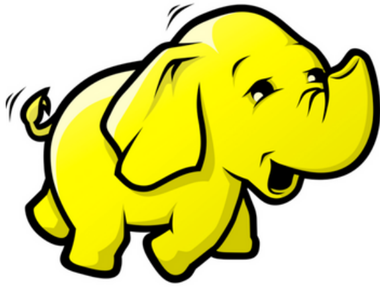
UNDERSTANDING MAPREDUCE IN HADOOP

Introduction

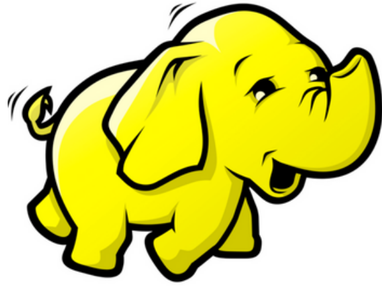
MapReduce is a component of the Apache Hadoop ecosystem, a framework that enhances massive data processing.



Introduction

There are two primary tasks in MapReduce: map and reduce.

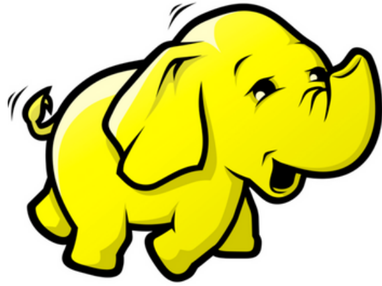
We perform the former task before the latter.



Introduction

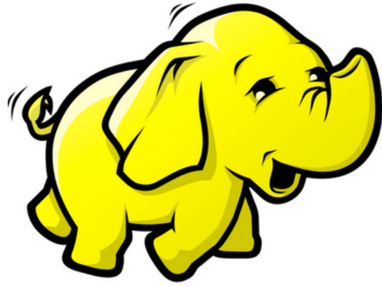
In the map job, we split the input dataset into chunks.

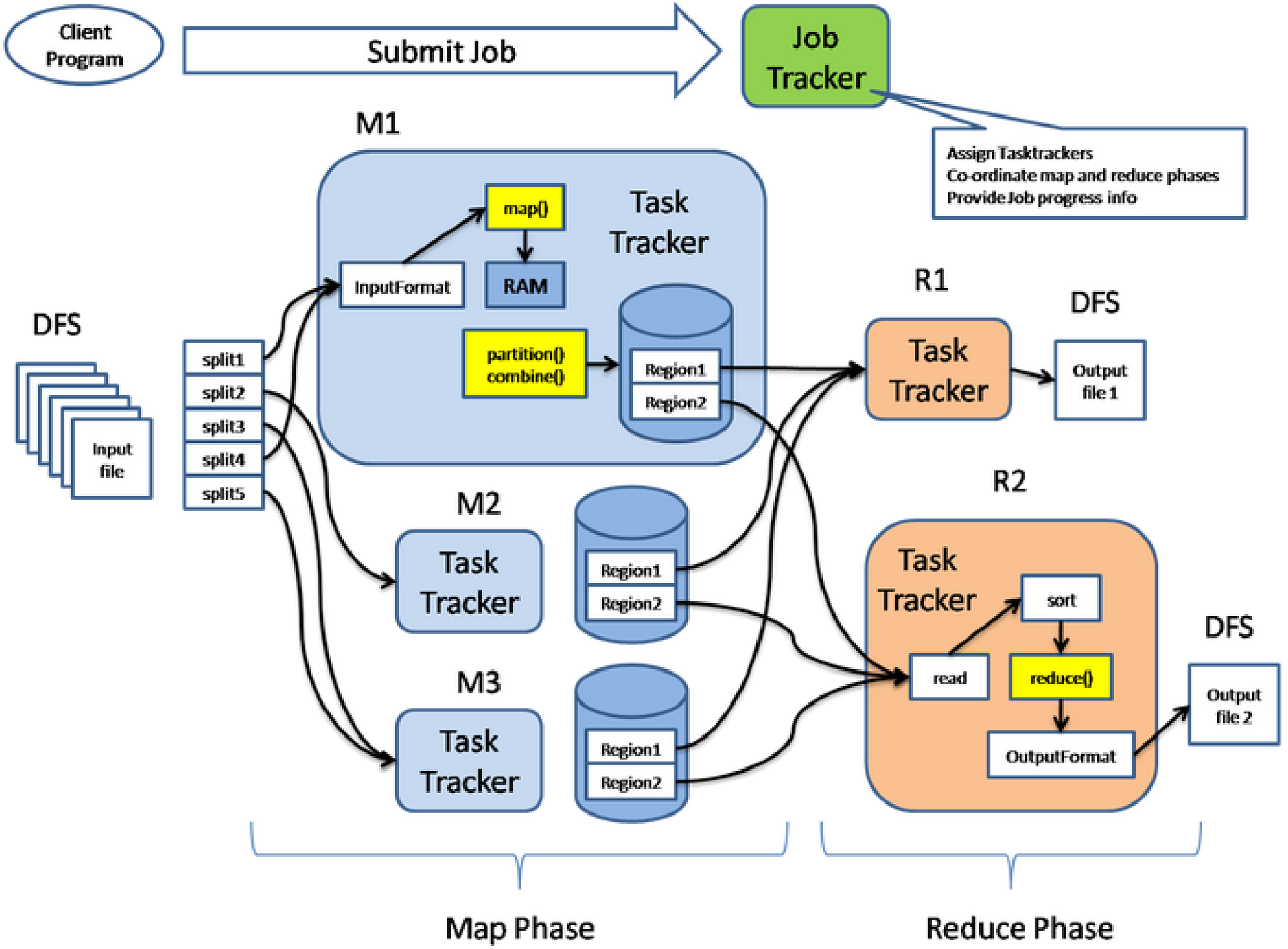
Reducers process the intermediate data from the maps into smaller tuples, that reduces the tasks, leading to the final output of the framework.



How MapReduce in Hadoop works

MapReduce architecture consists of various components.

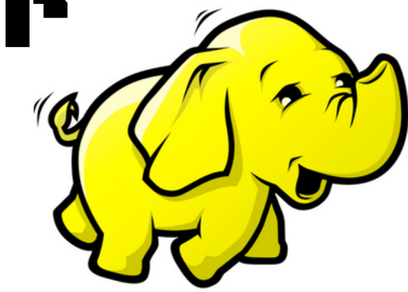



Client Program
Submit Job
Job Tracker
Assign Tasktrackers
Co-ordinate map and reduce phases
Provide Job progress info
M1
map()
Task Tracker
InputFormat
RAM
partition()
combine()
Region1
Region2
DFS
Input file
split1
split2
split3
split4
split5
R1
DFS
Task Tracker
Output file 1
M2
Task Tracker
Region1
Region2
R2
Task Tracker
sort
read
reduce()
OutputFormat
DFS
Output file 2
M3
Task Tracker
Region1
Region2
Map Phase
Reduce Phase

How MapReduce in Hadoop works

- Job: This is the actual work that needs to be executed or processed

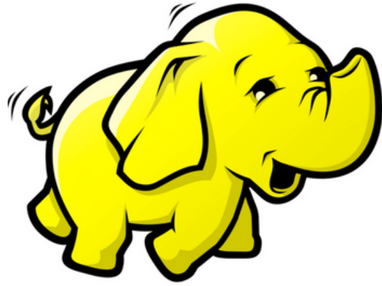
- Task: This is a piece of the actual work that needs to be executed or processed.



How MapReduce in Hadoop works

- Job Tracker: This tracker plays the role of scheduling jobs and tracking all jobs assigned to the task tracker.

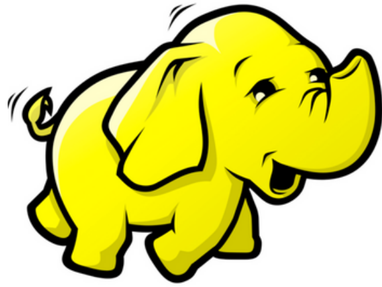
- Task Tracker: This tracker plays the role of tracking tasks and reporting the status of tasks to the job tracker.



How MapReduce in Hadoop works

- Input data: This is the data used to process in the mapping phase.

- Output data: This is the result of mapping and reducing.

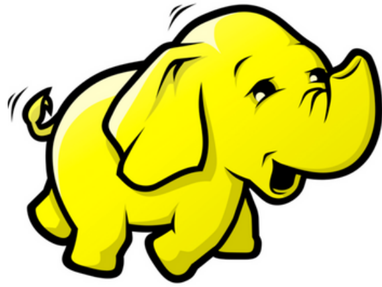


How MapReduce in Hadoop works

Client: This is a program that submits jobs to the MapReduce.

- Hadoop MapReduce Master: This plays the role of dividing jobs into job-parts.

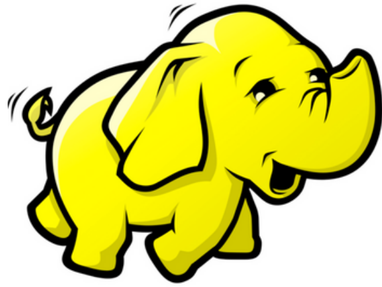
- Job-parts: These are sub-jobs that result from the division of the main job.



How MapReduce in Hadoop works

In the MapReduce architecture, clients submit jobs to the MapReduce Master.

This master will then sub-divide the job into equal sub-parts.

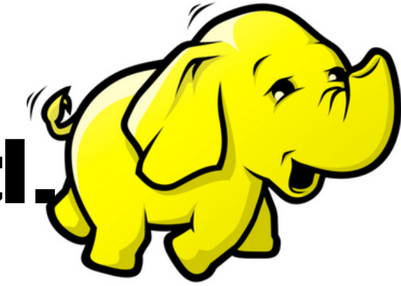


How MapReduce in Hadoop works

The job-parts will be used for the two main tasks in MapReduce: mapping and reducing.

The developer will write logic that satisfies the requirements of the organization or company.

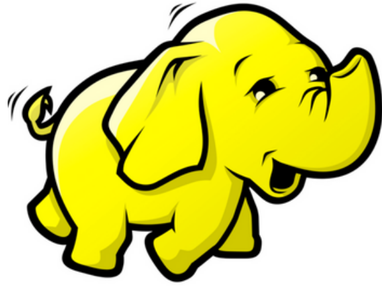
The input data will be split and mapped.



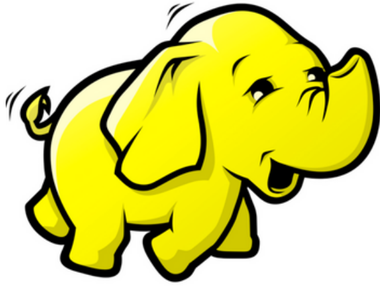
How MapReduce in Hadoop works

The intermediate data will then be sorted and merged.

The reducer that will generate a final output stored in the HDFS will process the resulting output.



How MapReduce in Hadoop works



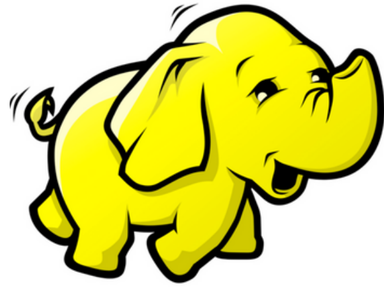
How job trackers and task trackers work

The job tracker acts as a master.

It ensures that we execute all jobs.

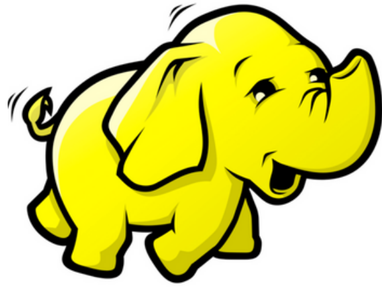
The job tracker schedules jobs that have been submitted by clients.

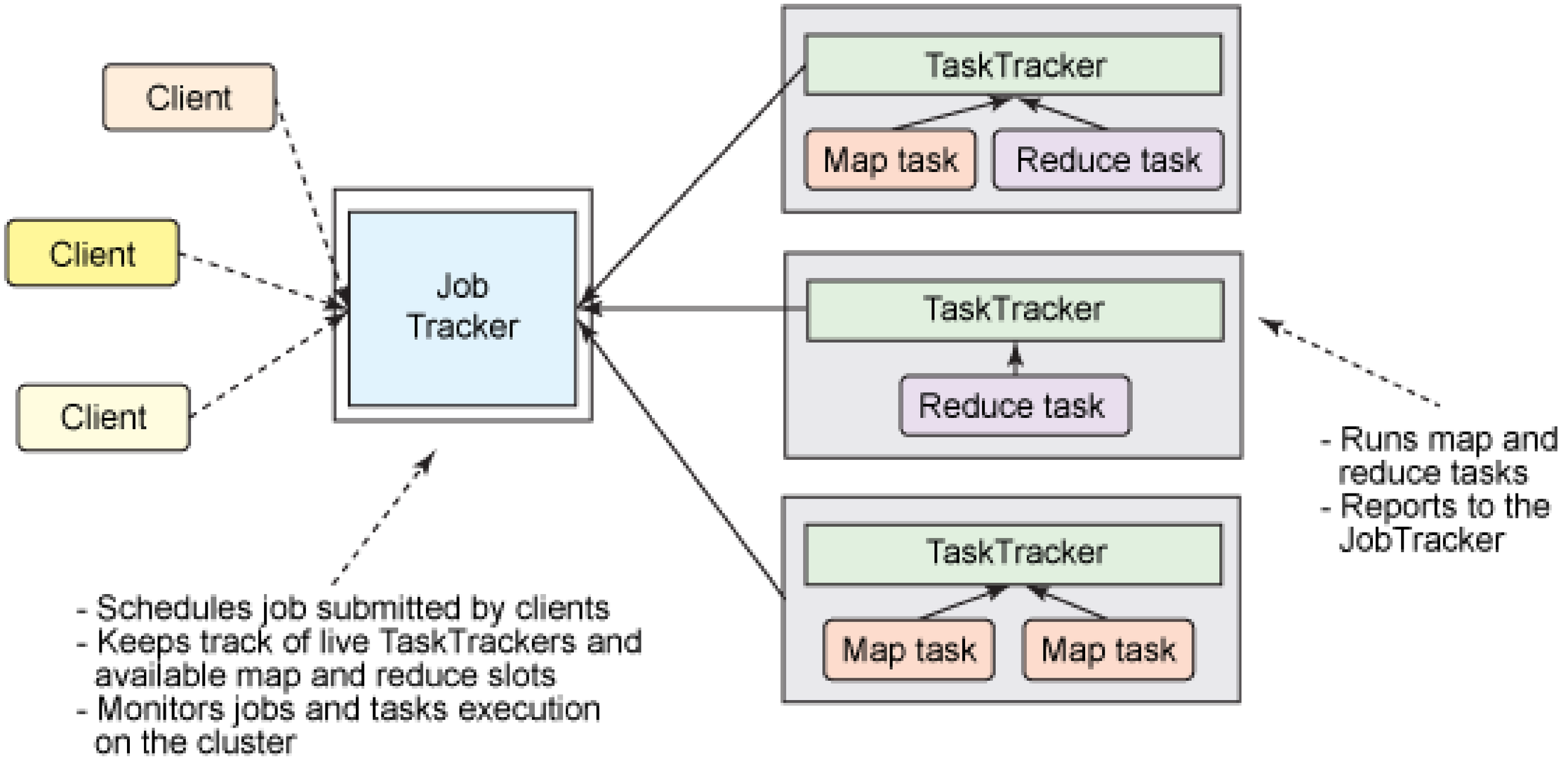
It will assign jobs to task trackers.



How job trackers and task trackers work

Task trackers report the status of each assigned job to the job tracker.

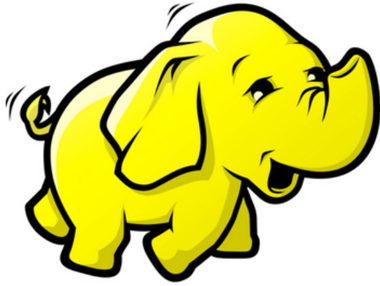



Client
Client
Client
Job Tracker
TaskTracker
Map task
Reduce task
TaskTracker
Reduce task
TaskTracker
Map task
Map task
- Schedules job submitted by clients
- Keeps track of live TaskTrackers and available map and reduce slots
- Monitors jobs and tasks execution on the cluster
- Runs map and reduce tasks
- Reports to the JobTracker

Phases of MapReduce

The MapReduce program is executed in three main phases: mapping, shuffling, and reducing.

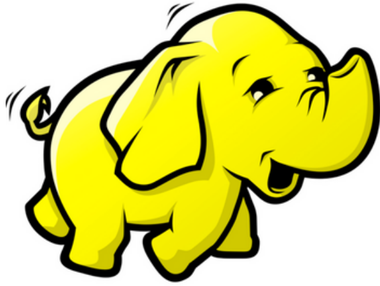
There is also an optional phase known as the combiner phase.



Mapping Phase

A dataset is split into equal units called chunks (input splits) in the splitting step.

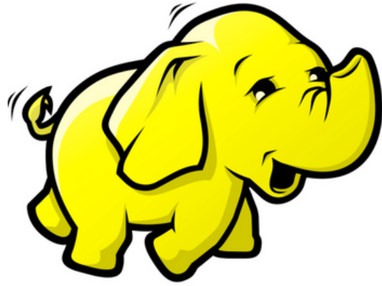
Hadoop consists of a RecordReader that uses TextInputFormat to transform input splits into key-value pairs.



Mapping Phase

The mapping step contains a coding logic that is applied to these data blocks.

In this step, the mapper processes the key-value pairs and produces an output of the same form (key-value pairs).

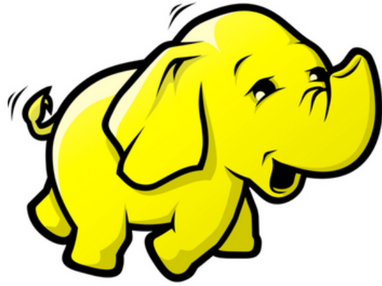


Shuffling phase

This is the second phase that takes place after the completion of the Mapping phase.

It consists of two main steps: sorting and merging.

In the sorting step, the key-value pairs are sorted using the keys.

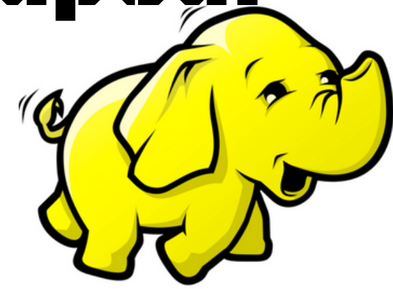


Shuffling phase

Merging ensures that key-value pairs are combined.

The shuffling phase facilitates the removal of duplicate values and the grouping of values.

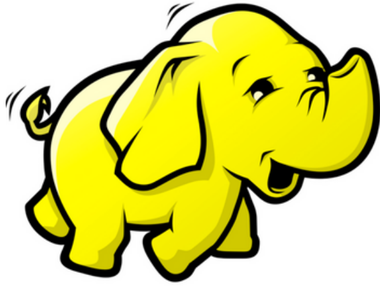
Different values with similar keys are grouped.



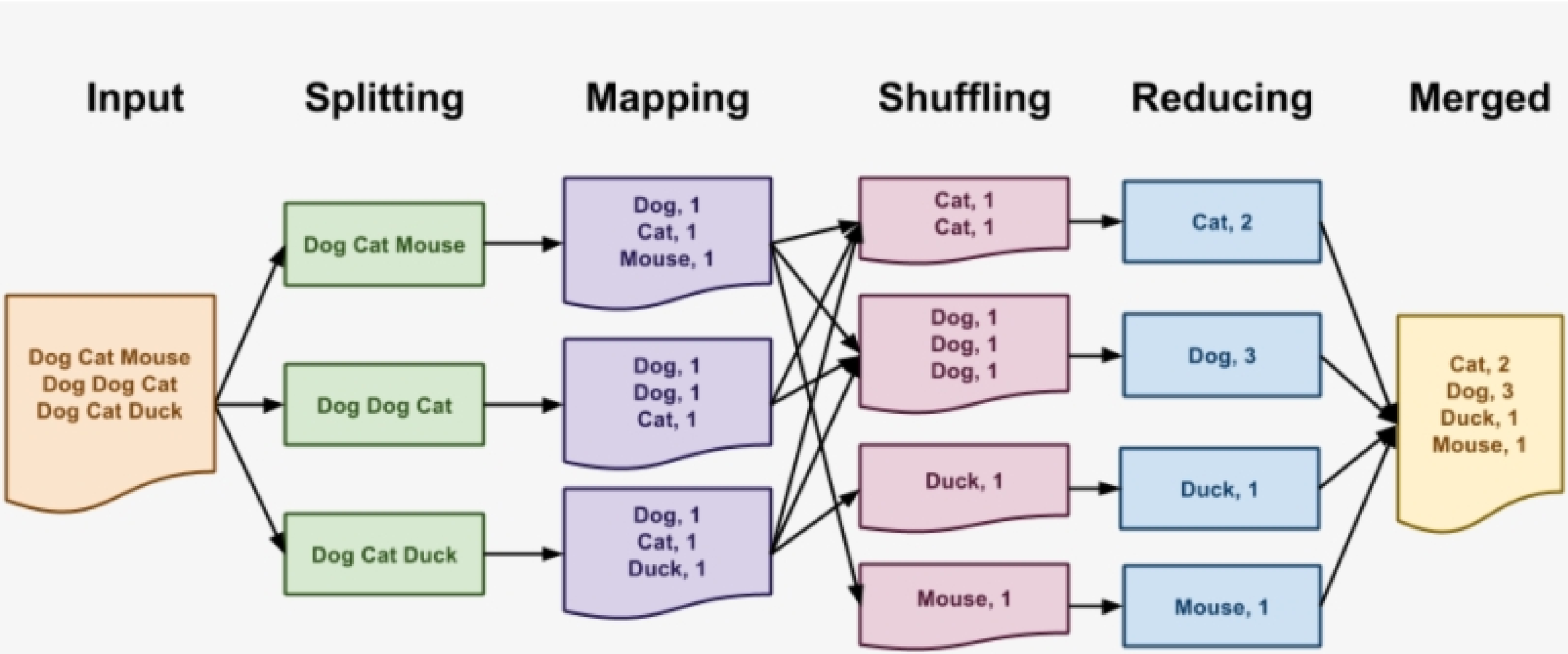
Reducer phase

In the reducer phase, the output of the shuffling phase is used as the input.

The reducer processes this input further to reduce the intermediate values into smaller values.



MR Word Count Process



On Summarizing