作业6 倒排索引

朱彤轩 191840376

1 代码编写思路

“单词忽略大小写，忽略标点符号（punctuation.txt），忽略停词（stop-word-list.txt），忽略数字，单词长度>=3”等要求在作业5中已经实现，在这里不做赘述。下面介绍整个Mapreduce程序的结构：

1.1 Mapper-InvertedIndexerMapper

此mapper接受默认的输入，输出`<word#filename, 1>`，与作业5中的实现思路类似。

1.2 Combiner-SumCombiner

使用Combiner将Mapper的输出结果中value部分的词频进行统计，输出`<word#filename, count>`。

1.3 Partitioner-NewPartitioner

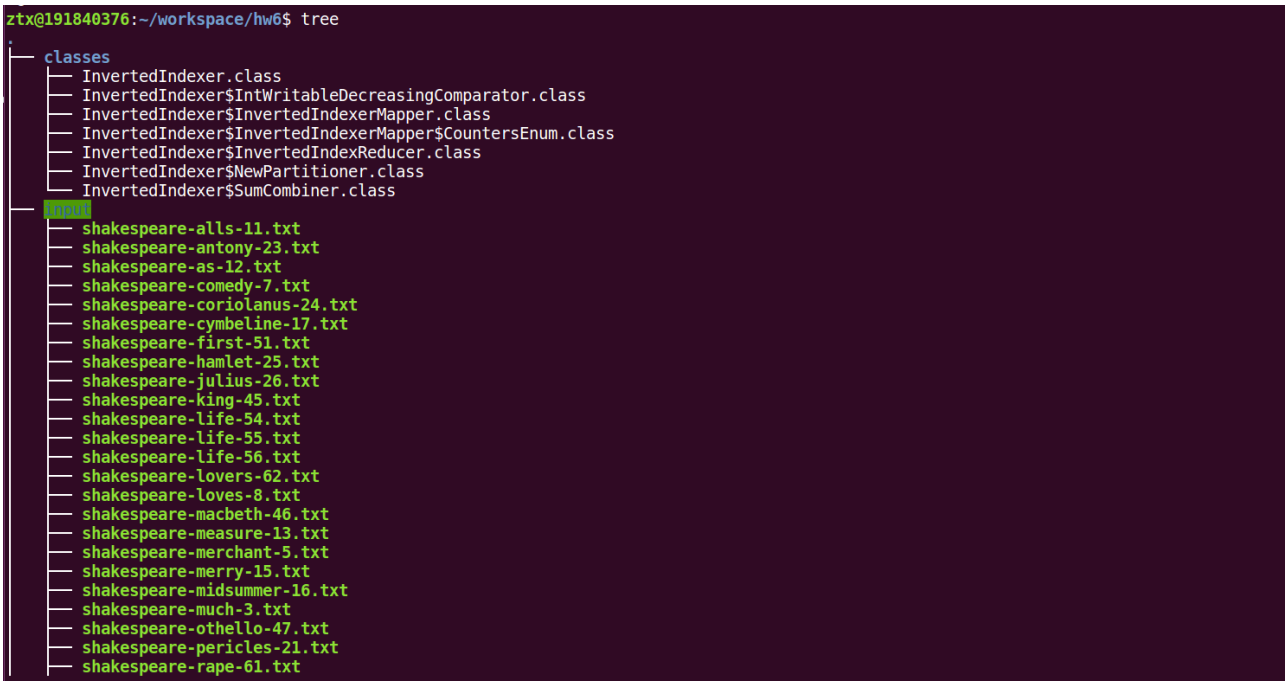
自定义HashPartitioner，保证`word#filename`格式的key值按照word分发给Reducer，保证同样的word在同一reducer下。接下来系统帮忙将所有的**key**按照字母序排序。

1.4-Reducer-InvertedIndexReducer

定义一个列表存储同一个word的filename和count。我选择按照`count#filename`的格式存储进列表，这样方便在访问完该单词的所有value之后，单词的索引按照单词在该文档中出现的次数从大到小排序。

逆序排序之后按照输出格式写入文件。

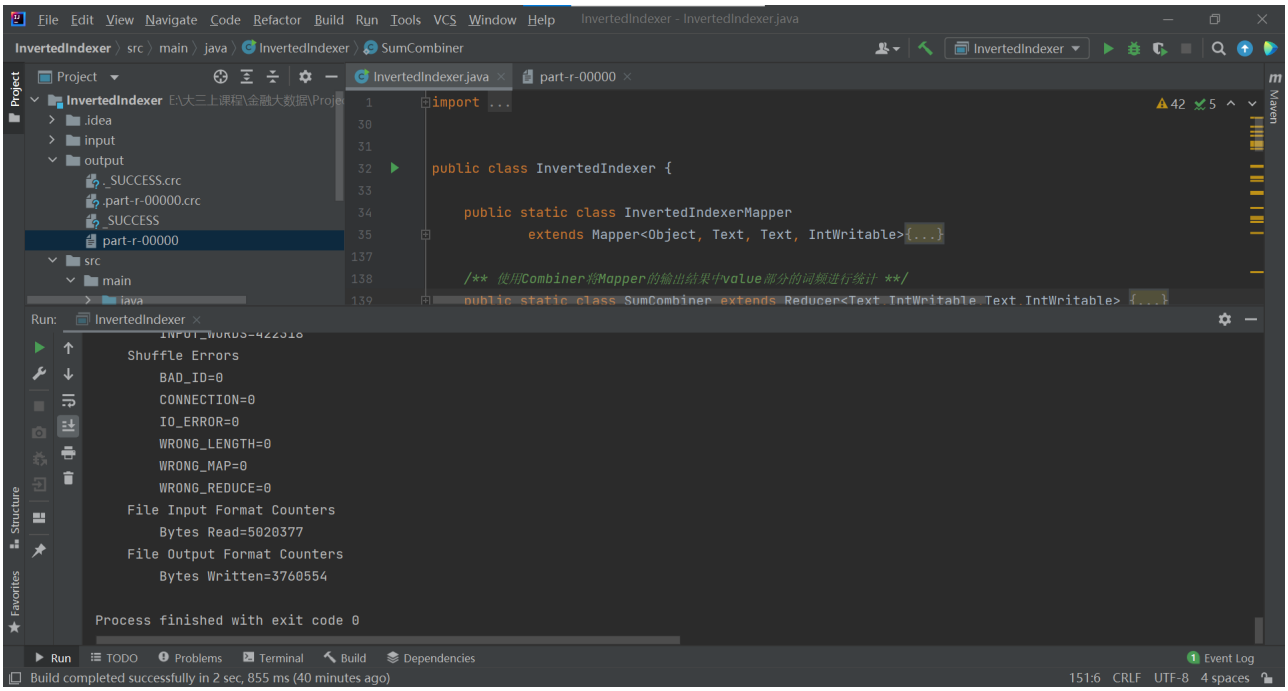
2 文件夹目录结构

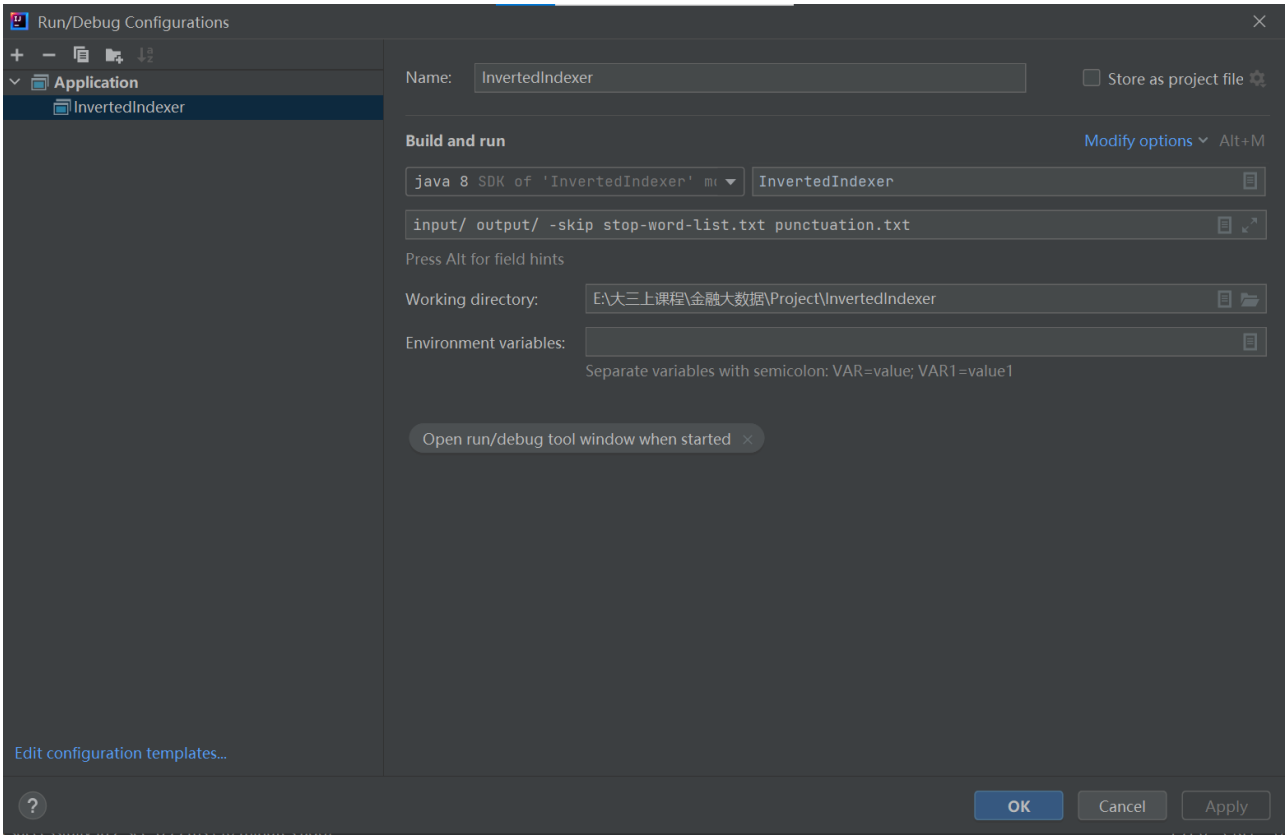


- classes：各个类
- input：40个莎士比亚作品txt
- output：倒排索引的结果输出
- skip：标点以及停用词存放
- src：源码

3 实验结果

3.1 windows系统下运行截图





3.2 Linux系统下运行截图

文件编译，文件上传至hdfs文件系统，在终端运行代码：

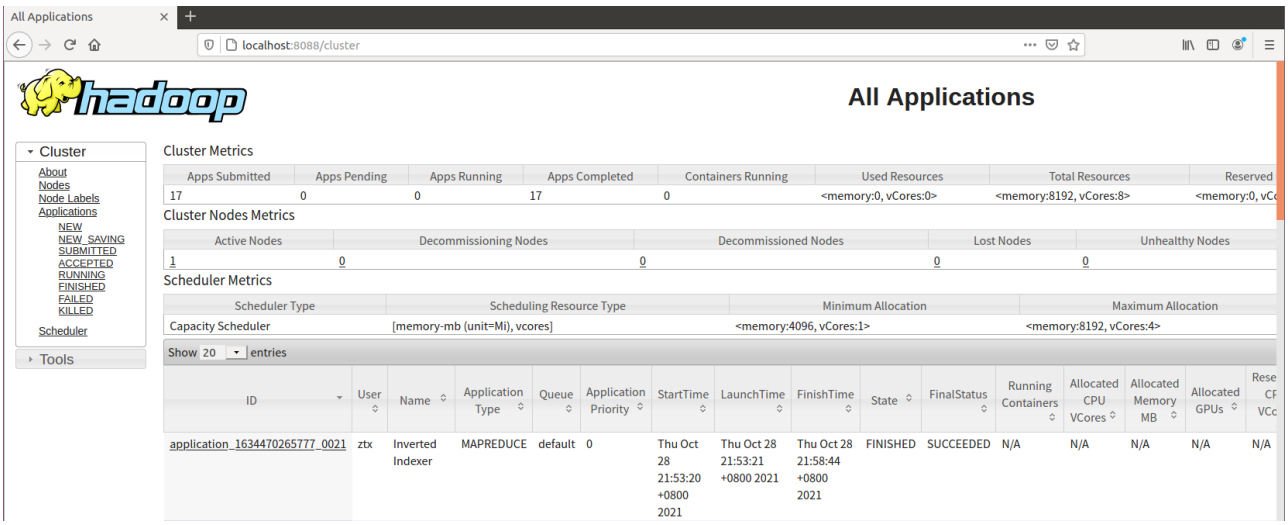
```
ztx@191840376:~/workspace/hw6$ hadoop jar invertedindexer.jar
InvertedIndexer /input /output -skip /skip/stop-word-list.txt
/skip/punctuation.txt
```

```
ztx@191840376:~/workspace/hw6$ javac -classpath /opt/hadoop-installs/hadoop-3.2.2/share/hadoop/common/hadoop-common-3.2.2.jar:/opt/hadoop-installs/hadoop-3.2.2/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.2.2.jar:/opt/hadoop-installs/hadoop-3.2.2/share/hadoop/common/lib/commons-cli-1.2.jar -d classes/ src/*.java -Xlint
/opt/hadoop-installs/hadoop-3.2.2/share/hadoop/common/hadoop-common-3.2.2.jar(org/apache/hadoop/fs/Path.class): warning: Cannot find annotation method 'value()' in type 'LimitedPrivate': class file for org.apache.hadoop.classification.InterfaceAudience not found
1 warning
ztx@191840376:~/workspace/hw6$ jar -cvf invertedindexer.jar classesadded manifest
adding: classes/(in = 0) (out= 0)(stored 0%)
adding: classes/InvertedIndexer$NewPartitioner.class(in = 997) (out= 514)(deflated 48%)
adding: classes/InvertedIndexer$InvertedIndexerMapper.class(in = 5365) (out= 2397)(deflated 55%)
adding: classes/InvertedIndexer$InvertedIndexReducer.class(in = 4013) (out= 1765)(deflated 56%)
adding: classes/InvertedIndexer$InvertedIndexerMapper$CountersEnum.class(in = 1123) (out= 512)(deflated 54%)
adding: classes/InvertedIndexer$SumCombiner.class(in = 1753) (out= 753)(deflated 57%)
adding: classes/InvertedIndexer$IntWritableDecreasingComparator.class(in = 639) (out= 352)(deflated 44%)
adding: classes/InvertedIndexer.class(in = 2806) (out= 1472)(deflated 47%)
ztx@191840376:~/workspace/hw6$ hdfs dfs -put skip /skip
ztx@191840376:~/workspace/hw6$ hdfs dfs -put input /input
ztx@191840376:~/workspace/hw6$ hadoop jar invertedindexer.jar InvertedIndexer /input /output -skip /skip/stop-word-list.txt /skip/punctuation.txt
2021-10-28 21:53:13,687 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-10-28 21:53:17,320 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ztx/.staging/job_1634470265777_0021
2021-10-28 21:53:18,430 INFO input.FileInputFormat: Total input files to process : 40
2021-10-28 21:53:18,745 INFO mapreduce.JobSubmitter: number of splits:40
2021-10-28 21:53:19,421 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634470265777_0021
2021-10-28 21:53:19,424 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-28 21:53:20,288 INFO conf.Configuration: resource-types.xml not found
2021-10-28 21:53:20,299 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-28 21:53:20,617 INFO impl.YarnClientImpl: Submitted application application_1634470265777_0021
2021-10-28 21:53:20,780 INFO mapreduce.Job: The url to track the job: http://11.111.64.120:8088/proxy/application_1634470265777_0021/
2021-10-28 21:53:20,789 INFO mapreduce.Job: Running job: job_1634470265777_0021
2021-10-28 21:53:40,488 INFO mapreduce.Job: Job job_1634470265777_0021 running in uber mode : false
2021-10-28 21:53:40,496 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-28 21:53:55,864 INFO mapreduce.Job:  map 3% reduce 0%
2021-10-28 21:54:10,116 INFO mapreduce.Job:  map 5% reduce 0%
2021-10-28 21:54:23,498 INFO mapreduce.Job:  map 8% reduce 0%
2021-10-28 21:54:30,792 INFO mapreduce.Job:  map 10% reduce 0%
2021-10-28 21:54:37,859 INFO mapreduce.Job:  map 13% reduce 0%
```

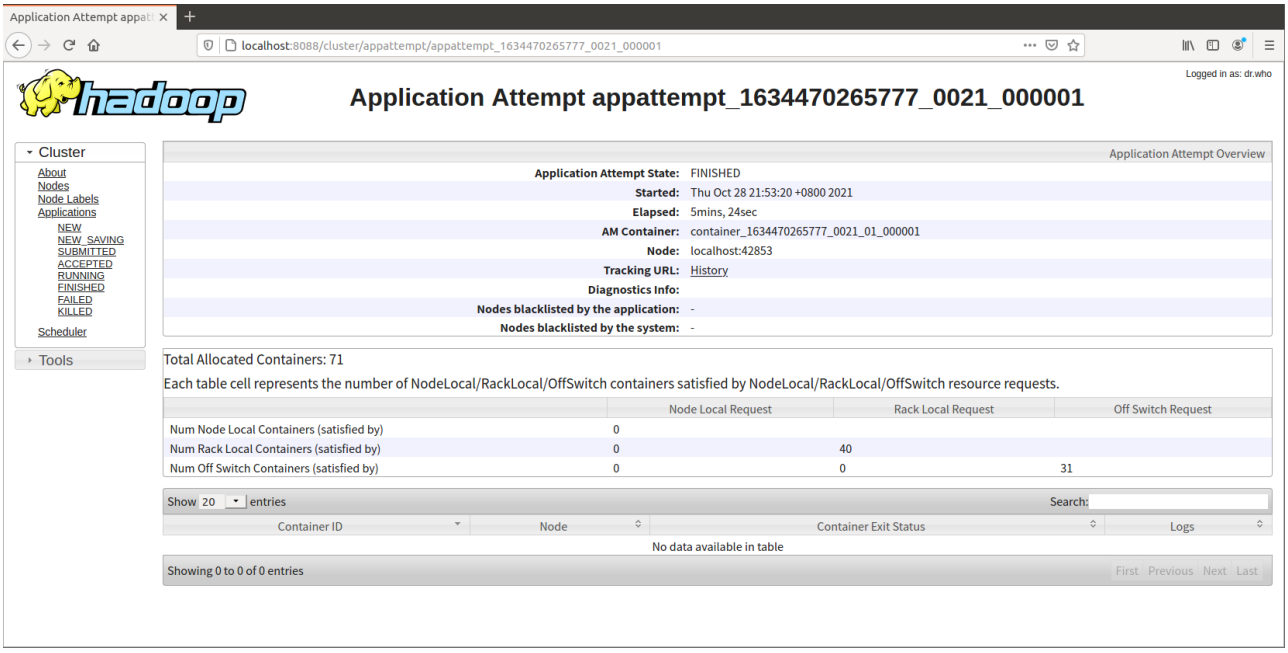
结果输出：

```
aaron: shakespeare-titus-50.txt#98
abaissiez: shakespeare-life-54.txt#1
abandon: shakespeare-as-12.txt#4, shakespeare-twelfth-20.txt#1, shakespeare-troilus-22.txt#1, shakespeare-timon-49.txt#1, shakespeare-third-53.txt#1, shakespeare-taming-2.txt#1, shakespeare-othello-47.txt#1
abandoned: shakespeare-titus-50.txt#1, shakespeare-alls-11.txt#1
abase: shakespeare-tragedy-58.txt#1
abash: shakespeare-troilus-22.txt#1
abate: shakespeare-life-54.txt#5, shakespeare-venus-60.txt#1, shakespeare-tragedy-58.txt#1, shakespeare-titus-50.txt#1, shakespeare-taming-2.txt#1, shakespeare-romeo-48.txt#1, shakespeare-midsummer-16.txt#1, shakespeare-merchant-5.txt#1, shakespeare-loves-8.txt#1, shakespeare-hamlet-25.txt#1, shakespeare-cymbeline-17.txt#1
abated: shakespeare-second-52.txt#1, shakespeare-king-45.txt#1, shakespeare-coriolanus-24.txt#1
abatement: shakespeare-twelfth-20.txt#1, shakespeare-king-45.txt#1, shakespeare-cymbeline-17.txt#1
abatements: shakespeare-hamlet-25.txt#1
abates: shakespeare-tempest-4.txt#1
abbess: shakespeare-comedy-7.txt#8
abbey: shakespeare-comedy-7.txt#9, shakespeare-life-56.txt#3, shakespeare-two-18.txt#2, shakespeare-life-55.txt#2, shakespeare-second-52.txt#1, shakespeare-romeo-48.txt#1
abbeys: shakespeare-life-56.txt#1
abbominable: shakespeare-loves-8.txt#1
abbot: shakespeare-tragedy-57.txt#8, shakespeare-life-55.txt#2
abbots: shakespeare-life-56.txt#1
abbreviated: shakespeare-loves-8.txt#1
abed: shakespeare-twelfth-20.txt#1, shakespeare-coriolanus-24.txt#1, shakespeare-as-12.txt#1, shakespeare-alls-11.txt#1
abel: shakespeare-tragedy-57.txt#1
abergavenny: shakespeare-life-55.txt#11
abet: shakespeare-tragedy-57.txt#1
abetting: shakespeare-comedy-7.txt#1
abettor: shakespeare-rape-61.txt#1
abhominable: shakespeare-loves-8.txt#1
abhor: shakespeare-othello-47.txt#3, shakespeare-timon-49.txt#2, shakespeare-sonnets-59.txt#2, shakespeare-rape-61.txt#2, shakespeare-measure-13.txt#2, shakespeare-life-55.txt#2, shakespeare-venus-60.txt#1, shakespeare-much-3.txt#1, shakespeare-merry-15.txt#1, shakespeare-loves-8.txt#1, shakespeare-life-56.txt#1, shakespeare-cymbeline-17.txt#1, shakespeare-coriolanus-24.txt#1, shakespeare-comedy-7.txt#1, shakespeare-as-12.txt#1
abhorr: shakespeare-timon-49.txt#2, shakespeare-coriolanus-24.txt#2, shakespeare-winters-19.txt#1, shakespeare-troilus-22.txt#1, shakespeare-tempest-4.txt#1, shakespeare-measure-13.txt#1, shakespeare-life-56.txt#1, shakespeare-king-45.txt#1, shakespeare-cymbeline-17.txt#1
```

All Applications截图：


All Applications
localhost:8088/cluster
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
17 | 0 | 0 | 17 | 0 | <memory:0, vCores:0> | <memory:8192, vCores:8> | <memory:0, vC
Cluster Nodes Metrics
Active Nodes | Decommissioning Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes
1 | 0 | 0 | 0 | 0
Scheduler Metrics
Scheduler Type | Scheduling Resource Type | Minimum Allocation | Maximum Allocation
Capacity Scheduler | [memory-mb (unit=Mi), vcores] | <memory:4096, vCores:1> | <memory:8192, vCores:4>
Show 20 entries
ID | User | Name | Application Type | Queue | Application Priority | StartTime | LaunchTime | FinishTime | State | FinalStatus | Running Containers | Allocated CPU VCores | Allocated Memory MB | Allocated GPUs | Rese CF VCc
application_1634470265777_0021 | ztx | Inverted Indexer | MAPREDUCE | default | 0 | Thu Oct 28 21:53:20 +0800 2021 | Thu Oct 28 21:53:21 +0800 2021 | Thu Oct 28 21:58:44 +0800 2021 | FINISHED | SUCCEEDED | N/A | N/A | N/A | N/A | N/A

job截图：


Application Attempt appatt
localhost:8088/cluster/appattempt/appattempt_1634470265777_0021_000001
Logged in as: dr.who
Application Attempt appattempt_1634470265777_0021_000001
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Application Attempt Overview
Application Attempt State: FINISHED
Started: Thu Oct 28 21:53:20 +0800 2021
Elapsed: 5mins, 24sec
AM Container: container_1634470265777_0021_01_000001
Node: localhost:42853
Tracking URL: History
Diagnostics Info:
Nodes blacklisted by the application: -
Nodes blacklisted by the system: -
Total Allocated Containers: 71
Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.
Node Local Request | Rack Local Request | Off Switch Request
Num Node Local Containers (satisfied by) | 0
Num Rack Local Containers (satisfied by) | 0 | 40
Num Off Switch Containers (satisfied by) | 0 | 0 | 31
Show 20 entries
Search:
Container ID | Node | Container Exit Status | Logs
No data available in table
Showing 0 to 0 of 0 entries
First Previous Next Last