20201030 regular meeting

Regular Meeting: Thursday, after the class. Eat dinner together and let’s have fun!

**Commit Rules**

[Implement] : when we implement something

[Fix] : when we need to fix something

[Delete] : when we delete something

[Typo] : when there is a typo

[Setting] : when the develop setting, gitignore, or something else changed

[Refactoring] : when we refactor the codes

[Document] : when we add or modify some documents

**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.

**Discussion**

How to implement several computers?

🡪 How many computers would we make? (How much data should one computer have?) 🡪 fixed

🡪 What is the size of RAM, and the total dataset? 🡪 need to know. The block size is 32MB.

🡪 how to implement the structure of master and worker? 🡪 gRPC

How to get datasets?

🡪 using gensort

How to divide data into workers?

🡪

Which sorting algorithm to use?

🡪 scala implicit sort function, or use heap sort

🡪 need to check how scala implements the implicit sort function

Which merge algorithm to use?

🡪 K-way merge sort, using tournament tree

1 TB data 🡪 divide into N computers

In N computers, sort and merge

How do we implement multi-threaded program?

**Question)**

What is the size of RAM, and the total dataset?

How to divide data into workers?

How do we implement multi-threaded program?