Final Presentation

Team Indigo

20200854 황찬기 / 20210210 이다민 / 20220019 안재영

Overall progress

Week	Progress
Week 1	Created Project Repository / Fixed Meeting Plan & Tools to Use
Week 2	Decided Git Convention / Researched Basic Concepts of Project
Week 3	Researched about Each Steps for Implementing / Basic Structure Design
Week 4	Detailed Structure Design
Week 5	Set gRPC & Basic Scala Project Files on Git
Week 6	Intermediate Presentation / Reset Project Structure Design
Week 7	Implemented gRPC / Sampling & Partitioning / Shuffling & Merging Each
Week 8	Tried to Merge All Implemented Features

Programming Environment

OS	Windows
Basic	JDK 22 / Scala 2.13.23 / sbt 1.10.5
Logging	Scala-Logging 3.9.5

Library	Version	Description
scalapb-runtime scalapb-runtime-grpc	0.11.13	Compile .proto Files
sbt-assembly	0.15.0	Scala Code -> .jar Files
grpc-netty grpc-stub grpc-protobuf	1.65.1	gRPC Communicating

Overall progress

- Before Progress Presentation

  initial setting → sorting → sampling → partitioning → shuffling → merging

- After

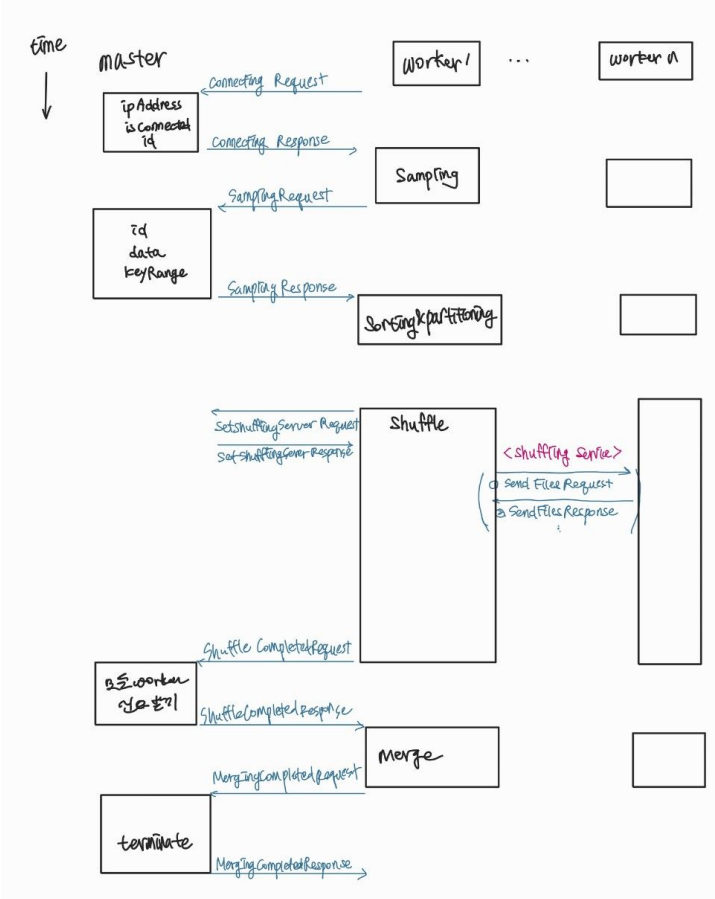
  initial setting → sampling → sorting & partitioning → shuffling → merging

- We changed our roles, according to each phases
  - took lots of time by connecting members' phases

Overall progress

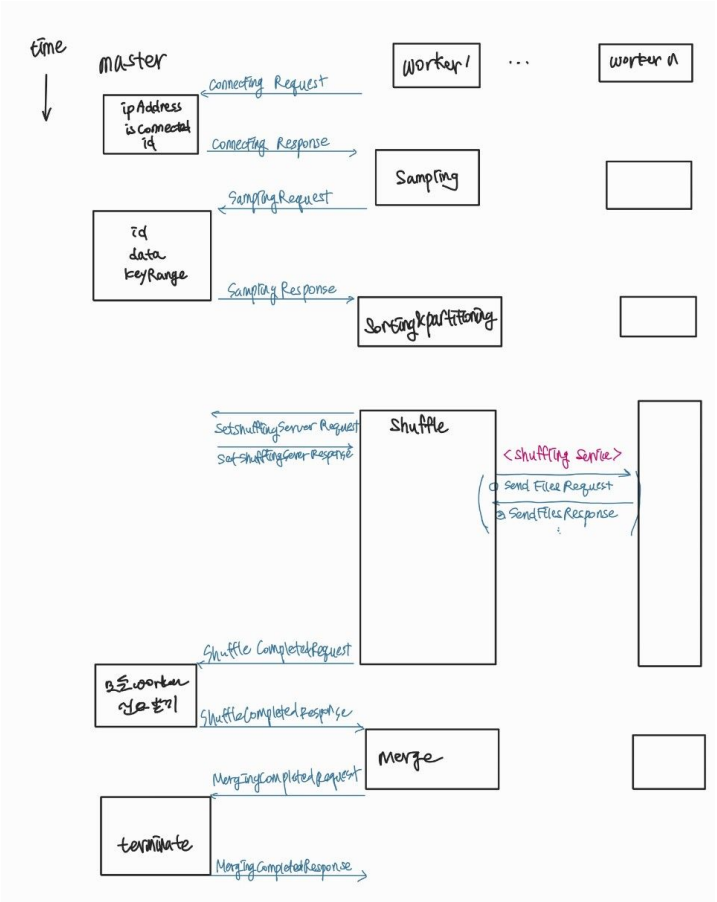
- Role

Member	Role
황찬기	Initial setting & gRPC Connection
이다민	Shuffling & Merging
안재영	Samping & Partitioning



Final architecture (master & worker)

- Changes in Phases
  - ordering of sorting
- Shuffling Grpc server
  - set shuffling server



How to Run

How to Compile and Install

1. Compile

```
sbt compile
```

332project 경로에서 명령어 실행을 통해 코드 컴파일

2. Install

```
sbt stage
```

컴파일 후 명령어 실행을 통해 실행 파일 생성 --> target/universal/stage/bin 경로에 master, worker 파일 생성됨

How to Run

컴파일과 실행 파일 생성 이후 실행 파일이 생성된 경로로 이동

```
cd target/universal/stage/bin
```

1. master

```
./master <# of workers>
```

연결할 worker의 개수를 인자로 전달, 입력 이후 연결할 ip 주소 출력됨

2. worker

```
./worker <Master IP>:<Master Port> <Input Number> <Input Directory Lists> <Output Directory>
```

연결할 모든 worker에 대해 각각 실행, 인자로 앞서 master에서 출력된 ip 주소 및 포트와 입력 데이터 경로의 수, 입력 데이터의 경로들 및 출력 데이터를 저장할 경로 전달. 아래와 같이 사용.

```
./worker 10.1.25.21:33465 1 /home/dataset/big /home/indigo/output
```

모든 worker에서 실행시 master 실행창에는 연결된 worker들의 ip 주소가 출력되고, 이후 실행이 완료될때까지 대기하면 worker 실행 파일에 전달한 출력 데이터 저장 경로에 정렬된 데이터 생성.

Final execution

- Successfully Connected Master - Workers
- Workers create sample and successfully send it to master
- Master calculates key range and successfully send it to worker
- We could get some sorted files during the sorting state of worker
- Master successfully prints changes of these sorting phases
- But after the sorting state is not completed…

```
[indigo@vm-1-master bin]$ cd ~/332project/target/universal/stage/bin ; ./master 4
2024-12-09 22:40:46 - Client Number : 4
2024-12-09 22:40:46 - Server started at 10.1.25.21:37053
2024-12-09 22:40:46 - Transition to ConnectingState.
2024-12-09 22:40:57 - Handshake from 2.2.2.101
2024-12-09 22:40:58 - Handshake from 2.2.2.102
2024-12-09 22:40:59 - Handshake from 2.2.2.103
2024-12-09 22:41:01 - Handshake from 2.2.2.104
2024-12-09 22:41:01 - 2.2.2.101, 2.2.2.102, 2.2.2.103, 2.2.2.104
2024-12-09 22:41:01 - Transition to SamplingState.
2024-12-09 22:41:04 - Received Data with length 2000.
2024-12-09 22:41:05 - Received Data with length 4000.
2024-12-09 22:41:05 - Received Data with length 6000.
2024-12-09 22:41:05 - Received Data with length 8000.
2024-12-09 22:41:05 - Received All Data from clients.
2024-12-09 22:41:05 - Check Key Range
 ([B@106893a7,[B@34a865b1)
2024-12-09 22:41:05 - Transition to SortingState.
```

```
^Cindigo@vm01:~/332project/target/universal/stage/bin$ cd ~/332project/tar
 /home/indigo/output_small
2024-12-09 22:40:56 - Try to connect with Master : 10.1.25.21
2024-12-09 22:41:01 - Connection : true
2024-12-09 22:41:01 - Worker ID : 0
```

Experiment

- We failed to implement complete sorting
  - each member tried to implement each part of sorting, but failed to merge & fix bugs
  - we ended up at sorting phase (**initial setting** → **sampling** → **sorting & partitioning** → *shuffling* → *merging (→ finish)*)

Analysis of Failures

- We wasted too much time on trivial things
  - After progress presentation, we had to make code but …
- We got stuck in file I/O
  - We should have tried to make sample code earlier
- Essentially, we had extremely lacking time
  - We did not have enough time to get familiar with code
- We were too optimistic
  - We did not consider communication overhead enough
  - Merging each part of members was hard

Lessons learned from the project

- Use time more efficiently & Use more time
- Run toy code first, then try to get more deeper