The main goal of your experiments is to study the performance, running time, and number of rounds of one of the MapReduce algorithms for the (k, ϵ)-coreset for the k-means problem that is presented in the class. We have learned a MPC algorithm that computes a (k, ϵ)-coreset for the k-means problem. The algorithm is for d-dimensional Euclidean spaces and needs $O(\log(n))$ rounds.

You will need to do several things:

- Decide on a more precise research question that you want to answer. Some examples for inspiration are listed below.
- Generate or find data for you experiment. Please choose datasets whose size are reasonably big, say having all these sizes 1000, 5000, 10000, 50000, 100000, 500000, and 1000000 and see how your algorithms behave when you increase the number of points. You can check how large could be the size of the point set that you can test your algorithm.
- Run experiments on your data sets using your implementation of the MPC algorithm for the (k, ϵ)-coreset for the k-means problem so you can compare the resulting clustering with the true clustering of your data and the efficiency of the algorithms.
- You present and discuss the results in your report and explain what conclusions can (or possibly cannot) be drawn from your results.
- Try to specify the number of logical machines that you use, the space per machine, and the number of rounds that your algorithm use. Make a plot of them.
- Please test your algorithm on at least 5 datasets, two of them should be real datasets or image segmentation/compression/processing.

This then goes into a report that we use to grade your experiments. On the next page you can find the rubric that we will use to grade the report so you have an idea of what to focus on.

Example research questions:

- Compare the coreset sizes and see how it effects the running time and the quality of clustering. For example, if you increase or decrease ϵ and/or k, what will happen to the size of the coreset and the quality of clustering.
- Instead of placing a grid inside the inner balls of each cluster or the annuli, you can sample x points inside each inner ball or annulus and take the union of the sample set as the coreset. The number of samples from each annuli can be a good research question. You can plot the number of samples against the quality of the clustering. You can also compare the coreset where you place grids inside annuli and the coreset where you sample points inside annuli and compare the running time and the quality of clustering for both coresets.

- You can test the coreset for image quantization, segmentation and compression. For example, you can test your coreset for Lena which is a famous picture in this scenario. See here: https://en.wikipedia.org/wiki/Lenna. For other examples, see this video: https://www.youtube.com/watch?v=oDLKN4qm_Ls. You can also look at the following papers:

    a. Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm, Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation

    b. There is a beautiful paper on k-means with lots of datasets and images segmentation. Here is the name of the paper: "An Efficient k-Means Clustering Algorithm: Analysis and Implementation". The paper is available at https://www.cs.umd.edu/~mount/Projects/KMeans/pami02.pdf. Table 1 and Figures 9 and 10 explain lots of examples. You can run your experiments on these datasets.

    c. Also, you can have a look at experiments in this paper: "Gereon Frahling, Piotr Indyk, Christian Sohler: Sampling in dynamic data streams and applications. SCG 2005: 142-149".

- Does your MapReduce algorithm converge very fast on your data sets and for what datasets do you think your MapReduce algorithm would perform better? Is there a heuristic or any other method that could reduce the communication rounds of your algorithm?
- Does your algorithm works well when the number of clusters is large (over 100)?

Data generation:

- The python library sklearn (https://scikit-learn.org/stable/datasets/index.html) has several functions that allow easy data generation. You can use this to generate many different types of data. It also contains some data-sets, but they are mainly higher-dimensional.
- I recommend using datasets at openml webpage: https://www.openml.org/search?type=data. They are labeled and there are APIs to download these datasets.

Assessment of the reports for 2IMA35 is based on an evaluation of various aspects, as listed below. The aspects are evaluated on a scale consisting of five possible scores, which can be interpreted as follows:

poor	almost satisfactory	decent	good	excellent
O	O	O	O	O

For each aspect there is a short explanation about what "weak" means, and what you need to do get the score "excellent". You can earn a maximum of 15 points for this reports based on scores for the listed aspects. Here's a rough indication of how the scores are translated to the final number of points:

- Several scores on the experiments are poor, other scores are mostly almost satisfactory ➔ final grade ≤ 5
- Most scores, in particular on the contents, are almost satisfactory ➔ final grade 6-7
- Most scores are decent, some may be almost satisfactory ➔ final grade 8-9
- Most score are good, some may be decent ➔ final grade 10-12
- Almost all scores are at least good, and most scores on the contents are excellent ➔ final grade ≥ 13

General

Appearance						
No use of sections or it is generally hard to find things in the report.	O	O	O	O	O	The report looks nice, figures fit with the text and the contents is divided well into the sections.
Writing						
The report is hard to read due to the many grammatical and spelling errors or very chaotic sentence structures.	O	O	O	O	O	The report is a pleasure to read: sentences are grammatically correct, easy to parse and follow each other naturally; there are hardly any spelling errors.

Setup

Research question						
Research question has an obvious answer or is impossible to answer.	O	O	O	O	O	Clear, non-trivial research question that fits the proposed area.
Data sets						
Chosen data sets do not help answer research question.	O	O	O	O	O	Data sets are varied and highly relevant for the research question.
Experimental Set up						
Experiments are too limited to provide any reasonable evidence towards answering research question.	O	O	O	O	O	Chosen experiments are varied and will be able to contribute strongly towards answering research question.

Experimental results

Presentation and discussion of results						
Results are missing or unreadable	O	O	O	O	O	Results are presented clearly and concisely and most important features are highlighted.
Conclusion						
Results are not used to make any conclusions about the research question	O	O	O	O	O	A conclusion is drawn from the results about the research question that is substantiated by the results.