Hadoop Basic Commands

Jps → to view the instance of the hdfs system

hadoop fs -usage ls → to view the commands in hdfs

hadoop fs -help ls → to view the entire documentation in hdfs

How to clone and get the data and import into hdfs system.

Get data files from GitHub to our Unix System
git clone https://github.com/sibaramKumar/dataFiles

#Rename the Folder
cd dataFiles

Unzip the Files
sudo apt install unzip
unzip SalesData.zip
ls –lrt
rm SalesData.zip

Create a Folder at HDFS
hadoop fs -mkdir -p practice/retail_db/

Copy the Files from Local to HDFS
hadoop fs -put dataFiles/* practice/retail_db/

Command - copyToLocal or get
hadoop fs -get practice/retail_db/orders .

Error if the destination path already exists. To overwrite use –f flag.
hadoop fs -get practice/retail_db/orders .
hadoop fs -get –f practice/retail_db/orders .

-p flag to preserves access and modification times, ownership and the mode.
hadoop fs -get -p practice/retail_db/orders .

To Only copy the files with out folder use a pattern.
hadoop fs -get practice/retail_db/orders/* .

###
When copying multiple files, the destination must be a directory.
mkdir copyHere
hadoop fs -get practice/retail_db/orders/* practice/sample.txt copyHere

Command – ls
hadoop fs -ls practice/retail_db

Using Pattern
hadoop fs -ls practice/retail_db/ord*

-R : Recursively list the contents of directories.
hadoop fs -ls -R practice/retail_db

-C : Display the paths of files and directories only.
hadoop fs -ls -C practice/retail_db

–r : Reverse the order of the sort.
hadoop fs -ls -r practice/retail_db

–S : Sort files by size.
hadoop fs -ls -S practice/retail_db

–t : Sort files by modification time (most recent first).
hadoop fs -ls -t practice/retail_db

head
hadoop fs -head practice/retail_db/orders/part-00000

tail
hadoop fs -tail practice/retail_db/orders/part-00000

cat
hadoop fs -cat practice/retail_db/orders/part-00000

HDFS cat Command with Unix head Command
hadoop fs -cat practice/retail_db/orders/part-00000 | head -10
hadoop fs -cat practice/retail_db/orders/part-00000 | head -5

HDFS cat Command with Unix tail Command
hadoop fs -cat practice/retail_db/orders/part-00000 | tail -10
hadoop fs -cat practice/retail_db/orders/part-00000 | tail -5

Stat Command
default or %y - Modification Time
hadoop fs -stat practice/retail_db/order_items/part-00000
hadoop fs -stat %y practice/retail_db/order_items/part-00000

%b - File Size in Bythadoop fs es
hadoop fs -stat %b practice/retail_db/order_items/part-00000

```
# %F - Type of object.
hadoop fs -stat %F practice/retail_db/order_items/part-00000
hadoop fs -stat %F practice/retail_db/order_items

# %o - Block Size
hadoop fs -stat %o practice/retail_db/order_items/part-00000

# %r - Replication
hadoop fs -stat %r practice/retail_db/order_items/part-00000

# %u - User Name
hadoop fs -stat %u practice/retail_db/order_items/part-00000

# %a - File Permission in Octal
hadoop fs -stat %a practice/retail_db/order_items/part-00000

# %A - File Permission in Symbolic
hadoop fs -stat %A practice/retail_db/order_items/part-00000


### df
hadoop fs -help df
hadoop fs -df
hadoop fs -df -h  #For Human Redable Format

### du
hadoop fs -help du
hadoop fs -du practice/retail_db

-v : Displays with Header
hadoop fs -du -v practice/retail_db

-h :Readable Format
hadoop fs -du -h practice/retail_db

-s : Summary of total size
hadoop fs -du -s practice/retail_db




###fsck Command Help
hadoop fsck -help

### Print a High Level Report.
hadoop fsck practice/retail_db
```

-files -->Print a detailed file level report.
hadoop fsck practice/retail_db -files

-files -blocks --> Print a detailed file and block report.
hadoop fsck practice/retail_db -files -blocks

-files -blocks -locations --> Print out locations for every block
hadoop fsck practice/retail_db -files -blocks –locations

-files -blocks -racks --> Print out Rack level Informatin

Chmod
#Octal Format
hadoop fs -chmod 755 practice/retail_db/orders/part-00000

#Symbolic Format
hadoop fs -chmod g+w practice/retail_db/orders/part-00000

1.
Change Properties in hdfs-site.xml or core-site.xml.

2.
Using -D option or --conf option.

Using -D Option.

Copy the file sample1.txt.
hadoop fs -put sample1.txt practice/retail_db

Check the statistics.
Check the Replication.
hadoop fs -stat %r practice/retail_db/sample1.txt
Check the Block Size
hadoop fs -stat %o practice/retail_db/sample1.txt

Copy the file sample1.txt using different replication and block size.
hdfs dfs -Ddfs.blocksize=64M -Ddfs.replication=3 -put -f sample1.txt practice/retail_db

Check the Statistics - Replication
hadoop fs -stat %r practice/retail_db/sample1.txt
Check the Statistics - Block Size
hadoop fs -stat %o practice/retail_db/sample1.txt

Using --conf
Copy a file sample2.txt.

hadoop fs -put sample2.txt practice/retail_db

Check the statistics
Check the Replication
hadoop fs -stat %r practice/retail_db/sample2.txt

Copy the File using --conf
hdfs dfs --conf hdfs-override.xml -put -f sample2.txt practice/retail_db

Check the Replication
hadoop fs -stat %r practice/retail_db/sample2.txt

3.
Change after copying the Files in HDFS (setRep)
Copy a file:
hadoop fs -put sample3.txt practice/retail_db

Check Replication.
hadoop fs -stat %r practice/retail_db/sample3.txt

Change Replication using setrep
hdfs dfs -setrep 2 practice/retail_db/sample3.txt

Check Replication
hadoop fs -stat %r practice/retail_db/sample3.txt

Ways to create RDD in Pyspark: