Spark Performance

Techniques for Performance

- Level of parallelism: (1) Provide as a parameter for shuffle (2) Resdistribute a RDD to have more or fewer partitions.
- Serialization format: Use Kryo
- Memory management
- Hardware provisioning: amount of memory, number of cores, total number of executors, and the number of local disks for scratch data