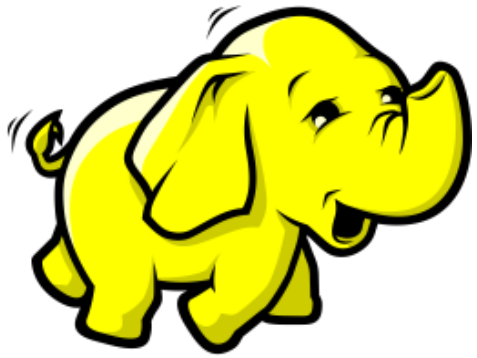


Big Data – Hadoop

Trainer: Mr. Nilesh Ghule.

Input/Output Format and Input Splits

- InputFormat – how to read data
  - FileInputFormat, TextInputFormat
  - KeyValueTextInputFormat, NLineInputFormat
  - DBInputFormat
- RecordReader – Logical division of record
- Number of mappers = Number of input splits
- Number of input splits ≈ Number of HDFS blocks
- Output format – how to write data
  - FileOutputFormat, TextOutputFormat
  - DBOutputFormat
- RecordWriter – Write individual record
- Number of output files = Number of reducers
- Output written on HDFS (replicated)

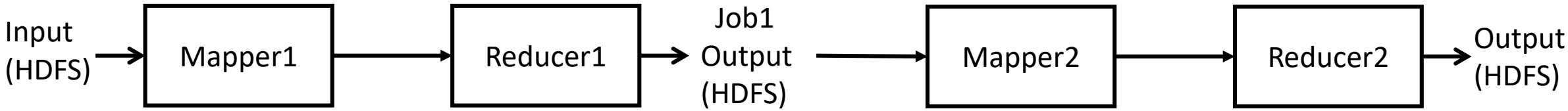
Uber job

- By default for each MR job separate MRAppMaster, Mapper(s) and Reducer(s) processes (containers) are created.
- For small data processing this process creation and their communication is overhead.
- Uber mode allows running such small jobs in single container i.e. MRAppMaster.
- Mapper(s) & Reducer runs in same process. As no IPC involves, these small jobs are executed quickly.
- It is configured using settings
  - mapreduce.job.ubertask.enable (default: false)
  - mapreduce.job.ubertask.maxmaps (default: 9)
  - mapreduce.job.ubertask.maxreduces (default: 1)

Chained Map Reduce Job

- Complex processing may not be completed in single MR job.
- Such processing can be done by feeding output of first job as input to second job. This way multiple jobs can be chained to each other.
- Need to update driver code. Run second job, only if first job is successful.

- Friends recommendation on Facebook or any social network site. Steps:
- Get count of common friends of X & Y (but X & Y are not friends of each other).
- For X, recommend top N Y's (based on common friends count).
- You may consider a minimal threshold for suggesting as friend.



Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>