

Distributed System Labwork 5



Group 1 - ICT

University of Science and Technology of Hanoi

January, 2022

Contents

1	Introduction	2
1.1	Overview	2
1.2	Protocol	2
1.3	Framework	2
2	Methodology	2
3	Result	3
	References	3

1 Introduction

1.1 Overview

Map-reduce is a programming model for distributed computing. It includes four main stages: splitting, mapping, shuffling, and reducing. Each stage is presented below:

- Splitting stage: Splitting is generally used during data processing in map-reduce programs. The input data is split equally based on user-defined. For example, a 100MB file can be split equally into four files; each file has a size of 25MB.
- Mapping stage: Each worker applies the map function to the data(usually in the file format). The mapper processes the data and creates several small chunks of data.
- Shuffling stage: Each worker nodes redistribute the data based on the output keys in a way such that all data with the same key belong to the same node
- Reducing stage: Each worker now processes the data after shuffling and produces the output

1.2 Protocol

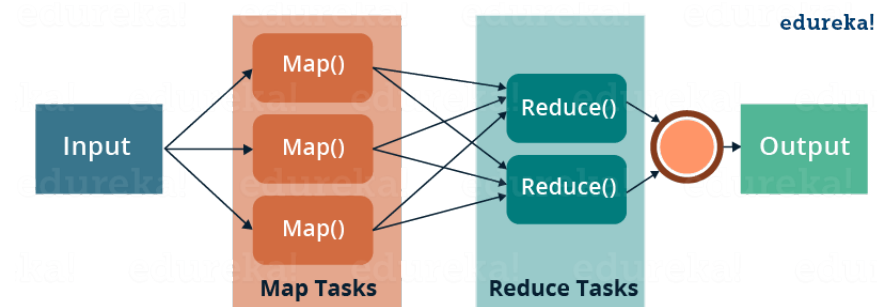


Figure 1: Map-reduce process[1]

1.3 Framework

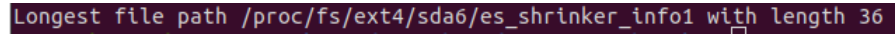
We use C++ in this project, so we have to implement the map-reduce framework ourselves.

2 Methodology

- We create one server and two slaves. First, we split the file in half and send the texts to each slave.
- Each slave will perform mapping and reducing.
- Afterwards, the server collects the slaver's results and prints out the results.

3 Result

This is a small part of the final result.

A terminal window with a dark background and light-colored text. The text reads: "Longest file path /proc/fs/ext4/sda6/es_shrinker_info1 with length 36".

```
Longest file path /proc/fs/ext4/sda6/es_shrinker_info1 with length 36
```

Figure 2: Map-reduce program results

References

- [1] <https://thirdeyedata.io/hadoop-mapreduce/>