5 APACHE SPARK BEST PRACTICES

SERIALIZATION IS KEY

Make sure that your programs can serialize, deserialize, and send objects across the wire quickly.

PROPER PARTITION RECOMMENDATIONS AND SIZING

Decide on the number of partitions in an RDD by equating the number of partitions to a multiple of the number of cores in the cluster. This way, all the partitions will process in parallel and the resources receive optimum utilization.

MONITOR EXECUTOR SIZE + YARN MEMORY OVERHEAD

When using Spark on YARN, keep an eye on both monitor executor size and YARN memory overhead. This is to prevent the YARN scheduler from killing an application that uses a large amount of NIO memory or other offhead memory areas.

GET THE MOST OUT OF DAG MANAGEMENT

Monitor your DAG, not just the overall complexity of the execution plan. Make sure each stage in your code is actually running in parallel.

MANAGE LIBRARY CONFLICTS

Ensure any external dependencies and classes you bring in are available in the environment you are using, and make sure that they don't conflict with internal libraries used by your version of Spark.