SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)
Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

Department of MCA

Topic: Hadoop Benchmarks

COURSE

16CA917

Big Data Analytics

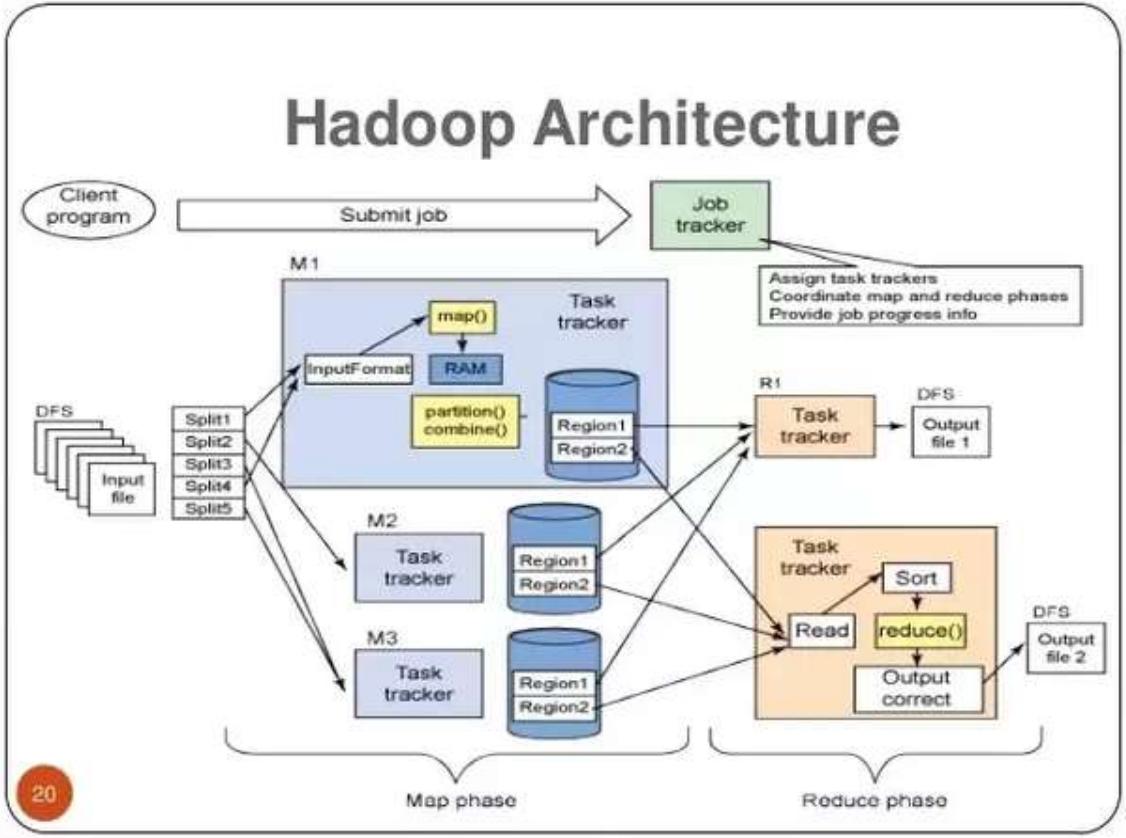
UNIT - III

Hadoop Environment

CLASS

V Semester / III MCA

Hadoop Architecture



Hadoop Benchmarks

- Good way to verify whether your HDFS cluster is set up properly and performs as expected
- Used to enhance the performance of processing on very large volume of data on clusters
- Benchmarks are packaged in the test JAR file
  - *hadoop jar $HADOOP_INSTALL/hadoop-*-test.jar*

TestDFSIO

- Tests the I/O (read & write) performance of HDFS
- Statistics are accumulated in the reduce to produce a summary
- Designed in such a way that it will use 1 map task per file
- Helpful for tasks such as
  - stress testing HDFS
  - to discover performance bottlenecks in network,
  - to shake out the hardware, OS and Hadoop setup of your cluster machines
  - how fast your cluster is in terms of I/O

Output of TestDFSIO

Generate 10 files of size 1 GB for a total of 10 GB:

```
$ hadoop jar hadoop-*test*.jar \
    TestDFSIO -write -nrFiles 10 -fileSize 1000
```

Typical output of write test

```
----- TestDFSIO ----- : write
           Date & time: Mon Oct 06 10:21:28 CEST 2014
       Number of files: 10
Total MBytes processed: 10000.0
     Throughput mb/sec: 12.874702111579893
Average IO rate mb/sec: 13.013071060180664
 IO rate std deviation: 1.4416050051562712
    Test exec time sec: 114.346
```

Output of TestDFSIO

Read 10 input files, each of size 1 GB:

```
$ hadoop jar hadoop-*test*.jar \
    TestDFSIO -read -nrFiles 10 -fileSize 1000
```

Typical output of read test

```
----- TestDFSIO ----- : read
            Date & time: Mon Oct 06 10:56:15 CEST 2014
        Number of files: 10
Total MBytes processed: 10000.0
     Throughput mb/sec: 402.4306813151435
Average IO rate mb/sec: 492.8257751464844
 IO rate std deviation: 196.51233829270575
    Test exec time sec: 33.206
```

MRBench (MapReduce Bench)

- It runs a small job a number of times
- It acts as a good counterpoint to sort, as it checks whether small job runs are responsive

NNBench(NameNode Bench)

- Useful for load testing namenode hardware
- Gridmix is a suite of benchmarks designed to model a realistic cluster workload, by mimicking a variety of data-access patterns seen in practice

UserJobs

- A few jobs that are representative of the jobs that users run, can be included as benchmarks.
- When our own jobs are used as benchmarks, we select a dataset for user jobs

Benchmarking with Sort

- Terasort Benchmark is used to test HDFS & MapReduce program by sorting some amount of data as quick as possible in order to measure the capabilities
- It includes three components:
  - TeraGen - Generate some random data,
  - TeraSort - Perform the sort, then
  - TeraValidate - validate the results

Challenges

- After running each benchmark test, HDFS should be cleaned so that the next benchmark can be run without any storage issue
- No two benchmark tests should be run simultaneously so as to avoid memory issues

References

- Tom White, " Hadoop: The Definitive Guide" Third Edition, O'reilly Media, 4th Edition, 2012

Web Resources

- https://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench/
- https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/Benchmarking.html

Thank You