



**SUB: BIG DATA ENGINEERING**  
**AY 2023-24 (Semester-VI)**

*Bhuvi Ghosh*  
*60009210191*

**Experiment No: 1**

**Batch: D22**

**Aim: Hadoop Installation & Map Reduce.**

The screenshot shows a Windows File Explorer window with the address bar set to 'This PC > Local Disk (C:) > hadoop-3.3.0 > bin'. The left sidebar shows 'Quick access' with 'This PC' selected. Overlaid on the File Explorer is an 'Administrator: Command Prompt - start-dfs' window. The Command Prompt displays the following text:

```
Microsoft Windows [Version 10.0.18363.1440]  
(c) 2019 Microsoft Corporation. All rights reserved.  
  
C:\WINDOWS\system32>cd..  
C:\Windows>cd..  
C:\>cd C:\hadoop-3.3.0\bin  
C:\hadoop-3.3.0\bin>cd..  
C:\hadoop-3.3.0>cd sbin  
C:\hadoop-3.3.0\sbin>start-dfs
```



Department of Computer Science and Engineering (Data Science)

[illegible]



Department of Computer Science and Engineering (Data Science)

The screenshot shows a Windows File Explorer window with the address bar set to 'This PC > Local Disk (C:) > hadoop-3.3.0 > bin'. The left sidebar shows 'This PC' selected. Overlaid on this is an 'Administrator: Command Prompt' window. The command prompt shows the following sequence of commands and their outputs:

```
Microsoft Windows [Version 10.0.18363.1440]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd..
C:\Windows>cd..
C:\>cd C:\hadoop-3.3.0\bin
C:\hadoop-3.3.0\bin>cd..
C:\hadoop-3.3.0\bin>cd sbin
C:\hadoop-3.3.0\bin>start-dfs
C:\hadoop-3.3.0\bin>start-yarn
starting yarn daemons
C:\hadoop-3.3.0\bin>jps
5072 NameNode
6464 ResourceManager
4148 DataNode
5572 Jps
5940 NodeManager
C:\hadoop-3.3.0\bin>
```

Below the command prompt, the File Explorer window shows the same path, and the command prompt is again overlaid, showing the following commands:

```
C:\hadoop-3.3.0\bin>start-dfs
C:\hadoop-3.3.0\bin>start-yarn
starting yarn daemons
C:\hadoop-3.3.0\bin>jps
5072 NameNode
6464 ResourceManager
4148 DataNode
5572 Jps
5940 NodeManager
C:\hadoop-3.3.0\bin>hadoop fs -mkdir /input
C:\hadoop-3.3.0\bin>
```

The taskbar at the bottom shows the system clock as 17:56 on 13-05-2021. An 'Activate Windows' watermark is visible in the bottom right corner of the screenshot.

**Creating data.txt**



Department of Computer Science and Engineering (Data Science)

The screenshot shows a Windows File Explorer window with the address bar set to 'This PC > Local Disk (C:)'. The left sidebar shows 'This PC' selected. The main pane displays a list of folders and files. A 'data' folder is highlighted. A Notepad window titled 'data - Notepad' is open, showing the following text:

```
File Edit Format View Help
Hello
Hi
Hello
hi
good
morning
good day
bye
hello
Hello
Hi
```

Below the File Explorer window, a Command Prompt window titled 'Administrator: Command Prompt' is open, showing the following commands and output:

```
C:\Windows>cd..
C:\>cd C:\hadoop-3.3.0\bin
C:\hadoop-3.3.0\bin>cd..
C:\hadoop-3.3.0>cd sbin
C:\hadoop-3.3.0\sbin>start-dfs
C:\hadoop-3.3.0\sbin>start-yarn
starting yarn daemons
C:\hadoop-3.3.0\sbin>jps
5072 NameNode
6464 ResourceManager
4148 DataNode
5572 Jps
5940 NodeManager
C:\hadoop-3.3.0\sbin>hadoop fs -mkdir /input
C:\hadoop-3.3.0\sbin>hadoop fs -put C:/data.txt /input
C:\hadoop-3.3.0\sbin>hadoop fs -ls /input/
Found 1 items
-rw-r--r-- 1 admin supergroup 69 2021-05-13 17:58 /input/data.txt
C:\hadoop-3.3.0\sbin>
```



Department of Computer Science and Engineering (Data Science)

The screenshot displays a Windows File Explorer window and a Command Prompt window. The File Explorer window shows the path `This PC > Local Disk (C:) > hadoop-3.3.0 > share > hadoop > mapreduce`. The Command Prompt window shows the following commands and output:

```
C:\hadoop-3.3.0\sbin>jps
5072 NameNode
6464 ResourceManager
4148 DataNode
5572 Jps
5940 NodeManager

C:\hadoop-3.3.0\sbin>hadoop fs -mkdir /input

C:\hadoop-3.3.0\sbin>hadoop fs -put C:/data.txt /input

C:\hadoop-3.3.0\sbin>hadoop fs -ls /input/
Found 1 items
-rw-r--r-- 1 admin supergroup 69 2021-05-13 17:58 /input/data.txt

C:\hadoop-3.3.0\sbin>hadoop dfs -cat /input/data.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Hello
Hi
Hello
hi
good
morning
good day
bye
hello
Hello
Hi
C:\hadoop-3.3.0\sbin>
```

The Command Prompt window also shows the following commands and output:

```
C:\hadoop-3.3.0\sbin>jps
5072 NameNode
6464 ResourceManager
4148 DataNode
5572 Jps
5940 NodeManager

C:\hadoop-3.3.0\sbin>hadoop fs -mkdir /input

C:\hadoop-3.3.0\sbin>hadoop fs -put C:/data.txt /input

C:\hadoop-3.3.0\sbin>hadoop fs -ls /input/
Found 1 items
-rw-r--r-- 1 admin supergroup 69 2021-05-13 17:58 /input/data.txt

C:\hadoop-3.3.0\sbin>hadoop dfs -cat /input/data.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Hello
Hi
Hello
hi
good
morning
good day
bye
hello
Hello
Hi
C:\hadoop-3.3.0\sbin>hadoop jar C:/hadoop-3.3.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0
```





Department of Computer Science and Engineering (Data Science)

```
Administrator: Command Prompt

Hello
Hi
C:\hadoop-3.3.0\sbin>hadoop jar C:/hadoop-3.3.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.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.
secondarysort: 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.
teragen: Generate data for the terasort
terasort: Run the terasort
teravalidate: Checking results of terasort
wordcount: A map/reduce program that counts the words in the input files.
wordmean: A map/reduce program that counts the average length of the words in the input files.
wordmedian: A map/reduce program that counts the median length of the words in the input files.
wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the words in the input files.

C:\hadoop-3.3.0\sbin>
C:\hadoop-3.3.0\sbin>hadoop jar C:/hadoop-3.3.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar wordcount /input /out
2021-05-13 18:03:10,870 INFO client.DefaultNoHARMFalloverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-05-13 18:03:12,848 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/admin/.staging/job_1620908816216_0001
2021-05-13 18:03:13,915 INFO input.FileInputFormat: Total input files to process : 1
2021-05-13 18:03:14,429 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-13 18:03:15,517 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1620908816216_0001
2021-05-13 18:03:15,517 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-13 18:03:16,093 INFO conf.Configuration: resource-types.xml not found
2021-05-13 18:03:16,096 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-13 18:03:17,085 INFO impl.YarnClientImpl: Submitted application application_1620908816216_0001
2021-05-13 18:03:17,255 INFO mapreduce.Job: The url to track the job: http://DESKTOP-7LC2104:8088/proxy/application_1620908816216_0001/
2021-05-13 18:03:17,257 INFO mapreduce.Job: Running job: job_1620908816216_0001
2021-05-13 18:03:40,831 INFO mapreduce.Job: Job job_1620908816216_0001 running in uber mode : false
2021-05-13 18:03:40,837 INFO mapreduce.Job: map 0% reduce 0%
2021-05-13 18:03:54,353 INFO mapreduce.Job: map 100% reduce 0%

Administrator: Command Prompt

Reduce input groups=8
Reduce shuffle bytes=93
Reduce input records=8
Reduce output records=8
Spilled Records=16
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=205
CPU time spent (ms)=3294
Physical memory (bytes) snapshot=449261568
Virtual memory (bytes) snapshot=615825408
Total committed heap usage (bytes)=286785536
Peak Map Physical memory (bytes)=294645760
Peak Map Virtual memory (bytes)=396242944
Peak Reduce Physical memory (bytes)=154615808
Peak Reduce Virtual memory (bytes)=219676672

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=69
File Output Format Counters
Bytes Written=55

C:\hadoop-3.3.0\sbin>
```



The screenshot shows a Windows desktop environment. On the left is the 'This PC' sidebar with various icons. In the center, an 'Administrator: Command Prompt' window is open, displaying the following output:

```
Physical memory (bytes) snapshot=449261568
Virtual memory (bytes) snapshot=615825408
Total committed heap usage (bytes)=286785536
Peak Map Physical memory (bytes)=294645760
Peak Map Virtual memory (bytes)=396242944
Peak Reduce Physical memory (bytes)=154615808
Peak Reduce Virtual memory (bytes)=219676672

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=69
File Output Format Counters
Bytes Written=55

C:\hadoop-3.3.0\sbin>hadoop fs -cat /out/*
Hello 3
Hi 2
bye 1
day 1
good 2
hello 1
hi 1
morning 1

C:\hadoop-3.3.0\sbin>
```

Below the Command Prompt, a web browser window is open to 'localhost:9870/explorer.html#/'. The browser's address bar shows the URL. The page has a green header with the title 'Hadoop' and several navigation tabs: Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled 'Browse Directory'.

## Browse Directory

/

Go!

Show

25

▼

entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		drwxr-xr-x		admin		supergroup		0 B		May 13 17:58		0		0 B		input	
<input type="checkbox"/>		drwxr-xr-x		admin		supergroup		0 B		May 13 18:04		0		0 B		tmp	
<input type="checkbox"/>		drwxr-xr-x		admin		supergroup		0 B		May 13 18:03		0		0 B		tmp	

Showing 1 to 3 of 3 entries

Previous

1

Next



All Applications x Browsing HDFS x +

localhost:9870/explorer.html#/input

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

### Browse Directory

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	admin	supergroup	69 B	May 13 17:58	1	128 MB	data.txt	

Showing 1 to 1 of 1 entries

Previous 1 Next

Download

Head the file (first 32K)

Tail the file (last 32K)

Block information — Block 0

Block ID: 1073741825

Block Pool ID: BP-1418600749-192.168.43.41-1620738523950

Generation Stamp: 1001

Size: 69

Availability:

- DESKTOP-7LC21O4

Close





Department of Computer Science and Engineering (Data Science)

Download      Head the file (first 32K)      Tail the file (last 32K)

Block information — Block 0 ▾

Block ID: 1073741825  
Block Pool ID: BP-1418600749-192.168.43.41-1620738523950  
Generation Stamp: 1001  
Size: 69  
Availability:  
• DESKTOP-7LC2104

File contents

```
hi  
good  
morning  
good day  
bye  
hello  
Hello  
Hi
```

Hadoop   Overview   Datanodes

Browse Directory

/out

Show 25 ▾ entries

Permission	Owner
-rw-r--r--	admin
-rw-r--r--	admin

Showing 1 to 2 of 2 entries

Block information — Block 0 ▾

Block ID: 1073741832  
Block Pool ID: BP-1418600749-192.168.43.41-1620738523950  
Generation Stamp: 1008  
Size: 55  
Availability:  
• DESKTOP-7LC2104

File contents

```
Hello 3  
Hi 2  
bye 1  
day 1  
good 2  
hello 1
```

Block Size   Name

MB	_SUCCESS
MB	part-r-00000

Previous 1 Next