



Hope Foundation's  
**Finolex Academy of Management and Technology, Ratnagiri**  
**Information Technology Department**

Subject name: Big Data Lab			Subject Code: ITL801
Class	BE IT	Semester – VIII (CBCGS)	Academic year: 2019-20
Name of Student	Kazi Jawwad A Rahim		QUIZ Score :
Roll No	28	Assignment/Experiment No.	05
Title: Run Hadoop wordcount program using Cloudera VM			

**1. Course objectives applicable**

**COB4.**Understand implementation, programming using MapReduce.

**COB3.**Understand tasks in HDFS.

**2. Course outcomes applicable:**

**CO3** –To understand Mapreduce technique

**CO6**-To install and run Map Reduce program

**3. Learning Objectives:**

1. Run Wordcount Program Using Mapreduce
2. Connect using Cloudera VM

**4. Practical applications of the assignment/experiment:****5. Prerequisites:**

1. Knowledge of Map Reduce
2. Internet Access
3. Cloudera Quickstart VM

**6. Hardware Requirements:**

1. Internet Access with Browser
2. PC to install Cloudera

**7. Software Requirements:**

Browser like Chrome, Internet Explorer Edge, Cloudera VM Setup

**8. Experiment/Assignment Evaluation:**

Sr. No.	Parameters	Marks obtained	Out of
1	Technical Understanding (Assessment may be done based on Q & A <u>or</u> any other relevant method.) Teacher should mention the other method used -		6
2	Neatness/presentation		2
3	Punctuality		2
<b>Date of performance (DOP)</b>		<b>Total marks obtained</b>	<b>10</b>
<b>Date of checking (DOC)</b>		<b>Signature of teacher</b>	

## 1. Starting Docker and pulling Cloudera quickstart vm

```
[root@lt121 ~]# systemctl start docker
[root@lt121 ~]# docker images
REPOSITORY          TAG      IMAGE ID      CREATED       SIZE
Trying to pull repository docker.io/cloudera/quickstart ...
sha256:f91be44d7a2c92ea3d52929a22f729d4d13fc038b00f120e630f91c941acb63: Pulling from docker.io/cloudera/quickstart
1d00052c2e734: Pull complete
Digest: sha256:191abe4cdfa2c92ea3d52929a22f729d4d13fc038b00f120e630f91c941acb63
Status: Downloaded newer image for docker.io/cloudera/quickstart:latest
[root@lt121 ~]# docker run --hostname=quickstart.cloudera --privileged=true -t -i /usr/bin/docker-quickstart
/usr/bin/docker-current: Error parsing reference: "/usr/bin/docker-quickstart": is not a valid repository/tag: invalid reference format.
See '/usr/bin/docker-current run -h'.
[root@lt121 ~]# docker images
REPOSITORY          TAG      IMAGE ID      CREATED       SIZE
docker.io/cloudera/quickstart latest   4239cd2958c6  3 years ago  6.34 GB
[root@lt121 ~]# docker run --hostname=quickstart.cloudera --privileged=true -t -i 4239cd2958c6 /usr/bin/docker-quickstart
Starting mysqld...
[ OK ]
if [ "$1" == "mysqld" ] ; then
  if [ "$EC2" == "true" ] ; then
    FIRST_BOOT_FLAG=/var/lib/cloudera-quickstart/.ec2-key-installed
    if [ ! -f $FIRST_BOOT_FLAG ] ; then
      METADATA_API=http://169.254.169.254/latest/meta-data
      KEY_URL=$METADATA_API/public-keys/0/openssh-key
      SSH_DIR=/home/cloudera/.ssh
      mkdir -p ${SSH_DIR}
      chmod 700 ${SSH_DIR}
      curl $KEY_URL >> ${SSH_DIR}/authorized_keys
      touch ${FIRST_BOOT_FLAG}
    fi
  fi
  if [ "$S(DOCKER)" != "true" ] ; then
    if [ -f /sys/kernel/mm/redhat_transparent_hugepage/defrag ] ; then
      echo never > /sys/kernel/mm/redhat_transparent_hugepage/defrag
    fi
  fi
cloudera-quickstart-ip
[root@lt121 ~]
```

## 2. Installing wget package

```
[root@quickstart ~]# sudo yum -y install wget
Loaded plugins: fastestmirror
Setting up Install Process
Determining fastest mirrors
  * base: centos.hbase.tifr.res.in
  * epel: mirrors.piconetw.webworks.in
  * extras: centos.hbase.tifr.res.in
  * updates: centos.hbase.tifr.res.in
base
base/primary_db
cloudera-accumulo
cloudera-accumulo/primary
cloudera-accumulo
cloudera-cdh5
cloudera-cdh5/primary
cloudera-cdh5
cloudera-gpl-extras
cloudera-gpl-extras/primary
cloudera-kafka
cloudera-kafka
cloudera-kafka
cloudera-manager
cloudera-manager/primary
cloudera-manager
```

## 3. Making directory temp and moving into it

```
[root@quickstart ~]# mkdir temp
[root@quickstart ~]# cd temp
[root@quickstart temp]#
```

#### 4.Create file with some content

```
Activities Terminal Mar 6 19:09 root@it121:~ root@it121:~/devops/mongo
root@it121:~#
[root@quickstart temp]# echo "This is wordcount program on docker and it is running successfully by taking input in wordcount file and giving output.">wordcount.txt
[root@quickstart temp]# clear
```

#### 5.Make input directory in HDFS system

```
Activities Terminal Mar 6 18:53 @quickstart:/temp Mar 6 18:56 root@it121:~/devops/mongo
@quickstart:/temp
root@it121:~/devops/mongo
[root@quickstart temp]# hdfs dfs -mkdir /user/cloudera/input
[root@quickstart temp]#
```

#### 6.Copy file from local directory to HDFS file system

```
Activities Terminal Mar 6 18:56 @quickstart:/temp Mar 6 18:57 root@it121:~/devops/mongo
@quickstart:/temp
root@it121:~/devops/mongo
[root@quickstart temp]# hdfs dfs -put /temp/wordcount.txt /user/cloudera/input/
[root@quickstart temp]#
```

#### 7.To check if your file is successfully or not

```
Activities Terminal Mar 6 18:57 @quickstart:/temp Mar 6 19:00 root@it121:~/devops/mongo
@quickstart:/temp
root@it121:~/devops/mongo
[root@quickstart temp]# hdfs dfs -ls /user/cloudera/input/
Found 1 items
-rw-r--r-- 1 root cloudera 120 2020-03-06 13:26 /user/cloudera/input/wordcount.txt
[root@quickstart temp]#
```

#### 8.To check Hadoop mapreduce examples

```
Activities Terminal Mar 6 19:12 root@it121:~ root@it121:~/devops/mongo
root@it121:~#
[root@quickstart temp]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar .
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondariesort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
```

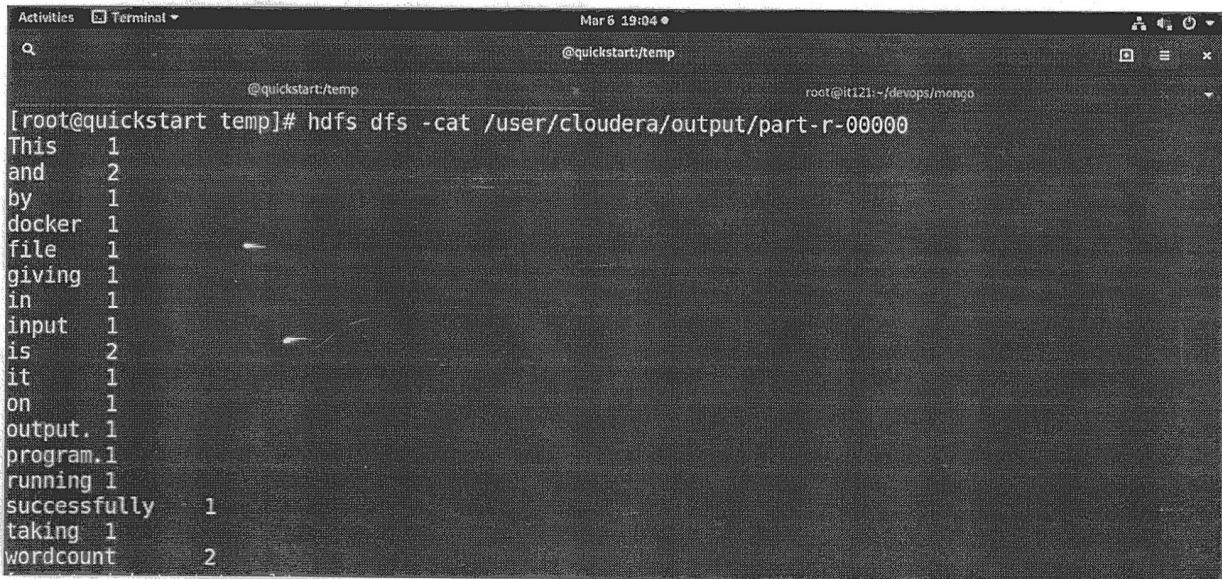
## 9.Run the wordcount program

```
Activities Terminal Mar 6 19:03 @quickstart/temp root@it121:~/devops/mongo  
[root@quickstart temp]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /user/cloudera/input/wordcount.txt /user/cloudera/output  
20/03/06 13:31:43 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032  
20/03/06 13:31:44 INFO input.FileInputFormat: Total input paths to process : 1  
20/03/06 13:31:44 INFO mapreduce.JobSubmitter: number of splits:1  
20/03/06 13:31:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1583500408903_0002  
20/03/06 13:31:46 INFO impl.YarnClientImpl: Submitted application application_1583500408903_0002  
20/03/06 13:31:46 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1583500408903_0002/  
20/03/06 13:31:46 INFO mapreduce.Job: Running job: job_1583500408903_0002  
20/03/06 13:31:59 INFO mapreduce.Job: Job job_1583500408903_0002 running in uber mode : false  
20/03/06 13:31:59 INFO mapreduce.Job: map 0% reduce 0%  
20/03/06 13:32:18 INFO mapreduce.Job: map 100% reduce 0%  
20/03/06 13:32:33 INFO mapreduce.Job: map 100% reduce 100%  
20/03/06 13:32:34 INFO mapreduce.Job: Job job_1583500408903_0002 completed successfully  
20/03/06 13:32:34 INFO mapreduce.Job: Counters: 49  
File System Counters  
FILE: Number of bytes read=211  
FILE: Number of bytes written=227973  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0
```

## 10.Check output directory

```
Activities Terminal Mar 6 19:03 @quickstart/temp root@it121:~/devops/mongo  
[root@quickstart temp]# hdfs dfs -ls /user/cloudera/output  
Found 2 items  
-rw-r--r-- 1 root cloudera 0 2020-03-06 13:32 /user/cloudera/output/_SUCCESS  
-rw-r--r-- 1 root cloudera 137 2020-03-06 13:32 /user/cloudera/output/part-r-00000  
[root@quickstart temp]#
```

## 11. Open part file to check output



A screenshot of a terminal window titled "Terminal". The window shows a command-line interface with the following text:

```
Activities Terminal Mar 6 19:04 • @quickstart/temp root@lt121:/devops/mongo
[root@quickstart temp]# hdfs dfs -cat /user/cloudera/output/part-r-00000
This      1
and       2
by        1
docker    1
file      1
giving    1
in        1
input     1
is        2
it        1
on        1
output.   1
program.  1
running.  1
successfully  1
taking     1
wordcount  2
```



### Learning Outcomes Achieved:

- 1) HDFS is discussed
- 2) Students downloaded cloudera quickstart VM.
- 3) Discussed how Map Reduce works.

### Conclusion:

1. Applications of the studied technique in industry.
  - a. Run the most basic example program Wordcount using your input data.
2. Engineering relevance
  - a. Perform some simple task in HDFS.
3. Skills Developed
  - a. Configuration of cloudera quickstart VM.