vs DWH

data model comparison

Cold storage vs Hot storage

some details about hot vs cold storage,

DWH Characteristics

some details about hot vs cold storage,

Distributed Systems Concepts

- Any Big Data solution working based distributed systems.

Distributed Systems Concepts

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

Hadoop Architecture

- Any Big Data solution working based distributed systems.

Hadoop Architecture

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

Storage

- Any Big Data solution working based distributed systems.

Storage

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

YARN

- Any Big Data solution working based distributed systems.

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

Hadoop I/O

- Any Big Data solution working based distributed systems.

Hadoop I/O

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

Processing

- Any Big Data solution working based distributed systems.

Processing

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

Map-Reduce

- Any Big Data solution working based distributed systems.

Map-Reduce

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

Map-Reduce Components

- Any Big Data solution working based distributed systems.

Map-Reduce Components

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

Word-Count Example

- Any Big Data solution working based distributed systems.

Word-Count Example

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

Hive

- Any Big Data solution working based distributed systems.

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

Function Programming

Spark Framework

Spark Framework: Spark Philosophy towards the Engine and the Programming languages

- Any Big Data solution working based distributed systems.

Spark Framework: Spark Philosophy towards the Engine and the Programming languages

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

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?

Building custom input and output connector using Spark

- Any Big Data solution working based distributed systems.

Building custom input and output connector using Spark

- 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

Massaging Systems

Elastic

NOSQL

Data Orchestration

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 Orchestration

- Any Big Data solution working based distributed systems.

Appendix E- Oozie Orchestration

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

Appendix F- DWH Concepts and Data Modeling Design

- Any Big Data solution working based distributed systems.

Appendix F- DWH Concepts and Data Modeling Design

- 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?