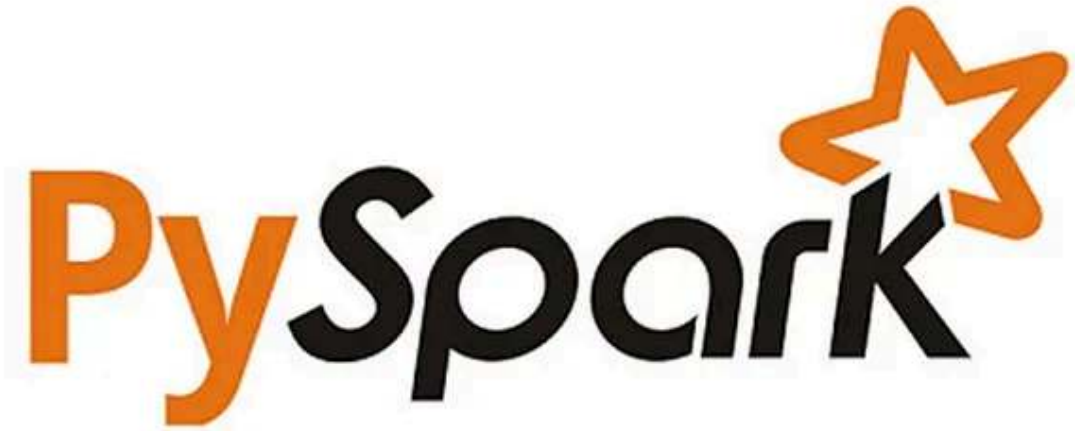
Complete Apache PySpark Learning Resources with Links — Data Engineering

Deepanshu tyagi · Follow

2 min read · Jan 23, 2024

171



Pyspark

Apache spark started as a research project at UC Berkeley AMP Lab in 2009, and it became open source in 2010. After this, Spark grew into a big developer community and moved to the Apache Software Foundation in 2013. Now, the project is used by various communities and various organizations.

> *At present, there is a growing trend for Pyspark in the data engineering.*

I've compiled a collection of educational materials for Apache Pyspark.

Let's start:

Introduction to Apache PySpark.

This blog will cover everything you need to know about the theory before learning Apache Pyspark.

Introduction to Apache PySpark RDD- A Practical Approach Part 2

This blog will cover everything you need to know about Apache Pyspark RDDs theoretically.

Apache PySpark RDD Transformation — A practical approach, Part 3

This blog will cover everything you need to know about Apache Pyspark RDDs practically.

Apache PySpark DataFrame–A practical approach, Part 4

PySpark Advance DataFrame — A practical approach, part 5

These blog will cover basics and advance transformation of DataFrame.

Introduction to Spark ML

PySpark Random Forest Regression Machine Learning — A practical approach, part 7

This blog will cover basics Spark ML.

Interview Spark Questions:

How to handle Bad Records in Apache Pyspark?

How we will do data validation?

How we will identify bad data?

How we will store bad data?

Solution: Advance interview question — Apache Pyspark [Handling bad data] -Part 1

How we can get all DataFrame available in that runtime in Pyspark?

How to get number of rows on each file in the DataFrame?

Advance interview question — Apache Pyspark -Part 2

How partitions are created?

How do you join two DataFrames in PySpark?

Why are partitions immutable?

Advance interview questions — Apache Pyspark -Part 3

When do you use Spark streaming?

What's different about SparkSQL from HQL and SQL?

Advance interview questions — Apache Pyspark -Part 4

Medium Lists

Deepanshu tyagi

PySpark Interview Questions

View list 5 stories

Todo Next:

Search Write

In the ever-changing world of data analytics, mastering cloud services has become critical for professionals seeking to...

brilliantprogrammer.medium.com

How to be a Data Engineer — 2023 Guide

Data engineering is a vital field that involves designing, building, and maintaining the systems and infrastructure...

blog.brilliantprogrammer.com

Google Cloud Roadmap for Data Engineering — Complete Guide — Part 1

Hello readers, I hope you are doing great in this blog. We are going to discuss data engineering tools available on...

blog.brilliantprogrammer.com

- Please reach out via *Linkedin* or Github in case of any questions!
- Subscribe to my free newsletter.
- Follow me on twitter.

> https://dataengineeringpy.substack.com/

Comment on whatever topic you want to blog about.

Data Engineering | Data Engineer | Technology | Data Science | Data Analytics