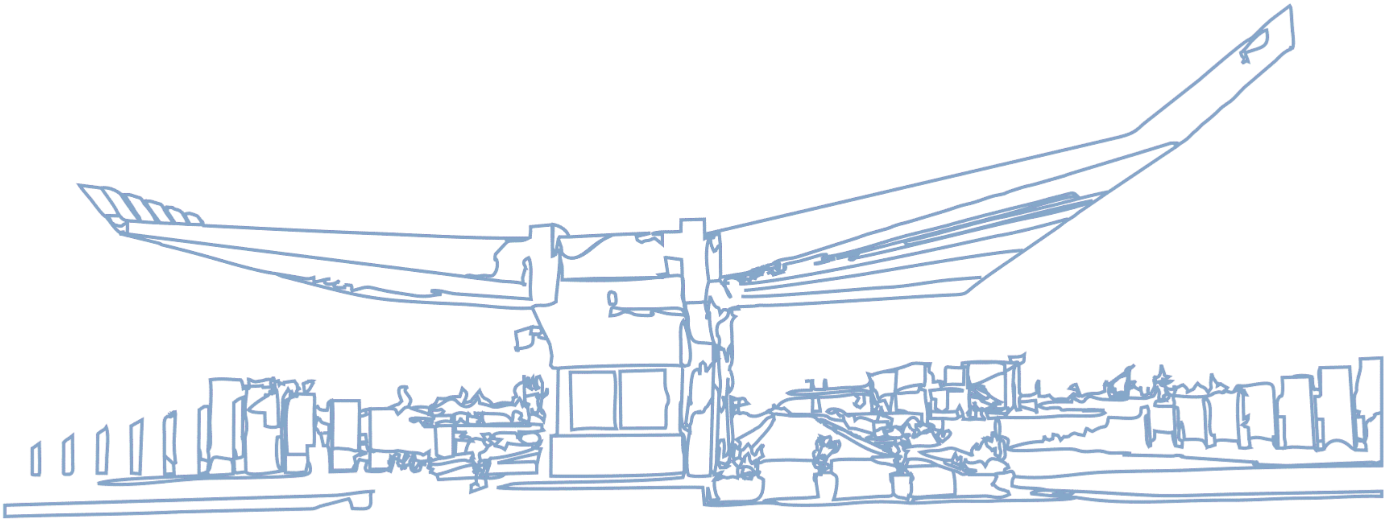
EPOKA UNIVERSITY | Computer Engineering Department

CEN 571 – Data Mining

Assignment 02 – Question 4



PREPARED:
Baftjar TABAKU

19.04.2020
Epoka University
Tirana, ALBANIA

ACCEPTED:
Prof.Dr. Arben Asllani

Question 4

We want to store a 500 MB file into a cluster with 12 nodes, which are located in three different tracks (4 nodes per rack).

1. If a data block can store 128 MB, how many data blocks are needed to split this file?
2. Use a replication factor of 3 and the *write* principles discussed in the chapter to allocate data blocks into this cluster.
3. Repeat steps 1 and 2 but with a block size of 256 MB

First of all, the data will be divided into 124MB blocks which means that 500MB file will be divided into blocks, 500 MB / 128 MB = ~4 Blocks, so 4 blocks and each of them with 128 MB and the last block will have 116MB, so the following schema build exactly for this problem will be shown as below.

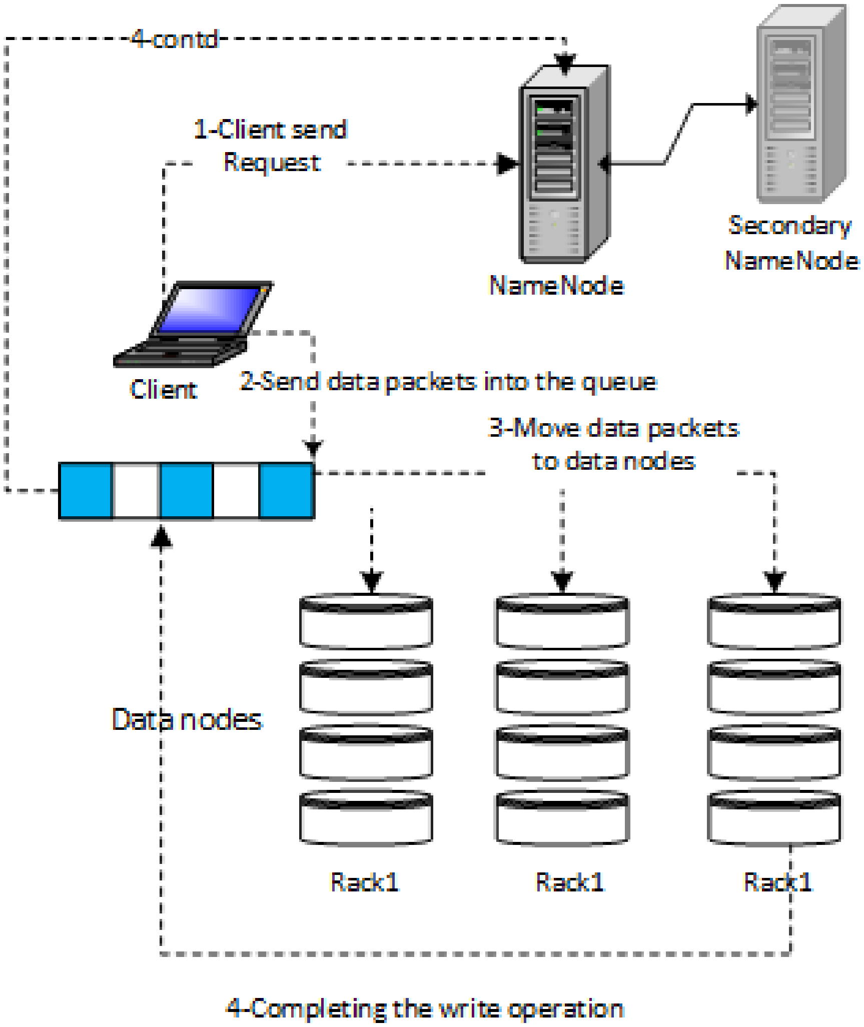

4-contd
1-Client send Request
NameNode
Secondary NameNode
Client
2-Send data packets into the queue
3-Move data packets to data nodes
Data nodes
Rack1
Rack1
Rack1
4-Completing the write operation

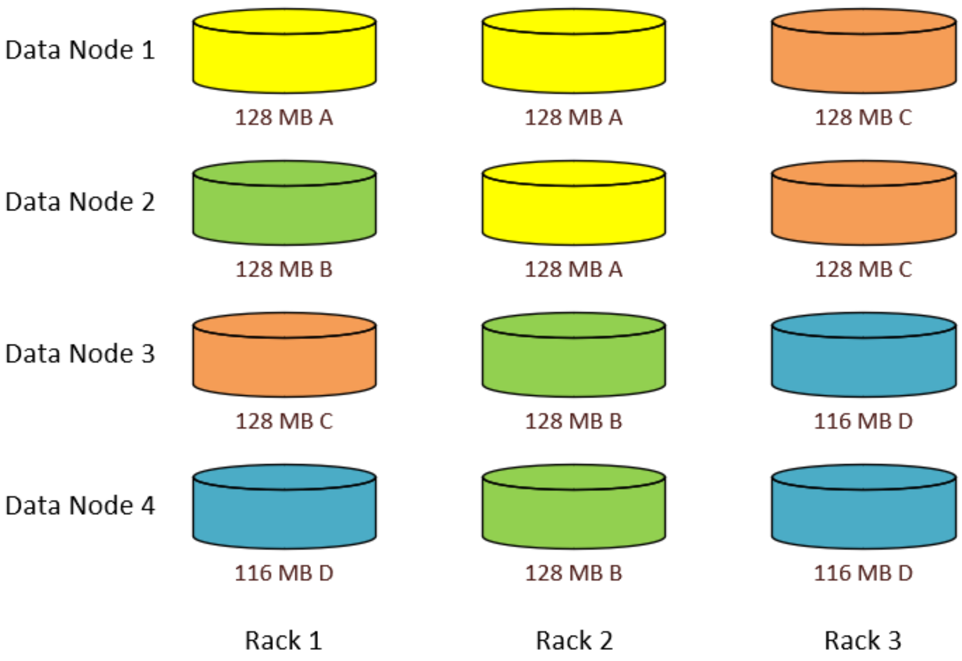
Figure 1-Writing files into DataNodes, according to the current exercise, made with Microssoft Visio

According to the replication factor 3 since this replication is 3 then each block will be replicated three times, so for a 500MB file /128MB blocks size = ~4 Blocks, but since the replication factor we have 4 Blocks * 3 (replication factor) = 12 Blocks in total where each of the 4 blocks is replicated three times, so 3 bocks will be of size 116MB and 9 Blocks of size 116MB, for simplicity we split the data for a better understanding, into 4 groups A, B, C, D according to corresponding blocks as shown in the table.

128 MB A	128 MB B	128 MB B	116 MB C
128 MB A	128 MB B	128 MB B	116 MB C
128 MB A	128 MB B	128 MB B	116 MB C
128 MB A	128 MB B	128 MB B	116 MB C

So, 4 groups, A, B, C, D (no matter their actual size, their maximum is 128)

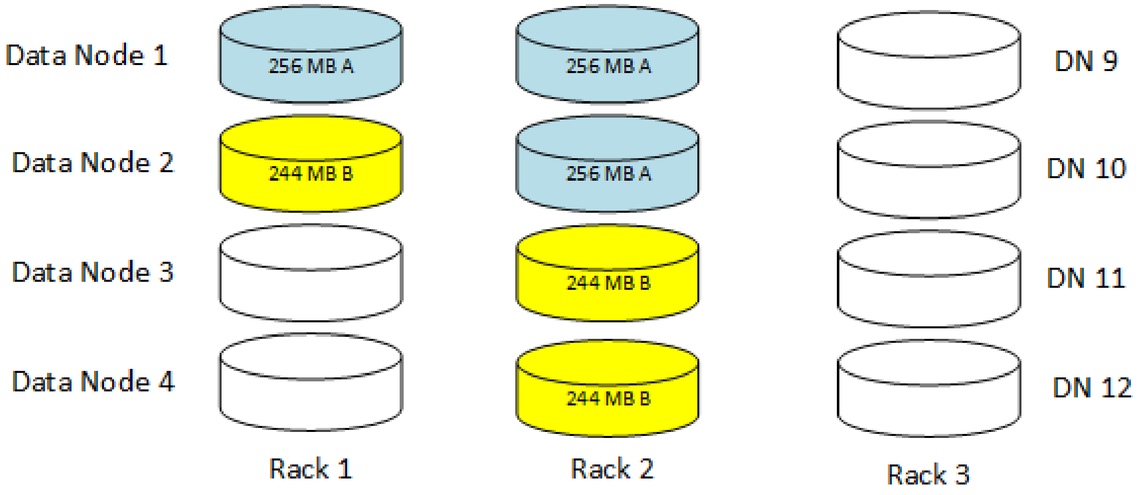
And their allocation would be as shown in the screenshot above,



According to the step 3, this time the block size would be 256MB , and for a 500 MB file we would have 500MB / 256 = ~2 so we would need two blocks of 256MB each so one block of 256MB will store 256MB and the other one will store 244MB, and according to the replication factor of 3 these two blocks would be replicated three times, so 2 Blocks * 3 (replication factor) = 6 Blocks in the total to be stored, in these 6 blocks, 3 blocks would be 256MB and 3 blocks would be 244 MB, as we did with the previous case, we assume that we have two groups A and B, so

256 MB A	244 MB B
256 MB A	244 MB B
256 MB A	244 MB B

And after assigning this data to the cluster, the data allocated would look like this as shown in the following screenshot.


Data Node 1
256 MB A
256 MB A
DN 9
Data Node 2
244 MB B
256 MB A
DN 10
Data Node 3
244 MB B
DN 11
Data Node 4
244 MB B
DN 12
Rack 1
Rack 2
Rack 3