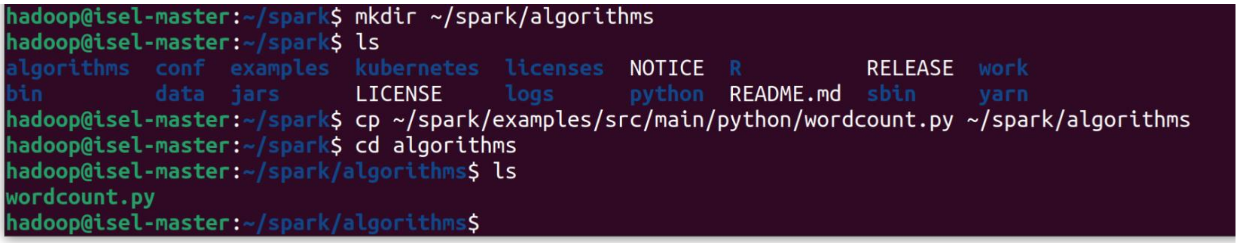
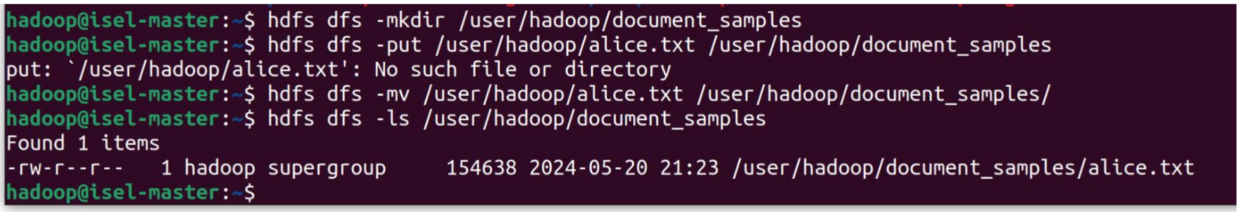
Submit Spark Jobs

I have created a directory for keeping the algorithms we submitted onto Spark
~/spark/algorithms/



```
hadoop@isel-master:~/spark$ mkdir ~/spark/algorithms
hadoop@isel-master:~/spark$ ls
algorithms  conf  examples  kubernetes  licenses  NOTICE  R          RELEASE  work
bin         data  jars      LICENSE     logs      python  README.md  sbin     yarn
hadoop@isel-master:~/spark$ cp ~/spark/examples/src/main/python/wordcount.py ~/spark/algorithms
hadoop@isel-master:~/spark$ cd algorithms
hadoop@isel-master:~/spark/algorithms$ ls
wordcount.py
hadoop@isel-master:~/spark/algorithms$
```

And also a directory for keeping the samples (just for now to test) in HDFS
/user/hadoop/document_samples/

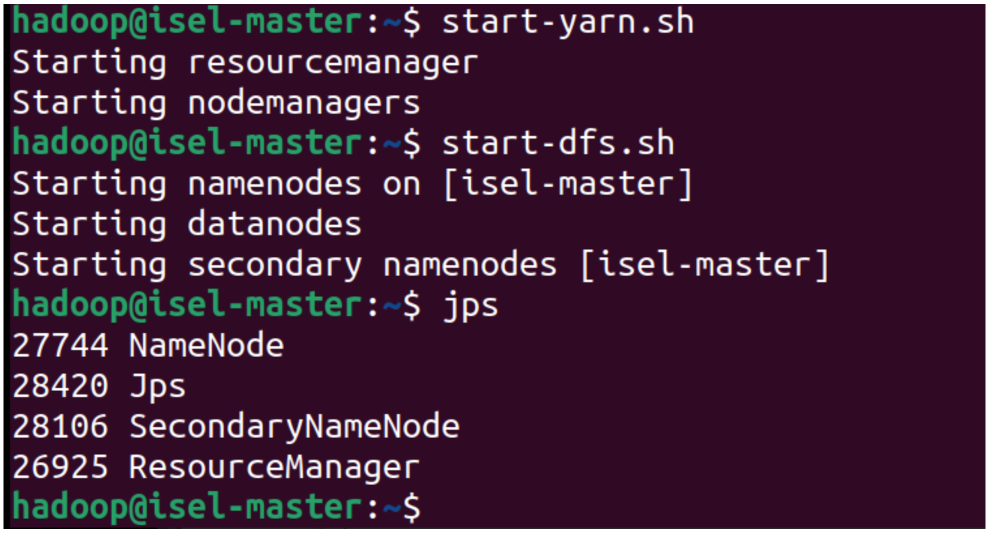


```
hadoop@isel-master:~$ hdfs dfs -mkdir /user/hadoop/document_samples
hadoop@isel-master:~$ hdfs dfs -put /user/hadoop/alice.txt /user/hadoop/document_samples
put: `/user/hadoop/alice.txt': No such file or directory
hadoop@isel-master:~$ hdfs dfs -mv /user/hadoop/alice.txt /user/hadoop/document_samples/
hadoop@isel-master:~$ hdfs dfs -ls /user/hadoop/document_samples
Found 1 items
-rw-r--r--   1 hadoop supergroup     154638 2024-05-20 21:23 /user/hadoop/document_samples/alice.txt
hadoop@isel-master:~$
```

Spark-submit can be run anywhere as the configuration settings

Starting Nodes

Start YARN, DFS and ensure name/datanodes are running



```
hadoop@isel-master:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@isel-master:~$ start-dfs.sh
Starting namenodes on [isel-master]
Starting datanodes
Starting secondary namenodes [isel-master]
hadoop@isel-master:~$ jps
27744 NameNode
28420 Jps
28106 SecondaryNameNode
26925 ResourceManager
hadoop@isel-master:~$
```

start-yarn.sh
start-dfs.sh
Jps

Analysis

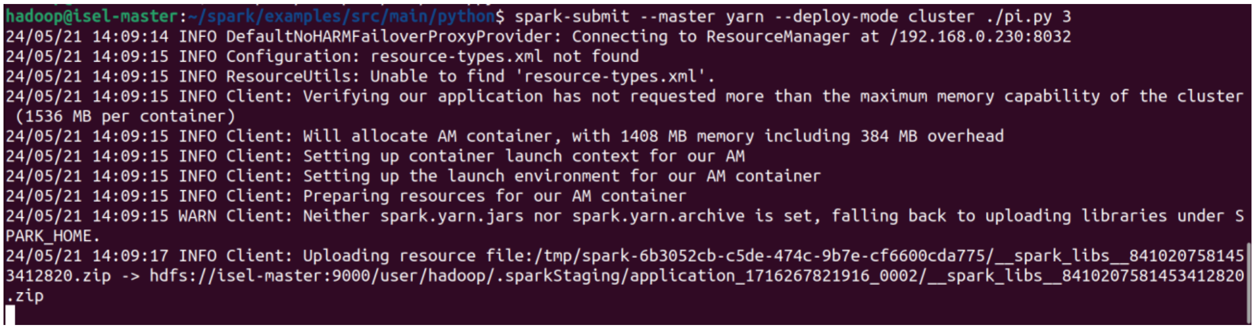
For non-documents related algorithms

Just for testing purposes

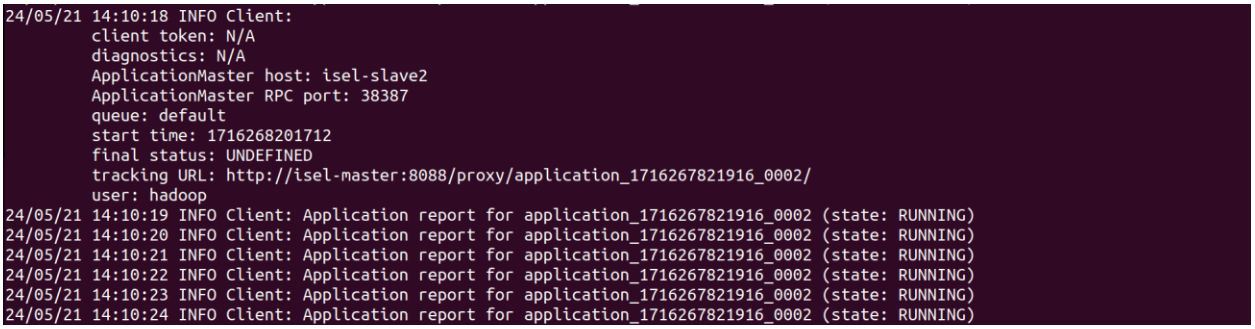
Run example:

spark-submit --master yarn --deploy-mode cluster ./pi.py 3

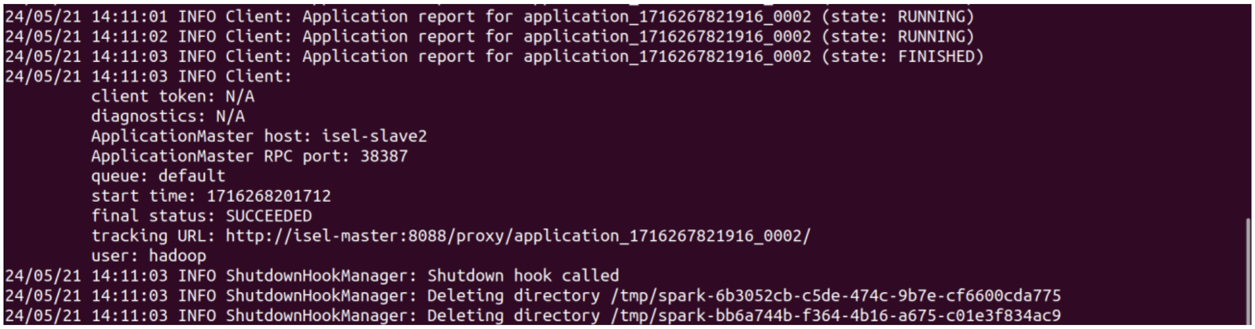
Put location of algorithm (~/spark/algorithms…) instead of ./pi.py example
Add arguments after if necessary



```
hadoop@isel-master:~/spark/examples/src/main/python$ spark-submit --master yarn --deploy-mode cluster ./pi.py 3
24/05/21 14:09:14 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /192.168.0.230:8032
24/05/21 14:09:15 INFO Configuration: resource-types.xml not found
24/05/21 14:09:15 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/05/21 14:09:15 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster
 (1536 MB per container)
24/05/21 14:09:15 INFO Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
24/05/21 14:09:15 INFO Client: Setting up container launch context for our AM
24/05/21 14:09:15 INFO Client: Setting up the launch environment for our AM container
24/05/21 14:09:15 INFO Client: Preparing resources for our AM container
24/05/21 14:09:15 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under S
PARK_HOME.
24/05/21 14:09:17 INFO Client: Uploading resource file:/tmp/spark-6b3052cb-c5de-474c-9b7e-cf6600cda775/__spark_libs__841020758145
3412820.zip -> hdfs://isel-master:9000/user/hadoop/.sparkStaging/application_1716267821916_0002/__spark_libs__8410207581453412820
.zip
```



```
24/05/21 14:10:18 INFO Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: isel-slave2
         ApplicationMaster RPC port: 38387
         queue: default
         start time: 1716268201712
         final status: UNDEFINED
         tracking URL: http://isel-master:8088/proxy/application_1716267821916_0002/
         user: hadoop
24/05/21 14:10:19 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
24/05/21 14:10:20 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
24/05/21 14:10:21 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
24/05/21 14:10:22 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
24/05/21 14:10:23 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
24/05/21 14:10:24 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
```

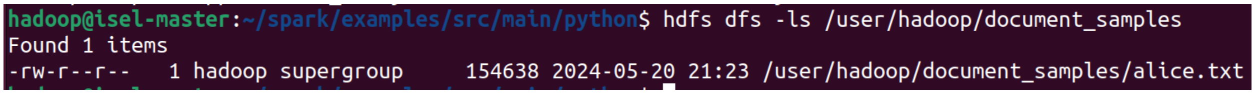


```
24/05/21 14:11:01 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
24/05/21 14:11:02 INFO Client: Application report for application_1716267821916_0002 (state: RUNNING)
24/05/21 14:11:03 INFO Client: Application report for application_1716267821916_0002 (state: FINISHED)
24/05/21 14:11:03 INFO Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: isel-slave2
         ApplicationMaster RPC port: 38387
         queue: default
         start time: 1716268201712
         final status: SUCCEEDED
         tracking URL: http://isel-master:8088/proxy/application_1716267821916_0002/
         user: hadoop
24/05/21 14:11:03 INFO ShutdownHookManager: Shutdown hook called
24/05/21 14:11:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-6b3052cb-c5de-474c-9b7e-cf6600cda775
24/05/21 14:11:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-bb6a744b-f364-4b16-a675-c01e3f834ac9
```

For document related analysis

First, check if document is successfully uploaded into HDFS

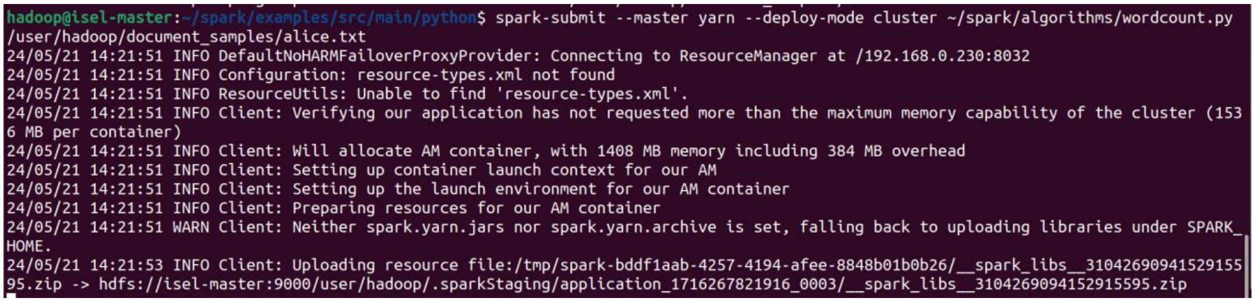
hdfs dfs -ls /user/hadoop/document_samples *directory path



```
hadoop@isel-master:~/spark/examples/src/main/python$ hdfs dfs -ls /user/hadoop/document_samples
Found 1 items
-rw-r--r--   1 hadoop supergroup     154638 2024-05-20 21:23 /user/hadoop/document_samples/alice.txt
```

Then run program and provide file path of the document in HDFS (hdfs://…) :

spark-submit --master yarn --deploy-mode cluster ~/spark/algorithms/wordcount.py /user/hadoop/document_samples/alice.txt

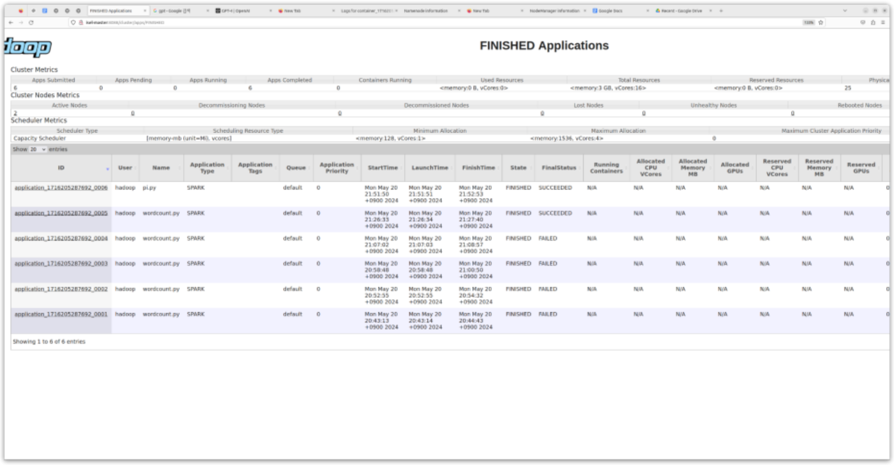


```
hadoop@isel-master:~/spark/examples/src/main/python$ spark-submit --master yarn --deploy-mode cluster ~/spark/algorithms/wordcount.py
/user/hadoop/document_samples/alice.txt
24/05/21 14:21:51 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /192.168.0.230:8032
24/05/21 14:21:51 INFO Configuration: resource-types.xml not found
24/05/21 14:21:51 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/05/21 14:21:51 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (153
6 MB per container)
24/05/21 14:21:51 INFO Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
24/05/21 14:21:51 INFO Client: Setting up container launch context for our AM
24/05/21 14:21:51 INFO Client: Setting up the launch environment for our AM container
24/05/21 14:21:51 INFO Client: Preparing resources for our AM container
24/05/21 14:21:51 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_
HOME.
24/05/21 14:21:53 INFO Client: Uploading resource file:/tmp/spark-bddf1aab-4257-4194-afee-8848b01b0b26/__spark_libs__31042690941529155
95.zip -> hdfs://isel-master:9000/user/hadoop/.sparkStaging/application_1716267821916_0003/__spark_libs__3104269094152915595.zip
```

View Results

Go to Hadoop web URI:

http://isel-master:8088/cluster



FINISHED Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physica
6	0	0	6	0	<memory:0 B, vCores:0>	<memory:3 GB, vCores:16>	<memory:0 B, vCores:0>	25

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
2	0	0	0	0	0

Scheduler Metrics

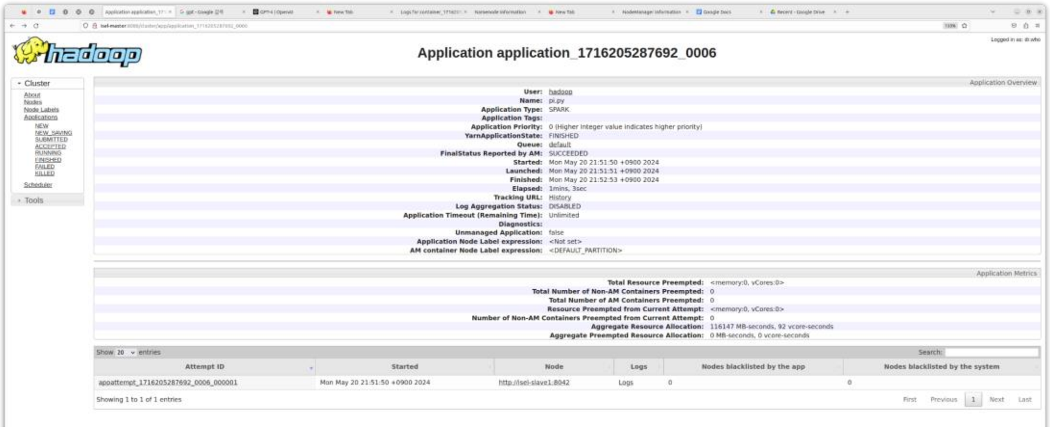
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:128, vCores:1>	<memory:1536, vCores:4>	0

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 Memory MB	Reserved GPUs	
application_1716205287692_0006	hadoop	pi.py	SPARK		default	0	Mon May 20 21:51:50 +0900 2024	Mon May 20 21:51:51 +0900 2024	Mon May 20 21:52:53 +0900 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0
application_1716205287692_0005	hadoop	wordcount.py	SPARK		default	0	Mon May 20 21:26:33 +0900 2024	Mon May 20 21:26:34 +0900 2024	Mon May 20 21:27:40 +0900 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0
application_1716205287692_0004	hadoop	wordcount.py	SPARK		default	0	Mon May 20 21:07:02 +0900 2024	Mon May 20 21:07:03 +0900 2024	Mon May 20 21:08:57 +0900 2024	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0
application_1716205287692_0003	hadoop	wordcount.py	SPARK		default	0	Mon May 20 20:58:48 +0900 2024	Mon May 20 20:58:48 +0900 2024	Mon May 20 21:00:50 +0900 2024	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0
application_1716205287692_0002	hadoop	wordcount.py	SPARK		default	0	Mon May 20 20:52:55 +0900 2024	Mon May 20 20:52:55 +0900 2024	Mon May 20 20:54:32 +0900 2024	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0
application_1716205287692_0001	hadoop	wordcount.py	SPARK		default	0	Mon May 20 20:43:13 +0900 2024	Mon May 20 20:43:14 +0900 2024	Mon May 20 20:44:43 +0900 2024	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0

Showing 1 to 6 of 6 entries

Click on the Spark application:



Application application_1716205287692_0006

Application Overview

User: hadoop
Name: pi.py
Application Type: SPARK
Application Tags:
Application Priority: 0 (Higher Integer value indicates higher priority)
YarnApplicationState: FINISHED
Queue: default
FinalStatus Reported by AM: SUCCEEDED
Started: Mon May 20 21:51:50 +0900 2024
Launched: Mon May 20 21:51:51 +0900 2024
Finished: Mon May 20 21:52:53 +0900 2024
Elapsed: 1mins, 3sec
Tracking URL: History
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

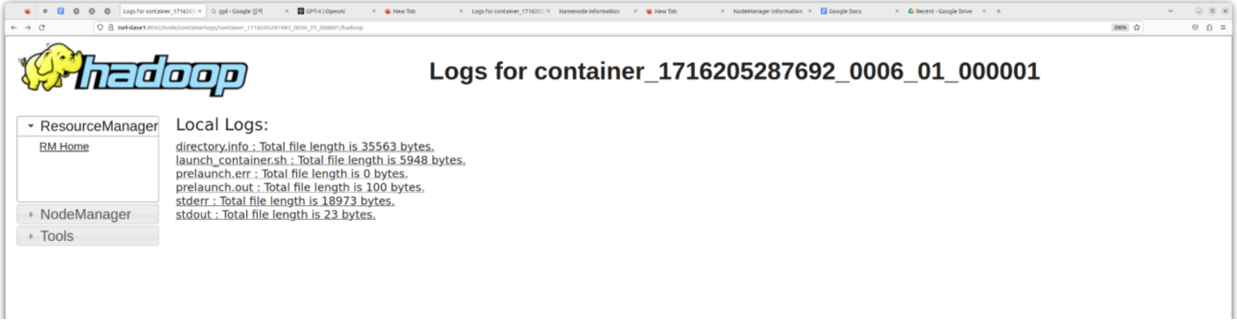
Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted from Current Attempt: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0
Aggregate Resource Allocation: 116147 MB-seconds, 92 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_1716205287692_0006_000001	Mon May 20 21:51:50 +0900 2024	http://isel-slave1:8042	Logs	0	0

Showing 1 to 1 of 1 entries

Click on Logs under the categories

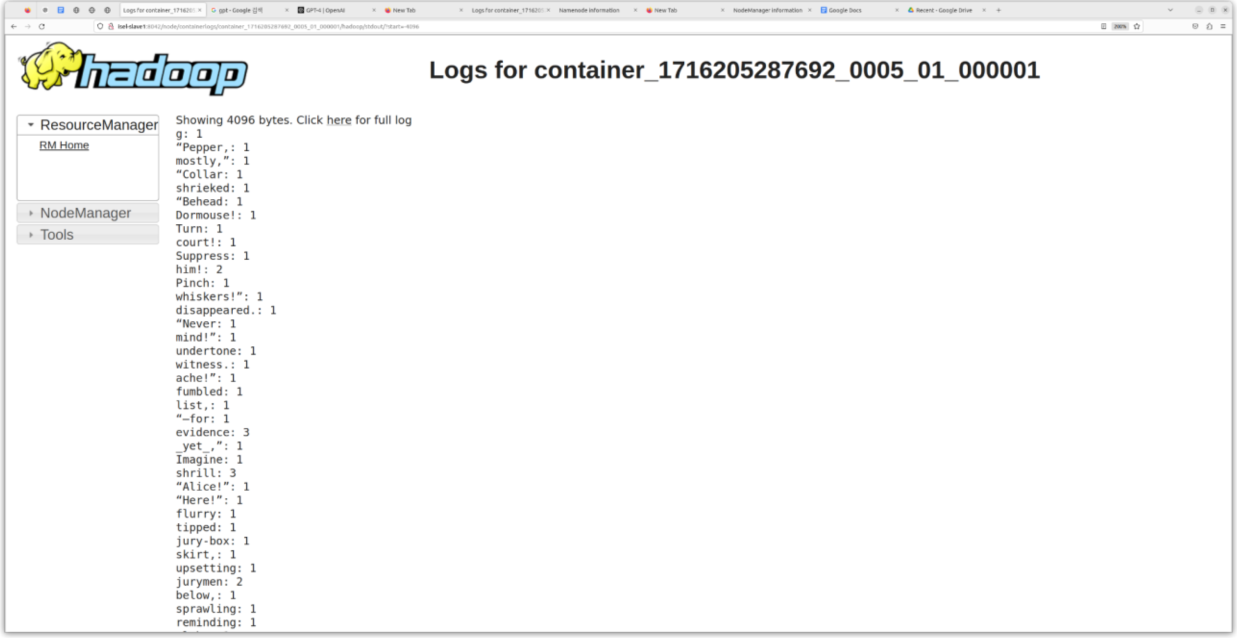


Logs for container_1716205287692_0006_01_000001

Local Logs:

directory.info : Total file length is 35563 bytes.
launch_container.sh : Total file length is 5948 bytes.
prelaunch.err : Total file length is 0 bytes.
prelaunch.out : Total file length is 100 bytes.
stderr : Total file length is 18973 bytes.
stdout : Total file length is 23 bytes.

Select stdout



Logs for container_1716205287692_0005_01_000001

Showing 4096 bytes. Click here for full log

```
g: 1
“Pepper,: 1
mostly,”: 1
“Collar: 1
shrieked: 1
“Behead: 1
Dormouse!: 1
Turn: 1
court!: 1
Suppress: 1
him!: 2
Pinch: 1
whiskers!”: 1
disappeared.: 1
“Never: 1
mind!”: 1
undertone: 1
witness.: 1
ache!”: 1
fumbled: 1
list,: 1
“—for: 1
evidence: 3
_yet_,”: 1
Imagine: 1
shrill: 3
“Alice!”: 1
“Here!”: 1
flurry: 1
tipped: 1
jury-box: 1
skirt,: 1
upsetting: 1
jurymen: 2
below,: 1
sprawling: 1
reminding: 1
```