Vežbe 5
MapReduce

MapReduce je framefork za procesiranje velike količine podataka.

Koristi se kao deo Apache Hadoop paketa. Apache Hadoop 3.2.1

Konfiguracija hadoopa na windows mašini je malo komplikovanija i može praviti nepredviđene greške pa ćemo na vežbama raditi u virtualnom Linux okruženju.

Da bi pokrenuli linux virtualnu mašinu potreban je Virtual Box i .ova fajl koji možete skinuti na ovom linku (pažljivo jer zauzima 4GB) na kom se nalazi već skinut i podešen hadoop, a možete koristiti i svoju linux mašinu po želji. **user:pass = admin123:admin123**

Da bismo koristili MapReduce framework potrebno je da dodamo 2 External JARs u naš projekat koji se nalaze u hadoop paketu, u share/hadoop/ folderu a to su:

1. hadoop-common-3.2.1.jar
2. hadoop-mapreduce-client-core-3.2.1.jar

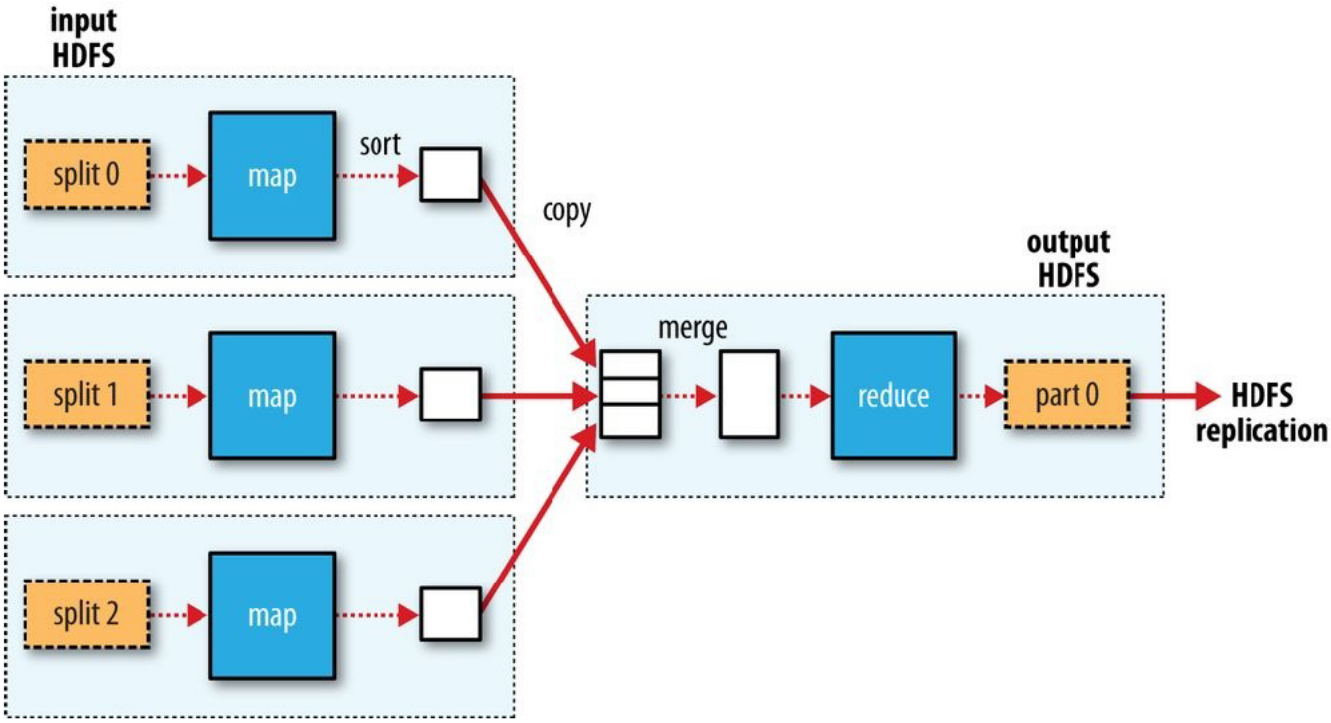
Kodove pokrećemo tako što prvo uradimo File -> Export -> Java, JAR File
da bismo dobili .jar file koji smestimo negde u hadoop folder, zatim koristimo komandu npr:

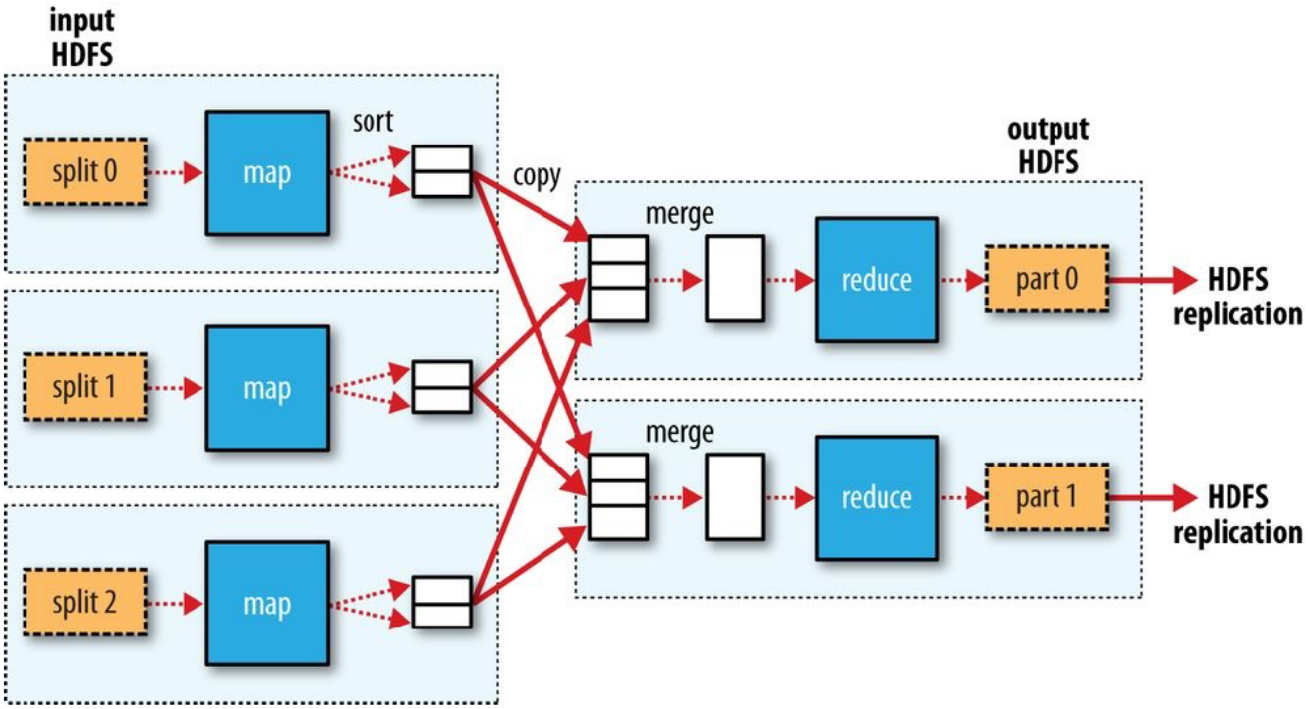
bin/hadoop jar myjar/wc.jar pds.WordCount input output

myjar/wc.jar je putanja do jar fajla koji pokrećemo

pds.WordCount je puno ime paketa(pds) i klase(WordCount) koju pokrećemo

input/output su imena direktorijuma ili fajlova koje koristimo za I/O


input
HDFS
split 0
map
sort
split 1
map
split 2
map
copy
output
HDFS
merge
reduce
part 0
HDFS
replication


input
HDFS
split 0
map
sort
split 1
map
split 2
map
copy
output
HDFS
merge
reduce
part 0
HDFS
replication
merge
reduce
part 1
HDFS
replication

Zadatak 1 Word Count

Koristeći MapReduce framework napisati program koji će prebrojati pojavljivanja svih reči u nekom fajlu.

Npr za tekst: **a b a c a b b a a a c a b a b c c a**

Program treba da ispiše: **a 9 b 5 c 4**

Uraditi prvo bez korišćenja Combinera, a onda sa istom Reducer klasom kao Combinerom.

Zadatak 2 Max Temp

U folderu MaxTempData nalaze se podaci sa izmerenim vrednostima temperatura i njihovim datumima u formatu: ITE00100554,**17630106**,TMAX,**13**,,,E,

Izdvojiti iz svih input fajlova koja je bila maksimalna temperatura za svaki mesec u godini

Druga vrednost u svakom redu je datum u formatu yyyyMMdd, a četvrta vrednost je izmerena temperatura

Zadatak 3 MinMaxCountTuple

Napraviti posebnu klasu MinMaxCountTuple koja implementira interface Writable i koju ćemo koristiti vrednost koju Mapper prosleđuje i koja dolazi do Reducera

Kao u prošlom primeru ispisati ne samo maksimalnu već i minimalnu temperaturu i count - broj zasebnih merenja izvršenih u svakom mesecu

Zadatak 4 Inverted Index

Razvrstati imena u 2 odgovarajuće grupe muških i ženskih. U svakom redu je dato slovo M ili F po kome znamo kojoj grupi pripada.