TUGAS 1

IF4044 TEKNOLOGI BIG DATA

MAP REDUCE DENGAN DATA SOSMED

DISUSUN OLEH:

MUHAMMAD GILANG RAMADHAN 13520137

PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

2022

Stack dan Implementasi *Hadoop MapReduce*

Apa itu Map Reduce ?

MapReduce adalah suatu *tools programming model* dan *software framework* yang digunakan untuk mengelola data yang jumlahnya besar agar dapat berjalan dengan lebih efisien. Adapun tahapan dari MapReduce itu sendiri terdiri dari dua fase, yaitu Map dan Reduce. Dimana tugas dari *map* ialah mengurus proses *splitting* dan *mapping* dari data. Sedangkan tugas dari *Reduce* ialah melakukan *shuffle* dan *reduce* terhadap data.

Hadoop mampu menjalankan program MapReduce yang ditulis dalam berbagai bahasa: Java, Ruby, Python, dan C++. Namun pada tugas kali ini saya mengimplementasikan *MapReduce* itu sendiri menggunakan bahasa pemrograman python dan bantuan kakas *jupyter notebook*. Adapun detail stack yang saya gunakan pada tugas 1 IF4044 Teknologi Big Data ialah sebagai berikut.

Stack yang digunakan:

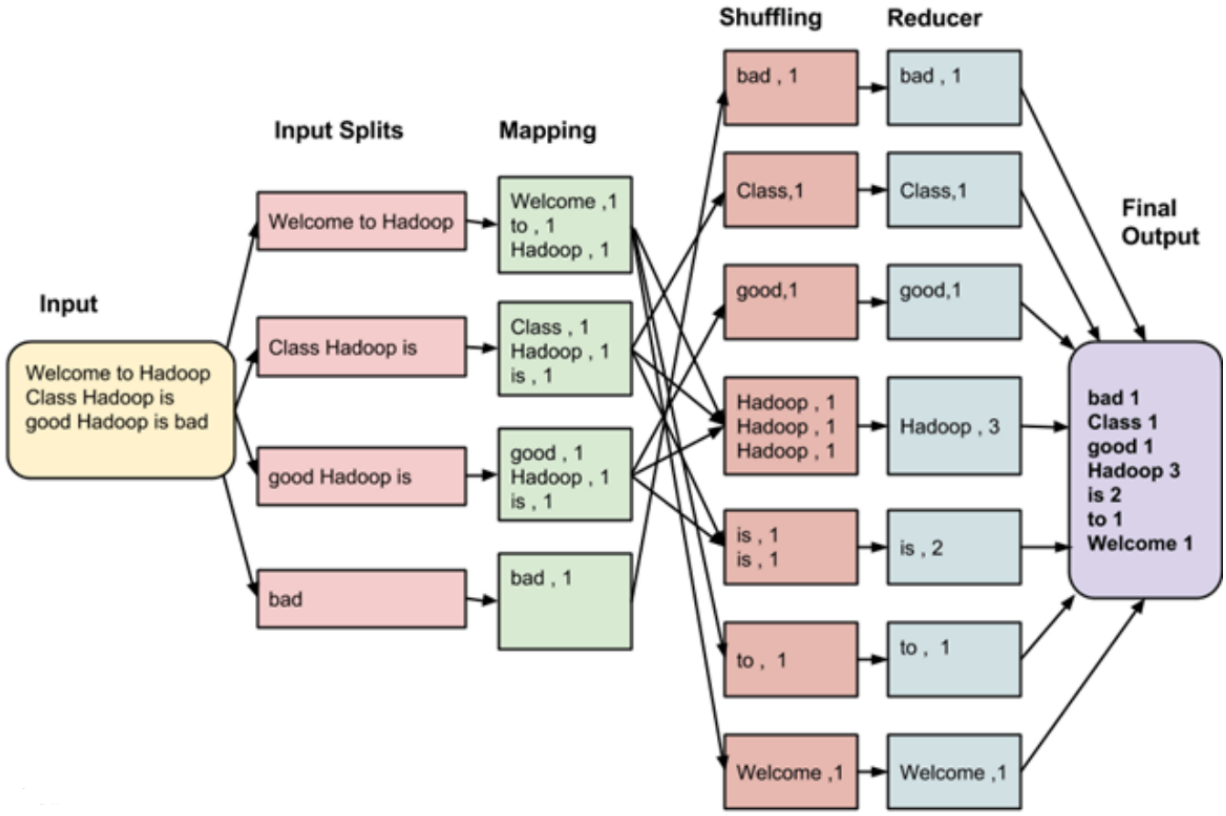
1. HDFS
2. JSON
3. CSV
4. Python
5. Jupyter Notebook
6. Sistem Operasi Linux (Ubuntu)

Library:

1. os
2. fnmatch
3. json
4. sys
5. datetime

Adapun secara teknis, program *MapReduce* bersifat parallel. Sehingga, kecepatan yang dihasilkan juga tentunya lebih cepat daripada program yang dieksekusi secara sekuensial. Oleh karena itu, melalui *MapReduce* ini banyak digunakan dalam pengelolaan dalam skala besar, karena kecepatannya komputasinya yang lebih cepat daripada pengelolaan data secara biasa.

Untuk setiap fase dari *MapReduce*, akan digunakan *pair of key-value* sebagai inputannya. Sehingga perlu dilakukan splitting terlebih dahulu untuk mendapatkan *pair of key-value* tersebut. Terkait dengan fase-fase pada *MapReduce* dapat dilihat melalui gambar 1 di bawah ini.



Gambar 1. Arsitektur *MapReduce*

Rincian fase-fase pada *MapReduce*:

1. ***Input Splits***

   Untuk memudahkan proses mapping, maka ada dilakukan split terlebih dahulu pada masing-masing line pada inputan.

2. ***Mapping***

   Data yang sudah displit akan dilakukan mapping untuk menghasilkan nilai-nilai output. Pada tugas ini mapping dilakukan dengan cara mengelompokkan komponen data sesuai dengan kesamaannya, kemudian akan dihitung jumlah kemunculan setiap kata dari input splits.

3. ***Shuffling***

Fase ini menerima output dari fase Mapping. Tugasnya ialah menggabungkan record-record yang relevan dari output fase Mapping. Pada tugas ini, kata yang sama dikumpulkan bersama dengan frekuensinya masing-masing.

4. ***Reducing***

   Dalam fase ini, nilai output dari fase Shuffling diaggregasikan (digabungkan). Fase ini merupakan fase yang mengkombinasikan nilai-nilai dari fase *Shuffling* dan mengembalikan suatu nilai output tunggal.

Terkait Implementasi *MapReduce Job* yang saya buat itu ialah memanfaatkan file .py yang sudah diimplementasikan *logic* dan *tools* dari *MapReduce* itu sendiri.

Adapun Fungsi *MapReduce* yang dibuat sesuai dengan spesifikasi pada tugas ini ialah sebagai berikut.

1. Social Media
2. Data
3. Count

Adapun rincian fungsionalitas dari masing-masing file yang sudah saya implementasikan pada repository IF4044_Social_Media_MapReduce ialah sebagai berikut.

1. *File* convert_json_to_txt.py merupakan implementasi untuk melakukan *convert* dari file json pada *crawling file* yaitu rawjson ke file berformat .txt.
2. *File* Socmed_Mapper merupakan implementasi dari proses *mapping* data pada *MapReduce.*
3. *File* Socmed_Reducer merupakan implementasi dari proses *reducing* sekaligus *shuffling* data pada *MapReduce*.
4. File test_json_to_txt.ipynb, test_mapreduce.ipynb, dan test_reducer.ipynb masing-masing merupakan berisikan *testing* dari setiap proses di atas pada file .json.

Link Github Repository:
https://github.com/gilanglahat22/IF4044_Social_Media_MapReduce/blob/master/src/test_reducer.ipynb