Hadoop:
Distributed Processing of Big Data

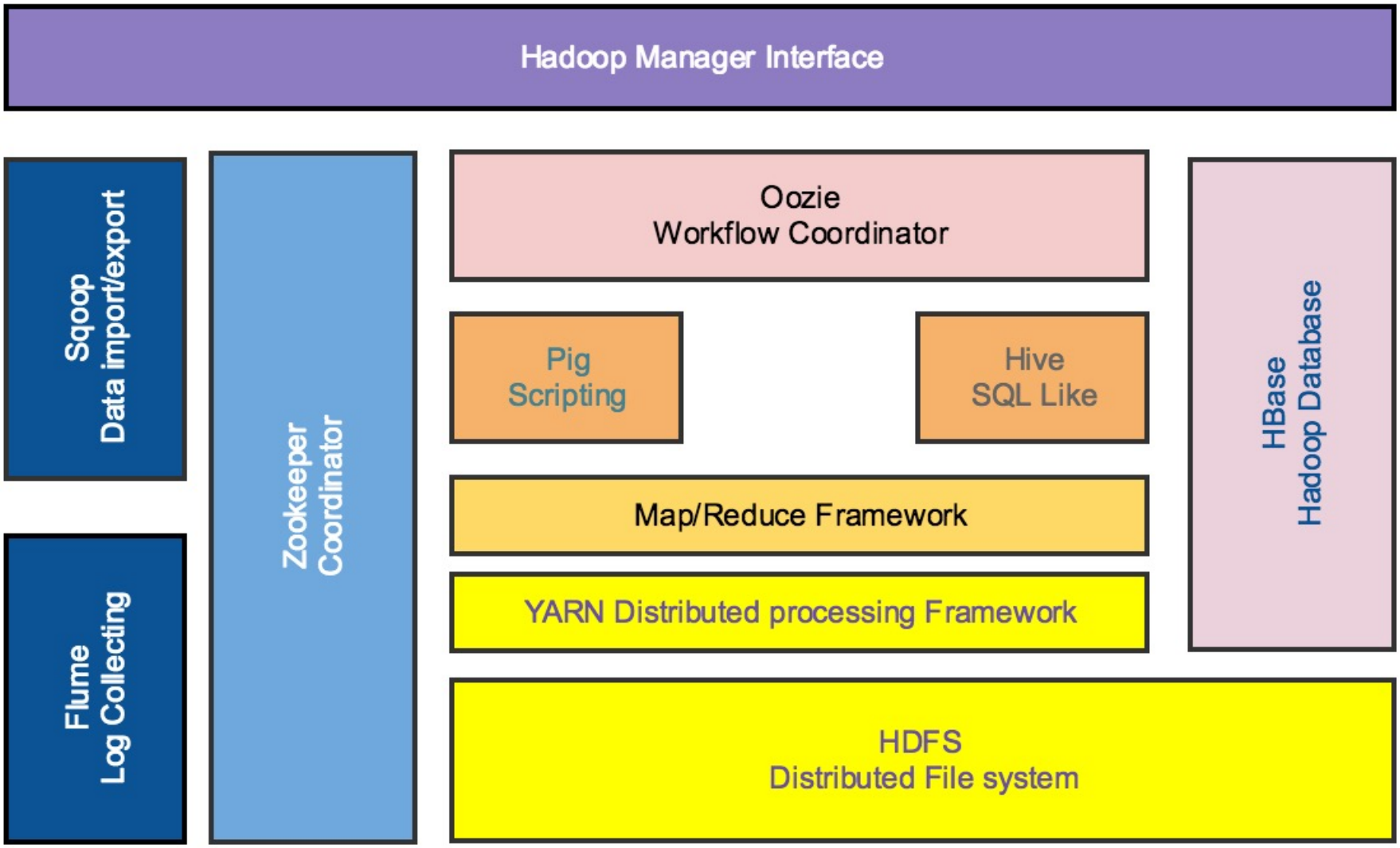
Lecture 02

Marilson Campos

UCSC Extension - 2022

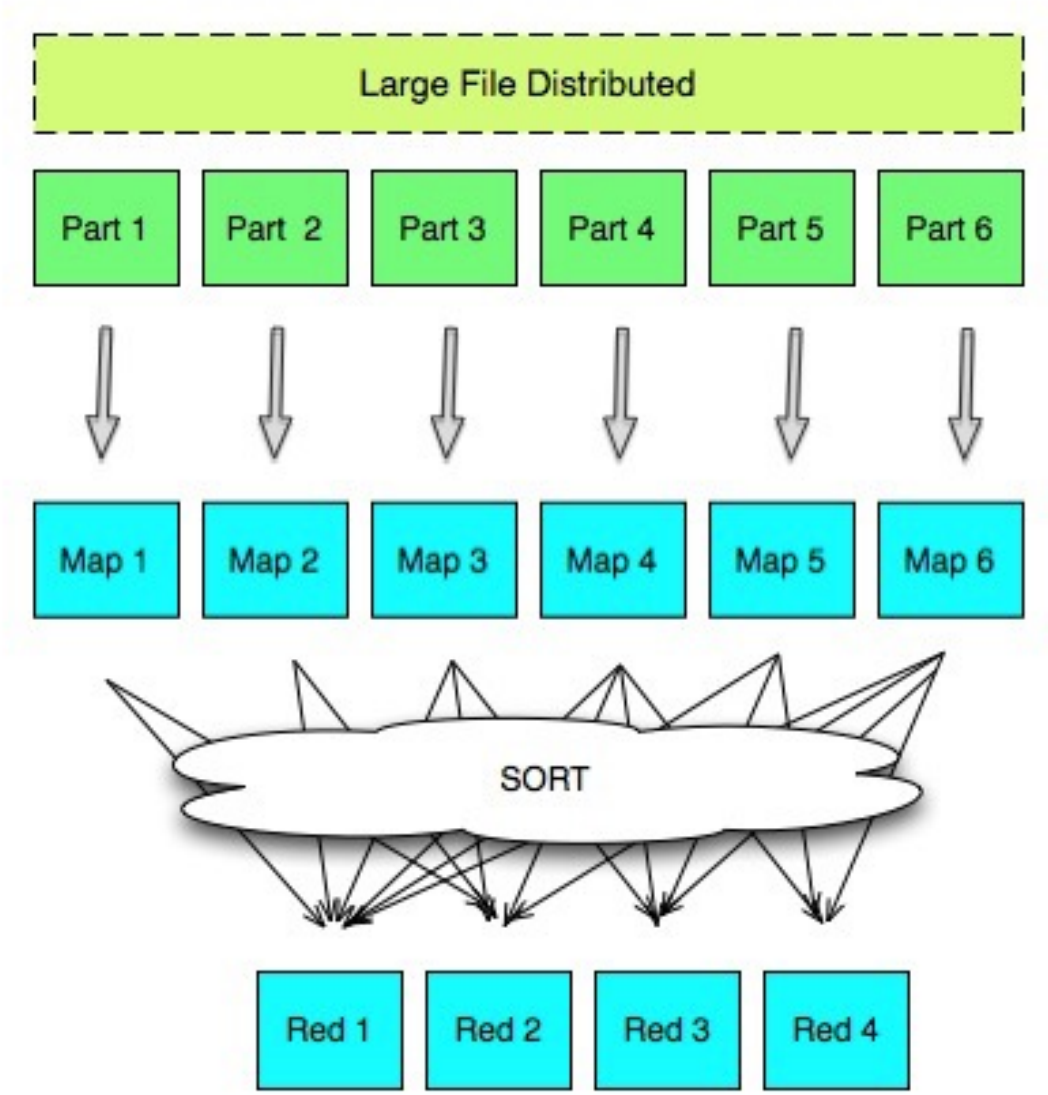
Hadoop is born

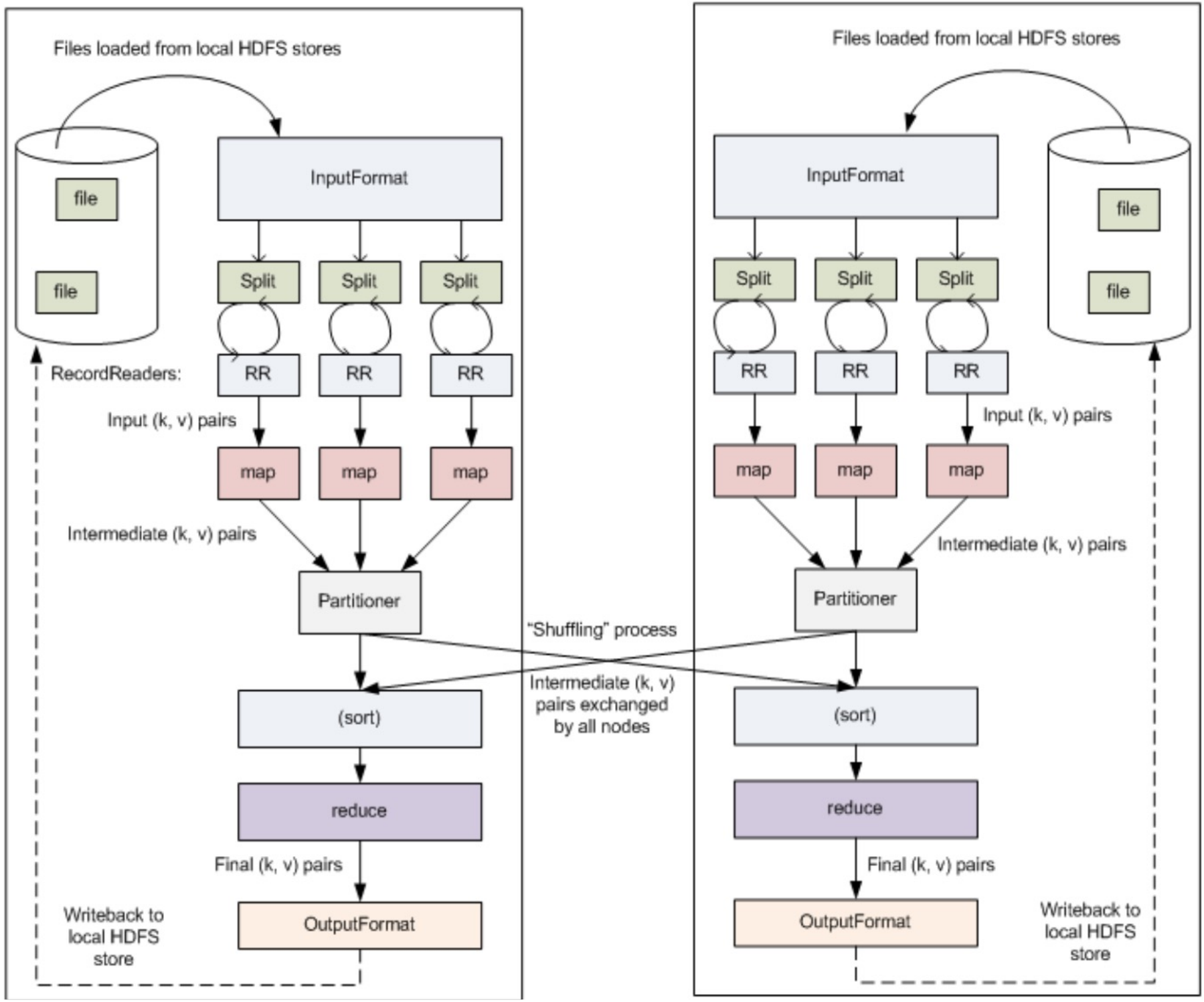
The Hadoop Stack



Map/Reduce

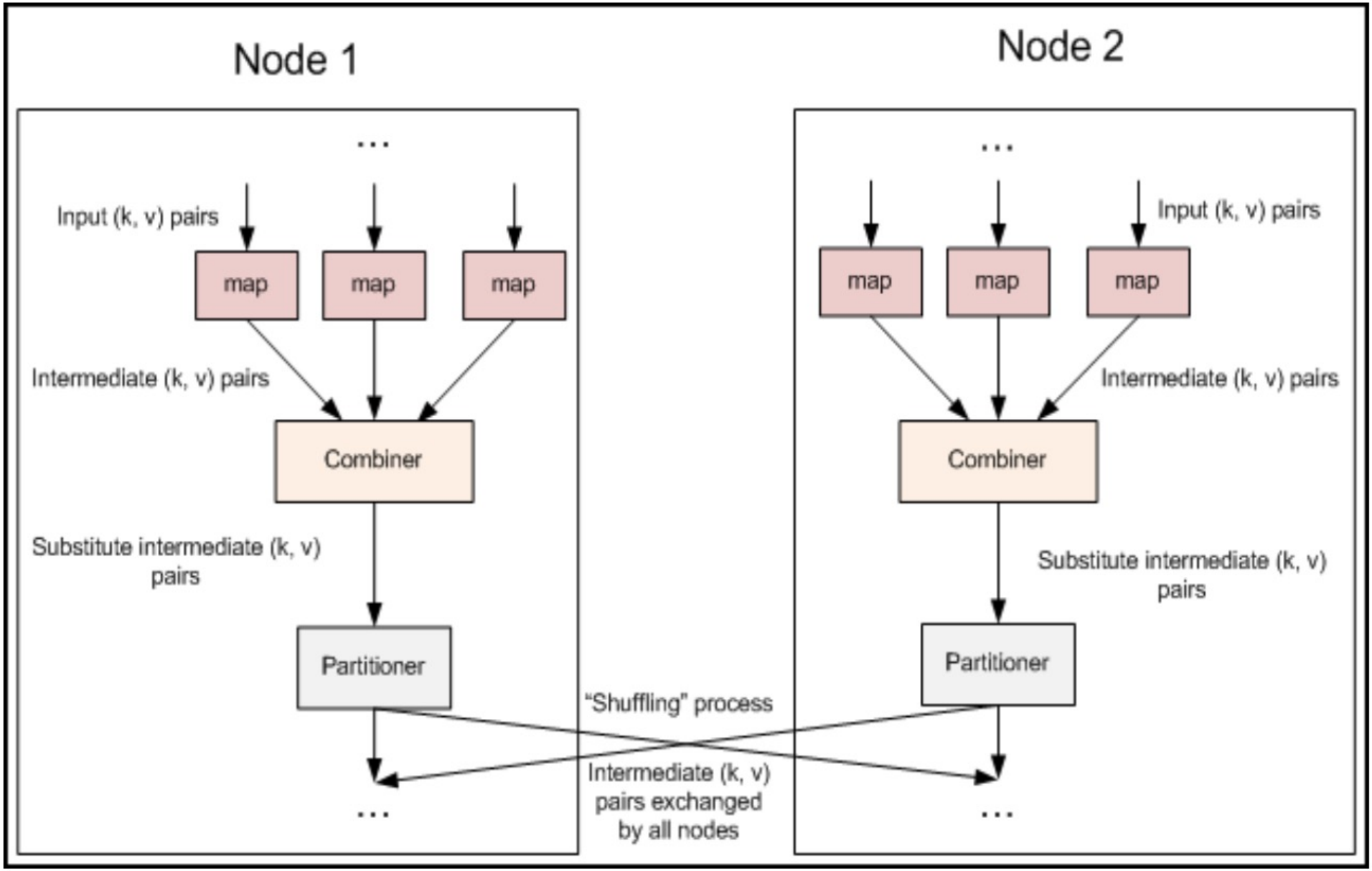
Map Reduce Review




Files loaded from local HDFS stores
file
file
InputFormat
Split
Split
Split
RecordReaders:
RR
RR
RR
Input (k, v) pairs
map
map
map
Intermediate (k, v) pairs
Partitioner
"Shuffling" process
Intermediate (k, v) pairs exchanged by all nodes
(sort)
reduce
Final (k, v) pairs
OutputFormat
Writeback to local HDFS store
Files loaded from local HDFS stores
InputFormat
file
file
Split
Split
Split
RR
RR
RR
Input (k, v) pairs
map
map
map
Intermediate (k, v) pairs
Partitioner
(sort)
reduce
Final (k, v) pairs
OutputFormat
Writeback to local HDFS store

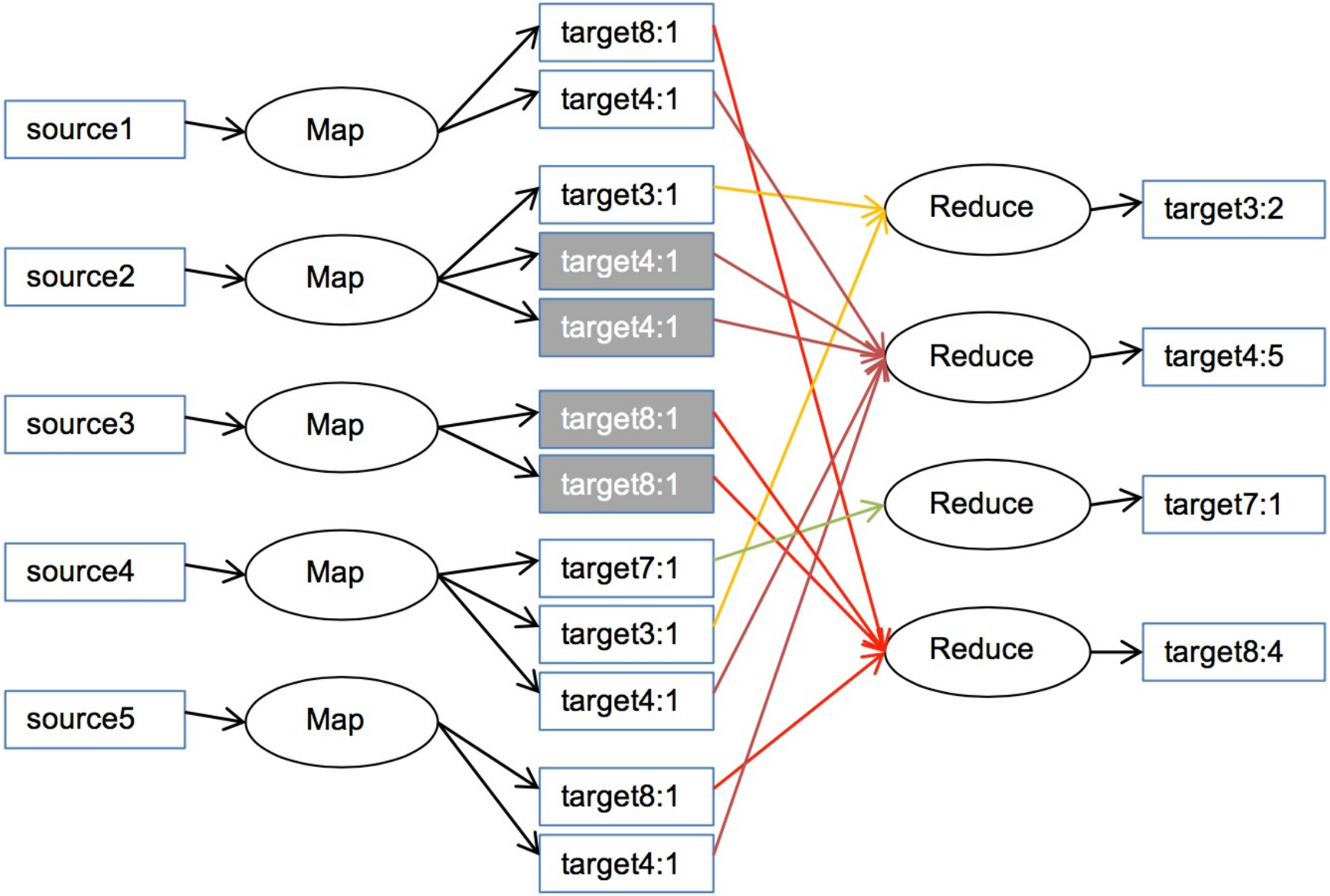
Map/Reduce

Partitioning



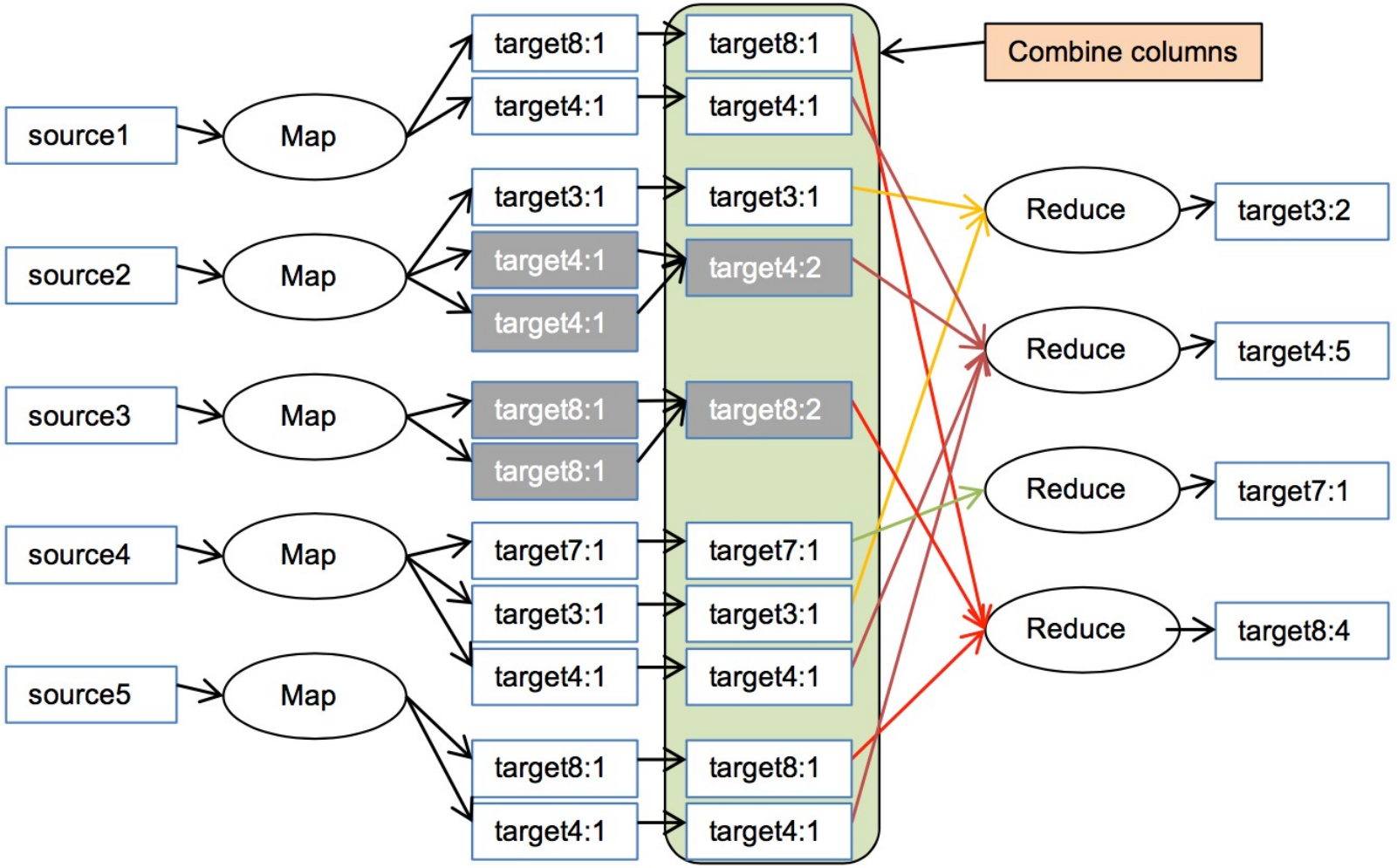
Map/Reduce

Partitioning / Reduce



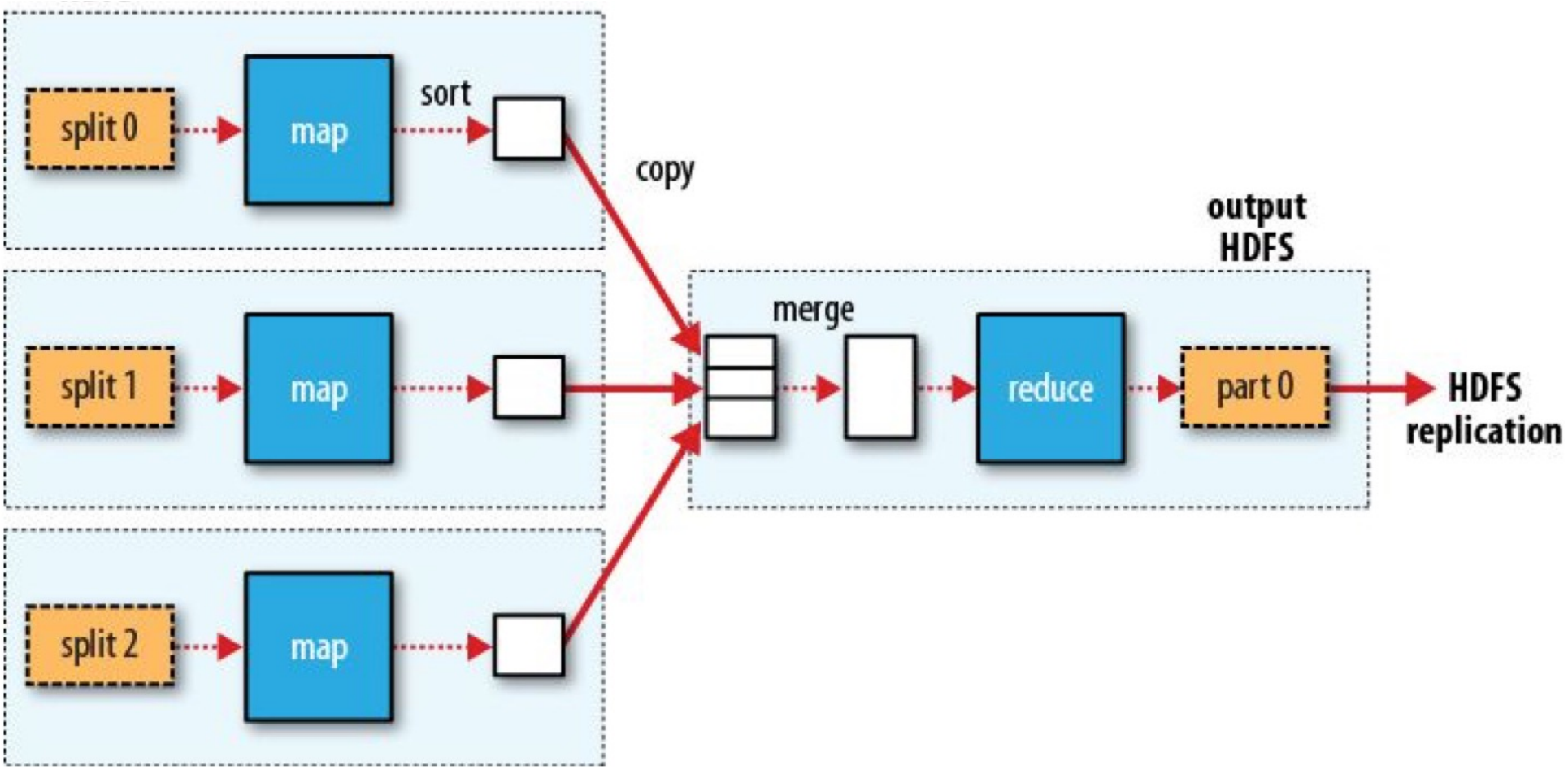
Map/Reduce

Partition with combiner



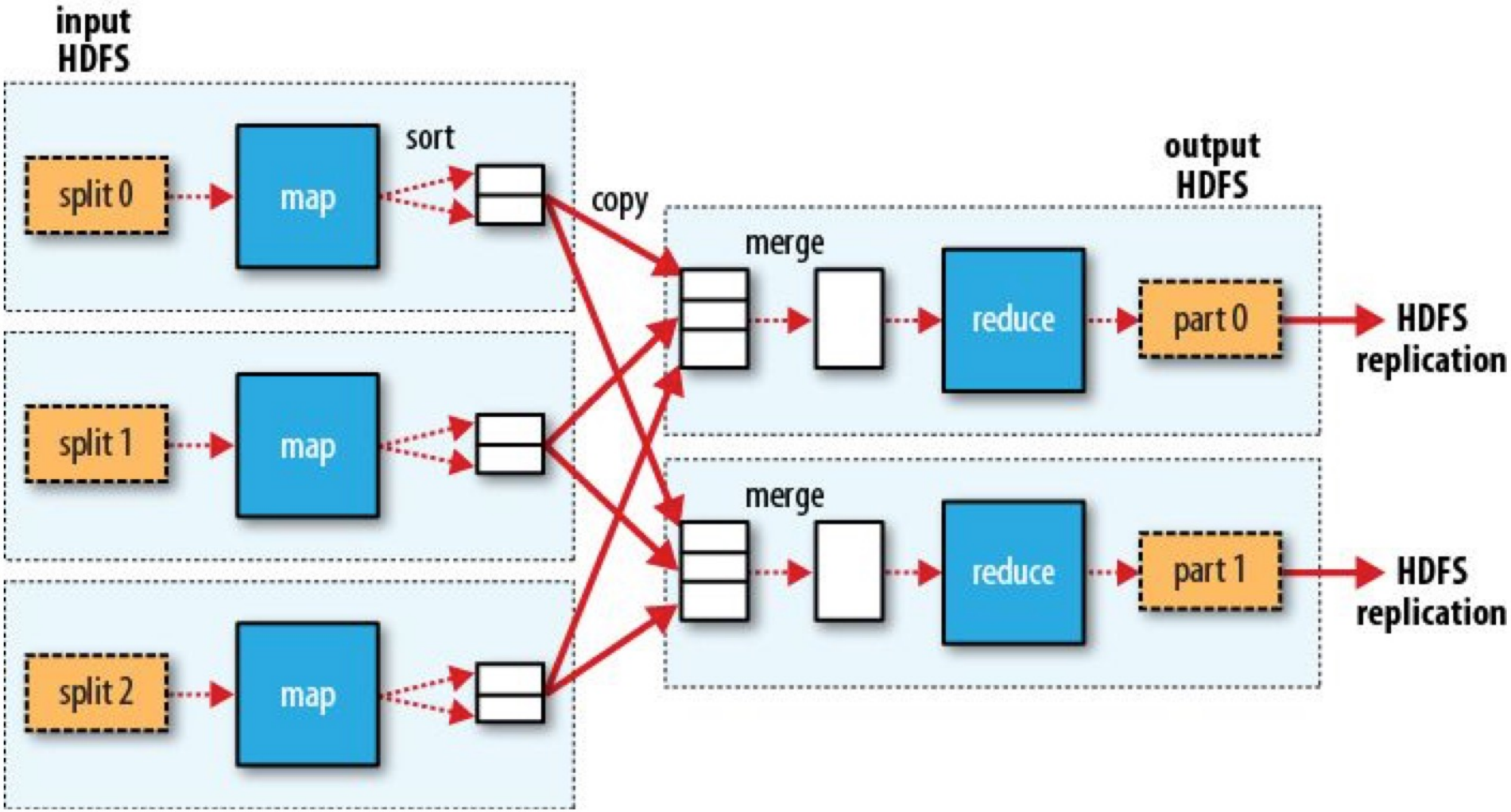
Hadoop Map/Reduce

MR Splits and dataflow 1 reducer

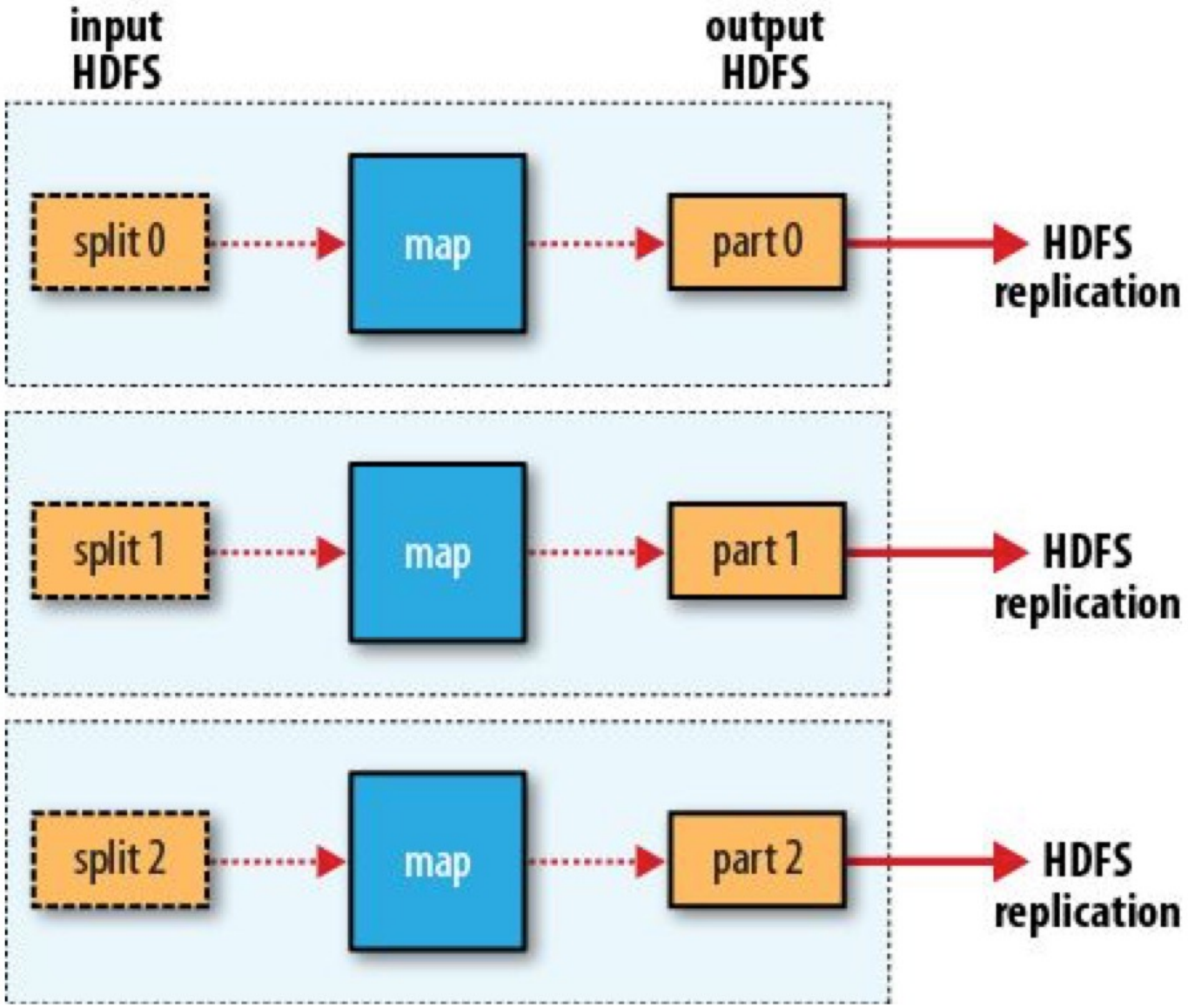


Hadoop Map/Reduce

MR Splits and dataflow multiple reducers



MR Splits and dataflow Zero reducers



Hadoop command line

- General command format for Apache Hadoop:

  <program> <cmd-group> [<options…]

Hadoop command line

Originally everything was under one program "hadoop"

As the codebase grew, the functionality has been moved into 3 programs.

"hadoop" – high level commands
"hdfs" – commands related with storage
"mapred" – commands related with M/R execution.

Hadoop command line

Hadoop – User command groups

* archive
* distcp
* fsck
* jar
* job
* queue
* version
* classpath

Hadoop command line

Hadoop archive

It allows you to store a whole sub-directory tree in a single file in HDFS.

> hadoop archive -archiveName <filename.har> -p <hdfs_dir> <dest_dir>

Hadoop command line

Hadoop distcp

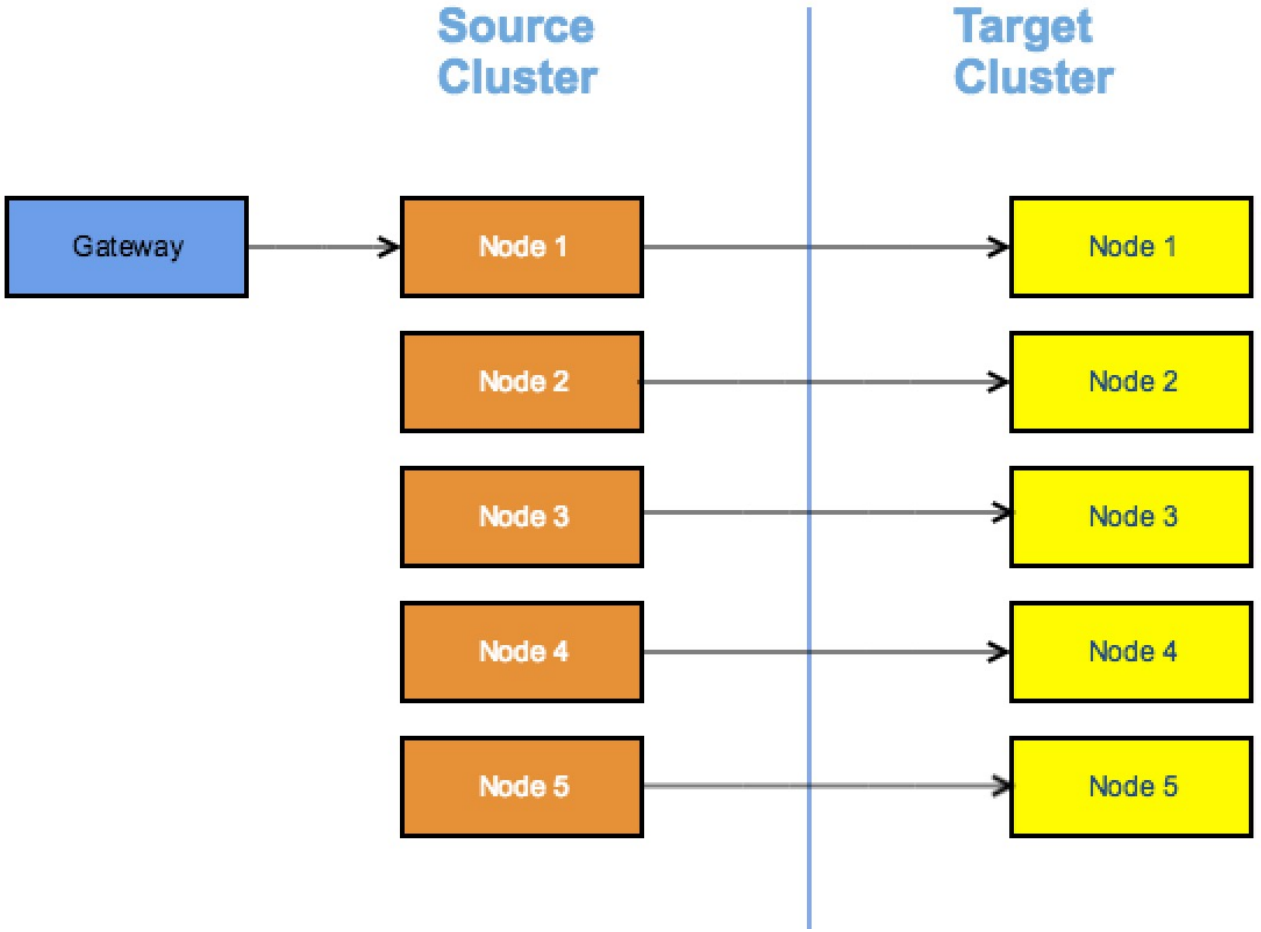
"Distributed copy" command allow parallel copy to different locations of cluster or to/from another cluster.

> hadoop distcp hdfs://<nn1>:8020/<source> hdfs://<nn2>:8020/<target>

Refer to:
https://hadoop.apache.org/docs/r2.6.5/hadoop-mapreduce-client/hadoop-mapreduce-client-core/DistCp.html

Hadoop command line

Hadoop distcp



Hadoop command line

Hadoop fsck

Deprecated: Use hdfs fsck

We will cover this command on the HDFS command section.

Hadoop command line

Hadoop jar

- Runs a jar file. Users can bundle their Map Reduce code in a jar file and execute it using this command.
- Used in streaming.

Usage: hadoop jar <jar> [mainClass] args...

Hadoop command line

Hadoop Job

Each Job is composed of tasks:

- Map tasks perform the map operation
- Reduce tasks perform the reduce operator

Hadoop give each task 3 chances to succeed before if fails the jobs.

Speculative execution:

If one task seems to be slow, Hadoop launches another task for the same work. It will pick the results of the one that finishes first and ignores the other(s).

Hadoop command line

Hadoop job

Deprecated. Use mapred job

```
Usage:
   mapred job
      -submit<job-file>
      -status<job-id>
      -counter<job-id><group-name><counter-name>
      -kill<job-id>
      -events<job-id><from-event-#><#-of-events>
      -history [all]
      -list
      -kill-task <task-id> -> Does not count against failed attempts.
      -fail-task <task-id>
      -set-priority <job-id> [VERY_HIGH, HIGH, NORMAL, LOW,
VERY_LOW]
```

Hadoop command line

Hadoop Job Queues

- Define resource allocations for groups of jobs.
- There are several types of queues.

We will cover them later in this class.

Example: I can define the following queues:

prod_etl -> “production etl” with 60% of resources.
dev -> “development” with 15% of resources.
default -> with 25% of resources.

Hadoop command line

Hadoop queue

Deprecated. Use mapred queue

Usage: mapred queue
-list
-info <job-queue-name>
-showJobs
-showacls

Hadoop command line

Hadoop - Administration

- daemonlog

Hadoop command line

What is a daemon

"A *daemon* is a type of program on Unix-like operating systems that runs unobtrusively in the background, rather than under the direct control of a user, waiting to be activated by the occurrence of a specific event or condition."

Hadoop command line

List the daemons in Hadoop

Use:

> sudo jps

‘sudo’ Unix command allows you to run another command as admin.

Hadoop command line

Hadoop daemonlog

Get/Set the log level for each daemon.

Usage:

hadoop daemonlog -getlevel <host:port> <name>

hadoop daemonlog -setlevel <host:port> <name> <level>

LEVEL = (ERROR, WARN, INFO)

HDFS command line

HDFS – User command groups

- dfs

- fsck

HDFS DFS command line

HDFS DFS options

cat	cp	getmerge	put	tail
chgrp	du	ls	rm	test
chmod	dus	lsr	**rmr (****)**	text
chown	expunge	mkdir	setfacl	touchz
copyFromLocal	get	moveFromLocal	setfattr	
CopyToLocal	getfactl	moveToLocal	setrep	
count	getfattr	mv	stat	

Hadoop command line

HDFS Access Rights

Control who can operate in files and folders.

```
                is dir?      owner     group    other
drwxrwxrwx -> [d]          [rwx]     [rwx]    [rwx]
```

Flags: ‘r’ = read, ‘w’ = write, ‘x’ = execute
Selector: ‘u’ = owner, ‘g’ = group, ‘o’ = other

o+wx = allow others to write and execute

u-x = disallow user to execute

```
drwxrwxrwx   - hdfs     supergroup          0 2017-07-19 05:34 /benchmarks
drwxr-xr-x   - cloudera supergroup          0 2018-01-10 17:27 /data
```

HDFS DFS – access rights

Access rights commands = [chgrp, chmod, chown]

chmod -> change access rights for user,group,other

Usage: hdfs dfs -chmod [-R] MODE URI

Ex: hdfs dfs –chmod a+w /tmp

Everybody can write to /tmp

HDFS DFS – access rights

Access rights commands = [chgrp, chmod, chown]

chown -> change the user that “owns” of the dir/file.

Usage: hdfs dfs -chown [-R] USER URI

Ex: hdfs dfs –chown joe /tmp/joe

HDFS DFS – access rights

Access rights commands = [chgrp, chmod, chown]

chgrp -> changes user-group that a path belongs to.

Usage: hdfs dfs -chgrp [-R] GROUP URI

Ex: hdfs dfs -chgrp operators /tmp/dir1

('operators' in this example is a group name)

HDFS DFS – Copying from cluster to local

Copying = [CopyToLocal, get]

CopyToLocal /get -> Get files our of the cluster

Usage:

hdfs dfs -copyToLocal URI <localdir>.
hdfs dfs -get URI <localdir>

HDFS DFS – Copying a cluster dir to local

Copying = [getmerge]

getmerge -> get data from a hdfs directory into a local file

Usage:

hdfs dfs –getmerge <src> <local-file>

HDFS DFS – copy from local to cluster

Copying = [put, copyFromLocal]

put, copyFromLocal -> copy files from local disk and loads into the cluster

Usage:

hdfs dfs –put <src> <hdfs-file>

You can use the –f option to force.

HDFS DFS – view & create directories

View/create = [ls, lsr, mkdir]

ls -> list files from a directories.

lsr -> list files recursively.

mkdir -> creates an HDFS directory.

HDFS DFS – delete data

delete = [rm, rmr, expunge]

rm -> Removes files from a folder.

rmr -> removes everything below the directory.

(Danger **)**

expunge -> delete the data on trash location.

HDFS DFS – get information

Get info = [du, dus]

du -> Displays sizes of files and directories contained in the given directory

```
[cloudera@quickstart shared]$ hdfs dfs -du /
0           0           /benchmarks
116497840   116497840   /data
7352        7352        /hbase
0           0           /solr
7149896     7266796     /tmp
858119738   858119738   /user
3337674     3337674     /var
[cloudera@quickstart shared]$
```

HDFS DFS – get information

Get info = [du, dus]

First column => file size

Second column => Disk space taken

```
[hadoop]$ hdfs dfs -du /
229867694088 6896030822658 /output
21297905593 63893716779 /tmp
6072184915396 1821655409976 /user
```

HDFS DFS – get information

Get info = [du, dus]

dus -> display results above in a summary.

```
[cloudera@quickstart shared]$ hdfs dfs -dus /
dus: DEPRECATED: Please use 'du -s' instead.
985112500  985229400  /
```

HDFS DFS – replication

Replication = [setrep]

setrep -> Sets the replication factor of a location in the cluster.

The -w flag makes the command to wait for the replication to complete.

If the value of the new factor is bigger than the current replication value. This will only apply to new data written after the command.

HDFS DFS – get info

Get info = [stat, test]

stat -> prints information about a file in the cluster.

test -> perform tests in hdfs files.

-e = exists, -z = zero size, -d=directory

returns zero if true.

HDFS DFS – see contents

Contents= [tail, text]

tail -> prints the end of the file.

text -> view inside of zip files.

HDFS DFS – marker files

markers= [touchz]

touchz -> Create a file of zero length.

Usage: hdfs dfs -touchz pathname

We can use this command to create “marker files” that will be explained later in the class.

HDFS balancer – Spread the data

hdfs hdfs balancer -> Makes sure that the files are spread across all nodes.

This is useful when you ‘add’ new nodes to the cluster as the new ’nodes’ have no data.

HDFS FSCK – File System check

hdfs fsck -> Verifies the state of the distributed file system.

Usage:

```
hdfs fsck [GENERIC_OPTIONS] <path>
  [-list-corruptfileblocks |
  [-move | -delete | -openforwrite]
```

Q & A