Apache Spark Configuration Cheat Sheet

Basic Spark Configs

spark.executor.memory = Memory allocated per executor (e.g., 4g)
spark.executor.cores = Number of cores per executor (e.g., 2)
spark.driver.memory = Memory allocated for driver (e.g., 4g)
spark.num.executors = Total number of executors
spark.sql.shuffle.partitions = Partitions after shuffle (default: 200)
spark.default.parallelism = Default number of tasks for RDDs

Advanced Tuning

spark.memory.fraction = Fraction of JVM heap for execution/caching (default: 0.6)
spark.memory.storageFraction = Memory for storage (default: 0.5 of memoryFraction)
spark.speculation = Enable speculative task execution (true/false)
spark.memory.offHeap.enabled = Enables off-heap memory (true/false)

Dynamic Allocation

spark.dynamicAllocation.enabled = Enables dynamic executor allocation
spark.dynamicAllocation.minExecutors = Minimum executors to allocate
spark.dynamicAllocation.maxExecutors = Maximum executors
spark.dynamicAllocation.initialExecutors= Initial executors

Performance Optimization Tips

- Avoid using collect() on large datasets.
- Tune shuffle partitions according to data volume.
- Use caching/persisting only when reuse is needed.
- Monitor Spark UI to understand job bottlenecks.
- Use G1GC for better garbage collection tuning.

AWS Glue & EMR

Apache Spark Configuration Cheat Sheet

- Glue: Set config in job parameters using --conf.
- EMR: Use configuration classifications to inject Spark properties.