Data Engineering
Interview Questions

Read or Repost for others ♻

PySpark

I-5

✅ Explain the difference between transformations and actions in PySpark.

✅ How does PySpark handle fault tolerance and data recovery?

✅ Can you describe how PySpark's DAG (Directed Acyclic Graph) scheduler works?

✅ What is the significance of partitioning in PySpark and how can it affect performance?

✅How would you perform iterative operations in PySpark and why are they challenging

6-10

✅Broadcast variables and accumulators play vital roles in PySpark.

✅ Narrow and wide transformations in PySpark have significant impacts on performance.

✅ PySpark integrates seamlessly with other components of the Hadoop ecosystem like HDFS and YARN.

✅The PySpark DataFrame API is preferred over RDDs in scenarios involving structured data and SQL-like operations.

✅ Managing and optimizing memory usage in PySpark applications is crucial for performance.

Follow

Asheesh .

Lead Data Engineer | Trainer 🌳 | 20k+ Followers | Mentor | AWS | Azure | Data Engineering | Data Analytics | Machine Learning

Top Data Engineering Voice

Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), Bhopal

for more

19 APRIL 2024

www.linkedin.com/in/asheeshlive