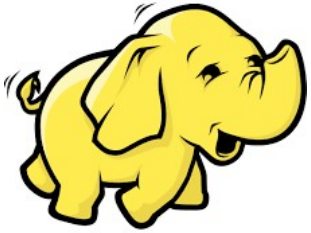
Map Only Jobs

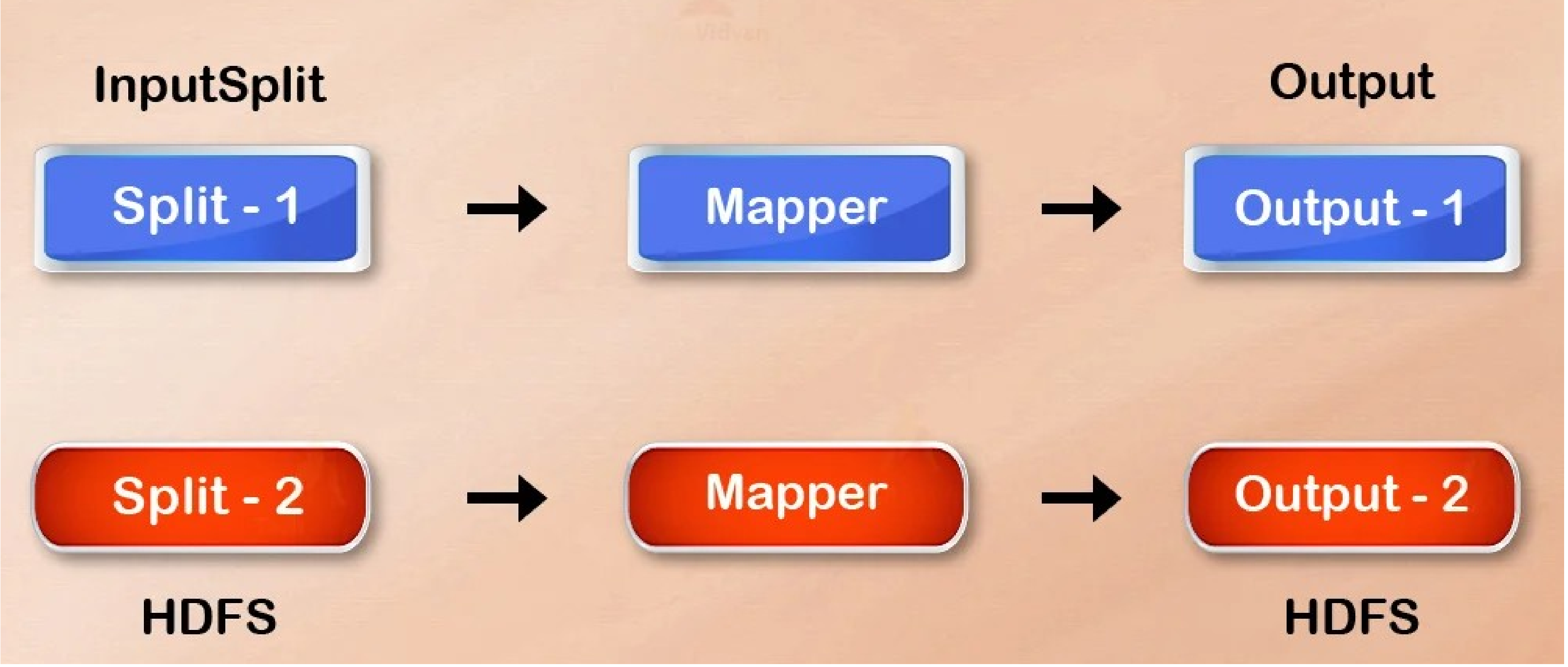
Tushar B. Kute,
http://tusharkute.com



Map-Only

- Map-Only job in the Hadoop is the process in which mapper does all tasks. No task is done by the reducer. Mapper’s output is the final output.
- MapReduce is the data processing layer of Hadoop. It processes large structured and unstructured data stored in HDFS. MapReduce also processes a huge amount of data in parallel.
- It does this by dividing the job (submitted job) into a set of independent tasks (sub-job). In Hadoop, MapReduce works by breaking the processing into phases: Map and Reduce.

Map-Only


InputSplit
Output
Split - 1
Mapper
Output - 1
Split - 2
Mapper
Output - 2
HDFS
HDFS

Map-Only

- How to avoid Reduce Phase in MapReduce?
  - By setting job.setNumreduceTasks(0) in the configuration in a driver we can avoid reduce phase.
  - This will make a number of reducer as 0. Thus the only mapper will be doing the complete task.

Advantages

- In MapReduce job execution in between map and reduces phases there is key, sort and shuffle phase.
- Shuffling –Sorting are responsible for sorting the keys in ascending order. Then grouping values based on the same keys. This phase is very expensive.
- If reduce phase is not required, we should avoid it. As avoiding reduce phase would eliminate sorting and shuffle phase as well. Therefore, this will also save network congestion.

Advantages

- The reason is that in shuffling, an output of the mapper travels to reduce. And when the data size is huge, large data needs to travel to the reducer.
- The output of the mapper is written to local disk before sending to reduce.
- But in map only job, this output is directly written to HDFS. This further saves time as well reduces cost.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com