Spark-perf Guide

1. Download

<https://github.com/databricks/spark-perf>

1. Configuration File

config/config.py.template to config/config.py

1. Running on YARN cluster

Vim config/config.py

#Spark home

SPARK\_HOME\_DIR = the path to installed YARN

#Spark URL for YARN

SPARK\_CLUSTER\_URL = “yarn-client”

#YARN mode

IS\_YARN\_MODE = True

IS\_MESOS\_MODE = False

#If have existing Spark cluster

USE\_CLUSTER\_SPARK = True

RSYNC\_SPARK\_HOME = False

#When fisrt run tests, have to compile them. So set PREP\_SPARK\_TESTS from True to False after first installation

PREP\_SPARK\_TESTS = False

#changes for YARN mode parameters

SPARK\_DRIVER\_MEMORY = “20g --num-executors 2 --executor-cores 1 --executor-memory 1640m”

#How mang times to run each experiment, if set IGNORED\_TRIALS to N, will remove the first N trials for each test,so default set to 0

IGNORED\_TRIALS = 0

OptionSet(“num-trials”,[1])

1. Change Scale

For controlling test time, have to set the scale factor.

#the SCALE\_FACTOR is for all test, use this variable to scale the values(e,g number of records in a generated dataset). It is a global variable.

SCALE\_FACTOR = 1.0

#For first test of spark core test-- scheduling-throughput, can set parameter about scale in SCHEDULING\_THROUGHPUT\_OPTS

# The number of tasks that should be launched in each job:

OptionSet("num-tasks", [10 \* 1000]),

# The number of jobs that should be run:

OptionSet("num-jobs", [1])

For other 7 tests in spark core test, add factor after SPARK\_TESTS , for example

SPARK\_TESTS += [("scala-agg-by-key", "spark.perf.TestRunner", SCALE\_FACTOR \* 0.01,

COMMON\_JAVA\_OPTS, [ConstantOption("aggregate-by-key")] + SPARK\_KV\_OPTS)]

The factor 0.01 is used to reduce the scale for the test-“scala-agg-by-key”

1. Driver Log

If change driver log from err File to std out:

Vim ./lib/sparkperf/testsuites.py

130

-cmd = "%s --class %s --master %s --driver-memory %s %s %s 1>> %s 2>> %s" % (

+ cmd = "%s --class %s --master %s --driver-memory %s %s %s 1>> %s" % (

133

-stdout\_filename, stderr\_filename)

+stdout\_filename)

1. Run

Bin/run

1. Result

bin/results/spark\_perf\_output\_\_timeStamp\_logs/test-name.out