2023-11-26

应用系统体系架构 — 作业9

学号：521030990006

姓名：VAHAGN GHAZARYAN

## *Your backend Spring The Boot project and MySQL database are deployed in containerized form in Compose .*

## 

## *Simple dockerfile for spring app*

## 

## *Docker compose combining mysql db, mongodb and spring boot app and binding*

## 

## 

## 

## I created book introduction files and a keyword list for my E-BookStore project. The book introductions were stored in multiple text files based on the book genre. For example, introductions of sceince books were in Science.txt.

## 

## 

## The keyword list contained relevant words like "novel", "about", "great" etc., which are significant in the context of these books.

## 

## My program is a MapReduce job designed to count the occurrences of each keyword in the book introductions. It was implemented in Java using the Hadoop framework. In the Mapper phase, the program reads the text files and maps each word to a key-value pair, where the key is the word and the value is one. In the Reducer phase, it aggregates these key-value pairs, summing the values for each key to get the total count of each word. I configured the program to ignore case by converting all words to lowercase, ensuring that words like "Novel" and "novel" are counted as the same word. I also set the input and output paths in the Hadoop job configuration, pointing to the respective HDFS directories.

## Mapper

## 

## Reducer

## 

## Main

## 

## RUN

## 

## 

## RESULTS

## 

## By default, Hadoop creates **one** Mapper task for each input split (which is a block of data, by default the size of an HDFS block). The reason behind one Mapper per block is to process the data **locally** where it resides, minimizing network traffic and allowing for **parallel** processing of data.

## 

## *From hadoop log*

## **# map tasks**

## Cause I have 4 separate input files, that each file was processed by a separate map task. Hadoop's default InputFormat, FileInputFormat, will create one map task for each input split, and input splits correspond to input files when the files are smaller than the block size.

## File size

## 

## because the input data size is much smaller than the block size, and Hadoop does not split files into blocks that are smaller than the file itself.

## **# reduce tasks**

## 

## 

## I have one node

## So its 1.

## The default behavior for most Hadoop installations is to run one reduce task if the job does not explicitly set a different number of reduce tasks.

## 