# PySpark Debugging Code Examples

1. 1. Enable Spark Logging

spark.sparkContext.setLogLevel("INFO")  
# Options: ALL, DEBUG, INFO, WARN, ERROR  
# Helps reduce or increase log verbosity for better debugging.

1. 2. Check Spark Configuration at Runtime

spark.sparkContext.getConf().getAll()  
# Useful to verify memory, cores, shuffle config.

1. 3. Executor Memory Debug

--conf spark.executor.memory=2g  
--conf spark.driver.memory=1g  
# Used when facing OutOfMemory errors.

1. 4. Shuffle Debugging - Too Much Shuffle

df.repartition(100, "key")  
# Before groupBy or join  
# Avoids default small partitions that cause skew.

1. 5. Handling Skewed Join Keys

from pyspark.sql.functions import broadcast  
result = big\_df.join(broadcast(small\_df), "key")  
# Fixes long tasks due to skew.

1. 6. Catch Read Failures

try:  
 df = spark.read.csv("path/does/not/exist")  
except Exception as e:  
 print("Failed to read:", e)  
# Good for detecting bad paths or file formats.

1. 7. Memory Usage via Executors Tab

Use Spark UI at localhost:4040 > Executors tab  
# Diagnoses executor memory issues visually.

1. 8. Monitor Job Execution Time

from time import time  
start = time()  
df.count()  
print("Execution Time:", time()-start)  
# Useful to measure slow operations.

1. 9. Use .persist() Wisely

df.filter("status = 'active'").persist()  
# Avoid caching raw or large unfiltered data.

1. 10. Analyze Task Duration

View task skew in Spark UI > Stages > Tasks  
# Useful to detect long-running tasks.