

28/02/25

WEEK-2

Aim:

Hadoop is used to store "BigData".

Hadoop
 { storage (HDFS)
 processing (Map Reduce)

HDFS — Name Node
 Data Node

using master slave architecture

Map Reduce (Process) :-
 ① Input :-
 { Bus car train
 ship ship train
 Bus ship Car }
 ② split
 { Bus car train
 ship ship train
 bus ship car }

③ Map Reduce :-

ship 1	bus 1
bus 1	ship 1
car 1	train 1
train 1	car 1

④ Shuffle & sort

Bus 1
Bus 1
Car 1
Car 1
ship 1
ship 1
ship 1
train 1
train 1

⑤

Reduce function :-
 Bus 2
 Car 2
 ship 3
 train 2

1. hadoop version

2. java -version

3. start-all.sh

4. jps

5. hadoop-3.4.0/bin/hdfs namenode -format

6. start-all.sh

7. Open browser http://localhost:9870

8. Open intelliJ idea click on new project give the name of the project.

9. Select maven goto advance give the group id
name org.dsba

10. remove main class

11. create dependencies in org.dsba copy the
dependencies code from the github.

12. go to maven click on the project name reload
all projects.

13. create the 3 java classes WC- Mapper, Reduce, Runner
copy the code from github.

14. create jar file --> click on maven clean enter and
maven enter.

15. target folder will be created which contains jar file.

16. goto ubuntu terminal create the textfile input2.txt

17. nano input2.txt

18. write some text with repeated words

19. ctrl + O enter ctrl + X

20. cat input.txt

21. create the folder on the localhost hadoop fs -mkdir/
input2

22. hadoop fs -put input2.txt/input2

23. back to the local host check the file system.
24. back to the IntelliJ go to terminal `hadoop jar`
`-target /week2-1.0-SNAPSHOT.jar org.dbbda.WCRunner`
`input2 / input 2.txt / output2`
25. back to the local host check the file system.
26. click on output2 file
27. back to the main terminal
28. `hadoop fs -cat /output2/part-00000`

→ open VM Ware Workstation.
[ubuntu 64-bit]

→ power on.

→ show applications - open terminal

→ check hadoop version [3.4.0]

→ check java version [1.8.0_432]

→ To start hadoop file system, check start-all.sh
[Here all data nodes will start]

→ To see what kind of nodes are there in hadoop system.
check jps. (java virtual machine processing status).

→ for fresh node processing check: `hadoop -3.4.0/bin/hdfs` namenode
-format

→ again start-all.sh

- open Firefox now
- type : localhost:9870/dfshealth.html # tab-overview
 - [this is. hadoop file system, changes done in command prompt are seen here]
- go to show application and open IntelliJ IDEA Community Edition.
- go to to file → new → project
- give name as → Lab2.
- select as Maven, used for automation
- go to advance settings :- keep groupId :- org.data science
- click on create
- click on this Window.
- Right click on Main and Delete → select ok
- org.data science is highlighted.
- Under it, under `</dependencies>` type `<dependencies>`
`</dependencies>`
- copy paste code from github `dependencies.txt` of `github.com/sushikumar1992`
- paste it inside the `<dependencies>` `</dependencies>` only.
- change the version in it from 3.3.3 to 3.4.0
- To reload; click on symbol on right
- Tap on Lab2 project, click on Refresh.
- first sync then reload.

→ Now we need to create the classes, so go to github repository.

→ copy the WC-Mapper.txt.

→ right click on org.datascience - new - java class

↳ write name as WC-Mapper. ~~txt~~


→ Remove line 3 of public class

→ paste the copied text from github WC-Mapper.txt.


→ Similarly create other two classes :- WC-Reducer

→ In order to create jar file :- WC-Runner.

click on  Maven - M.

→ click on  Create main goal

→ click on mvn clean and click enter.

→ again click same  and click on mvn install

and enter.

→ On left the ~~target~~ folder is created.

→ The ~~lab2-1.0-SNAPSHOT.jar~~ should be created under target folder.

→ To create input file if type nano data.csd.txt.

→ In file type :- This is DSBDA lab. 1

This is second week in DSBDA lab
to save it ctrl + c and enter and ctrl + x

→ then type cat csd.txt

→ Next Create a directory :- `hadoop fs -mkdir /csdc.`
in terminal

→ In hadoop folder we can be able to see csdc. [utilities → browser Directory]

→ push the file into folder: - hadoop fs -put csd.txt /csdc

→ under csd folder csdc file is there. Hence file is in hadoop file system.

→ Now program and input are ready; just need to run one command.

→ goto IntelliJ IDEA in the bottom terminal [] click on it

and type: - hadoop jar target/Lab2-1.0-SNAPSHOT.jar org.datascience

WC-Runner /csdc/csd.txt /week2output

→ click Enter

→ after 100%. & check browser click on week2 output
map 100% reduce 100% [head the file] 32K. part-00000

→ hadoop space jar space target/Lab2-1.0-SNAPSHOT.jar
space org.datascience.WC-Runner space /csdc/csd.txt
space /week2output.

OUTPUT:-

File information - part - 00000

Download - had the file (first 32K) Tail the file (last 32K)

Block information - - **Block 0**

Block ID : 10737418234.

Block pool ID : BP-603145226-122.

Generation stamp : 1008.

Size : 42

availability :

Standard - worked in getdata.

File Contents

DSBDA 2

Thus 2

in 1

~~Sp~~ 2

Lab 2

Second 1

weeks 1

close