Distributed System Labwork 4

USTH

VIETNAM FRANCE UNIVERSITY

Group 1 - ICT

University of Science and Technology of Hanoi

January, 2022

Contents

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

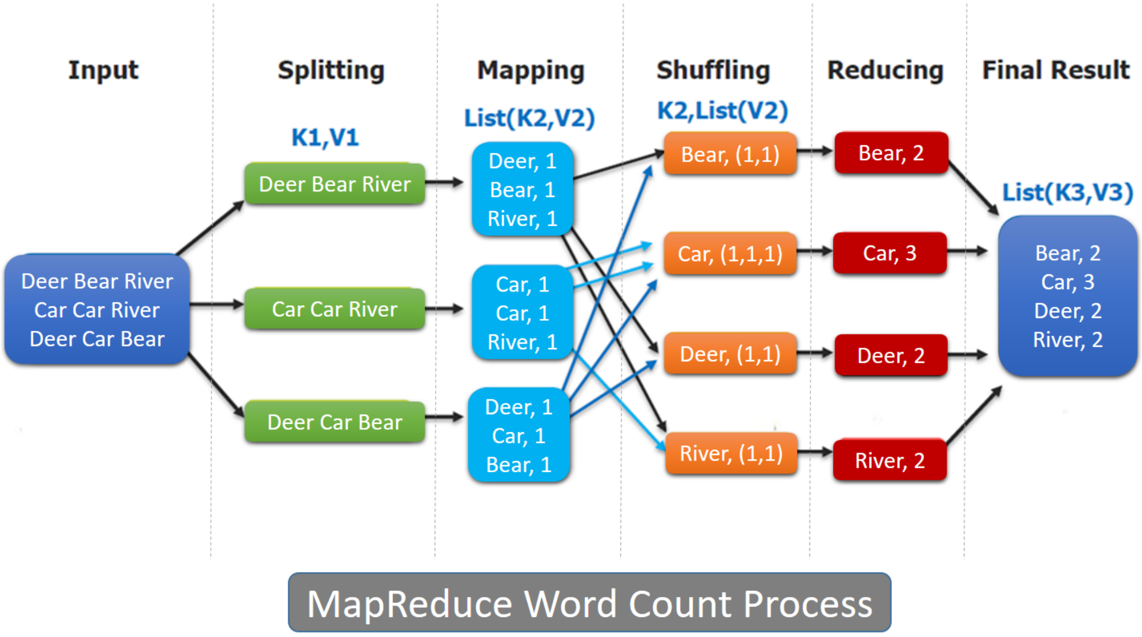


Figure 1: Map-reduce process[1]

1.3 Framework

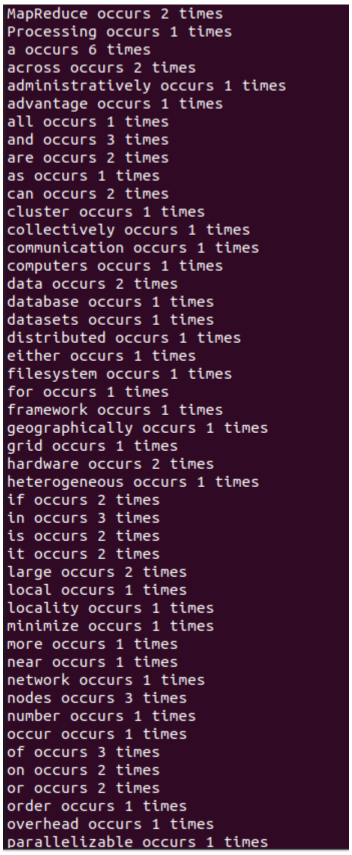
We use C++ in this project, so we have to implement the map-reduce framework ourselves.

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.



```
MapReduce occurs 2 times
Processing occurs 1 times
a occurs 6 times
across occurs 2 times
administratively occurs 1 times
advantage occurs 1 times
all occurs 1 times
and occurs 3 times
are occurs 2 times
as occurs 1 times
can occurs 2 times
cluster occurs 1 times
collectively occurs 1 times
communication occurs 1 times
computers occurs 1 times
data occurs 2 times
database occurs 1 times
datasets occurs 1 times
distributed occurs 1 times
either occurs 1 times
filesystem occurs 1 times
for occurs 1 times
framework occurs 1 times
geographically occurs 1 times
grid occurs 1 times
hardware occurs 2 times
heterogeneous occurs 1 times
if occurs 2 times
in occurs 3 times
is occurs 2 times
it occurs 2 times
large occurs 2 times
local occurs 1 times
locality occurs 1 times
minimize occurs 1 times
more occurs 1 times
near occurs 1 times
network occurs 1 times
nodes occurs 3 times
number occurs 1 times
occur occurs 1 times
of occurs 3 times
on occurs 2 times
or occurs 2 times
order occurs 1 times
overhead occurs 1 times
parallelizable occurs 1 times
```

Figure 2: Map-reduce program results

References

[1] https://www.oreilly.com/library/view/distributed-computing-in/9781787126992/5fef6ce5-20d7-4d7c-93eb-7e669d48c2b4.xhtml