20221103 regular meeting

**Flow**

1. load 🡪 Divide data into N workers.
2. sort 🡪 In each worker, execute the sorting algorithm. Need to use multi-thread.
3. sampling 🡪 Since every machine should know its keys’ range, get the range and send it to the master.
4. partition 🡪 Divide data using the unit block. Need to use multi-thread.
5. shuffle 🡪 Shuffle the blocks so that every worker has their own blocks well.
6. merge 🡪 Combine the workers so that we are ready to print the output.
7. exit 🡪 Print the output.

We need multiple machines 🡪 VM실이 존재하지 않음. 대책 아직 없음

Multithreaded program 🡪 다음주 concurrent 개념 공부하고 난 뒤 추가적인 논의가 필요.

How many workers? How much data at one worker?

🡪 Need to know the size of RAM, the size of total dataset // don’t know yet

🡪 Number of workers is given in input

Input specification

🡪 1 master (maybe with IP address)

🡪 fixed number of workers with their IP addresses

🡪 Input blocks on each worker (block size 32MB)

Output specification

🡪 Order of workers

🡪 Order of output blocks on each worker

🡪 Sorted output blocks (any size)

Are Master and workers already implemented in Scala? (PPT 21p)

In Input specification, what is the form of input blocks?