Spark Configurations

DSC 232R - Spring 2025

Cores and Memory

Dataset size = 20 GB

Assume 2x more memory is required due to overhead.

Memory required = 2 * 20 GB = 40 GB

Assume 2GB per core

Cores = 40 GB / 2 GB = 20 Cores

Total Memory = 40 GB

Executors

Assume each executor has 4 cores

Memory per executor = cores per executor * memory per core

Memory per executor = 4 cores * 2GB

= 8GB

No. of executors = Total Memory / Memory per executor

= 40GB / 8GB = 5 executors

Driver

Usually driver memory is assigned same as one executor memory.

So in our case driver memory = 8GB.

Ideally, this much memory is not required so try to use only as much as required. This depends on how much data you are shuffling, collecting to the driver.

Code

```
from pyspark.sql import SparkSession


spark = SparkSession.builder \
    .appName("MySparkJob") \
    .config("spark.executor.instances", "5") \
    .config("spark.executor.cores", "4") \
    .config("spark.executor.memory", "8g") \
    .config("spark.driver.memory", "8g") \
    .getOrCreate()
```

Spark RDD Optimizations

1. Prefer reduceByKey() over groupByKey()

```
#BAD

rdd.groupByKey().mapValues(sum)


#GOOD

rdd.reduceByKey(lambda a, b: a + b)
```

Spark RDD Optimizations

2. Persist RDDs that are reused multiple times and unpersist when they are no longer needed.

3. When joining with small datasets, use broadcast variables instead of joins.

4. Filter as much data as possible and only broadcast the data that is actually required.

5. Use coalesce instead of repartition for decreasing partitions.