**GROUP NUMBER:** 007

Recall that groups with odd (resp. even) number can use the cluster only in odd (resp. even), days.

**GROUP MEMBERS:** Alberto Pasqualetto, Michela Schibuola, Michele Sprocatti

**AVAILABLE INPUTS:** All input files are stored in the hdfs directory /data/BDC2324/ of the cluster. In particular, file artificialXM\_9\_100.csv contains a dataset of (approximately) X million points, for X=1, 10, 20, 50, 100. The points have been chosen in a random fashion, but they are roughly partitioned into 9 clusters and 100 outliers.

**TEST 1:** The goal of this test is to assess the scalability of the proposed strategy for outlier detection with respect to the number of executors. **You must fill in the following table**. *In case one run with 2 executors takes more than 10 minutes first try to optimize your code, and if it remains slow use the next smaller dataset until the run falls within the 10 minutes bound*.

Our code takes more than 10 minutes for 100M dataset (more information about this is in the next page), but we are below 10 minutes for the 50M dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SCALABILITY WITH RESPECT TO NUMBER OF EXECUTORS**  **Dataset: Artificial50M\_9\_100.csv,** **M: 10, K: 110, L: 16** | | | | |
| **Number of executors** | **Running time of R1 of MRFFT in ms**  **(average of 3 runs)** | **Running time of R2 of MRFFT in ms**  **(average of 3 runs)** | **Running time of R3 of MRFFT in ms**  **(average of 3 runs)** | **Running time of MRApproxOutliers in ms**  **(average of 3 runs)** |
| **2** | 98627.2 | 93.84 | 302881.5 | 33308.44 |
| **4** | 52479.99 | 79.98 | 153672.96 | 16990.64 |
| **8** | 26908.417 | 82.27 | 78499.81 | 9007.74 |
| **16** | 13090.33 | 108.48 | 39614.39 | 4754.29 |

**TEST 2:** The goal of this test is to assess how effective is the approximate strategy to detect the outliers when the values M and K are suitably set and the true outliers are well separated from the other points. **You must fill in the following table**.

|  |  |  |  |
| --- | --- | --- | --- |
| **EFFECTIVENESS OF THE STRATEGY**  **Dataset: Artificial10M\_9\_100.csv,** **M: 3, L: 16, Executors: 16** | | | |
| **K** | **Radius** | **Number of sure outliers** | **Number of uncertain points** |
| **50** | 8.831669208026307 | 15 | 70 |
| **70** | 6.501931097758571 | 40 | 52 |
| **90** | 3.7329465573458154 | 87 | 13 |
| **110** | 2.21580256340677 | 98 | 2 |
| **130** | 1.53291128249485 | 100 | 0 |

**GENERAL HINTS:**

* Do not include the reading of the input in your running times
* In your program, after defining the Spark Configuration variable “conf”, add the line

conf.set("spark.locality.wait", "0s");

which should force Spark to use all required executors even for small datasets.

**COMMENTS:**

In the scope of test 1, we executed the code with the 100M input, but the execution time is more than 10 minutes with 2 executors (about 15), we report here another table with such results as told us by the professor:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SCALABILITY WITH RESPECT TO NUMBER OF EXECUTORS**  **Dataset: Artificial100M\_9\_100.csv,** **M: 10, K: 110, L: 16** | | | | |
| **Number of executors** | **Running time of R1 of MRFFT in ms**  **(average of 3 runs)** | **Running time of R2 of MRFFT in ms**  **(average of 3 runs)** | **Running time of R3 of MRFFT in ms**  **(average of 3 runs)** | **Running time of MRApproxOutliers in ms**  **(average of 3 runs)** |
| **2** | 204375 | 91 | 609098 | 61857 |
| **4** | 109065 | 78 | 313098 | 31541 |
| **8** | 55052 | 67 | 154623 | 16021 |
| **16** | 26597 | 95 | 79067 | 8493 |