Laporan Hadoop-Streaming with Rust

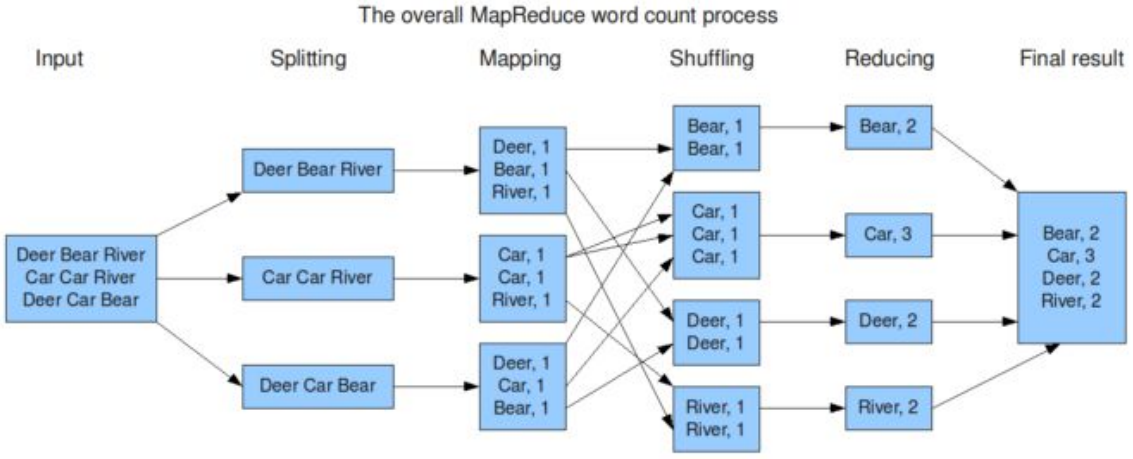
Source code utama dari tugas ini berasal dari:
https://github.com/d-unseductable/rust_hadoop_streaming

Hadoop

Hadoop adalah sebuah framework open source yang digunakan untuk menyimpan dan memproses big data menggunakan pemodelan cluster computer menggunakan model bahasa pemrograman sederhana.

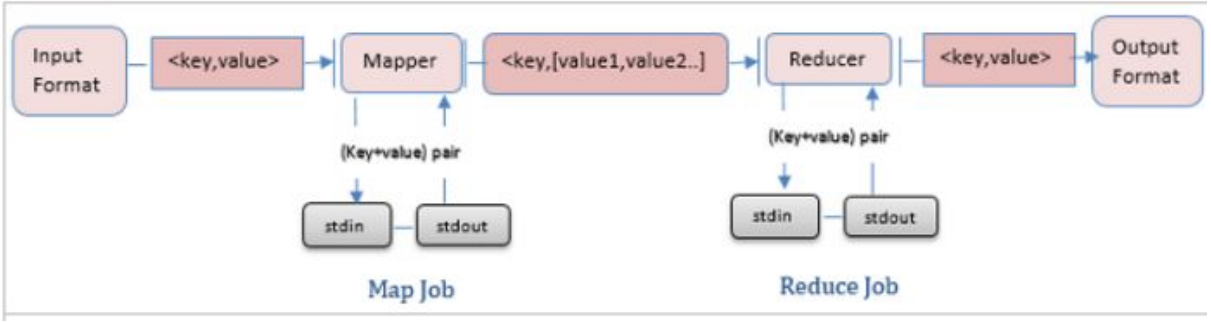
Mapreduce

Dalam hadoop digunakan algoritma Mapreduce sebagai solusi pemrosesan suatu big data. Mapreduce sendiri adalah sebuah model pemrograman yang didesain untuk dapat melakukan pemrosesan data dengan jumlah yang sangat besar dengan cara membagi pemrosesan tersebut ke beberapa tugas yang independen satu sama lain.



Hadoop streaming

Hadoop streaming adalah sebuah utilitas yang disediakan oleh hadoop berfungsi untuk membebaskan user membuat dan menjalankan Mapreduce menggunakan script/program user sendiri. Script/program tersebut bisa berfungsi sebagai Mapper, Reducer atau keduanya.



Instalasi hadoop

1. Install Java Development Kit dengan perintah

```
adiguna7@adiguna-Suryo ~ sudo apt-get install default-jdk
```

Setelah itu untuk cek apakah java dan jdk sudah terpasang

```
adiguna7@adiguna-Suryo ~ java --version ✓ 6289 14:52:57
openjdk 11.0.5 2019-10-15
OpenJDK Runtime Environment (build 11.0.5+10-post-Ubuntu-2ubuntu116.04)
OpenJDK 64-Bit Server VM (build 11.0.5+10-post-Ubuntu-2ubuntu116.04, mixed mode,
 sharing)
```

Apabila sudah tampil versi java dan jdk, maka jdk sudah terpasang dengan benar

2. Ambil hadoop versi binary menggunakan perintah

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming wget https://downloads.ap
ache.org/hadoop/common/hadoop-3.1.4/hadoop-3.1.4.tar.gz --no-check-certificate
```

Lalu extract menggunakan perintah

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming tar -xzvf hadoop-3.1.4.ta
r.gz
```

3. Cek lokasi JAVA_HOME menggunakan perintah

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming readlink -f /usr/bin/java
 | sed "s:bin/java::"
/usr/lib/jvm/java-11-openjdk-amd64/
```

Lalu set file hadoop-env.sh, dalam kasus ini terletak di

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming sudo gedit /opt/hadoop-ru
st-streaming/hadoop-3.1.4/etc/hadoop/hadoop-env.sh
```

Set JAVA_HOME pada file tersebut

```
# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=$(readlink -f /usr/bin/java | sed "s:bin/java::")
```

4. Test hadoop dengan menggunakan perintah

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming /opt/hadoop-rust-streamin
g/hadoop-3.1.4/bin/hadoop
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
 or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
  where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:

buildpaths                       attempt to add class files from build tree
--config dir                     Hadoop config directory
--debug                          turn on shell script debug mode
--help                           usage information
hostnames list[,of,host,names]   hosts to use in slave mode
hosts filename                   list of hosts to use in slave mode
loglevel level                   set the log4j level for this command
workers                          turn on worker mode

  SUBCOMMAND is one of:
```

5. Buat alias untuk menjalankan hadoop melalui file .zshrc, agar langsung dapat menjalankan hadoop hanya dengan menulis hadoop

```
alias hadoop=/opt/hadoop-rust-streaming/hadoop-3.1.4/bin/hadoop
```

Lalu source terminal menggunakan perintah source ~/.zshrc

6. Setelah memastikan hadoop terpasang, selanjutnya clone repository dari **https://github.com/d-unseductable/rust_hadoop_streaming** dengan perintah

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming git clone https://github.
com/d-unseductable/rust_hadoop_streaming
```

7. Lalu pada directory repository compile mapper dan reducer menggunakan cargo rust

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming/rust_hadoop_streaming master • cargo build --release
```

8. Download hadoop streaming sesuai versi pada link berikut https://jar-download.com/artifacts/org.apache.hadoop/hadoop-streaming pada kasus ini menggunakan hadoop versi 3.1.4 dan letakkan pada folder hadoop

9. Lalu run menggunakan perintah

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming/rust_hadoop_streaming master • hadoop jar ../hadoop-3.1.4/hadoop-streaming-3.1.4.jar -input ncdc_data -output
output -mapper target/release/mapper -reducer target/release/reducer
```

Apabila berhasil akan muncul folder output

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming/rust_hadoop_streaming m
aster • ls ✓ 6298 15:27:27
Cargo.lock  Cargo.toml  ncdc_data  output  README.md  samples  src  target
```

Lalu apabila file didalam output tersebut dibuka menggunakan perintah cat akan muncul sebagai berikut:

```
adiguna7@adiguna-Suryo /opt/hadoop-rust-streaming/rust_hadoop_streaming m
aster • cat output/* ✓ 6299 15:27:32
1901    317
1902    244
1903    289
1904    256
1905    283
```

Output tersebut berasal dari input data 1901.all, 1902.all, 1903.all, 1904.all, 1905.all pada directory ncdc_data.