实验4：K-Means聚类算法

实验内容与要求

- 实验内容：请实现课堂上介绍的“K-Means聚类算法”。
  - 给定的数据集为要进行聚类的向量。
  - 给定初始的聚类中心和聚类的簇数。
  - 要得到的输出为K-Means聚类算法收敛后的聚类中心。
  - 为了实现该算法，需要在 Driver 中控制 MapReduce 任务的迭代。
- 输出文件格式要求：聚类中心ID和聚类中心向量用制表符（\TAB）分隔，向量的各分量以西文逗号（,）分隔。如下图所示：

```
0   4.49367,4.501744,4.49628,4.49701,4.4969,4.500036,4.500506,4.491124,4.499306,4.503578
1   14.50264,14.500742,14.495192,14.506466,14.495812,14.50276,14.497252,14.501858,14.496668,14.497848
2   24.501768,24.50761,24.50257,24.495495,24.497847,24.506466,24.502674,24.507177,24.501446,24.507666
```

- 选做内容：将所属不同聚类的数据划分到不同的文件中，即产生 k 个输出文件。每个文件对应一个簇，文件中存储属于该簇的数据。
- 其他具体要求请见“本科教学支撑平台”。

实验数据

- 输入文件内容：本次实验的数据包括一个存储要聚类向量的数据集和一个存储初始聚类中心的文件，如下图所示。

```
0: 6 4 5 1 8 2 3 8 5 7 3 5 8 5 6
1: 2 1 0 5 1 4 0 3 7 8 1 2 1 1 1
2: 4 4 8 1 2 0 5 2 1 8 7 4 2 2 0
3: 9 9 7 1 5 9 6 4 6 8 0 5 2 9 8
4: 0 5 1 9 4 2 5 1 9 3 1 7 4 9 4
5: 3 0 6 3 3 7 8 1 2 5 4 0 7 1 3
6: 2 1 8 5 6 5 7 6 7 3 8 6 6 7 3
7: 1 6 0 9 7 7 2 1 4 2 9 1 7 3 4
8: 4 1 0 9 9 8 4 1 8 4 4 6 9 7 2
9: 8 2 4 4 7 5 9 3 6 2 7 0 1 1 4
```

```
0: 4,6,4,11,4,7,5,7,3,11,5,3,9,8,3
1: 14,19,13,17,14,16,15,22,17,21,13,21,18,17,22
2: 24,23,32,31,24,31,30,25,30,27,24,31,32,26,25
3: 37,40,36,33,41,34,38,41,36,35,40,33,37,39,38
4: 52,47,45,45,50,43,44,45,51,51,43,44,49,44,52
5: 62,57,57,58,60,55,60,61,53,55,56,57,62,62,55
6: 64,67,67,64,63,72,63,65,65,67,69,71,63,63,63
7: 74,82,74,73,75,78,76,74,73,77,80,81,77,78,81
8: 90,85,83,84,84,90,86,88,86,92,87,85,85,86,89
9: 98,93,100,96,97,101,99,102,96,95,97,98,95,94,98
```

- 其中，数据集中包含 2,000,000 条数据，数据集中每条数据表示一个向量。每个向量的维度为15，而“:”之前为该数据条目的ID。
- 单机测试样例：提供部分数据作为单机测试样例，可在“本科教学支撑平台”中下载。该数据集主要供本地调试使用。
- 全部数据集：全部数据集位于集群的 HDFS 存储上，HDFS 存储位置为：hdfs://master001:9000/data/2022s/kmeans
- 注意：HDFS 上数据集的文件名为 dataset.data，与测试样例中的文件名不同。提交前请注意修改。

实验报告提交要求

- 实验报告要求提交一个压缩包，除了包含源代码、JAR包、JAR包执行方式说明，还需要包含一个实验报告。实验报告中包含：
  - Map 和 Reduce 的设计思路（含 Key、Value 类型）。
  - MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
  - 输出结果文件的部分截图。并请指明输出结果文件在 HDFS 上的路径。
  - 试分析算法可能存在的不足和可能的改进之处（如性能、可扩展性等）。
  - 请在报告中包含在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容。每个MapReduce Job对应一个执行报告，报告提交要求与Hive Join试验相同。

实验报告提交要求

- 实验报告文件命名规则： MPLab4-小组编号-组长姓名.docx
- 实验报告提交至“本科教学支撑平台”： http://cslabcms.nju.edu.cn/
- 实验提交截止日期：5月31日（包含当天）