MapReduce

- Abhay Dandekar

Agenda

1. **What is** parallel programming?
2. **What is** map-reduce?
3. **What are** other paradigms apart from map-reduce?
4. **Why** map-reduce?
5. **What is** Hadoop Architecture
6. Definitions: Mapper, Reducer, Combiner
7. **What is** Resource manager?
8. **What are** the different processes?
9. **How to** execute HelloWorld of BigData?
10. **How** Single Reducer MR works?
11. **How** Multi Reducer MR works?
12. **How** Shuffle Sort Magic takes place?
13. **Log time!!!**
14. Q n A?

What is parallel programming?

1. Scale-out v/s Scale-up
2. Resource utilization
    a. CPU utilization
    b. Memory
    c. Hard-disk IOPS
3. CPU idle time
4. Disk Input Output bottlenecks
5. Map-Reduce benefits
    a. Independent of resources
    b. Near linear increase in throughput
    c. Lesser context changing overhead.

What is Map-Reduce?

1. Programing Paradigm OR Framework OR Concept ?
2. Programing Paradigm ?
3. Framework ?
4. Concept ?

What are the other paradigms of MR?

1. Alternatives to the MR framework
    a. HT Condor
    b. Spark
    c. Hive
    d. Pig
2. Internally most use the concept of MR

Why MapReduce

1. Simple logic
2. Can easily handle huge amounts of data
3. Parallel execution
4. Linear growth in scale-out

Architecture

Let’s get started !!!

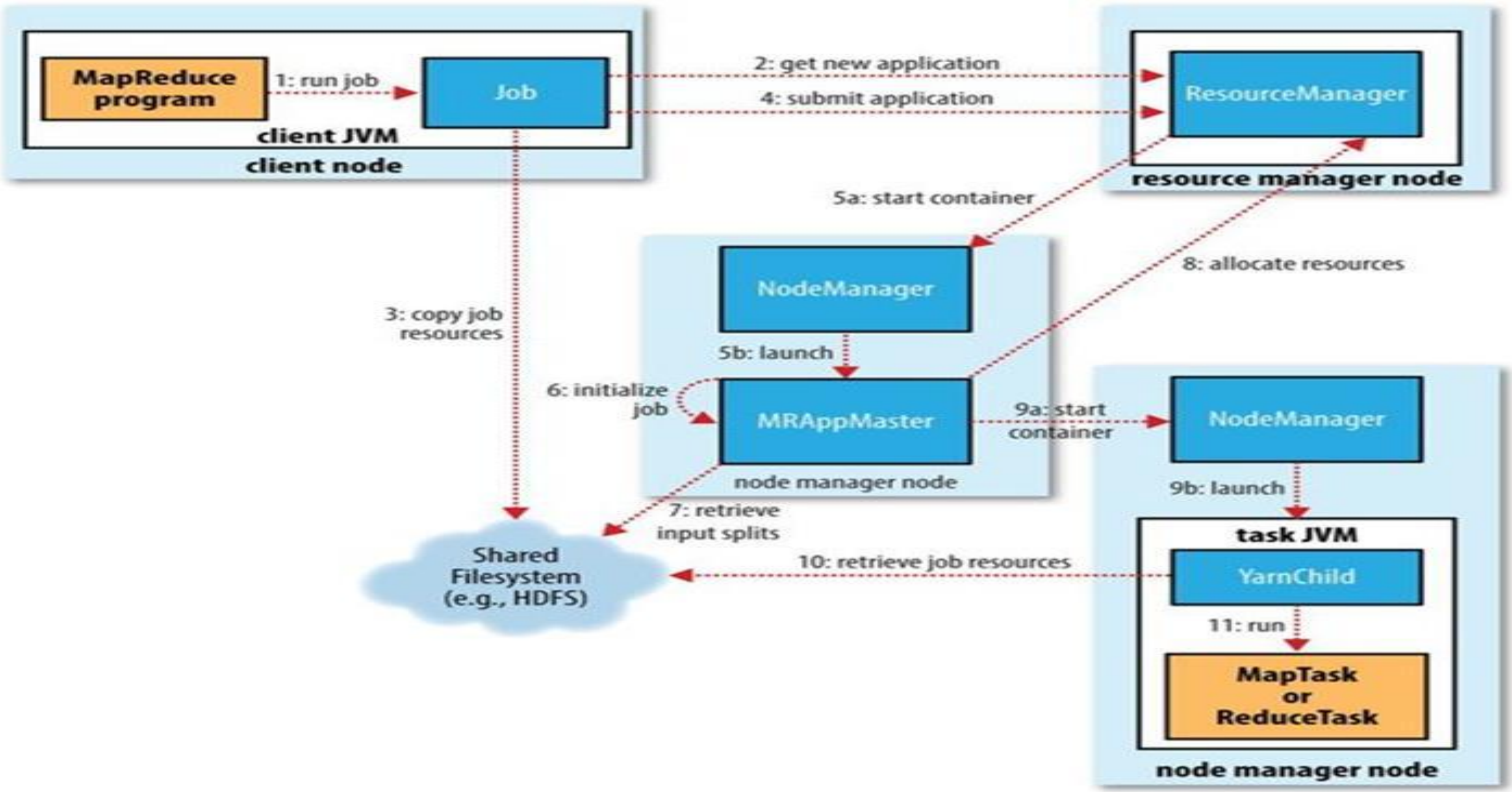
What are the different processes?

Start HDFS and YARN from console

Different processes

1. ResourceManager
2. NameNode
3. SecondaryNameNode
4. DataNode
5. NodeManager

Hadoop Architecture (YARN)



MR Execution

Live Action !!!

Definitions

1. Mapper
    a. Runs directly on the input from HDFS
2. Combiner (a.k.a local reducer)
    a. Runs on the individual output of Mapper (locally). Framework may or may not run Combiner over the map output.
    b. Also known as Local Reducer
3. Reducer
    a. Runs on the “grouped by key” output of Mapper

All the above processes are nothing but ***YarnChilds*** getting spawned onto cluster

Hello World of BigData

1. Single reducer
2. Multiple reducer
3. With combiner

Single Reducer DFD

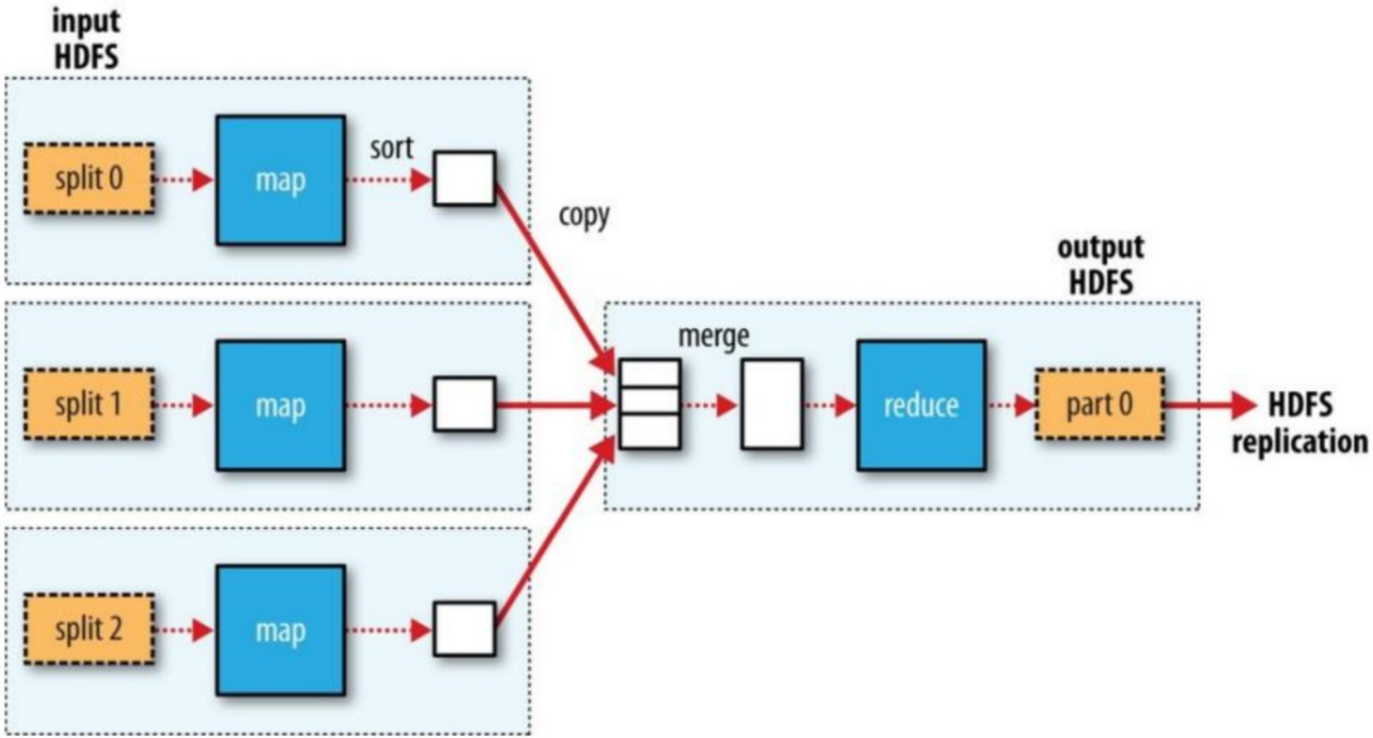


Figure 2-3. MapReduce data flow with a single reduce task

Multi Reducer DFD

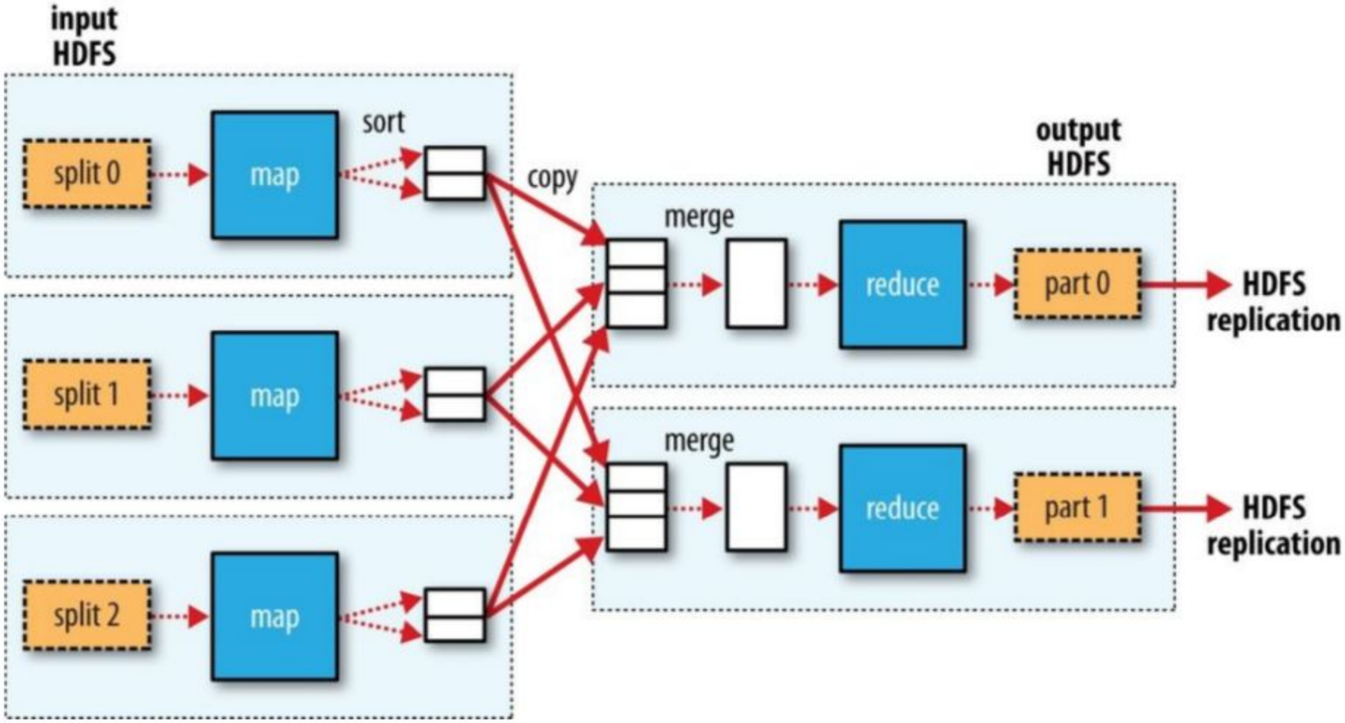


Figure 2-4. MapReduce data flow with multiple reduce tasks

Shuffle - Sort Magic

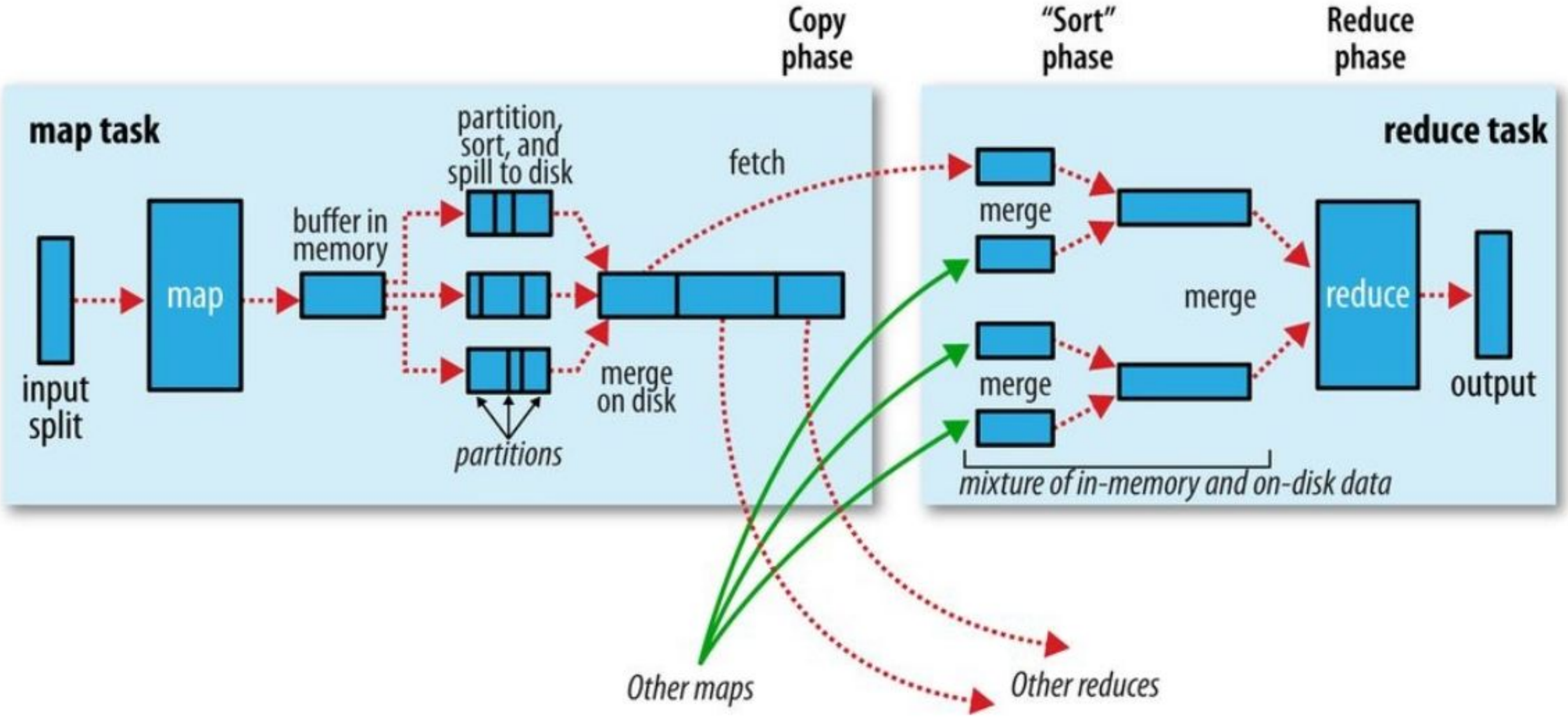


Figure 7-4. Shuffle and sort in MapReduce

DataTypes in Hadoop

1. Writables
   a. Text
   b. BooleanWritable
   c. DoubleWritable
   d. FloatWritable
   e. IntWritable
   f. LongWritable
   g. ShortWritable
   h. ArrayWritable
   i. VIntWritable
   j. VLongWritable
2. Comparable Interface

Execution

Log Time !!!

Questions and Answers / Practical