Solution – Exercise 01

Hadoop, HDFS, YARN, MapReduce Example

Solution

Prerequisites:

- install Ubuntu 18.04
- Install Java JDK 1.8.0
- Create Hadoop user
- Install and Setup SSH (*public/private key authentication, authorized_keys, …*)
- Install and Configure Hadoop 3.1.3 (*pseudo-distributed mode*)
- Start HDFS and YARN
- Clone Git Repo:

```
git clone https://github.com/marcelmittelstaedt/BigData.git
```

Solution

Exercise 2:

1. Copy sample file from GIT repo to HDFS user directory:

```
hadoop fs -put BigData/exercises/winter_semester_2019-2020/01_hadoop/sample_d
ata/Faust_1.txt /user/hadoop/Faust_1.txt
```

2. Use and run default MapReduce Jar (*hadoop-mapreduce-examples-3.1.2.jar*) to calculate **wordcount** for text file „*Faust_1.txt*“.

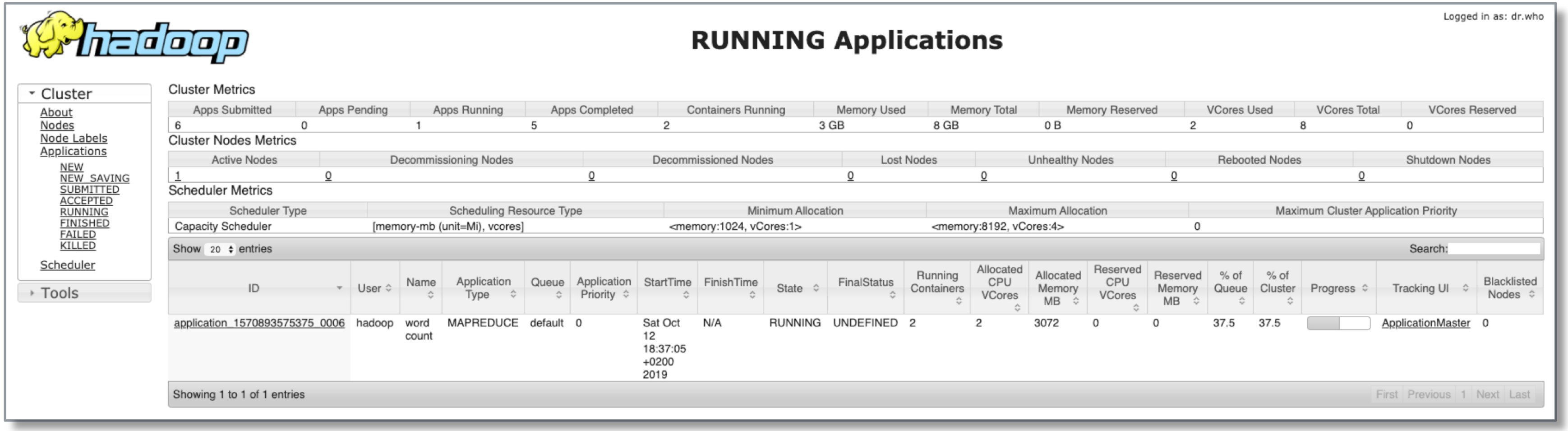
```
hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount /user/
hadoop/Faust_1.txt /user/hadoop/Faust_1_Output

[…]
2019-10-12 16:37:06,033 INFO mapreduce.Job: Running job: job_1570893575375_0006
2019-10-12 16:37:12,157 INFO mapreduce.Job: Job job_1570893575375_0006 running in uber mode : false
2019-10-12 16:37:12,158 INFO mapreduce.Job:  map 0% reduce 0%
2019-10-12 16:37:17,236 INFO mapreduce.Job:  map 100% reduce 0%
2019-10-12 16:37:22,279 INFO mapreduce.Job:  map 100% reduce 100%
2019-10-12 16:37:23,295 INFO mapreduce.Job: Job job_1570893575375_0006 completed successfully
[…]
```

Solution

Exercise 2:

3. Take a look at Ressource Manager for Job Execution (http://XXX.XXX.XXX.XXX:8088/cluster/apps/RUNNING):



Logged in as: dr.who

RUNNING Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - 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
6	0	1	5	2	3 GB	8 GB	0 B	2	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1570893575375_0006	hadoop	word count	MAPREDUCE	default	0	Sat Oct 12 18:37:05 +0200 2019	N/A	RUNNING	UNDEFINED	2	2	3072	0	0	37.5	37.5		ApplicationMaster	0

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

www.marcel-mittelstaedt.com

Solution

Exercise 2:

4. a) Copy MapReduce output file back to ubuntu local filesystem (using bash):

```
hadoop fs -get /user/hadoop/Faust_1_Output/part-r-00000 Faust_1_Output.csv
```

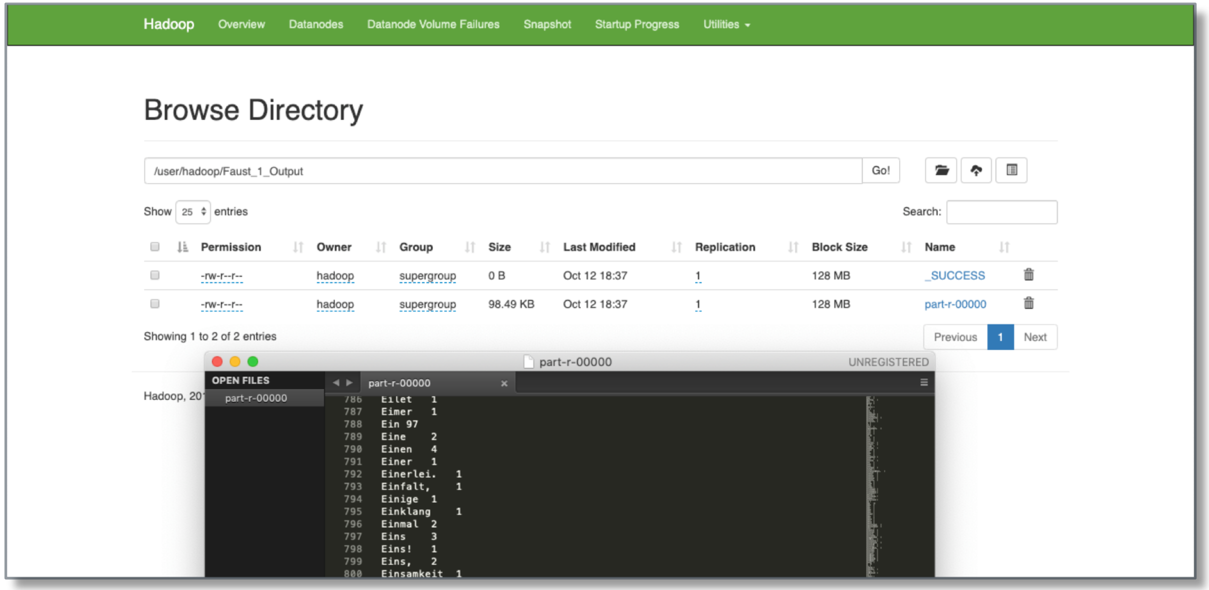
```
shuf -n 10 Faust_1_Output.csv

Phantasie,          1
unwanden. 1
winden,   1
Offenbarung,        2
Undene!   1
Winternächte        1
derweil   1
wiederholten        1
tun       3
Gestalten.          1
```

Solution

Exercise 2:

4. b) Copy MapReduce output file back to ubuntu local filesystem (using Web Filebrowser):



Browse Directory

/user/hadoop/Faust_1_Output

Show 25 entries   Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Oct 12 18:37	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	98.49 KB	Oct 12 18:37	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Hadoop, 20

part-r-00000

```
Eilet	1
Eimer	1
Ein 97
Eine	2
Einen	4
Einer	1
Einerlei.	1
Einfalt,	1
Einige	1
Einklang	1
Einmal	2
Eins	3
Eins!	1
Eins,	2
Einsamkeit	1
```

Solution

Exercise 3:

1. Copy sample file from GIT repo to HDFS user directory:

```
hadoop fs -put BigData/exercises/winter_semester_2019-2020/01_hadoop/sample_d
ata/Faust_1.txt /user/hadoop/Faust_1.txt
```

2. Use and run default MapReduce Jar (*hadoop-mapreduce-examples-3.1.2.jar*) to **grep** for string „Faust“ in text file „*Faust_1.txt*“ and count appearances of string.

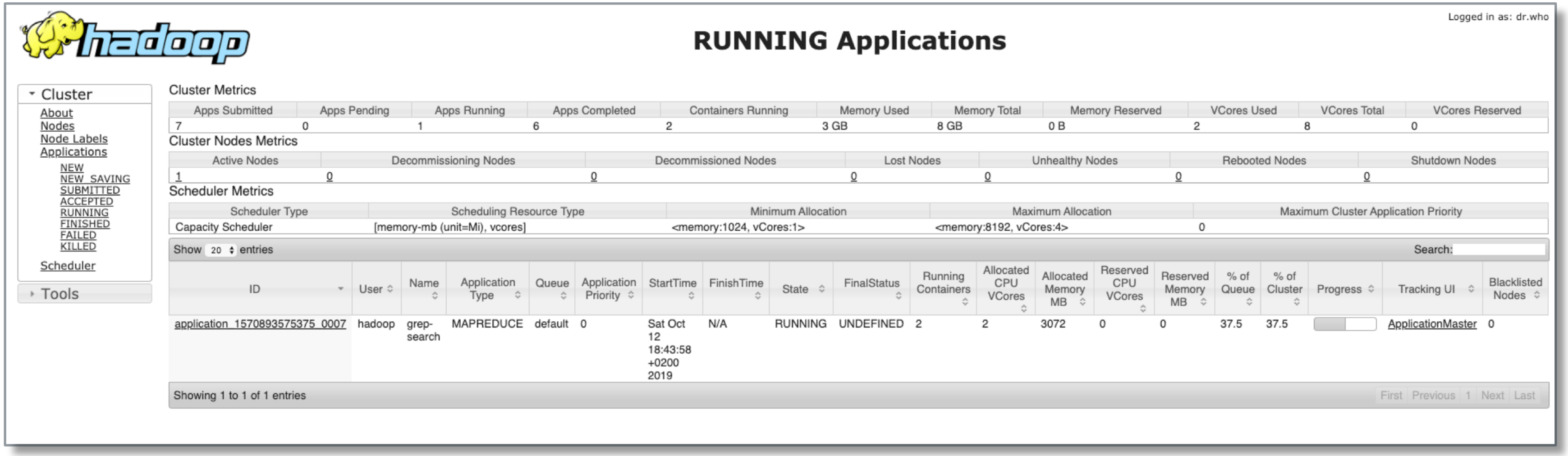
```
hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.1.jar grep /user/hadoo
p/Faust_1.txt /user/hadoop/Faust_1_Count_Output 'Faust`

[…]
2019-10-12 16:44:16,517 INFO mapreduce.Job: Running job: job_1570893575375_0008
2019-10-12 16:44:27,637 INFO mapreduce.Job: Job job_1570893575375_0008 running in uber mode : false
2019-10-12 16:44:27,638 INFO mapreduce.Job:  map 0% reduce 0%
2019-10-12 16:44:31,678 INFO mapreduce.Job:  map 100% reduce 0%
2019-10-12 16:44:36,717 INFO mapreduce.Job:  map 100% reduce 100%
2019-10-12 16:44:37,735 INFO mapreduce.Job: Job job_1570893575375_0008 completed successfully
[…]
```

Solution

Exercise 3:

3. Take a look at Ressource Manager for Job Execution (http://XXX.XXX.XXX.XXX:8088/cluster/apps/RUNNING):


Logged in as: dr.who

RUNNING Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - 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
7	0	1	6	2	3 GB	8 GB	0 B	2	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1570893575375_0007	hadoop	grep-search	MAPREDUCE	default	0	Sat Oct 12 18:43:58 +0200 2019	N/A	RUNNING	UNDEFINED	2	2	3072	0	0	37.5	37.5		ApplicationMaster	0

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Solution

Exercise 2:

4. a) Copy MapReduce output file back to ubuntu local filesystem (using bash):

```
hadoop fs -get /user/hadoop/Faust_1_Count_Output/part-r-00000 Faust_1_Count_O
utput.csv
```

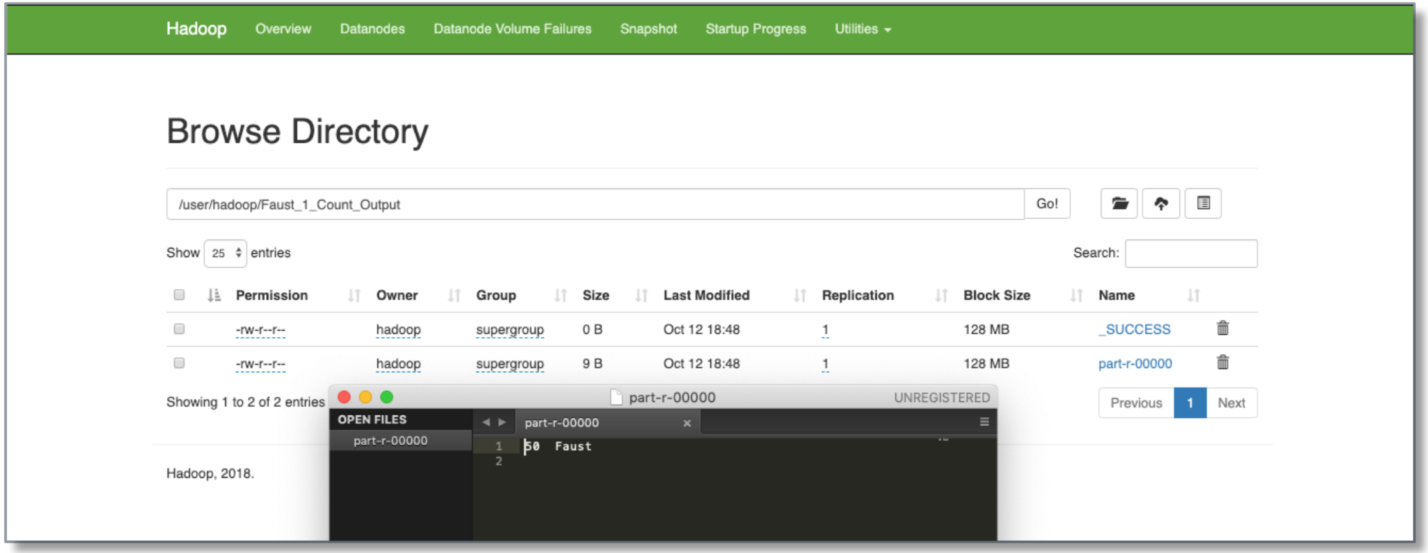
```
cat Faust_1_Count_Output.csv

50          Faust
```

Solution

Exercise 2:

4. b) Copy MapReduce output file back to ubuntu local filesystem (using Web Filebrowser):



Browse Directory

/user/hadoop/Faust_1_Count_Output

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
	-rw-r--r--	hadoop	supergroup	0 B	Oct 12 18:48	1	128 MB	_SUCCESS
	-rw-r--r--	hadoop	supergroup	9 B	Oct 12 18:48	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Hadoop, 2018.

part-r-00000

```
50  Faust
```

Help

MapReduce Examples within *hadoop-mapreduce-examples-3.1.1.jar*:

aggregatewordcount:	An Aggregate based mapreduce program that counts the words in the input files.
aggregatewordhist:	An Aggregate based mapreduce program that computes the histogram of the words in the input files.
bbp:	A mapreduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
dbcount:	An example job that counts the pageview logs stored in a database.
distbbp:	A mapreduce program that uses a BBP-type formula to compute exact bits of Pi.
grep:	A mapreduce program that counts the matches of a regex in the input.
join:	A job that performs a join over sorted, equally partitioned datasets.
multifilewc:	A job that counts words from several files.
pentomino:	A mapreduce tile laying program to find solutions to pentomino problems.
pi:	A mapreduce program that estimates Pi using a quasi-Monte Carlo method.
randomtextwriter:	A mapreduce program that writes 10 GB of random textual data per node.
randomwriter:	A mapreduce program that writes 10 GB of random data per node.
secondarysort:	An example defining a secondary sort to the reduce phase.
sort:	A mapreduce program that sorts the data written by the random writer.
sudoku:	A sudoku solver.
teragen:	Generate data for the terasort.
terasort:	Run the terasort.
teravalidate:	Checking results of terasort.
wordcount:	A mapreduce program that counts the words in the input files.
wordmean:	A mapreduce program that counts the average length of the words in the input files.
wordmedian:	A mapreduce program that counts the median length of the words in the input files.
wordstandarddeviation:	A mapreduce program that counts the standard deviation of the length of the words in the input files.