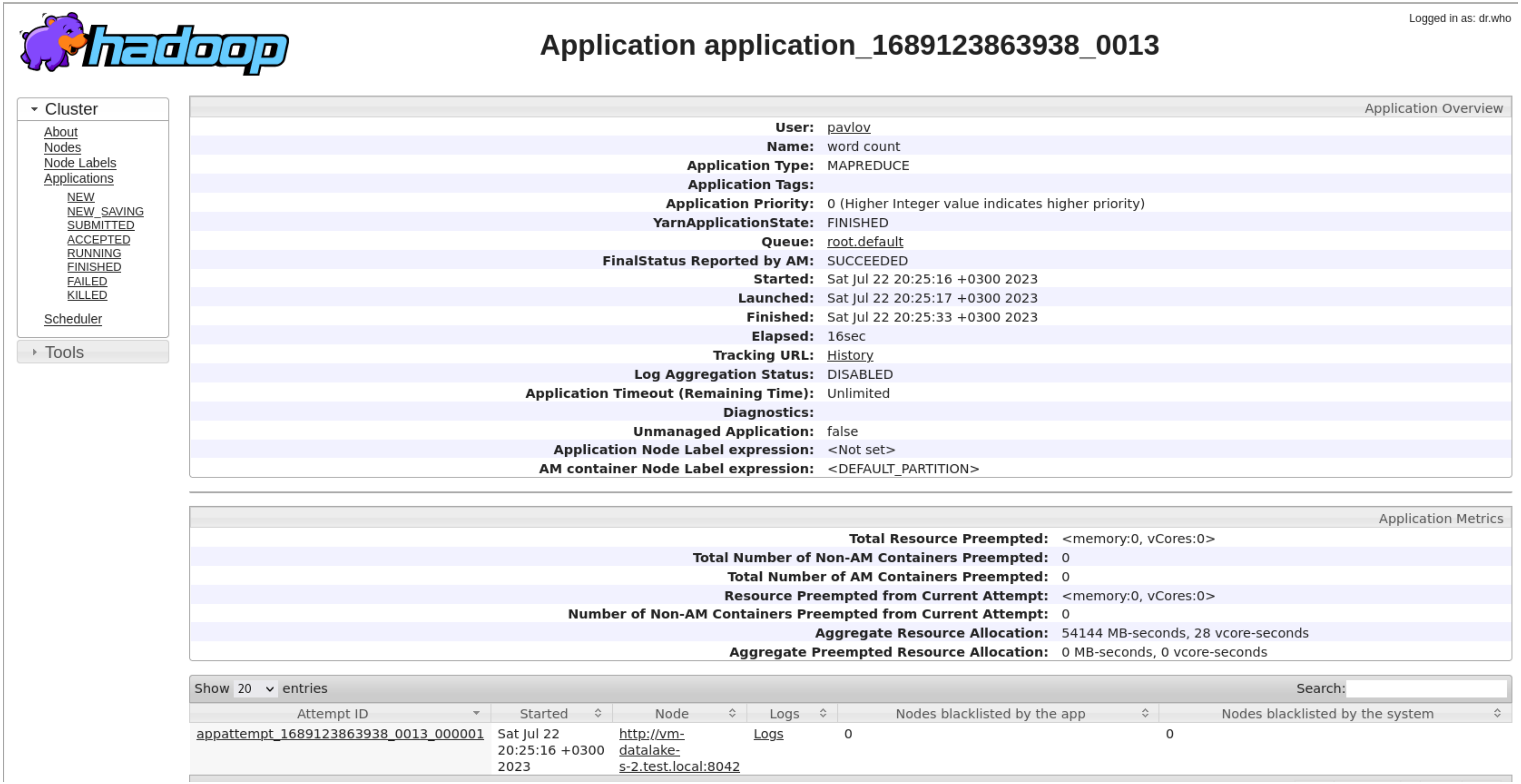
Лабораторная работа №3

Задание №1

Файл https://github.com/sic-rus-ai/stepik-dl-nlp/raw/master/datasets/nyt-ingredients-snapshot-2015.csv

Из списка выполненных Application Master задач, выберите свою выполненную задачу и приложите результат (в виде скриншота)



Logged in as: dr.who

Application application_1689123863938_0013

Application Overview

User:	pavlov
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Jul 22 20:25:16 +0300 2023
Launched:	Sat Jul 22 20:25:17 +0300 2023
Finished:	Sat Jul 22 20:25:33 +0300 2023
Elapsed:	16sec
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>

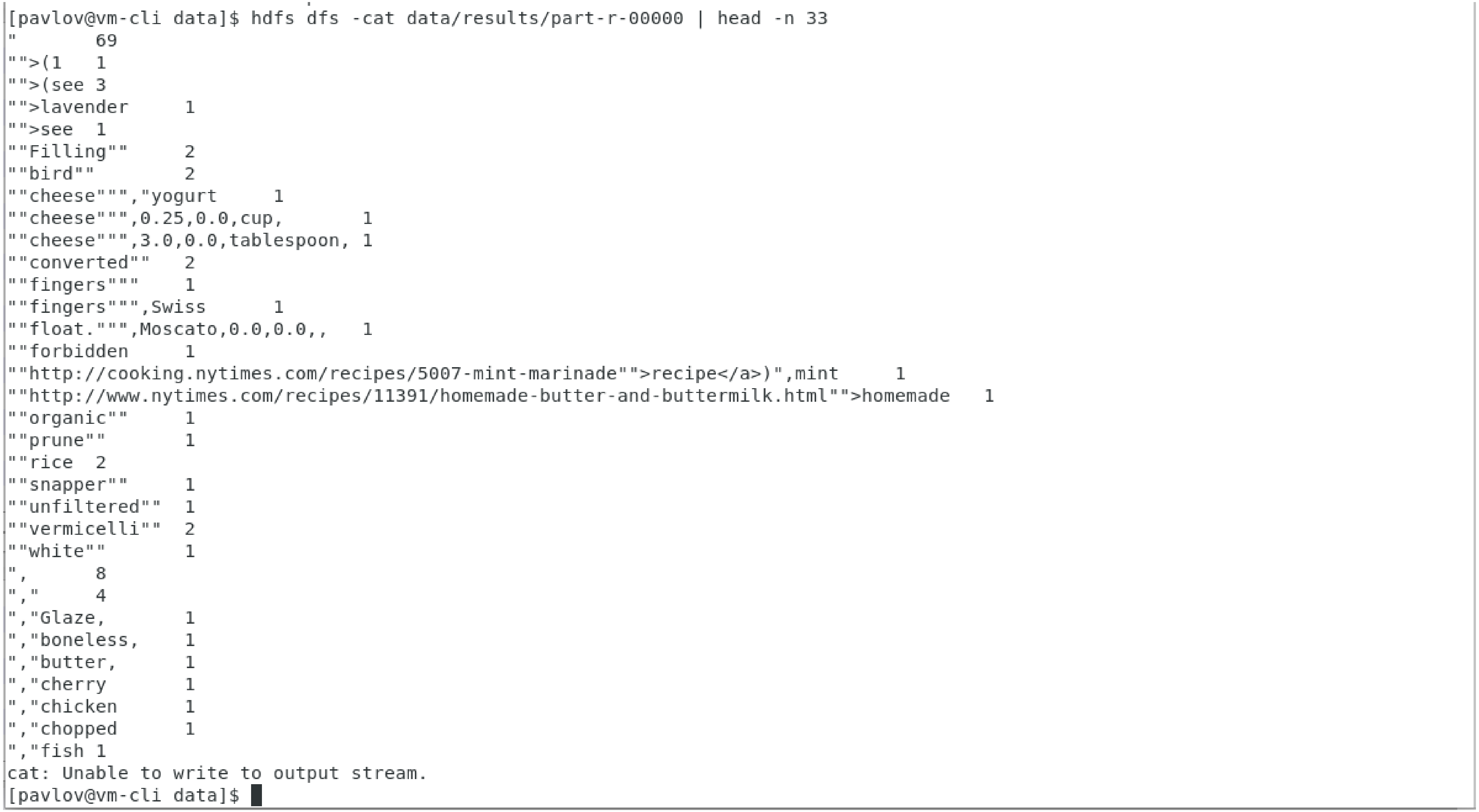
Application Metrics

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:	54144 MB-seconds, 28 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_1689123863938_0013_000001	Sat Jul 22 20:25:16 +0300 2023	http://vm-datalake-s-2.test.local:8042	Logs	0	0

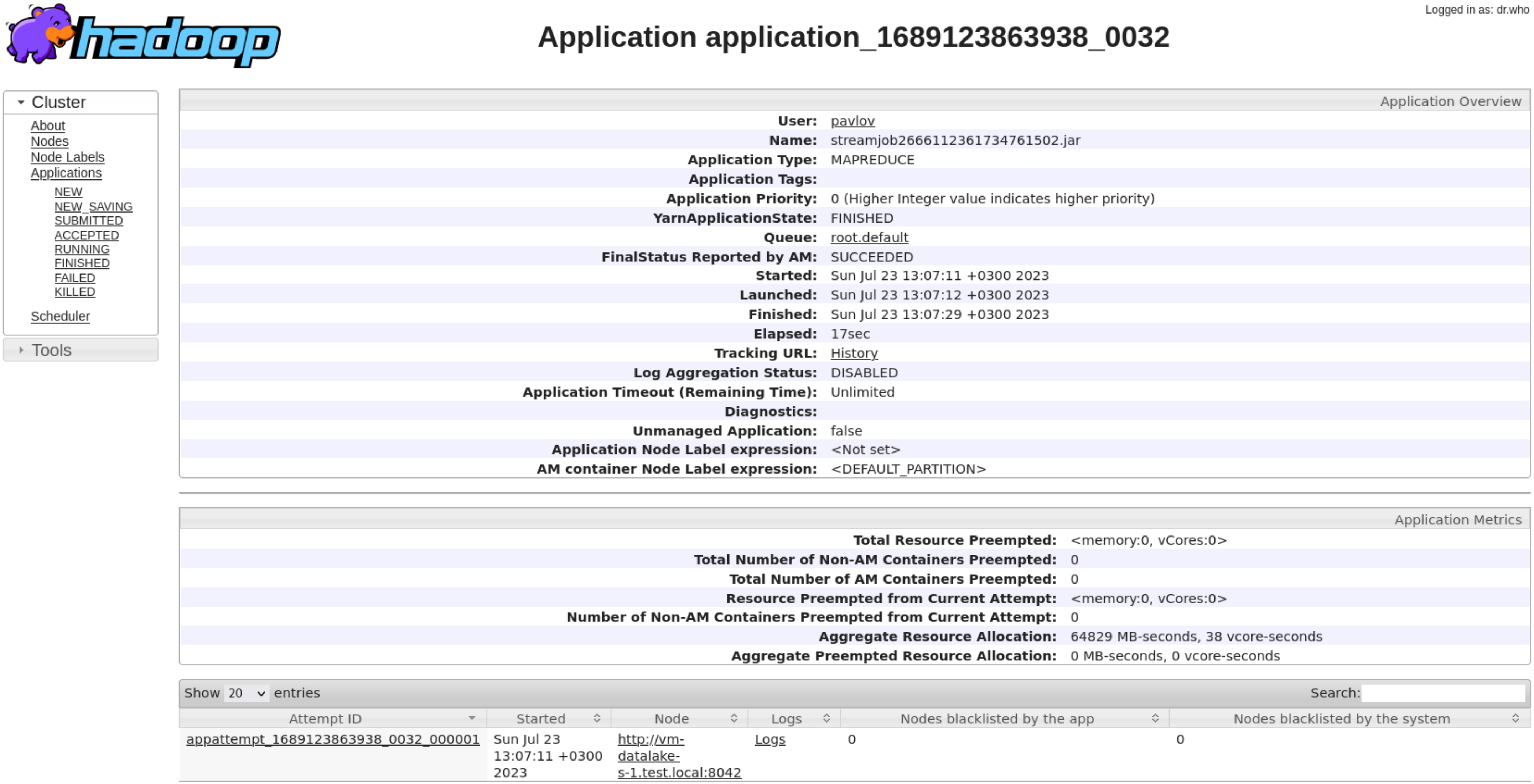
Посмотрите результат выполнения алгоритма **wordcount,** выведя на экран первых 100 строк файла из папки **./data/results**
Вывел 33 строки (иначе скриншот не помещается на лист):



```
[pavlov@vm-cli data]$ hdfs dfs -cat data/results/part-r-00000 | head -n 33
"       69
"">(1   1
"">(see 3
"">lavender     1
"">see  1
""Filling""     2
""bird""        2
""cheese""","yogurt     1
""cheese""",0.25,0.0,cup,       1
""cheese""",3.0,0.0,tablespoon, 1
""converted""   2
""fingers"""    1
""fingers""",Swiss      1
""float.""",Moscato,0.0,0.0,,   1
""forbidden     1
""http://cooking.nytimes.com/recipes/5007-mint-marinade"">recipe</a>)",mint     1
""http://www.nytimes.com/recipes/11391/homemade-butter-and-buttermilk.html"">homemade   1
""organic""     1
""prune""       1
""rice  2
""snapper""     1
""unfiltered""  1
""vermicelli""  2
""white""       1
",      8
","     4
","Glaze,       1
","boneless,    1
","butter,      1
","cherry       1
","chicken      1
","chopped      1
","fish 1
cat: Unable to write to output stream.
[pavlov@vm-cli data]$ █
```

Задание №2

Из списка выполненных Application Master задач, выберите свою выполненную задачу и приложите результат (в виде скриншота)



Logged in as: dr.who

Application application_1689123863938_0032

Application Overview

User: pavlov
Name: streamjob2666112361734761502.jar
Application Type: MAPREDUCE
Application Tags:
Application Priority: 0 (Higher Integer value indicates higher priority)
YarnApplicationState: FINISHED
Queue: root.default
FinalStatus Reported by AM: SUCCEEDED
Started: Sun Jul 23 13:07:11 +0300 2023
Launched: Sun Jul 23 13:07:12 +0300 2023
Finished: Sun Jul 23 13:07:29 +0300 2023
Elapsed: 17sec
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>

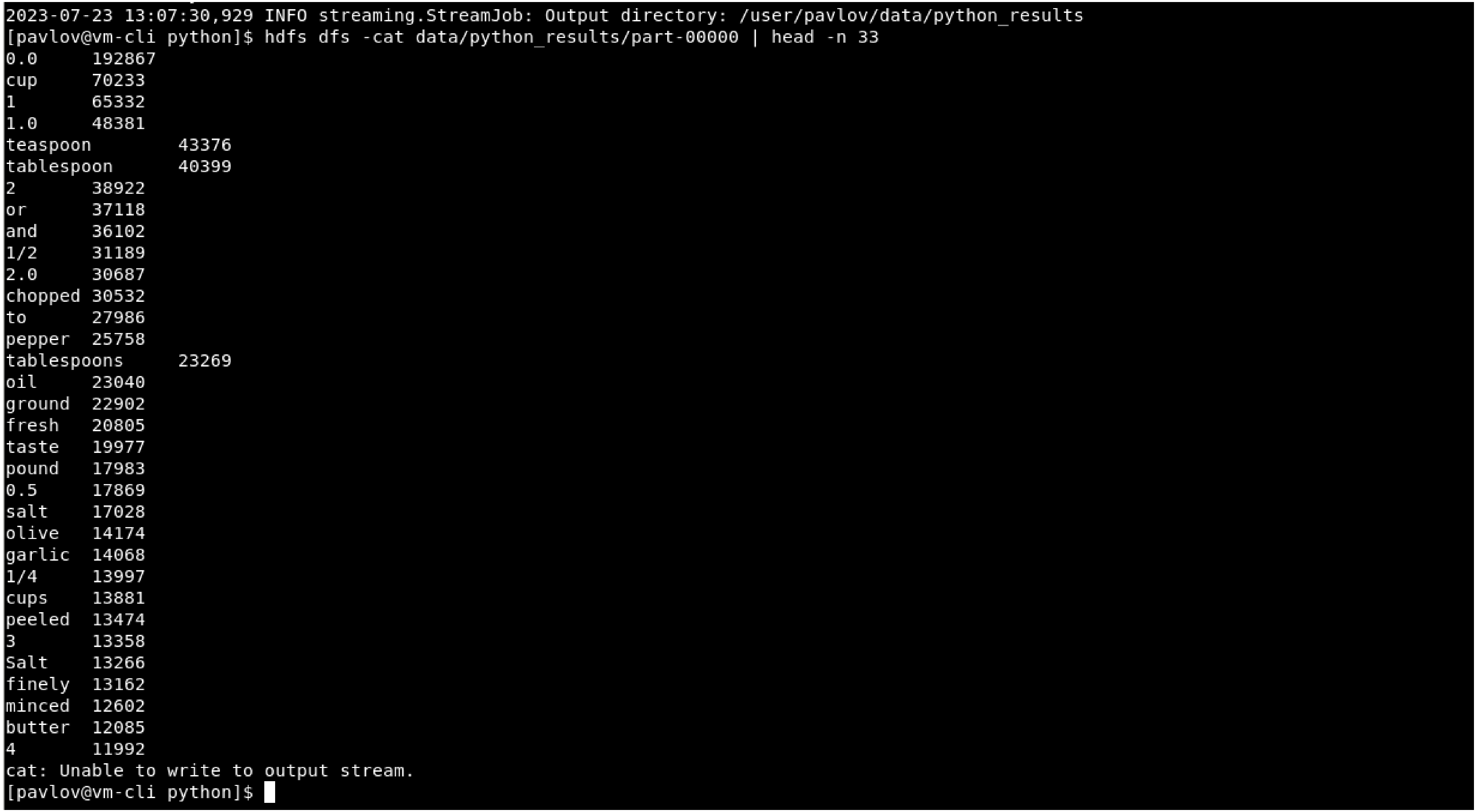
Application Metrics

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: 64829 MB-seconds, 38 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_1689123863938_0032_000001	Sun Jul 23 13:07:11 +0300 2023	http://vm-datalake-s-1.test.local:8042	Logs	0	0

Посмотрите результат выполнения вашего алгоритма**,** выведя на экран первых 100 строк файла из папки **./data/python_results**
Вывел 33 строки (иначе скриншот не помещается на лист):



```
2023-07-23 13:07:30,929 INFO streaming.StreamJob: Output directory: /user/pavlov/data/python_results
[pavlov@vm-cli python]$ hdfs dfs -cat data/python_results/part-00000 | head -n 33
0.0     192867
cup     70233
1       65332
1.0     48381
teaspoon        43376
tablespoon      40399
2       38922
or      37118
and     36102
1/2     31189
2.0     30687
chopped 30532
to      27986
pepper  25758
tablespoons     23269
oil     23040
ground  22902
fresh   20805
taste   19977
pound   17983
0.5     17869
salt    17028
olive   14174
garlic  14068
1/4     13997
cups    13881
peeled  13474
3       13358
Salt    13266
finely  13162
minced  12602
butter  12085
4       11992
cat: Unable to write to output stream.
[pavlov@vm-cli python]$
```