Apache Spark Configuration Cheat Sheet

Core Execution Configs

spark.master = Master URL (e.g., local[*], yarn)
spark.app.name = Name of the Spark application
spark.executor.instances = Number of executor instances (when dynamic allocation is off)
spark.executor.memory = Memory per executor (e.g., 4g)
spark.executor.cores = Cores per executor
spark.driver.memory = Memory for the Spark driver
spark.driver.cores = Cores for the driver (mainly in cluster mode)

Dynamic Resource Allocation

spark.dynamicAllocation.enabled = Enable dynamic executor allocation (true/false)
spark.dynamicAllocation.minExecutors = Minimum executors
spark.dynamicAllocation.maxExecutors = Maximum executors
spark.dynamicAllocation.initialExecutors = Initial number of executors

Memory Management

spark.memory.fraction = Fraction of JVM heap for execution/storage (default: 0.6)
spark.memory.storageFraction = Fraction of memory.fraction reserved for storage (default: 0.5)
spark.memory.offHeap.enabled = Enable off-heap memory (true/false)
spark.memory.offHeap.size = Size of off-heap memory if enabled

Shuffle & Parallelism

spark.sql.shuffle.partitions = Partitions for DataFrame shuffle operations (default: 200)
spark.default.parallelism = Default number of RDD partitions (usually 2 * total cores)
spark.reducer.maxSizeInFlight = Max size for a single shuffle block in memory (e.g., 48m)
spark.shuffle.compress = Enable compression for shuffle (default: true)
spark.shuffle.file.buffer = Buffer size for shuffle files (e.g., 32k)

Speculative Execution & Fault Tolerance

Apache Spark Configuration Cheat Sheet

spark.speculation = Enable speculative execution (default: false)
spark.speculation.multiplier = Threshold for speculative task (default: 1.5x median)
spark.speculation.quantile = Fraction of tasks that must be completed before speculation
spark.task.maxFailures = Max failures allowed per task (default: 4)
spark.yarn.maxAppAttempts = Max number of app attempts on YARN

Serialization & Compression

spark.serializer = Serialization library (e.g., org.apache.spark.serializer.KryoSerializer)
spark.kryo.registrationRequired = Require class registration for Kryo (true/false)
spark.rdd.compress = Compress serialized RDD partitions (default: false)
spark.io.compression.codec = Compression codec (e.g., lz4, snappy, zstd)

Logging & Debugging

spark.eventLog.enabled = Enable Spark event logging (default: false)
spark.eventLog.dir = Directory to store event logs
spark.history.fs.logDirectory = Location of Spark History logs
spark.ui.showConsoleProgress = Show progress bars in console (default: true)

GC & JVM Tuning

spark.executor.extraJavaOptions = JVM options for executors (e.g., GC tuning)
spark.driver.extraJavaOptions = JVM options for the driver
-XX:+UseG1GC or -XX:+UseZGC = Common GC flags for tuning