This book is for anyone who wants to learn quickly how to use pyspark to effectively load, process, and transform large volumes of data using Python.

In more detail, this is a quick and introductory book about pyspark, which is the Python API for Apache Spark. Apache Spark is the de facto standard engine for big-data analytics. It is largely used to build data processing, data ingestion, and machine learning applications that process very large volumes of data.

One of the many reasons why Apache Spark became popular is because of its APIs. You can build Spark applications using different programming languages, such as Python, R, and Scala. But this book focuses solely on the Python API.

In this book, you will learn about:

- How an Apache Spark application works?
- What are Spark DataFrames?
- How to build, transform and model your Spark DataFrame.
- How to import data into (or export data out of) Apache Spark.
- How to work with SQL inside pyspark.
- Tools for manipulating specific data types (e.g. strings, dates and datetimes).
- How to use window functions.

Introduction to pyspark 1st. **edition** **Pedro Duarte Faria**

Introduction to pyspark

First Edition

Pedro Duarte Faria