MyKMeans

- 实现基于Hadoop的KMeans算法

运行说明

- 输入参数
  - cluster_number：聚类数量
  - iterate_number：迭代次数
  - input_path：输入路径，该路径下可以有多个文件
  - output_path：输出路径（再次执行时要确保该目录尚不存在）

  ```
    hadoop jar target/mykmeans-1.0.jar cluster_number iterate_number
  input_path output_path
  ```

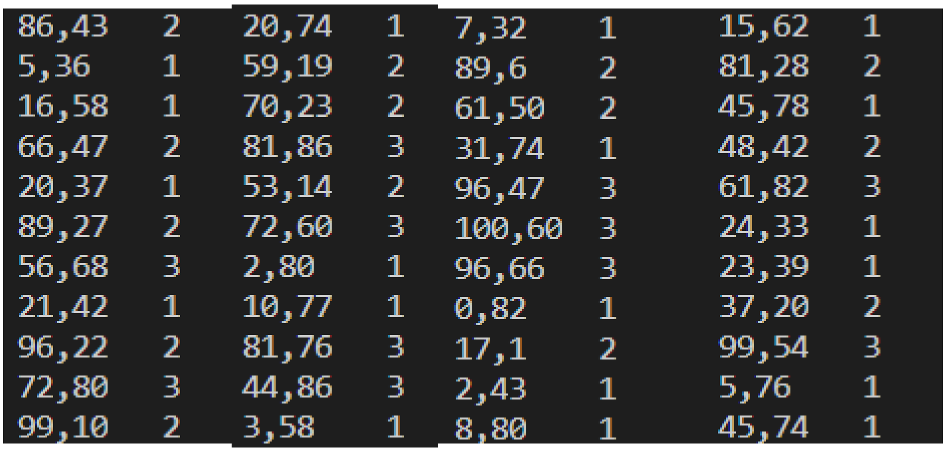
设计思路

- 每个map节点读取上一次迭代生成的cluster centers，判断自己节点上的数据归属于哪个cluster
- reduce节点计算每个cluster的数据点，计算出新的cluster centers
- 项目结构设计
  - Instance.java：以ArrayList存放数据点的各个分量，对应文件中原始数据点的格式。边写加法、乘法、除法函数用于计算簇中心。
  - Cluster.java：记录簇的信息，包括id、数据点个数、簇中心
  - KMeans.java：实现KMeans算法。mapper读取每个数据点，通过计算欧氏距离，选择距离最小的簇中心，并输出分类结果；combiner计算新的簇中心；reducer将计算结果进行汇总，计算全局的簇中心。
  - KMeansCluster.java：在最终产生结果后，再对输入文件中的所有实例进行分簇，最后把实例按照（实例，簇id）的方式写入结果文件
  - KMeansDriver.java：启动MapReduce，读取参数
  - RandomClusterGenerator.java：随机生成簇中心
  - Utils：计算距离

运行情况

输出结果

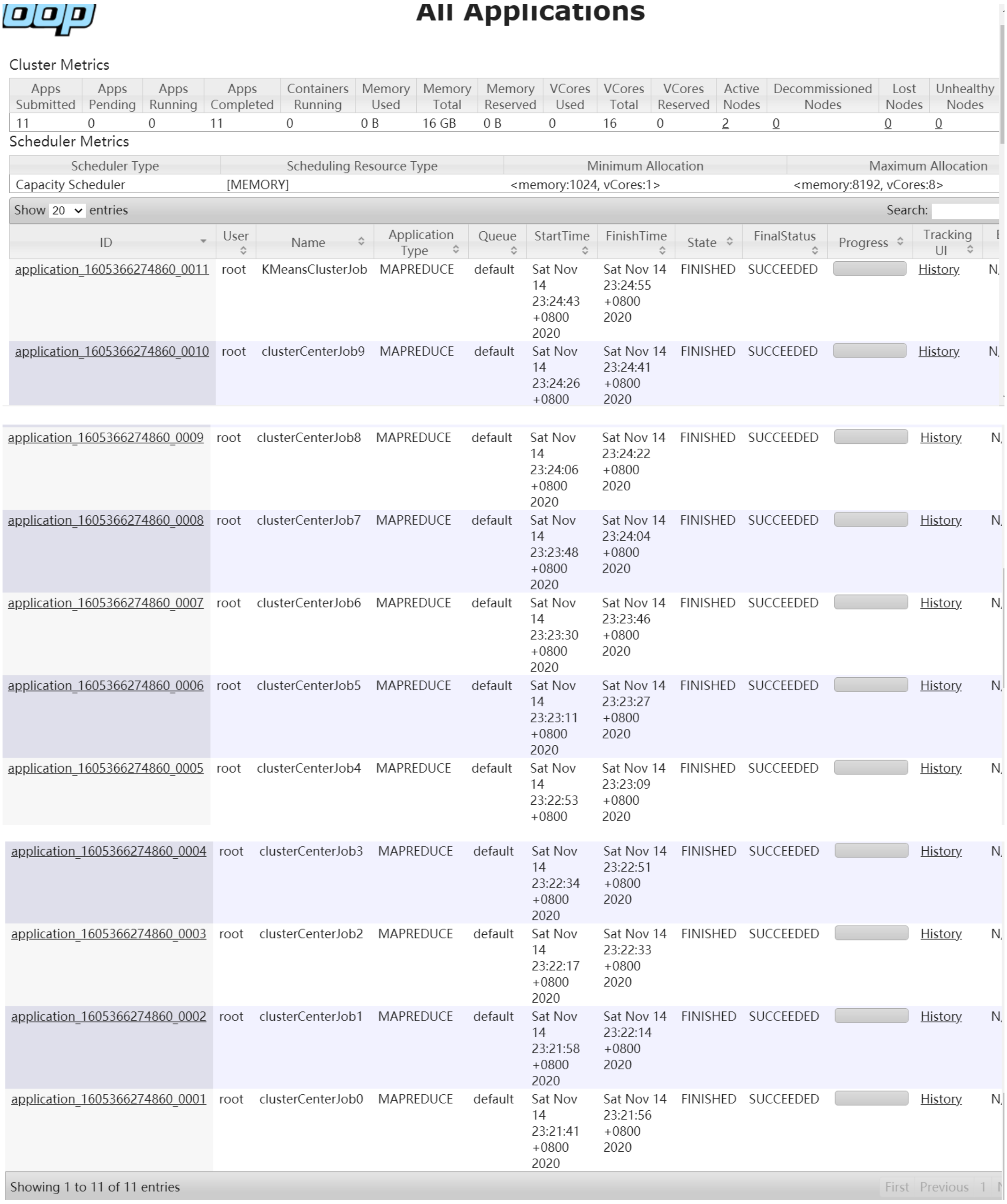
（仅截取部分）



```
86,43    2    20,74    1    7,32     1        15,62    1
5,36     1    59,19    2    89,6     2        81,28    2
16,58    1    70,23    2    61,50    2        45,78    1
66,47    2    81,86    3    31,74    1        48,42    2
20,37    1    53,14    2    96,47    3        61,82    3
89,27    2    72,60    3    100,60   3        24,33    1
56,68    3    2,80     1    96,66    3        23,39    1
21,42    1    10,77    1    0,82     1        37,20    2
96,22    2    81,76    3    17,1     2        99,54    3
72,80    3    44,86    3    2,43     1        5,76     1
99,10    2    3,58     1    8,80     1        45,74    1
```

监控

Yarn



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes
11	0	0	11	0	0 B	16 GB	0 B	0	16	0	2	0	0	0

Scheduler Metrics

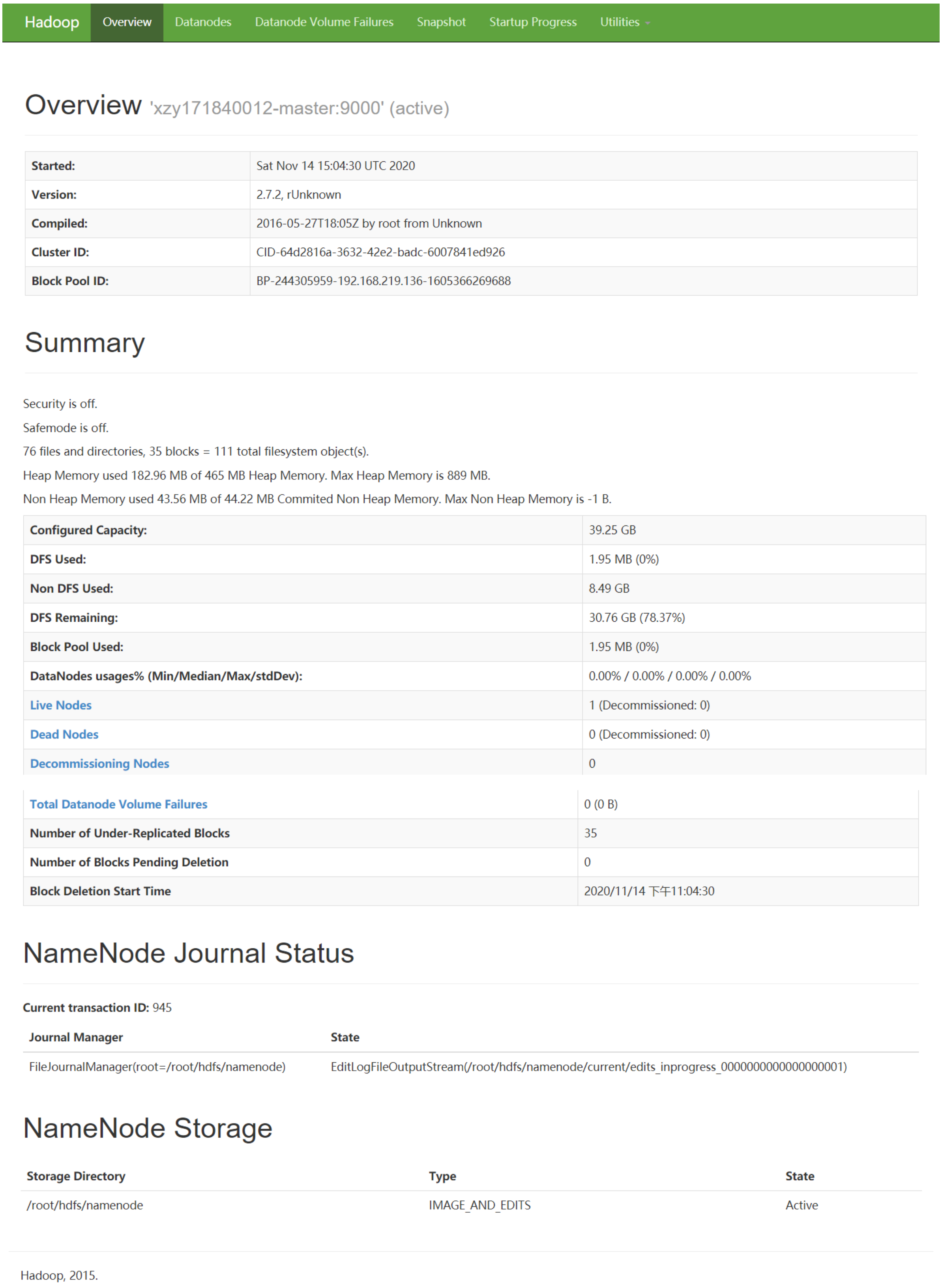
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:8>

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	
application_1605366274860_0011	root	KMeansClusterJob	MAPREDUCE	default	Sat Nov 14 23:24:43 +0800 2020	Sat Nov 14 23:24:55 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0010	root	clusterCenterJob9	MAPREDUCE	default	Sat Nov 14 23:24:26 +0800	Sat Nov 14 23:24:41 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0009	root	clusterCenterJob8	MAPREDUCE	default	Sat Nov 14 23:24:06 +0800 2020	Sat Nov 14 23:24:22 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0008	root	clusterCenterJob7	MAPREDUCE	default	Sat Nov 14 23:23:48 +0800 2020	Sat Nov 14 23:24:04 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0007	root	clusterCenterJob6	MAPREDUCE	default	Sat Nov 14 23:23:30 +0800 2020	Sat Nov 14 23:23:46 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0006	root	clusterCenterJob5	MAPREDUCE	default	Sat Nov 14 23:23:11 +0800 2020	Sat Nov 14 23:23:27 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0005	root	clusterCenterJob4	MAPREDUCE	default	Sat Nov 14 23:22:53 +0800	Sat Nov 14 23:23:09 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0004	root	clusterCenterJob3	MAPREDUCE	default	Sat Nov 14 23:22:34 +0800 2020	Sat Nov 14 23:22:51 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0003	root	clusterCenterJob2	MAPREDUCE	default	Sat Nov 14 23:22:17 +0800 2020	Sat Nov 14 23:22:33 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0002	root	clusterCenterJob1	MAPREDUCE	default	Sat Nov 14 23:21:58 +0800 2020	Sat Nov 14 23:22:14 +0800 2020	FINISHED	SUCCEEDED		History	N,
application_1605366274860_0001	root	clusterCenterJob0	MAPREDUCE	default	Sat Nov 14 23:21:41 +0800 2020	Sat Nov 14 23:21:56 +0800 2020	FINISHED	SUCCEEDED		History	N,

Showing 1 to 11 of 11 entries

HDFS



Overview 'xzy171840012-master:9000' (active)

Started:	Sat Nov 14 15:04:30 UTC 2020
Version:	2.7.2, rUnknown
Compiled:	2016-05-27T18:05Z by root from Unknown
Cluster ID:	CID-64d2816a-3632-42e2-badc-6007841ed926
Block Pool ID:	BP-244305959-192.168.219.136-1605366269688

Summary

Security is off.

Safemode is off.

76 files and directories, 35 blocks = 111 total filesystem object(s).

Heap Memory used 182.96 MB of 465 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 43.56 MB of 44.22 MB Commited Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	39.25 GB
DFS Used:	1.95 MB (0%)
Non DFS Used:	8.49 GB
DFS Remaining:	30.76 GB (78.37%)
Block Pool Used:	1.95 MB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	35
Number of Blocks Pending Deletion	0
Block Deletion Start Time	2020/11/14 下午11:04:30

NameNode Journal Status

Current transaction ID: 945

Journal Manager	State
FileJournalManager(root=/root/hdfs/namenode)	EditLogFileOutputStream(/root/hdfs/namenode/current/edits_inprogress_0000000000000000001)

NameNode Storage

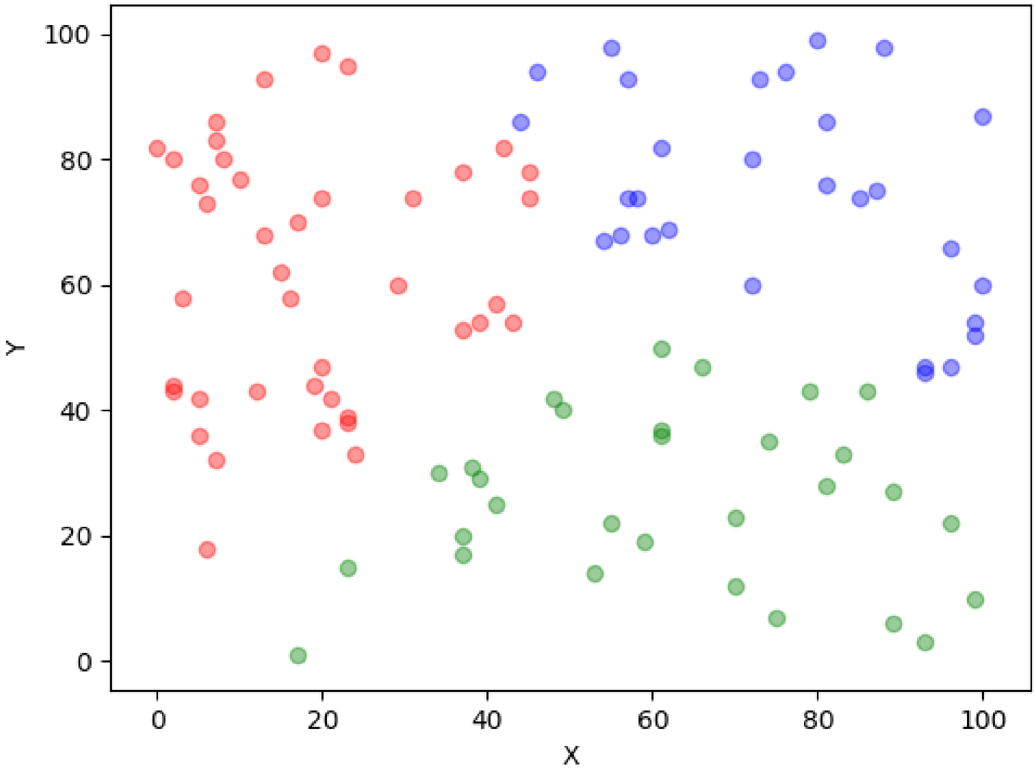
Storage Directory	Type	State
/root/hdfs/namenode	IMAGE_AND_EDITS	Active

Hadoop, 2015.

可视化

通过python的matplotlib包进行可视化，代码见visualization文件夹。

对于3个簇，10次迭代的聚类，可视化结果如下：


100
80
60
40
20
0
Y
0
20
40
60
80
100
X