Hadoop

Una volta installato Hadoop e configurato su *single-node* in modalità *pseudo-distributed* (ogni daemon esegue su un processo Java separato), tramite il seguente comando avviene l'esecuzione dei vari servizi, visualizzabili poi tramite *jps*

In [22]:
```
! /usr/local/cellar/hadoop/3.3.1/libexec/sbin/start-all.sh
```

```
WARNING: Attempting to start all Apache Hadoop daemons as gabrielesavoia in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: /Users/gabrielesavoia/.bashrc: line 1: pyenv: command not found
Starting datanodes
localhost: /Users/gabrielesavoia/.bashrc: line 1: pyenv: command not found
Starting secondary namenodes [MacBook-Air-di-Gabriele.local]
MacBook-Air-di-Gabriele.local: /Users/gabrielesavoia/.bashrc: line 1: pyenv: command not found
2021-11-29 19:21:23,917 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
g builtin-java classes where applicable
Starting resourcemanager
Starting nodemanagers
localhost: /Users/gabrielesavoia/.bashrc: line 1: pyenv: command not found
```

Elenco dei servizi attivi

Alcuni si riferiscono all'**HDFS** (NameNode, SecondaryNameNode e DataNode), mentre altri a **YARN** (ResourceManager e NodeManager).

In [25]:
```
! jps
```

```
27028 SecondaryNameNode
28039 Jps
27224 ResourceManager
26889 DataNode
27326 NodeManager
26783 NameNode
```

HDFS

Trasferisco il file di nome 'documents.txt' nel HDFS in locazione '/'.

In [144…
```
! hdfs dfs -put ./data/documents.txt /
```

```
2021-11-30 10:18:22,203 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
g builtin-java classes where applicable
```

In [145…
```
! hdfs dfs -ls /
```

```
2021-11-30 10:18:29,585 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
g builtin-java classes where applicable
Found 4 items
-rw-r--r--   1 gabrielesavoia supergroup        299 2021-11-30 10:18 /documents.txt
drwxr-xr-x   - gabrielesavoia supergroup          0 2021-11-30 10:16 /output
drwx------   - gabrielesavoia supergroup          0 2021-11-29 23:00 /tmp
drwxr-xr-x   - gabrielesavoia supergroup          0 2021-11-29 23:00 /user
```

Interfaccia web per il monitoraggio dell'HDFS :

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Overview 'localhost:9000' (✔active)

Started:	Mon Nov 29 19:23:19 +0100 2021
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2
Compiled:	Tue Jun 15 07:13:00 +0200 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	CID-8e25594c-0ce1-41dc-8e9c-e9fc6062e280
Block Pool ID:	BP-1692426983-192.168.0.112-1638210165615

Summary

Security is off.

Safemode is off.

57 files and directories, 30 blocks (30 replicated blocks, 0 erasure coded block groups) = 87 total filesystem object(s).

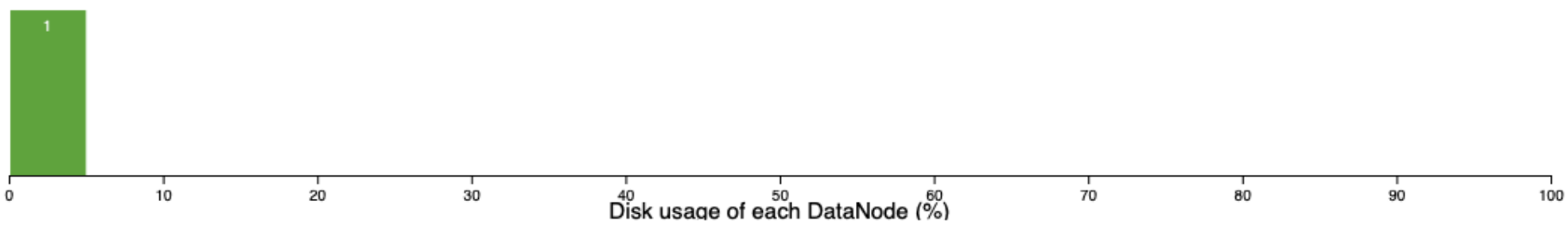
Heap Memory used 89.94 MB of 318.5 MB Heap Memory. Max Heap Memory is 1.78 GB.

Non Heap Memory used 74.13 MB of 76.03 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	931.32 GB

DataNode monitoraggio :

Datanode usage histogram



In operation

DataNode State: All | Show 25 entries | Search:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
✔/default-rack/192.168.0.112:9866 (127.0.0.1:9866)	http://192.168.0.112:9864	2s	321m	4.65 MB	91 GB	931.32 GB	30	4.65 MB (0%)	3.3.1

Showing 1 to 1 of 1 entries

Previous | 1 | Next

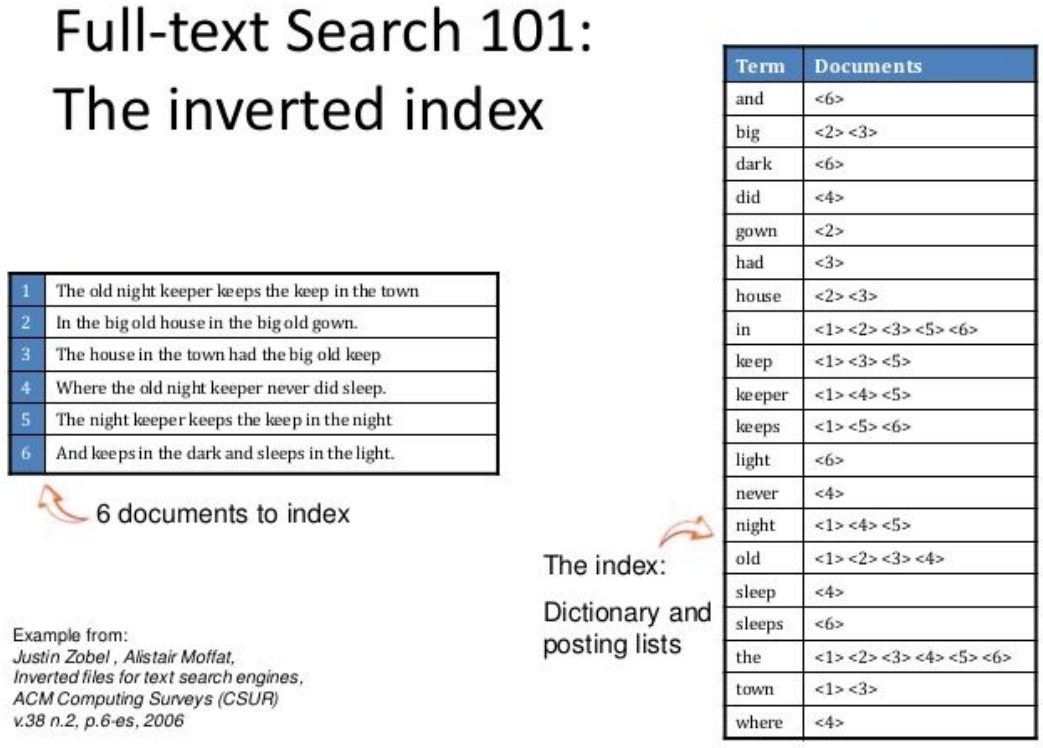
MapReduce

Come esempio dimostrativo, è stata implementata una versione basilare di MapReduce per la creazione di un **InvertedIndex** (ad ogni parola è associata la corrispondente posting list con i documenti in cui questa si presenta).
In particolare è stata utilizzata la libreria **MRJob** mediante la quale è possibile scrivere applicazioni MapReduce in Python per poi eseguerle sia in locale che in un cluster Hadoop. Nonostante Hadoop sia scritto principalmente di Java, mette a disposizione un particolare modulo definito **HadoopStreaming** mediante il quale è possibile interagire con MapReduce anche da altri linguaggi in quanto la comunicazione avviene tramite *stdin* e *stdout*.

Di seguito è riportato il codice relativo al **job MapReduce** *InvertedIndexMR*, nel quale sono definite principalmente due funzioni :

- **map**: si occupa di leggere in input riga per riga del file 'documents.txt'. Questo file infatti contiene per ciasuna riga l'id del documento separato poi con un ':' dal relativo testo. La map ritorna quindi, senza considerare le stopword, coppie del tipo : (word, doc_id);
- **reduce**: elabora il risultato della map e ritorna in output, per ogni parola, la corrispondente posting list contenente, senza duplicati, i documenti a cui fa parte.

Di seguito è riportato ciò che si vuole ottenere.



In [161…

```
%%file inverted_index.py
from mrjob.job import MRJob

# Non utilizzo NLTK dal momento che si tratta di codice di esempio
stop_words = ['il', 'con', 'i', 'a', 'e', 'al', 'con', '.', ',', 'nella', 'nei', 'nel', 'per', 'di', 'la', 'va', '

class InvertedIndexMR(MRJob):

    def mapper(self, _, line):
        """
        Input : righe del file.
        Return : coppie (word, doc_id)
        """
        doc_id, doc_text = line.split(':')
        doc_id = doc_id.strip()
        for word in doc_text.split():
            word = word.lower()
            if word not in stop_words:
                yield word, doc_id

    def reducer(self, word, doc_list):
        """
        Input : coppie (word, [doc_id_1, doc_id_1, ... , doc_id_n] )
        Return : coppie (word, [doc_id_1, ... , doc_id_n] )            --> senza duplicati di documenti
        """
        unique_doc = list(set(doc_list))

        yield word, unique_doc

if __name__ == '__main__':
    InvertedIndexMR.run()
```

```
Overwriting inverted_index.py
```

Documento da elaborare

Il documento su cui viene eseguita la funzione di MapReduce è salvato in *./data/documents.txt*.

In [162…
```
! cat ./data/documents.txt
```

```
doc1 : La storia iniziò nella città di Berlino nel lontano 1790
doc2 : Oggi Luca va con tutti i suoi amici a giocare a calcetto
doc3 : Questo fine settimana esco con alcuni amici di amici
doc4 : Durante la notte alcuni animali escono per cercare cibo
doc5 : Matteo e Luca torneranno a casa alle 23
```

Run in locale

In questo caso ci si riferisce al documento in locale presente in *./data*.

In [163…
```
! python inverted_index.py ./data/documents.txt
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /var/folders/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/inverted_index.gabrielesavoia.20211130.17
1917.118742
Running step 1 of 1...
job output is in /var/folders/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/inverted_index.gabrielesavoia.20211130.171917.11
8742/output
Streaming final output from /var/folders/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/inverted_index.gabrielesavoia.2021113
0.171917.118742/output...
"oggi"  ["doc2"]
"questo"        ["doc3"]
"settimana"     ["doc3"]
"storia"        ["doc1"]
"cercare"       ["doc4"]
"cibo"  ["doc4"]
"citt\u00e0"    ["doc1"]
"durante"       ["doc4"]
"esco"  ["doc3"]
"escono"        ["doc4"]
"fine"  ["doc3"]
"giocare"       ["doc2"]
"inizi\u00f2"   ["doc1"]
"lontano"       ["doc1"]
"luca"  ["doc5", "doc2"]
"matteo"        ["doc5"]
"notte" ["doc4"]
"animali"       ["doc4"]
"berlino"       ["doc1"]
"calcetto"      ["doc2"]
"casa"  ["doc5"]
"suoi"  ["doc2"]
"torneranno"    ["doc5"]
"tutti" ["doc2"]
"1790"  ["doc1"]
"23"    ["doc5"]
"alcuni"        ["doc4", "doc3"]
"amici" ["doc2", "doc3"]
Removing temp directory /var/folders/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/inverted_index.gabrielesavoia.20211130.17
1917.118742...
```

Run in Hadoop

In questo caso invece, ci si riferisce al file presente nell'HDFS. Per l'esecuzione è necessario inoltre definire la posizione del file *hadoop-streaming-3.3.1.jar*.

In [158…
```
! hdfs dfs -rm -R /output
```

```
2021-11-30 11:35:14,467 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
g builtin-java classes where applicable
Deleted /output
```

In [159…
```
! python inverted_index.py -r hadoop hdfs:///documents.txt --hadoop-streaming-jar hadoop-streaming-3.3.1.jar --out
```

```
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/local/bin/hadoop
Using Hadoop version 3.3.1
Creating temp directory /var/folders/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/inverted_index.gabrielesavoia.20211130.10
3519.667336
uploading working dir files to hdfs:///user/gabrielesavoia/tmp/mrjob/inverted_index.gabrielesavoia.20211130.103519
.667336/files/wd...
```

```
Copying other local files to hdfs:///user/gabrielesavoia/tmp/mrjob/inverted_index.gabrielesavoia.20211130.103519.6
67336/files/
Running step 1 of 1...
  Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
  packageJobJar: [/var/folders/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/hadoop-unjar5602148106066165414/] [] /var/folde
rs/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/streamjob8707569005565255841.jar tmpDir=null
  Connecting to ResourceManager at /127.0.0.1:8032
  Connecting to ResourceManager at /127.0.0.1:8032
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/gabrielesavoia/.staging/job_1638210220692_0007
  Total input files to process : 1
  number of splits:2
  Submitting tokens for job: job_1638210220692_0007
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1638210220692_0007
  The url to track the job: http://192.168.0.112:8088/proxy/application_1638210220692_0007/
  Running job: job_1638210220692_0007
  Job job_1638210220692_0007 running in uber mode : false
   map 0% reduce 0%
   map 33% reduce 0%
   map 67% reduce 0%
   map 83% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1638210220692_0007 completed successfully
  Output directory: hdfs:///output
Counters: 50
        File Input Format Counters
                Bytes Read=449
        File Output Format Counters
                Bytes Written=529
        File System Counters
                FILE: Number of bytes read=579
                FILE: Number of bytes written=832616
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=623
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=529
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=11
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=2
                Launched map tasks=2
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=131346432
                Total megabyte-milliseconds taken by all reduce tasks=50506752
                Total time spent by all map tasks (ms)=128268
                Total time spent by all maps in occupied slots (ms)=128268
                Total time spent by all reduce tasks (ms)=49323
                Total time spent by all reduces in occupied slots (ms)=49323
                Total vcore-milliseconds taken by all map tasks=128268
                Total vcore-milliseconds taken by all reduce tasks=49323
        Map-Reduce Framework
                CPU time spent (ms)=0
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=326
                Input split bytes=174
                Map input records=5
                Map output bytes=509
                Map output materialized bytes=585
                Map output records=32
                Merged Map outputs=2
                Physical memory (bytes) snapshot=0
                Reduce input groups=28
                Reduce input records=32
                Reduce output records=28
                Reduce shuffle bytes=585
                Shuffled Maps =2
                Spilled Records=64
                Total committed heap usage (bytes)=682622976
                Virtual memory (bytes) snapshot=0
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///output
Removing HDFS temp directory hdfs:///user/gabrielesavoia/tmp/mrjob/inverted_index.gabrielesavoia.20211130.103519.6
67336...
Removing temp directory /var/folders/z3/yl30gjf55_x2qqgttnn1gjjc0000gn/T/inverted_index.gabrielesavoia.20211130.10
```

```
3519.667336...
```

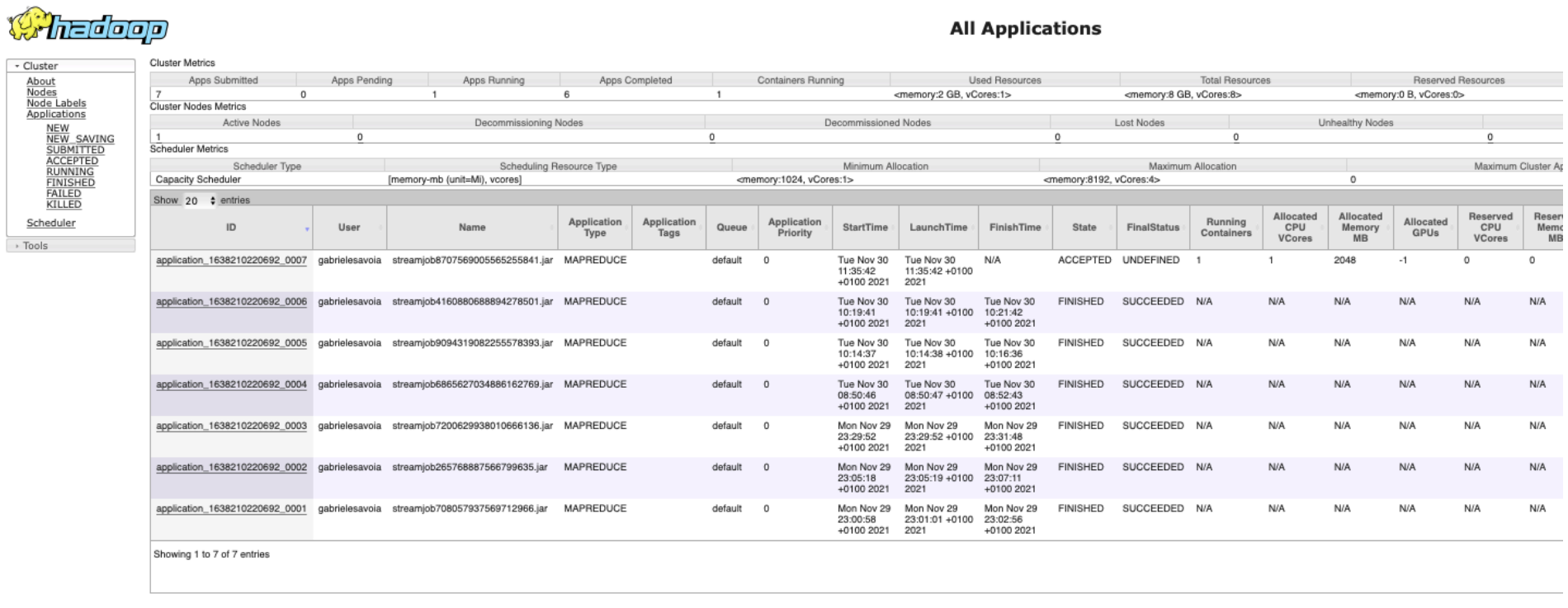
In [160…

```
! hadoop fs -cat hdfs:///output/part-00000
```

```
2021-11-30 11:38:18,857 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
g builtin-java classes where applicable
"1790"  ["doc1"]
"23"    ["doc5"]
"alcuni"        ["doc3", "doc4"]
"amici" ["doc2", "doc3"]
"animali"       ["doc4"]
"berlino"       ["doc1"]
"calcetto"      ["doc2"]
"casa"  ["doc5"]
"cercare"       ["doc4"]
"cibo"  ["doc4"]
"citt\u00e0"    ["doc1"]
"durante"       ["doc4"]
"esco"  ["doc3"]
"escono"        ["doc4"]
"fine"  ["doc3"]
"giocare"       ["doc2"]
"inizi\u00f2"   ["doc1"]
"lontano"       ["doc1"]
"luca"  ["doc2", "doc5"]
"matteo"        ["doc5"]
"notte" ["doc4"]
"oggi"  ["doc2"]
"questo"        ["doc3"]
"settimana"     ["doc3"]
"storia"        ["doc1"]
"suoi"  ["doc2"]
"torneranno"    ["doc5"]
"tutti" ["doc2"]
```

YARN

Durante l'esecuzione del job MapReduce di prima, tramite l'interfaccia web di YARN, è stata monitorata l'esecuzione dell'applicazione. In particolare nella figura di seguito sono riportate tutte le applicazioni eseguite (la prima è quella in esecuzione non ancora terminata):



All 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	Used Resources	Total Resources	Reserved Resources
7	0	1	6	1	<memory:2 GB, vCores:1>	<memory:8 GB, vCores:8>	<memory:0 B, vCores:0>

Cluster Nodes Metrics

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

Scheduler Metrics

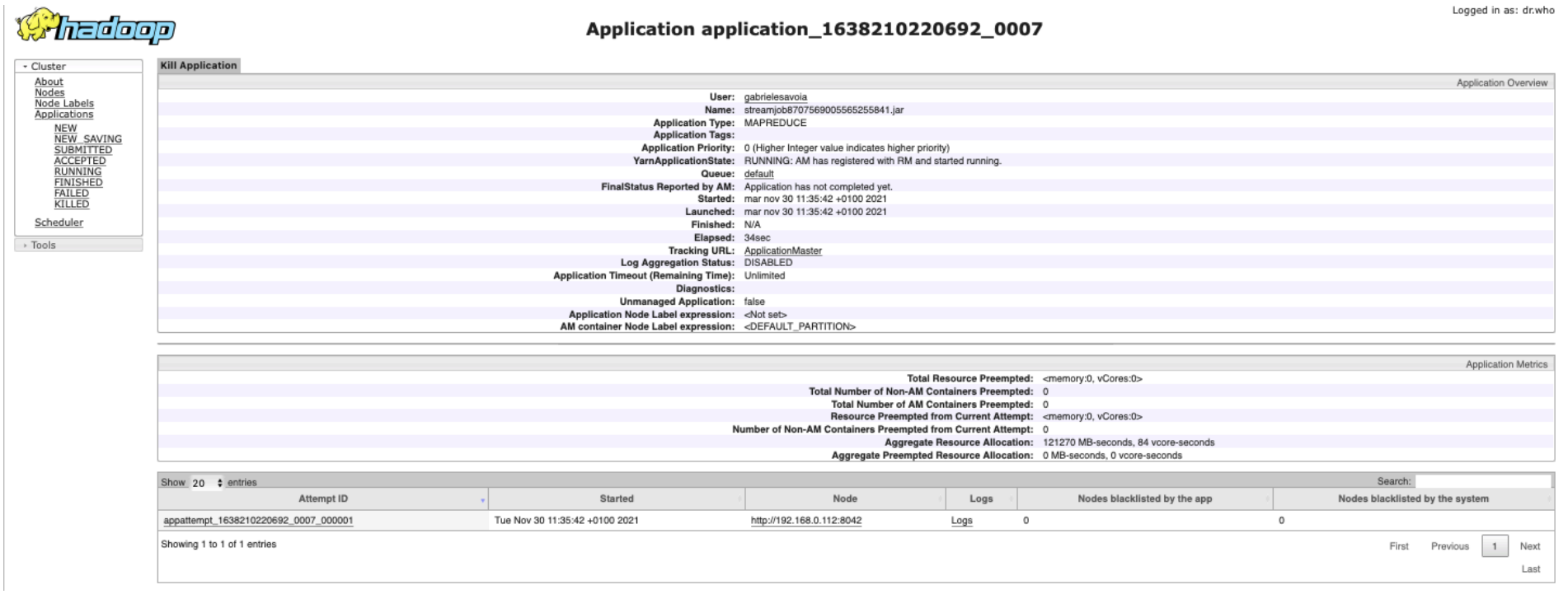
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Ap
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memo MB
application_1638210220692_0007	gabrielesavoia	streamjob8707569005565255841.jar	MAPREDUCE		default	0	Tue Nov 30 11:35:42 +0100 2021	Tue Nov 30 11:35:42 +0100 2021	N/A	ACCEPTED	UNDEFINED	1	1	2048	-1	0	0
application_1638210220692_0006	gabrielesavoia	streamjob4160880688894278501.jar	MAPREDUCE		default	0	Tue Nov 30 10:19:41 +0100 2021	Tue Nov 30 10:19:41 +0100 2021	Tue Nov 30 10:21:42 +0100 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1638210220692_0005	gabrielesavoia	streamjob9094319082255578393.jar	MAPREDUCE		default	0	Tue Nov 30 10:14:37 +0100 2021	Tue Nov 30 10:14:38 +0100 2021	Tue Nov 30 10:16:36 +0100 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1638210220692_0004	gabrielesavoia	streamjob6865627034886162769.jar	MAPREDUCE		default	0	Tue Nov 30 08:50:46 +0100 2021	Tue Nov 30 08:50:47 +0100 2021	Tue Nov 30 08:52:43 +0100 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1638210220692_0003	gabrielesavoia	streamjob7200629938010666136.jar	MAPREDUCE		default	0	Mon Nov 29 23:29:52 +0100 2021	Mon Nov 29 23:29:52 +0100 2021	Mon Nov 29 23:31:48 +0100 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1638210220692_0002	gabrielesavoia	streamjob2657688875667996635.jar	MAPREDUCE		default	0	Mon Nov 29 23:05:18 +0100 2021	Mon Nov 29 23:05:19 +0100 2021	Mon Nov 29 23:07:11 +0100 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1638210220692_0001	gabrielesavoia	streamjob7080579375697129 66.jar	MAPREDUCE		default	0	Mon Nov 29 23:00:58 +0100 2021	Mon Nov 29 23:01:01 +0100 2021	Mon Nov 29 23:02:56 +0100 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A

Showing 1 to 7 of 7 entries

E' inoltre possibile avere dettagli maggiori riferiti alle singole applicazioni :



Application application_1638210220692_0007

Logged in as: dr.who

Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools

Kill Application

Application Overview

User:	gabrielesavoia
Name:	streamjob8707569005565255841.jar
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	RUNNING: AM has registered with RM and started running.
Queue:	default
FinalStatus Reported by AM:	Application has not completed yet.
Started:	mar nov 30 11:35:42 +0100 2021
Launched:	mar nov 30 11:35:42 +0100 2021
Finished:	N/A
Elapsed:	34sec
Tracking URL:	ApplicationMaster
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

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:	121270 MB-seconds, 84 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1638210220692_0007_000001	Tue Nov 30 11:35:42 +0100 2021	http://192.168.0.112:8042	Logs	0	0

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