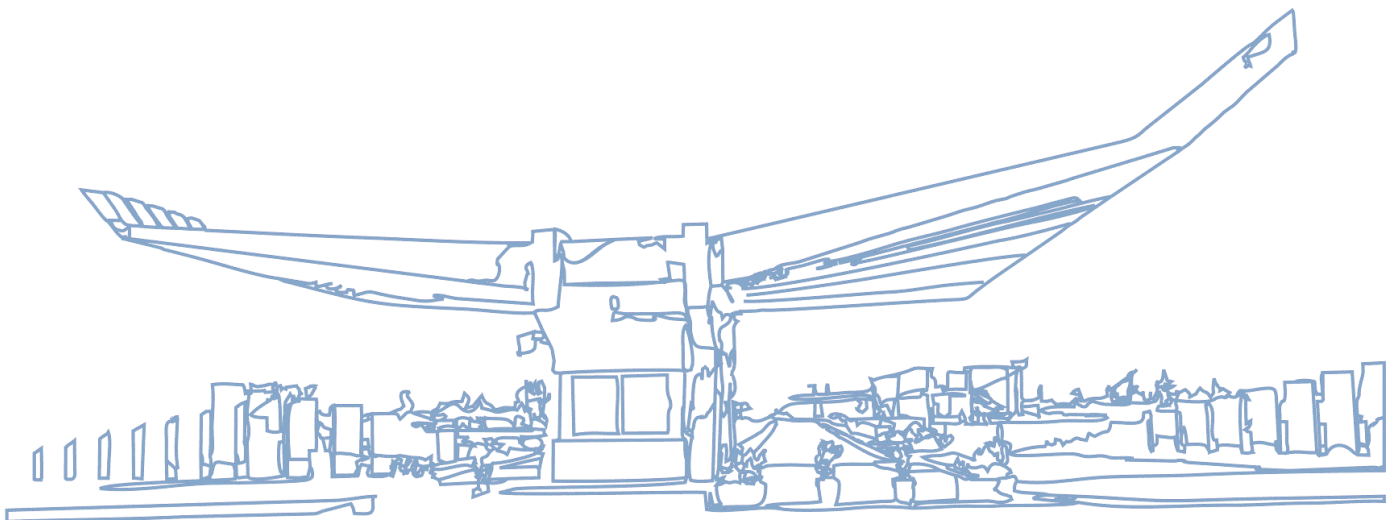# CEN 571 – Data Mining

Assignment 02 – Question 3

PREPARED:

Baftjar TABAKU

19.04.2020
Epoka University
Tirana, ALBANIA

ACCEPTED:

Prof.Dr. Arben Asllani

Question 3

Imagine that we want to analyze one terabyte (1TB) of data that is residing in a single machine with 8 input/output channels where each channel has a reading speed of 150 megabytes per second (MB/s).

1. Calculate the time it takes for the reader to read the entire file.
2. To speed up the reading operation, consider adding more machines and creating a distributed cluster. What is the minimum number of machines to be installed in the cluster so the entire read time is less than 10 seconds?

For an Entire file of 1TB (1024GB) we have a total data size of 1024GB * 1024MB = 1,048,576MB in total.

By knowing that for a single machine we have the reading speed per min 150MB/s * 8 channels * 60 seconds = 72,000 MB/min.

And The reading time for a single file would be 1,048,576MB / 72,000 MB/min =

14.56 min to read a file of 1TB, using a single machine of 150MB/s reading speed and 8 input/output channels.
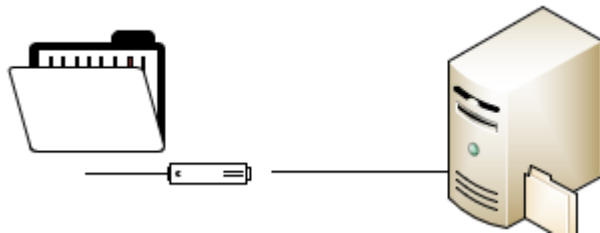
But for a limited time, amount of 10 seconds then we would add more machines, making a few calculations, we know from previous calculation that:

for a single machine we have the reading speed per min 150MB/s * 8 channels * 60 seconds = 72,000 MB/min, and for 88 machines would be 88 * 72,000 MB/min = 6,336,000 MB/min,

and for 1,048,576 MB (1TB) would be 1,048,576 MB / 6,336,000 MB/min = 0.1654 min, and converting them to seconds would be 0.1654 min * 60 seconds = 9.924.

And the more efficient solution would be with 88 machines, but this may vary on conditions and needs, in this case we need it transferred within 10 seconds.

By repeating the same calculation for 87 machines would be 10.043 seconds, for 89 machines would be 9.81 seconds.