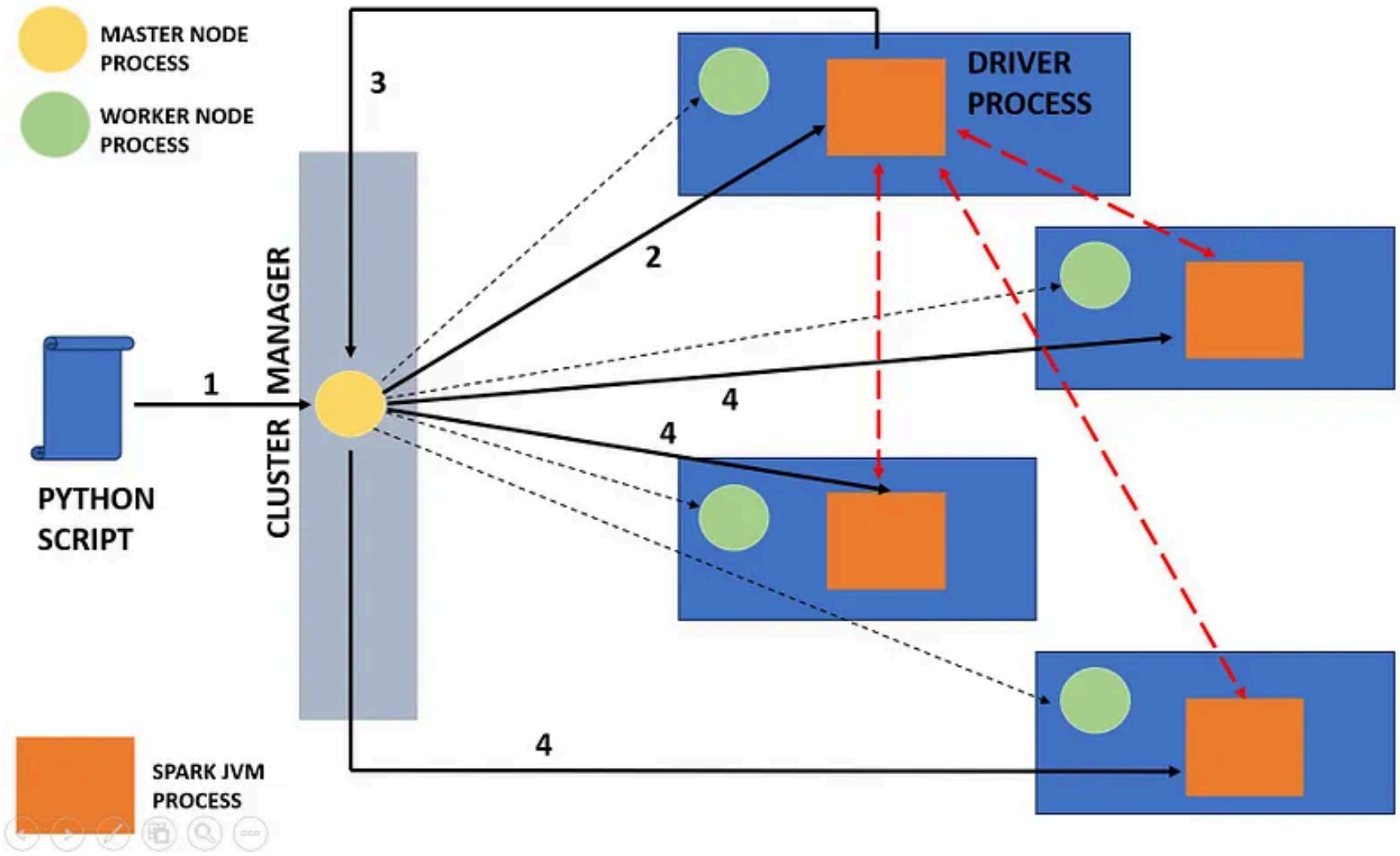
Pyspark Inner Workings : How

Yash Kothari · Follow

2 min read · Apr 24, 2024

Ever wondered what happens behind the scenes when you submit a Python script for execution in a PySpark application?



Let's break it down step by step.

Step 1: Client Request
The process kicks off with the client submitting the PySpark application. This submission, often a Python script containing the application logic, triggers a request to the cluster manager's driver node. Upon acceptance by the cluster manager, resources are allocated, and the driver process is placed onto a node within the cluster. With the client's task completed, the PySpark application seamlessly transitions to running on the cluster.

Step 2: Launch
With the driver process now deployed, it begins executing the user's Python code. This code initializes a SparkSession, setting up the Spark cluster comprising the driver and executor processes.

Step 3: Start-up Communication

The SparkSession communicates with the cluster manager to request the launch of executor processes across the cluster. User-specified configurations, such as the number of executors, are passed via command-line arguments during the spark-submit call.

Step 4 : Spark Cluster Established

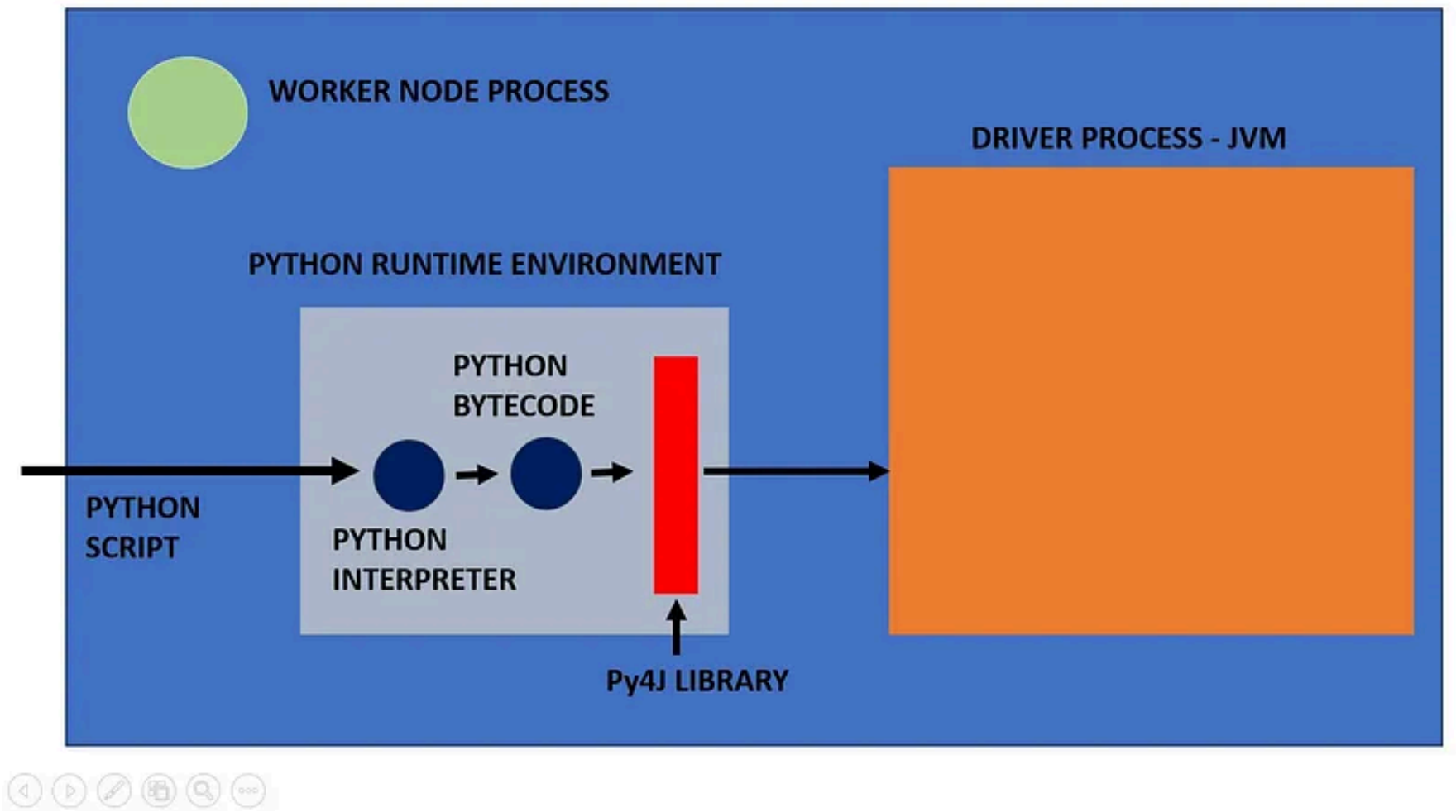
Upon receiving the request, the cluster manager responds by launching the executor processes, provided everything proceeds smoothly. It also communicates relevant information about the executor locations back to the driver process. This step establishes the complete "Spark Cluster" setup for the application.

Step 5: Execution

The PySpark application swings into action, executing its code across the Spark Cluster. Driver and executor processes collaborate, executing code and managing data flow. The driver schedules tasks onto each executor, while the executors report back with task execution status updates, indicating success or failure.

Step 6: Completion
As the PySpark application concludes its tasks, the driver process exits, signaling either success or failure. Subsequently, the cluster manager shuts down the executor processes associated with the Spark cluster. Users can then query the cluster manager for status updates to determine the outcome of the Spark application.

How the Execution Happens at the Executor End?



At the executor end, the Python script is received by the Python Runtime Environment, where the Python Interpreter stands ready to process it. The Python Interpreter first converts the script into

Python bytecode. This bytecode is then executed on the Java Virtual Machine (JVM) with the help of the Py4J library, facilitating seamless communication between the Python runtime environment and the JVM.

Pyspark Spark

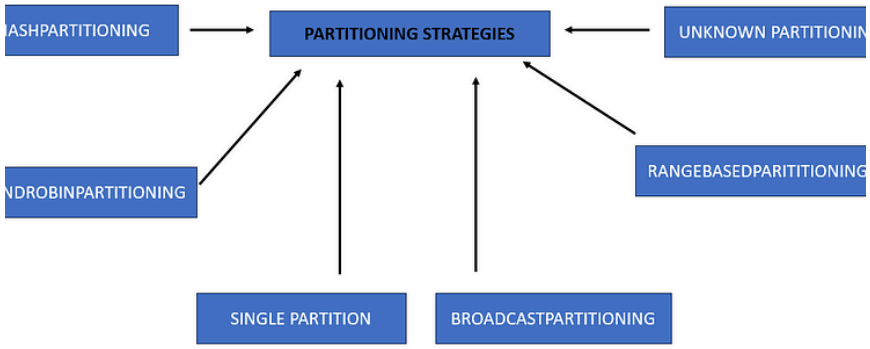
Written by Yash Kothari

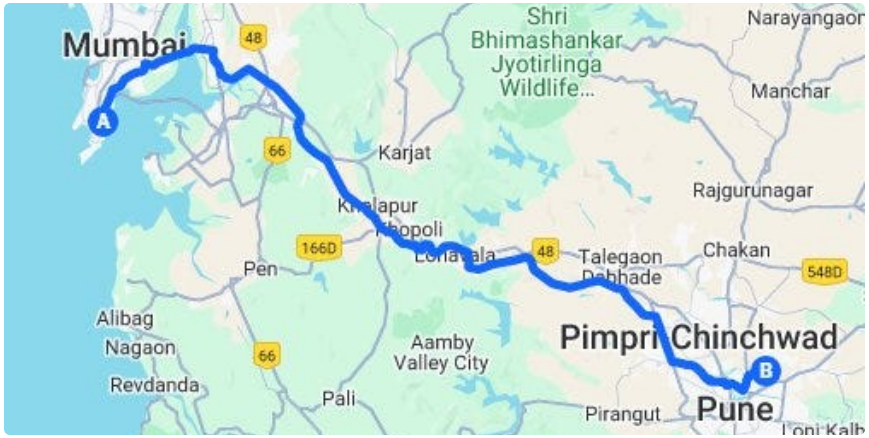
Follow

17 Followers

Data Engineer | Azure | Databricks | Problem Solver | Unleashing C++ Wizardry on YouTube - @theycallmecoder | https://www.linkedin.com/in/ykothari99/

More from Yash Kothari





Yash Kothari

Partitioning Strategies in Spark

Apache Spark, with its distributed computing model, excels at processin...

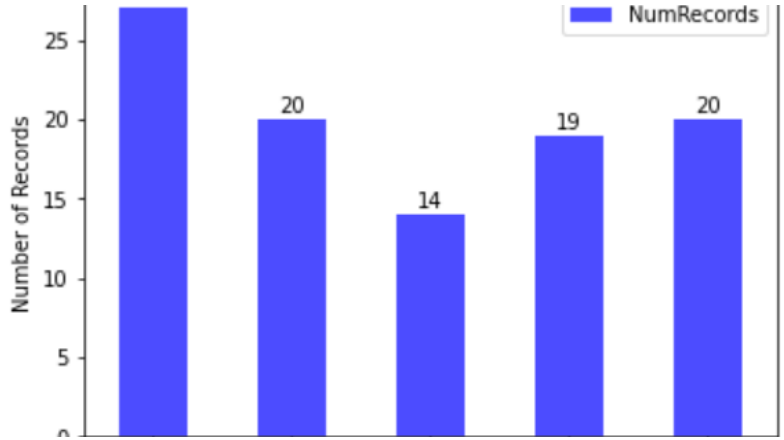
Apr 21 7

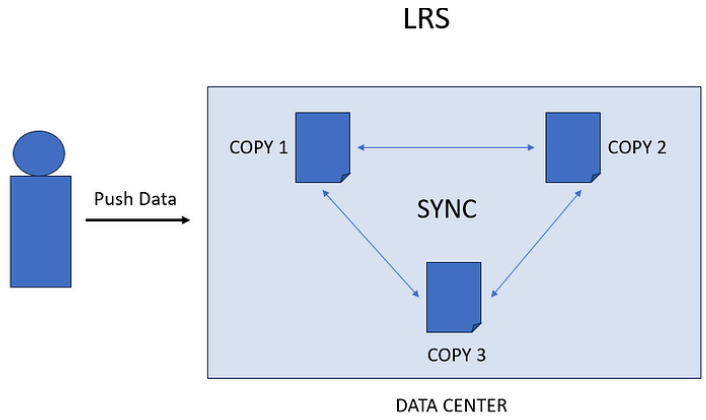
Yash Kothari

Coalesce in Spark, the internal working

Coalesce in spark is mainly used to reduce the number of partitions. Why i...

Apr 30 5 1





Yash Kothari

Custom Partitioning in Pyspark

In Apache Spark, the partitioner plays a crucial role in determining how data is...

May 6 2

Yash Kothari

Azure Data Redundancy Choices

Selecting the appropriate redundancy choice is crucial for optimizing...

Jun 26

See all from Yash Kothari

Recommended from Medium


Vishal Barvaliya

How Many Partitions Will Be Created for a 10 GB File?

Now, let's talk about data partitioning. When you work with large files, they ar...


Feruz Urazaliev

Mastering Spark Performance: Top Secrets for Turbochargin...

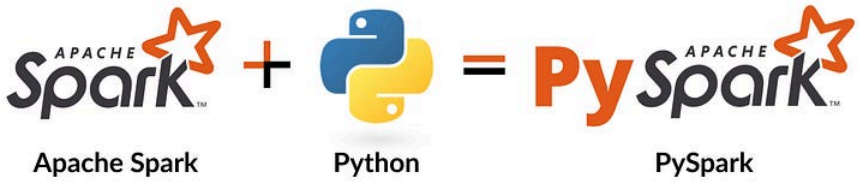
Apache Spark is a robust and scalable engine for large-scale data processing,...

Aug 19 82 1 Aug 20 3

Lists

Natural Language Processing

1681 stories · 1254 saves



Mohammed Azarudeen Bilal

PySpark Interview Questions

60+ PySpark Coding Questions Every Data Engineer Should Know

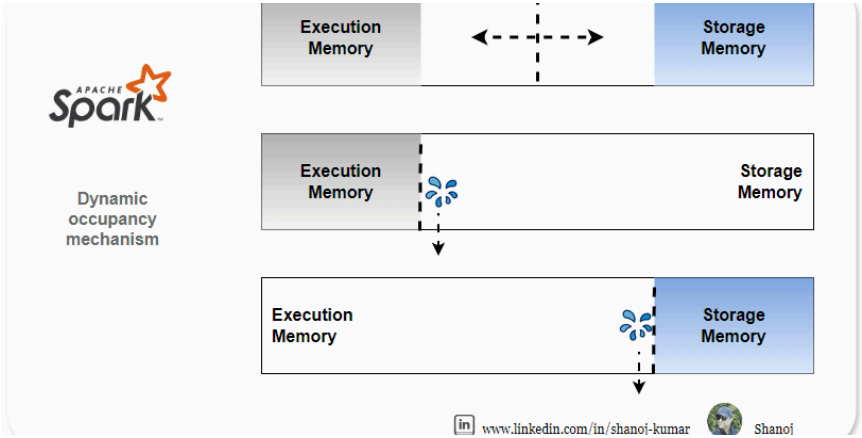
Aug 17 115 2

Subham Khandelwal

PySpark—Spark Streaming Error and Exception Handling

Understand How to handle Spark Streaming Errors and Exceptions

Mar 20 17



Rubihali

Spark Mastery—Lets explore data skewness interview...

We will go through most asked questions in spark interviews along wit...

May 19 53

Shanoj in Stackademic

Understanding Memory Spills in Apache Spark

Memory spill in Apache Spark is the process of transferring data from RAM...

Mar 11 208

See more recommendations