Script	Companion should detect …	Spark-UI / symptom to verify
taxi_skewed_join_small.py	Skewed shuffle & broadcast disabled; single-file write	One long straggler task; shuffle read skew; 1 output file
taxi_collect_driver_small.py	Excessive collect() to driver	Driver GC / memory spike; executors idle
taxi_too_many_partitions_small.py	Excessive shuffle partitions & duplicate shuffle	Two 10 000-task stages; high scheduler delay
taxi_cache_no_unpersist_small.py	Large DF cached but never unpersisted	Storage tab shows DF cached; later jobs GC/spill
taxi_python_udf_small.py	Python (exec) UDF disables whole-stage codegen	Task CPU time dominated by Python; no codegen
taxi_cartesian_join_small.py	Cartesian join	Plan shows CartesianProduct; huge shuffle read
taxi_many_small_files_small.py	Writes thousands of tiny files	5 000 FileOutputCommitter tasks; many small files in GCS
taxi_no_compression_small.py	Output Parquet with compression disabled	codec=none in Env tab; output size unusually large
taxi_multi_cache_small.py	Multiple large caches not released	Several cached DFs; executor memory near cap
taxi_rdd_conversion_small.py	Unnecessary DF → RDD → DF conversion	Plan loses Tungsten/columnar; extra serialize stage
taxi_autoscaling_backlog_small.py	initialExecutors too low → backlog at start	Executor count jumps sharply after launch; Task backlog queue grows ear
taxi_broadcast_threshold_small.py	Broadcast join blocked by tiny 1 MB threshold	No BroadcastHashJoin node; shuffle join instead
taxi_memory_spill_small.py	Low spark.memory.fraction causes spill	Shuffle spill (disk) metrics high; Execution memory extremely low
taxi_pandas_udf_small.py	Row-wise pandas UDF instead of built-in agg	Pandas UDF node; ‘Arrow optimization’ absent; higher task deserializatio
taxi_gc_heavy_rdd_small.py	RDD flatMap generates many small objects → GC	Executors tab: GC Time high (red); CPU util low