
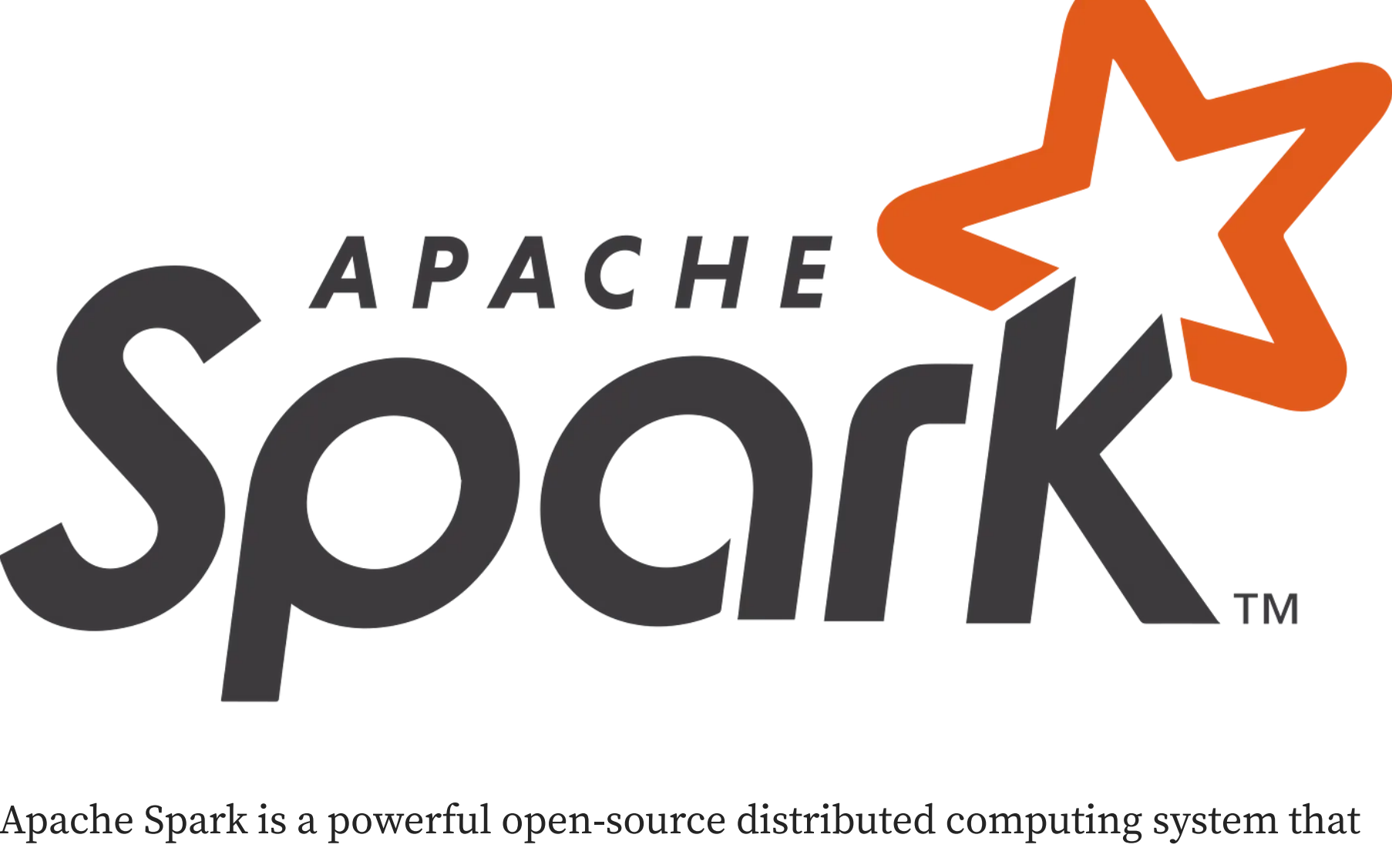
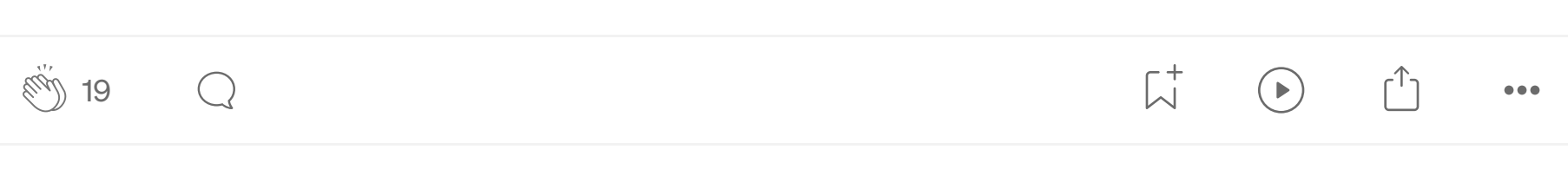


A Step-by-Step Guide to Installing Pyspark on Windows

Introduction

 **Deepak Rawat** · Follow

3 min read · Jan 5, 2024

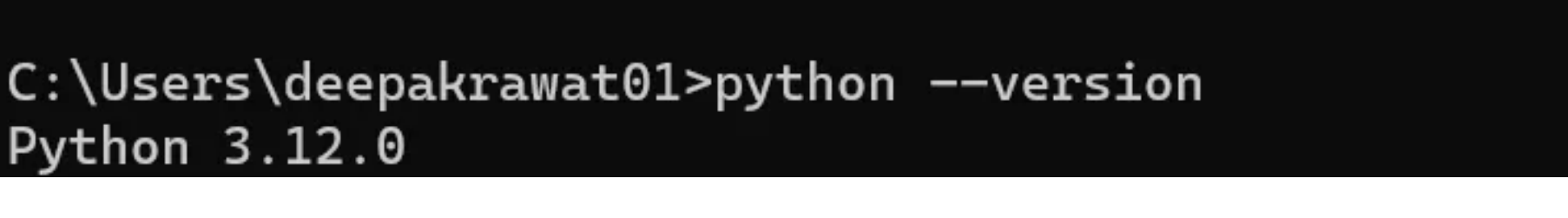


Apache Spark is a powerful open-source distributed computing system that has gained immense popularity for its speed and ease of use. Pyspark, the Python API for Apache Spark, allows developers to harness the capabilities of Spark using the Python programming language. While Spark is commonly associated with Linux environments, setting up Pyspark on Windows can be a bit challenging. In this guide, we'll walk through the process of installing Pyspark on a Windows machine.

Now, let's dive into the step-by-step process of installing Pyspark on Windows:

Step 1 : Install Python in local system

- a. Download the latest version of python from - <https://www.python.org/downloads/>
- b. Once the installer is downloaded, double-click on the executable file (e.g., `python-3.x.x.exe`) to start the installation, make sure to check the box that says “*Add Python to PATH*” during installation.
- c. Open a command prompt type `python --version` or `python -V` and press enter to check if Python is installed.



Step 2 : Install Java Development Kit (JDK)

- a. Download the latest JDK installer from the Oracle website.
- b. Run the installer and follow the on-screen instructions to complete the installation.
- c. Set the `JAVA_HOME` environment variable. Add the JDK installation path to the system environment variables.

Java path configuration in environmental variable

- d. Open a command prompt type `java --version` and press enter to check if Java is installed.

Step 3: Install Apache Spark

- a. Visit spark download(<https://spark.apache.org/downloads.html>)
- b. Set `SPARK_HOME` in environment variables where spark is extracted — to be edited

Step 4 : Setup Hadoop

- a. Create a hadoop folder and create a bin folder inside it
- b. Download winutils file from - <https://github.com/steveloughran/winutils> and put in bin folder
- c. Set `HADOOP_HOME` path in environment variables

Step 5 : Install Pyspark

- a. open Command prompt run command — `pip install pyspark`
- b. set `PYSPARK_HOME` path in environment variables

Step 6 :Verify installation

- a. Add all the *variables* in the path
- b. Open a *Python shell* or a *Jupyter notebook* and run the following code to verify the installation:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.master("local").appName("PySpark Installation Test")
df = spark.createDataFrame([(1, "Hello"), (2, "World")], ["id", "message"])
df.show()
```

Output :

- Key Considerations for PySpark Installation on Windows :**
- a. Ensure Compatibility Between Apache Spark and PySpark Versions
 - b. Restart Command Prompt After Environment Variable Setup
 - c. Rely on Official Documentation and Stay Updated on PySpark and Dependencies

Conclusion:
Congratulations! You have successfully installed PySpark on your Windows machine. Now, you can leverage the power of PySpark to perform distributed data processing and analysis using the simplicity of the Python programming language. Enjoy exploring the vast capabilities that PySpark has to offer on your Windows environment.

Spark

Pyspark

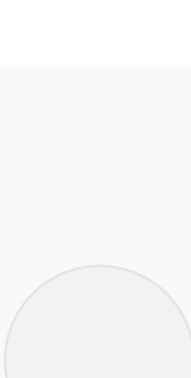
Windows

Big Data

Python

19 likes · 0 comments

Bookmark · Share · More

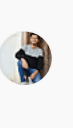


Written by Deepak Rawat
18 Followers

Follow

Share

More from Deepak Rawat

 Deepak Rawat

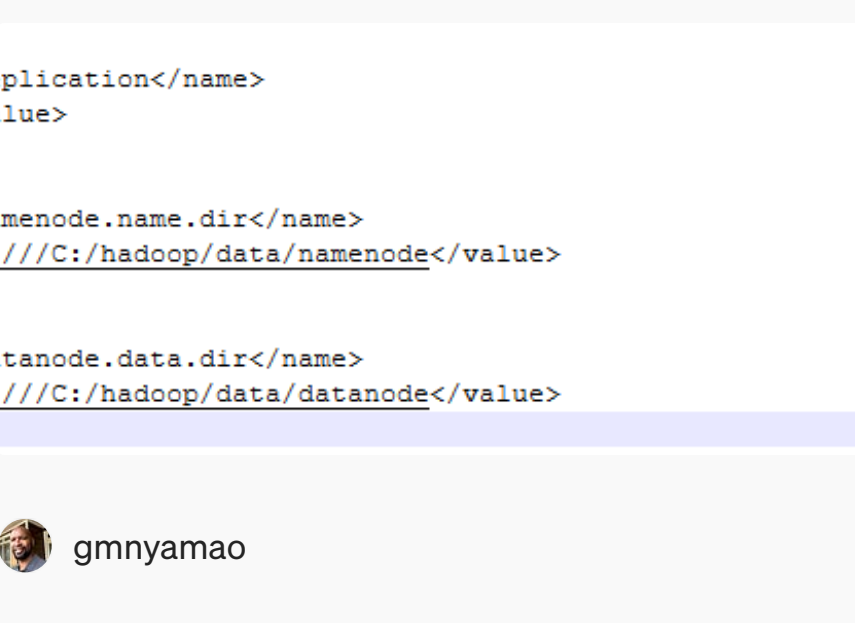
Databricks Data Engineer Associate Exam Made Easy—A Comprehensive Guide
Mastering the Databricks Certified Data Engineer Associate exam has been a pivotal achievement in my journey as a data...
Feb 7 · 1 like · 1 share

Bookmark


More

See all from Deepak Rawat

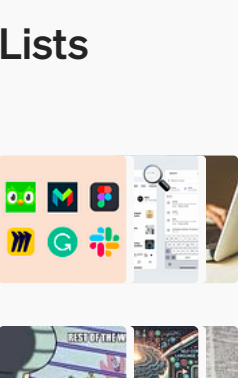
Recommended from Medium

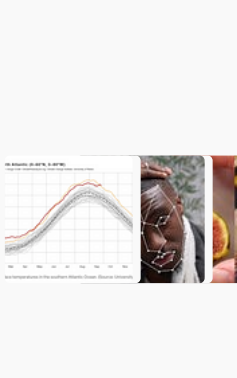


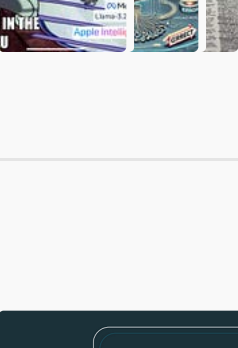
How to Install Hadoop on Windows 10/11: A Step-by-Step Guide
Hadoop is a powerful framework for processing and storing large datasets across...
Jul 6 · 1 like · 1 share

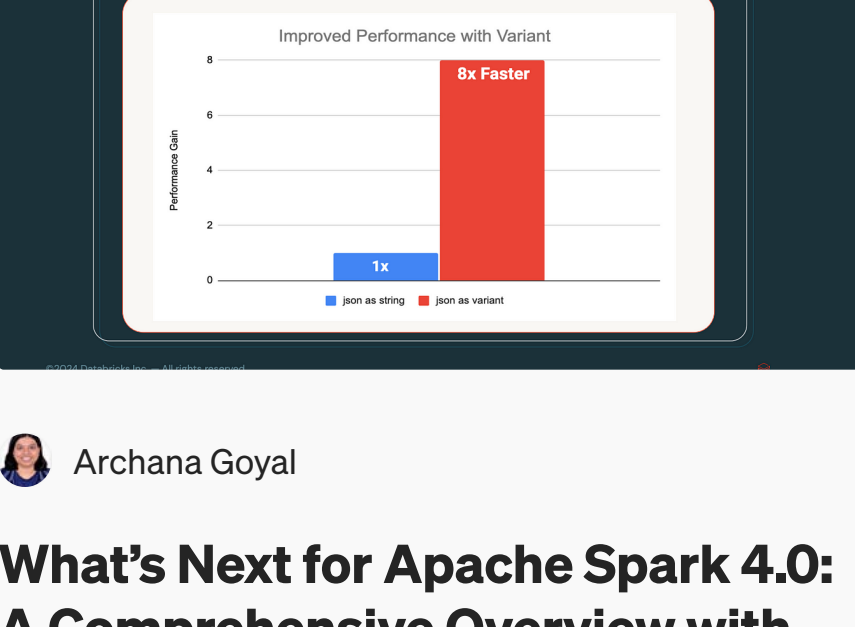


Optimizing API Data Ingestion with PySpark
Leveraging Pagination for Performance
Sep 8 · 22 likes · 2 shares

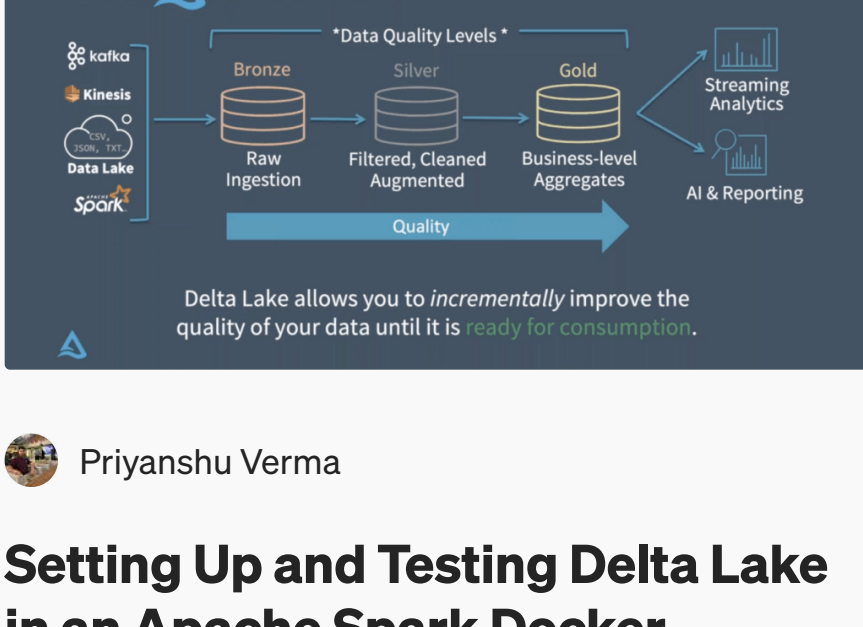
 **Interesting Design Topics**
257 stories · 823 saves

 **Staff Picks**
747 stories · 1356 saves

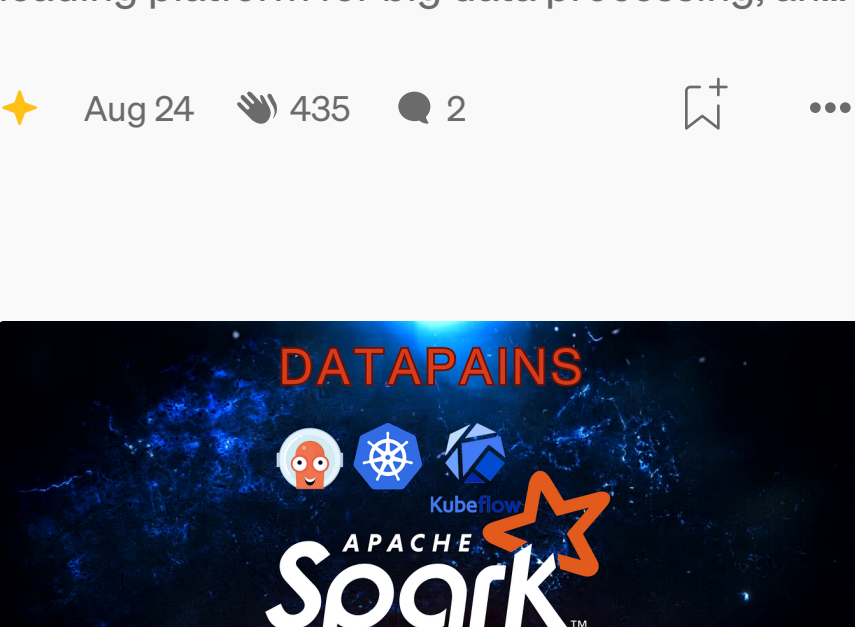
 **Natural Language Processing**
1747 stories · 1337 saves




What's Next for Apache Spark 4.0: A Comprehensive Overview with...
Apache Spark has established itself as a leading platform for big data processing, an...
Aug 24 · 435 likes · 2 shares



Setting Up and Testing Delta Lake in an Apache Spark Docker...
Delta Lake Setup in Apache Spark Docker
Aug 11



Master Apache Spark on Kubernetes and Beyond!
Apache Spark is a powerful open-source distributed computing system, and...
Jul 14 · 21 likes · 1 share



How Many Partitions Will Be Created for a 10 GB File?
Access this blog for free...
Aug 18 · 182 likes · 3 shares

See more recommendations

Help

Status

About

Careers

Press

Blog

Privacy

Terms

Text to speech

Teams