

MAPREDUCE TUTORIAL

Hands-on Session

by Suchitra Jayaprakash

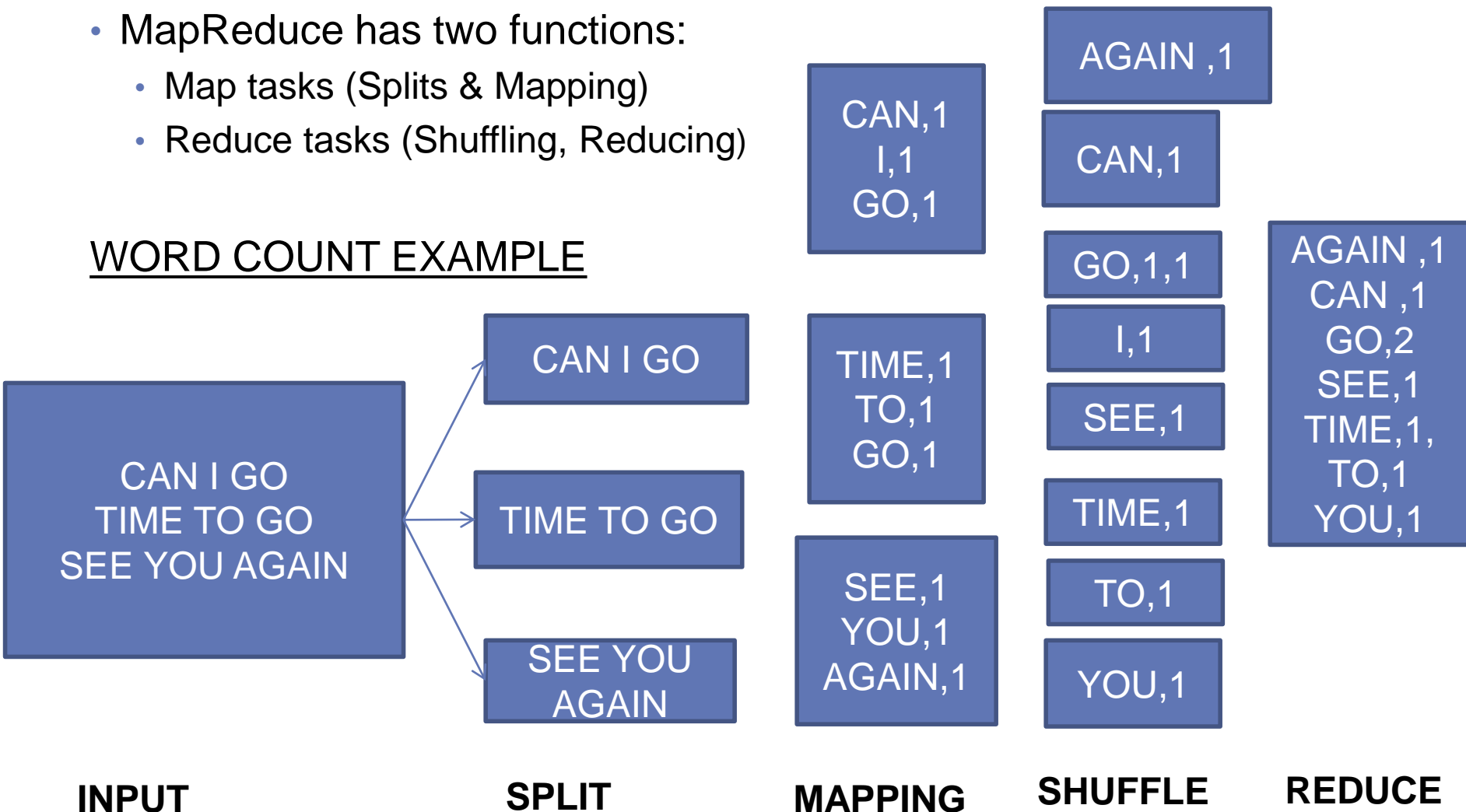
MAPREDUCE

- MapReduce is a programming model. It is a specialization of the *split-apply-combine* strategy for data analysis.
- Hadoop's MapReduce implementation is based on Google's paper : MAPREDUCE: SIMPLIFIED DATA PROCESSING ON LARGE CLUSTERS" by Jeffery Dean and Sanjay Ghemawat in 2004.
- Google's proprietary MapReduce system ran on the Google File System (GFS).
- MapReduce divides work into smaller tasks. Each task is executed parallelly on cluster server to create individual output. Individual output is processed & consolidated to create final output.

How MapReduce Works?

- MapReduce has two functions:
 - Map tasks (Splits & Mapping)
 - Reduce tasks (Shuffling, Reducing)

WORD COUNT EXAMPLE



MAPREDUCE - Run WordCount

- Start cloudera quick start

```
docker run --hostname=quickstart.cloudera --privileged=true -t -i --  
publish-all=true -p 8888:8888 -p 8080:80 -p 50070:50070 -p 8088:8088 -p  
50075:50075 -p 8032:8032 -p 8042:8042 -p 19888:19888  
cloudera/quickstart /usr/bin/docker-quickstart
```

Port	Purpose
8088	Yarn (MRv2) - job tracker
8032	ResourceManager
50070	Name node web interface
50075	Data node
8042	NodeManager
19888	JobHistory Server

MAPREDUCE - Run WordCount

- Copy text file to HDFS.

```
hadoop fs -mkdir DATA
```

```
docker cp c:/tmp/sample1.txt <containerid>:/tmp/sample1.txt
```

```
docker cp c:/tmp/sample2.txt <containerid>:/tmp/sample2.txt
```

```
hadoop fs -copyFromLocal /tmp/sample1.txt DATA/sample1.txt
```

```
hadoop fs -copyFromLocal /tmp/sample2.txt DATA/sample2.txt
```

- Copy WordCount java to Docker .

```
docker cp c:/tmp/WordCount.java
```

```
<containerid>:/tmp/WordCount.java
```

- https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount1_source.html

MAPREDUCE - Run WordCount

- Compile java class & create jar

mkdir -p build

**javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/*
/tmp/WordCount.java -d build -Xlint**

```
Using CATALINA_PID: /var/run/solr/solr.pid
Started Impala Catalog Server (catalogd) : [ OK ]
Started Impala Server (impalad): [ OK ]
[root@quickstart /]# hadoop fs -mkdir DATA
[root@quickstart /]# hadoop fs -copyFromLocal /tmp/sample1.txt DATA/sample1.txt
[root@quickstart /]# hadoop fs -copyFromLocal /tmp/sample2.txt DATA/sample2.txt
[root@quickstart /]# mkdir -p build
[root@quickstart /]# javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* /tm
p/WordCount1.java -d build -Xlint
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb-api.jar": no su
ch file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/activation.jar": no
such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jsr173_1.0_api.jar":
no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb1-impl.jar": no
such file or directory
4 warnings
[root@quickstart /]#
```

jar -cvf wordcount.jar -C build/ .

```
[root@quickstart /]# jar -cvf wordcount.jar -C build/ .
added manifest
adding: WordCount1$Map.class(in = 2192) (out= 982)(deflated 55%)
adding: WordCount1$Reduce.class(in = 1630) (out= 687)(deflated 57%)
adding: WordCount1.class(in = 1959) (out= 986)(deflated 49%)
[root@quickstart /]#
```

MAPREDUCE - Run WordCount

- Execute Mapreduce Word Count job .

hadoop jar wordcount.jar WordCount DATA DATA2

```
[root@quickstart /]# hadoop jar wordcount.jar WordCount1 DATA DATA2
19/12/25 14:37:10 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/12/25 14:37:11 INFO input.FileInputFormat: Total input paths to process : 2
19/12/25 14:37:12 INFO mapreduce.JobSubmitter: number of splits:2
19/12/25 14:37:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1577284264785_0001
19/12/25 14:37:13 INFO impl.YarnClientImpl: Submitted application application_1577284264785_0001
19/12/25 14:37:13 INFO mapreduce.Job: The url to track the job: http://quickstart.t.cloudera:8088/proxy/application_1577284264785_0001/
19/12/25 14:37:13 INFO mapreduce.Job: Running job: job_1577284264785_0001
19/12/25 14:37:26 INFO mapreduce.Job: Job job_1577284264785_0001 running in uber mode : false
19/12/25 14:37:26 INFO mapreduce.Job: map 0% reduce 0%
19/12/25 14:37:43 INFO mapreduce.Job: map 100% reduce 0%
19/12/25 14:37:52 INFO mapreduce.Job: map 100% reduce 100%
19/12/25 14:37:53 INFO mapreduce.Job: Job job_1577284264785_0001 completed successfully
19/12/25 14:37:53 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=96
  FILE: Number of bytes written=340703
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=286
  HDFS: Number of bytes written=39
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
Job Counters
  Number of write operations=2
  Launched reduce tasks=1
  Total time spent by all maps in occupied slots (ms)=27727
  Total time spent by all map tasks (ms)=27727slots (ms)=6612
  Total vcore-seconds taken by all map tasks=27727
  Total megabyte-seconds taken by all map tasks=28392448
Map-Reduce Framework
  Map output records=8
  Map output materialized bytes=102
```

```

  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
Job Counters
  Number of write operations=2
  Launched reduce tasks=1
  Total time spent by all maps in occupied slots (ms)=27727
  Total time spent by all map tasks (ms)=27727slots (ms)=6612
  Total vcore-seconds taken by all map tasks=27727
  Total megabyte-seconds taken by all map tasks=28392448
Map-Reduce Framework
  Map output records=8
  Map output materialized bytes=102
  Input split bytes=246
  Combine input records=0
  Combine output records=0
  Reduce input groups=5
  Reduce shuffle bytes=102
  Reduce input records=8
  Reduce output records=5
  Spilled Records=16
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=278
  CPU time spent (ms)=3250
  Physical memory (bytes) snapshot=720171008
  Virtual memory (bytes) snapshot=4096000192
  Total committed heap usage (bytes)=624427008
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=40
File Output Format Counters
  Bytes Written=39
[root@quickstart /]#
```

MAPREDUCE - Run WordCount


- Open Yarn web page in browser

<http://192.168.99.100:8088/cluster>

<http://localhost:8088/cluster>

← → ↻ 🏠 ⓘ Not secure | 192.168.99.100:8088/cluster ☆

Apps 🔄



All Applications

▼ Cluster

[About](#)
[Nodes](#)
[Applications](#)
[NEW](#)
[NEW SAVING](#)
[SUBMITTED](#)
[ACCEPTED](#)
[RUNNING](#)
[FINISHED](#)
[FAILED](#)
[KILLED](#)
[Scheduler](#)

► Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes
1	0	0	1	0	0 B	8 GB	0 B	0	8	0	1	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used
0	0	0	1	0	0	0	0 B	0 B	0 B	0

Show 20 ▼ entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCo
application_1577284264785_0001	root	wordcount	MAPREDUCE	root.root	Wed Dec 25 20:07:12	Wed Dec 25 20:07:51	FINISHED	SUCCEEDED	N/A	N/A

MAPREDUCE - Run WordCount

- Click on JOB ID

Cluster

[About](#)
[Nodes](#)
[Applications](#)
NEW
NEW SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
[Scheduler](#)

Tools

Application Overview

User: root

Name: wordcount

Application Type: MAPREDUCE

Application Tags:

State: FINISHED

FinalStatus: SUCCEEDED

Started: Wed Dec 25 14:37:12 +0000 2019

Elapsed: 38sec

Tracking URL: [History](#)

Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>

Total Number of Non-AM Containers Preempted: 0

Total Number of AM Containers Preempted: 0

Resource Preempted from Current Attempt: <memory:0, vCores:0>

Number of Non-AM Containers Preempted from Current Attempt: 0

Aggregate Resource Allocation: 137655 MB-seconds, 86 vcore-seconds

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Wed Dec 25 14:37:12 +0000 2019	quickstart.cloudera:8042	logs

- To View Node Manager
- <http://localhost:8042/node/node>

MAPREDUCE - Run WordCount

- View Job history

<http://localhost:19888/jobhistory>

▸ Application

▾ Job

Overview

Counters

Configuration

Map tasks

Reduce tasks

▸ Tools

Job Name:

wordcount

User Name:

root

Queue:

root.root

State:

SUCCEEDED

Uberized:

false

Submitted:

Wed Dec 25 14:37:12 UTC 2019

Started:

Wed Dec 25 14:37:25 UTC 2019

Finished:

Wed Dec 25 14:37:51 UTC 2019

Elapsed:

25sec

Diagnostics:

Average Map Time

13sec

Average Shuffle Time

5sec

Average Merge Time

0sec

Average Reduce Time

0sec

ApplicationMaster

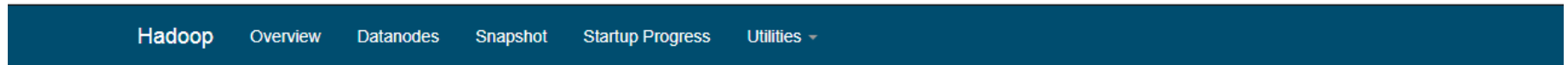
Attempt Number	Start Time	Node	Logs
1	Wed Dec 25 14:37:18 UTC 2019	quickstart.cloudera:8042	logs

Task Type	Total	Complete	
Map	2	2	
Reduce	1	1	
Attempt Type	Failed	Killed	Successful
Maps	0	0	2
Reduces	0	0	1

MAPREDUCE - Run WordCount

- Open DATA2 folder in Namenode to view mapreduce output

<http://localhost:50070/explorer.html#/user/root/DATA2>



Browse Directory

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	0 B	Wed Dec 25 20:07:51 +0530 2019	1	128 MB	_SUCCESS
-rw-r--r--	root	supergroup	39 B	Wed Dec 25 20:07:50 +0530 2019	1	128 MB	part-r-00000

MAPREDUCE - Run WordCount

- Input file

This is sample1 text

This is sample2 text

- Final Output

1	This	2
2	is	2
3	sample1	1
4	sample2	1
5	text	2
6		

MRJOB – Python package

Try writing MapReduce jobs in Python

- <https://mrjob.readthedocs.io/en/latest/>

THANK YOU