

LAB REPORT

Exercise-3

By

Ambuj Mishra

202116003

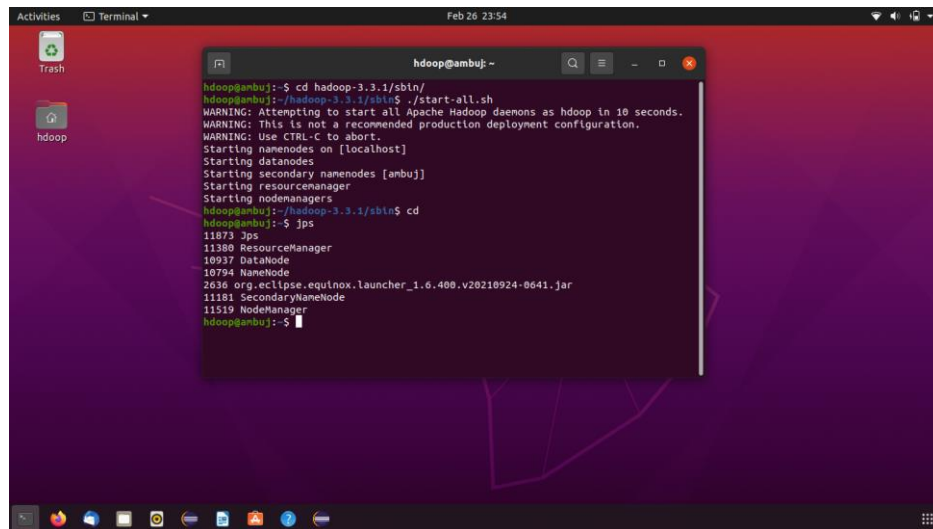
Big Data & Large-Scale Computing

DA-IICT, Gandhinagar

31-March-2022

Environment setup:

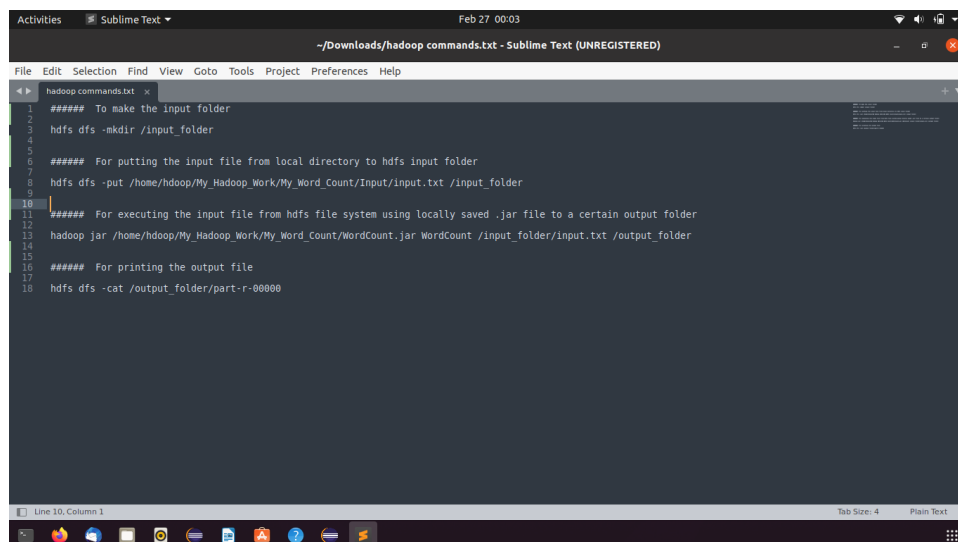
Before moving to the questions, we must first set up the hdfs environment to make all the nodes and resource managers running.



```
hadoop@ambuj:~$ cd hadoop-3.3.1/sbin/
hadoop@ambuj:~/hadoop-3.3.1/sbin$ ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ambuj]
Starting resourcemanager
Starting nodemanagers
hadoop@ambuj:~/hadoop-3.3.1/sbin$ cd
hadoop@ambuj:~$ jps
11873 Jps
11380 ResourceManager
10937 DataNode
10794 NameNode
2616 org.eclipse.equinox.launcher_1.6.400.v20210924-0641.jar
11181 SecondaryNameNode
11519 NodeManager
hadoop@ambuj:~$
```

Figure 1: Starting the HADOOP environment

To run the java projects using .jar files on input data in hdfs file system, following commands are used:



```
File Edit Selection Find View Goto Tools Project Preferences Help
hadoop commands.txt x
1 ##### To make the input folder
2 hdfs dfs -mkdir /input_folder
3
4
5
6 ##### For putting the input file from local directory to hdfs input folder
7 hdfs dfs -put /home/hadoop/My_Hadoop_Work/My_Word_Count/Input/input.txt /input_folder
8
9
10 ##### For executing the input file from hdfs file system using locally saved .jar file to a certain output folder
11 hadoop jar /home/hadoop/My_Hadoop_Work/My_Word_Count/WordCount.jar WordCount /input_folder/input.txt /output_folder
12
13
14
15 ##### For printing the output file
16 hdfs dfs -cat /output_folder/part-r-00000
17
18
```

Figure 2: Commands to run a .jar file in hdfs file system

1st Question:

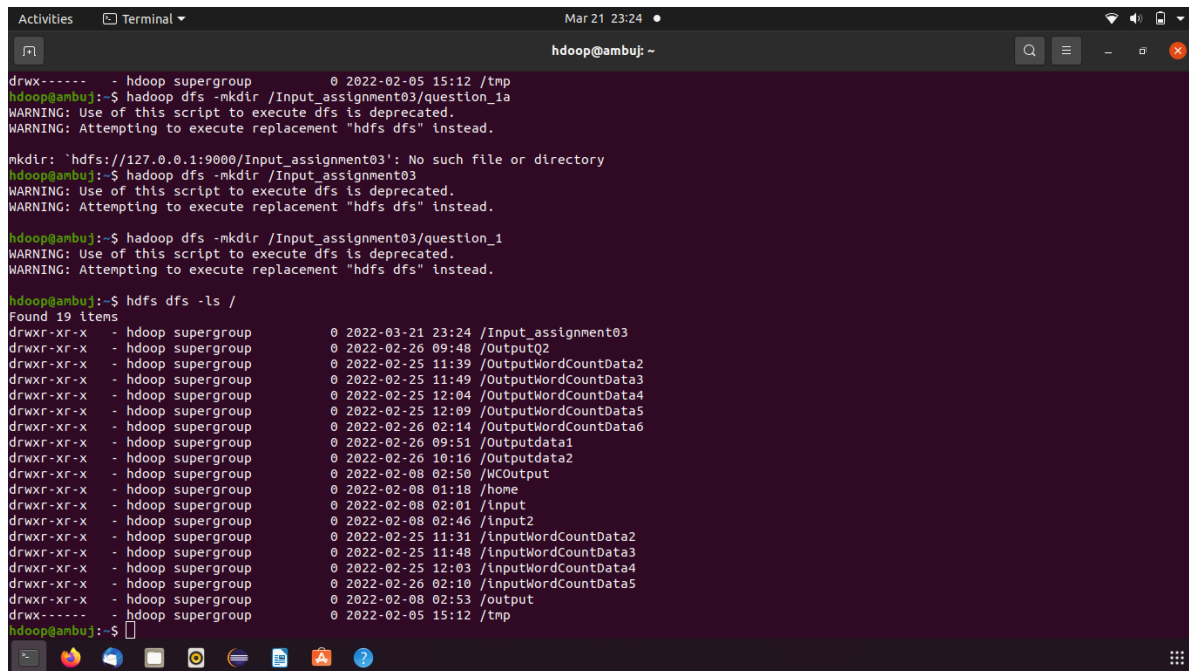
Part-(a)

Environment setup -

We have already set up the environment for Hadoop to work in pseudo-distributed mode. Now we'll run the following commands:

a.) mkdir -

First, we have listed the HDFS files and then we have created a new folder named "Input_assignment03" and inside this folder, we further created a subfolder 'question_1'.

A terminal window titled 'Terminal' with a dark background. The prompt is 'hadoop@ambuj: ~'. The user enters 'hadoop dfs -ls /' and the output shows a list of 19 items in HDFS, including directories like /Input_assignment03 and /Output, and files like /OutputWordCountData*. The user then enters 'hadoop dfs -mkdir /Input_assignment03/question_1' and receives a warning that the command is deprecated and is replaced by 'hdfs dfs'. The user then enters 'hadoop dfs -mkdir /Input_assignment03/question_1' and receives the same warning. Finally, the user enters 'hadoop dfs -ls /' and the output shows the updated list of items, including the newly created directory structure.

```
drwx----- - hadoop supergroup          0 2022-02-05 15:12 /tmp
hadoop@ambuj:~$ hadoop dfs -mkdir /Input_assignment03/question_1a
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

mkdir: 'hdfs://127.0.0.1:9000/Input_assignment03': No such file or directory
hadoop@ambuj:~$ hadoop dfs -mkdir /Input_assignment03
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

hadoop@ambuj:~$ hadoop dfs -mkdir /Input_assignment03/question_1
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

hadoop@ambuj:~$ hdfs dfs -ls /
Found 19 items
drwxr-xr-x - hadoop supergroup          0 2022-03-21 23:24 /Input_assignment03
drwxr-xr-x - hadoop supergroup          0 2022-02-26 09:48 /Outputq2
drwxr-xr-x - hadoop supergroup          0 2022-02-25 11:39 /OutputWordCountData2
drwxr-xr-x - hadoop supergroup          0 2022-02-25 11:49 /OutputWordCountData3
drwxr-xr-x - hadoop supergroup          0 2022-02-25 12:04 /OutputWordCountData4
drwxr-xr-x - hadoop supergroup          0 2022-02-25 12:09 /OutputWordCountData5
drwxr-xr-x - hadoop supergroup          0 2022-02-26 02:14 /OutputWordCountData6
drwxr-xr-x - hadoop supergroup          0 2022-02-26 02:51 /Outputdata1
drwxr-xr-x - hadoop supergroup          0 2022-02-26 10:16 /Outputdata2
drwxr-xr-x - hadoop supergroup          0 2022-02-08 02:50 /Output
drwxr-xr-x - hadoop supergroup          0 2022-02-08 01:18 /home
drwxr-xr-x - hadoop supergroup          0 2022-02-08 02:01 /Input
drwxr-xr-x - hadoop supergroup          0 2022-02-08 02:46 /Input2
drwxr-xr-x - hadoop supergroup          0 2022-02-25 11:31 /InputWordCountData2
drwxr-xr-x - hadoop supergroup          0 2022-02-25 11:48 /InputWordCountData3
drwxr-xr-x - hadoop supergroup          0 2022-02-25 12:03 /InputWordCountData4
drwxr-xr-x - hadoop supergroup          0 2022-02-26 02:10 /InputWordCountData5
drwxr-xr-x - hadoop supergroup          0 2022-02-08 02:53 /output
drwx----- - hadoop supergroup          0 2022-02-05 15:12 /tmp
hadoop@ambuj:~$
```

Fig: Using mkdir to create folder structure

b.) copyFromLocal -

We have used 'copyFromLocal' command to put the file in HDFS from our local file system.

```
Activities Terminal Mar 21 23:30
hadoop@ambuj:~$ hadoop dfs -mkdir /Input_assignment03
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

hadoop@ambuj:~$ hadoop dfs -mkdir /Input_assignment03/question_1
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

hadoop@ambuj:~$ hdfs dfs -ls /
Found 19 items
drwxr-xr-x - hadoop supergroup 0 2022-03-21 23:24 /Input_assignment03
drwxr-xr-x - hadoop supergroup 0 2022-02-26 09:48 /OutputQ2
drwxr-xr-x - hadoop supergroup 0 2022-02-25 11:39 /OutputWordCountData2
drwxr-xr-x - hadoop supergroup 0 2022-02-25 11:49 /OutputWordCountData3
drwxr-xr-x - hadoop supergroup 0 2022-02-25 12:04 /OutputWordCountData4
drwxr-xr-x - hadoop supergroup 0 2022-02-25 12:09 /OutputWordCountData5
drwxr-xr-x - hadoop supergroup 0 2022-02-26 02:14 /OutputWordCountData6
drwxr-xr-x - hadoop supergroup 0 2022-02-26 09:51 /Outputdata1
drwxr-xr-x - hadoop supergroup 0 2022-02-26 10:16 /Outputdata2
drwxr-xr-x - hadoop supergroup 0 2022-02-08 02:50 /WCOuput
drwxr-xr-x - hadoop supergroup 0 2022-02-08 01:18 /home
drwxr-xr-x - hadoop supergroup 0 2022-02-08 02:01 /input
drwxr-xr-x - hadoop supergroup 0 2022-02-08 02:46 /input2
drwxr-xr-x - hadoop supergroup 0 2022-02-25 11:31 /inputWordCountData2
drwxr-xr-x - hadoop supergroup 0 2022-02-25 11:48 /inputWordCountData3
drwxr-xr-x - hadoop supergroup 0 2022-02-25 12:03 /inputWordCountData4
drwxr-xr-x - hadoop supergroup 0 2022-02-26 02:10 /inputWordCountData5
drwxr-xr-x - hadoop supergroup 0 2022-02-08 02:53 /output
drwxr-xr-x - hadoop supergroup 0 2022-02-05 15:12 /tmp

hadoop@ambuj:~$ hdfs dfs -ls /Input_assignment03/
Found 1 items
drwxr-xr-x - hadoop supergroup 0 2022-03-21 23:24 /Input_assignment03/question_1

hadoop@ambuj:~$ hdfs dfs -copyFromLocal /home/My Hadoop Work/Assignment03/Input/data.txt /Input_assignment03/question_1/input.txt
copyFromLocal: /home/My Hadoop Work/Assignment03/Input/data.txt: No such file or directory

hadoop@ambuj:~$ hdfs dfs -copyFromLocal /home/hadoop/My_Hadoop_Work/Assignment03/Input/data.txt /Input_assignment03/question_1/input.txt
hadoop@ambuj:~$
```

Fig: Using copyFromLocal to put the file in HDFS

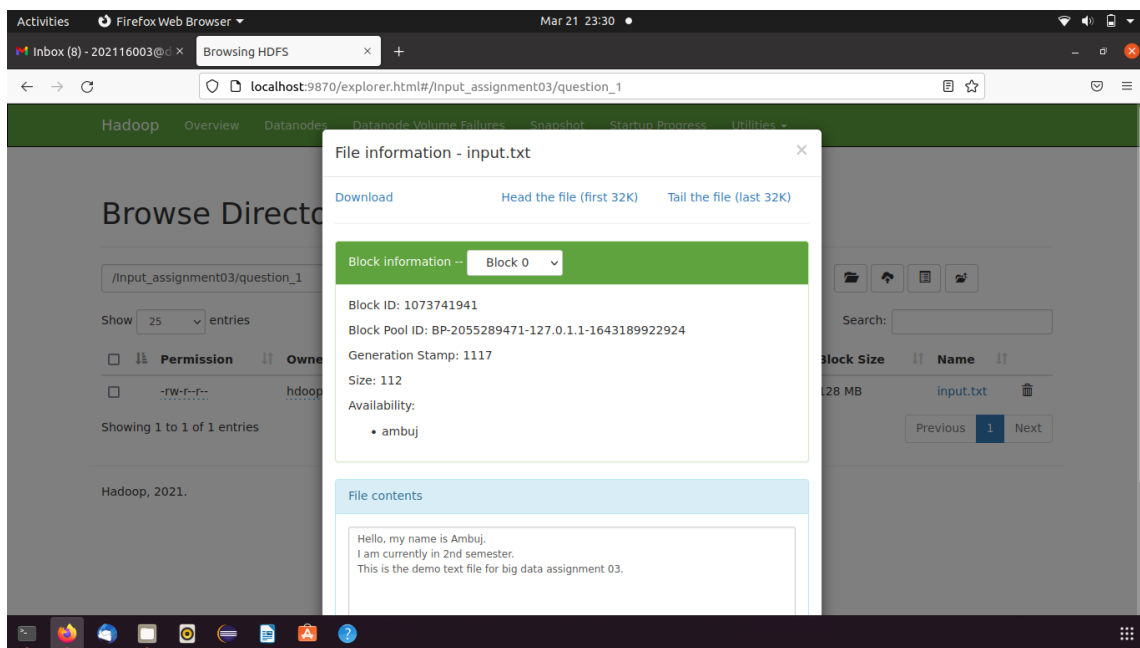
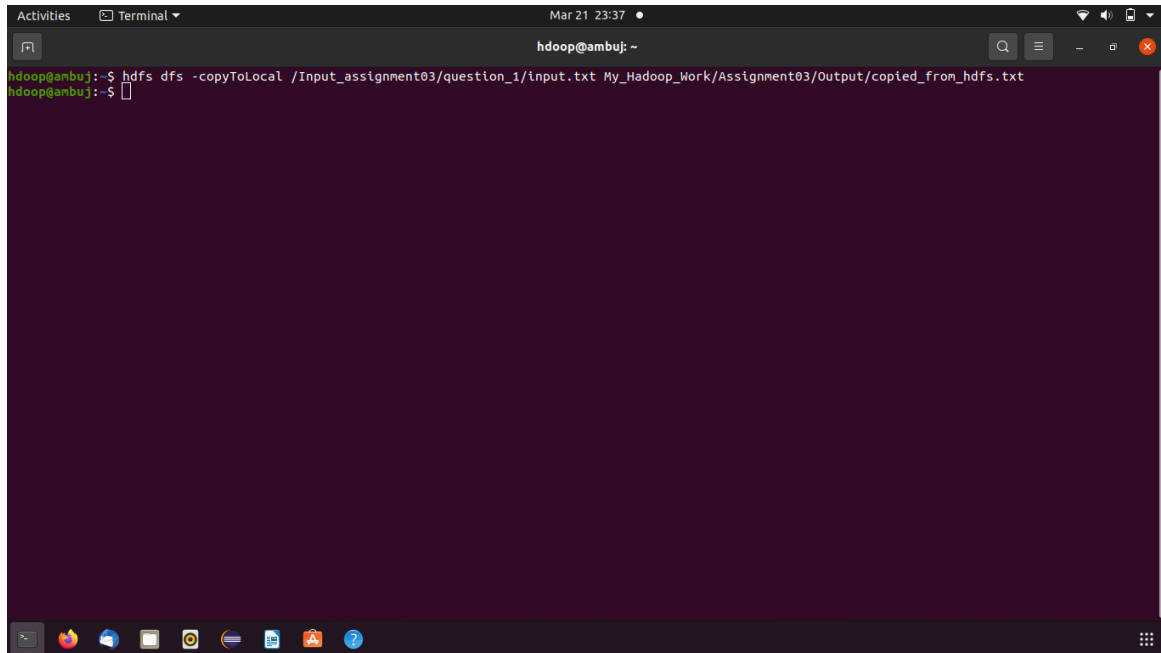


Fig: Checking the newly created file in HDFS

c.) copyToLocal -

We have used 'copyToLocal' command to put the file from HDFS from our local file system.



```
hadoop@anbu1:~$ hdfs dfs -copyToLocal /Input_assignment03/question_1/input.txt My_Hadoop_Work/Assignment03/Output/copied_from_hdfs.txt
hadoop@anbu1:~$
```

Fig: Using copyToLocal to put the file in local from HDFS

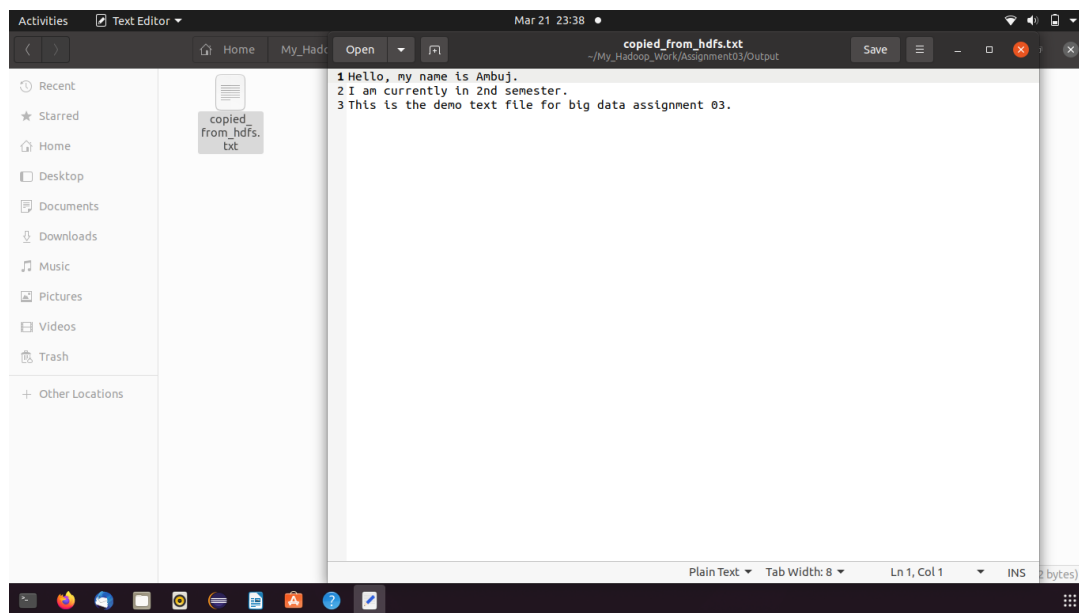
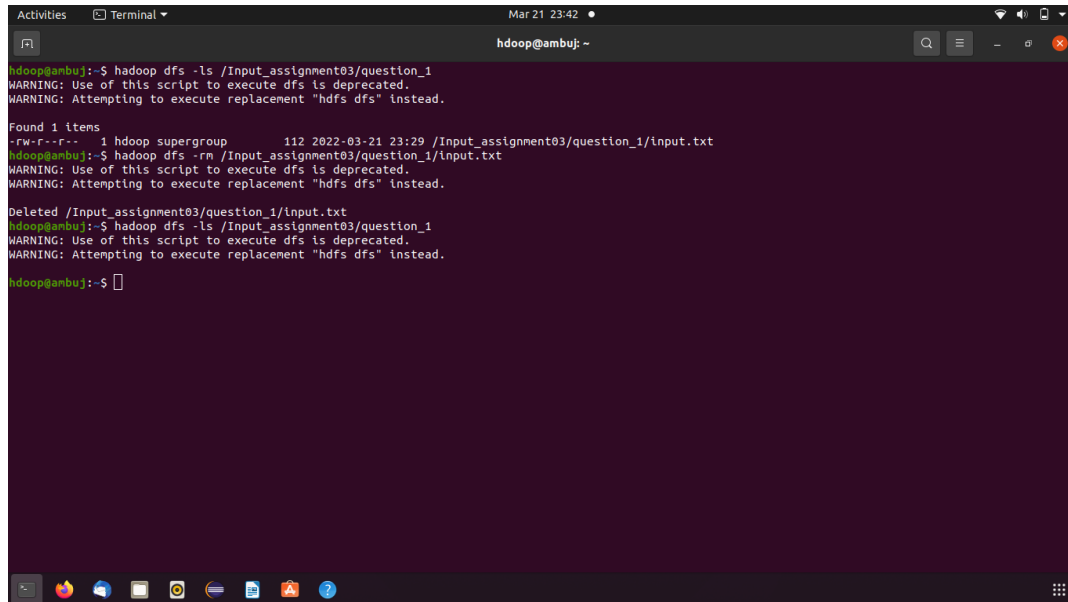


Fig: Checking the newly created file in local

d.) rm -

We have finally used 'rm' command to remove the 'input.txt' file from the question_1 folder.



```
hadoop@ambuj:~$ hadoop dfs -ls /Input_assignment03/question_1
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

Found 1 items
-rw-r--r-- 1 hadoop supergroup      112 2022-03-21 23:29 /Input_assignment03/question_1/input.txt
hadoop@ambuj:~$ hadoop dfs -rm /Input_assignment03/question_1/input.txt
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

Deleted /Input_assignment03/question_1/input.txt
hadoop@ambuj:~$ hadoop dfs -ls /Input_assignment03/question_1
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

hadoop@ambuj:~$
```

Fig: Deleting the file from HDFS using rm command

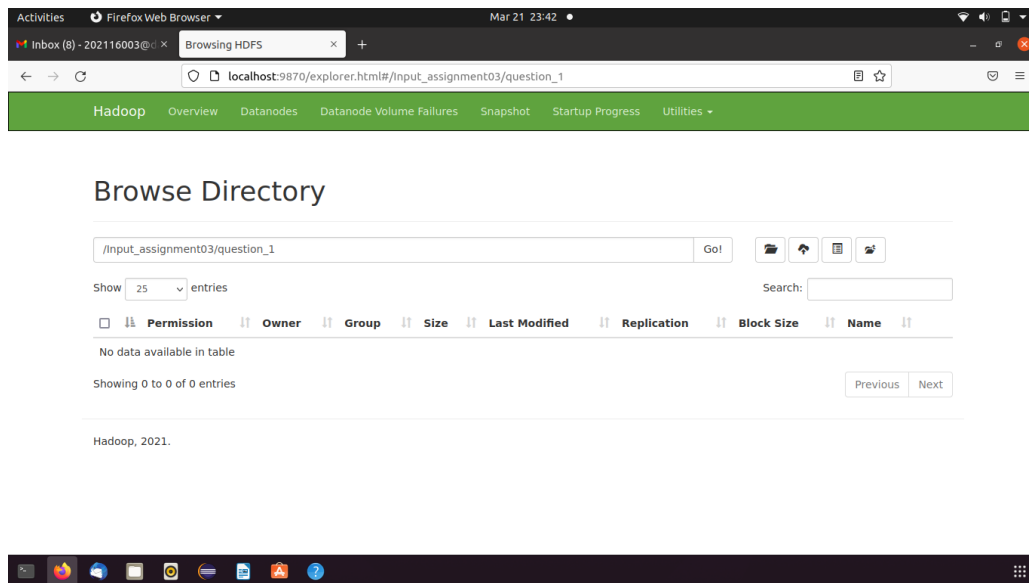


Fig: Confirming the results from interface

Part-(b)

After reading “Managing Files with the Hadoop File System Commands” in chapter 5 from Tom White’s book, we’ll be working on following 5 commands -

- a.) *'ls'*
- b.) *'lsr'*
- c.) *'put'*
- d.) *'get'*
- e.) *'mv'*

a.) 'ls' -

'ls' command is used to list all the files and folders present in the path given.

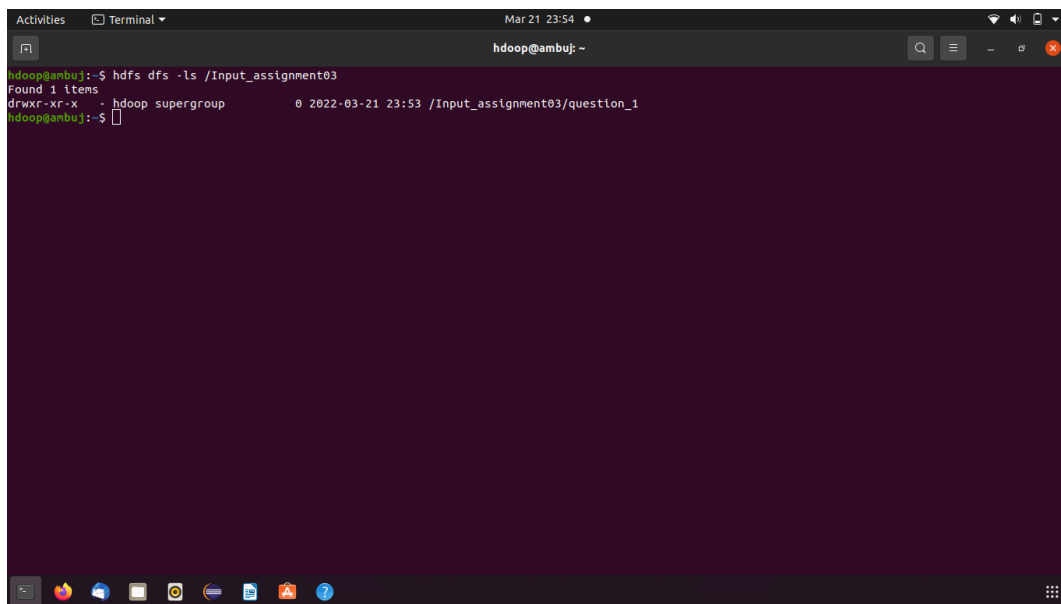
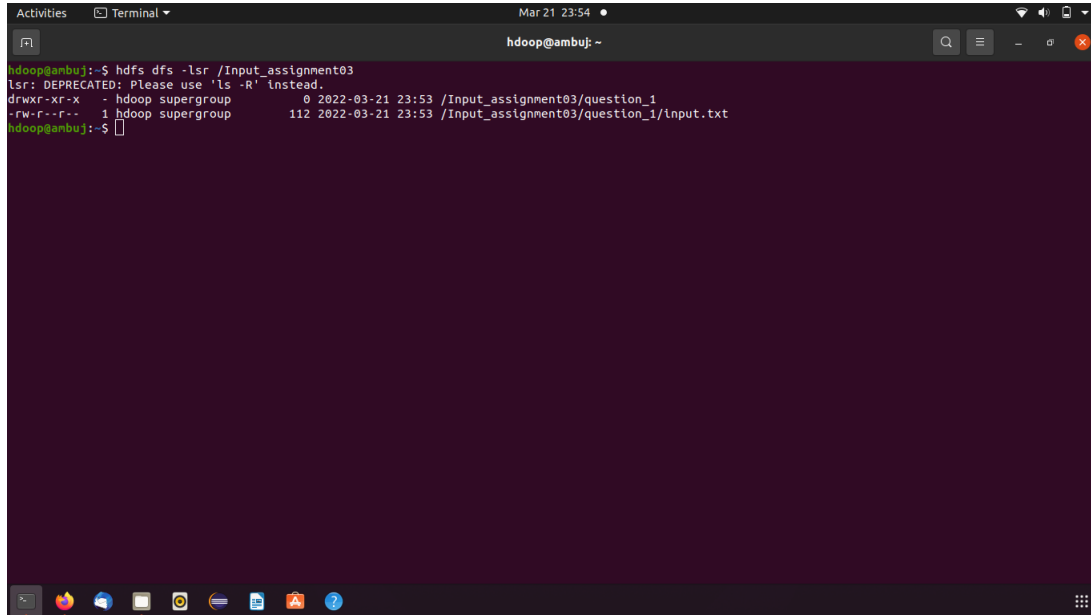
A terminal window titled 'Terminal' with a dark background. The prompt is 'hadoop@ambuj: ~'. The command entered is 'hdfs dfs -ls /Input_assignment03'. The output shows 'Found 1 items' followed by a line of file details: 'drwxr-xr-x - hadoop supergroup 0 2022-03-21 23:53 /Input_assignment03/question_1'. The prompt returns to 'hadoop@ambuj: ~'.

Fig: listing all the files and folders

b.) 'lsr' -

'lsr' command is used to recursively list all the files and folders present in the path given.

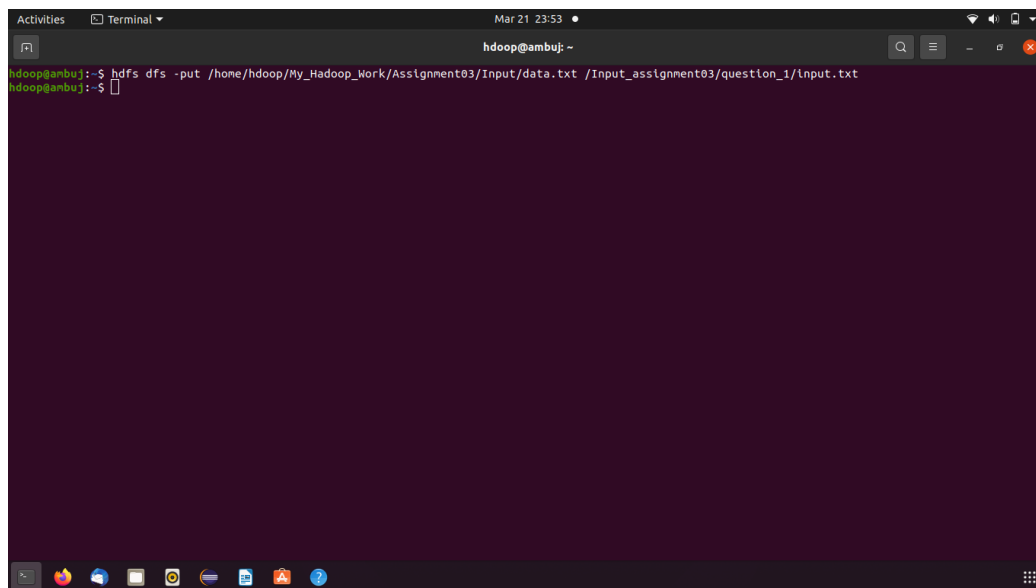


```
hadoop@anbu:~$ hdfs dfs -lsr /Input_assignment03
lsr: DEPRECATED: Please use 'ls -R' instead.
drwxr-xr-x  - hadoop supergroup          0 2022-03-21 23:53 /Input_assignment03/question_1
-rw-r--r--  1 hadoop supergroup        112 2022-03-21 23:53 /Input_assignment03/question_1/Input.txt
hadoop@anbu:~$
```

Fig: Recursively listing all the files and folders

c.) 'put' -

Similar to 'copyFromLocal' command, 'put' command is also used to transfer the file from local to HDFS.



```
hadoop@anbu:~$ hdfs dfs -put /home/hadoop/My_Hadoop_Work/Assignment03/Input/data.txt /Input_assignment03/question_1/input.txt
hadoop@anbu:~$
```

Fig: using 'put' command to transfer files from local to HDFS

d.) 'get' -

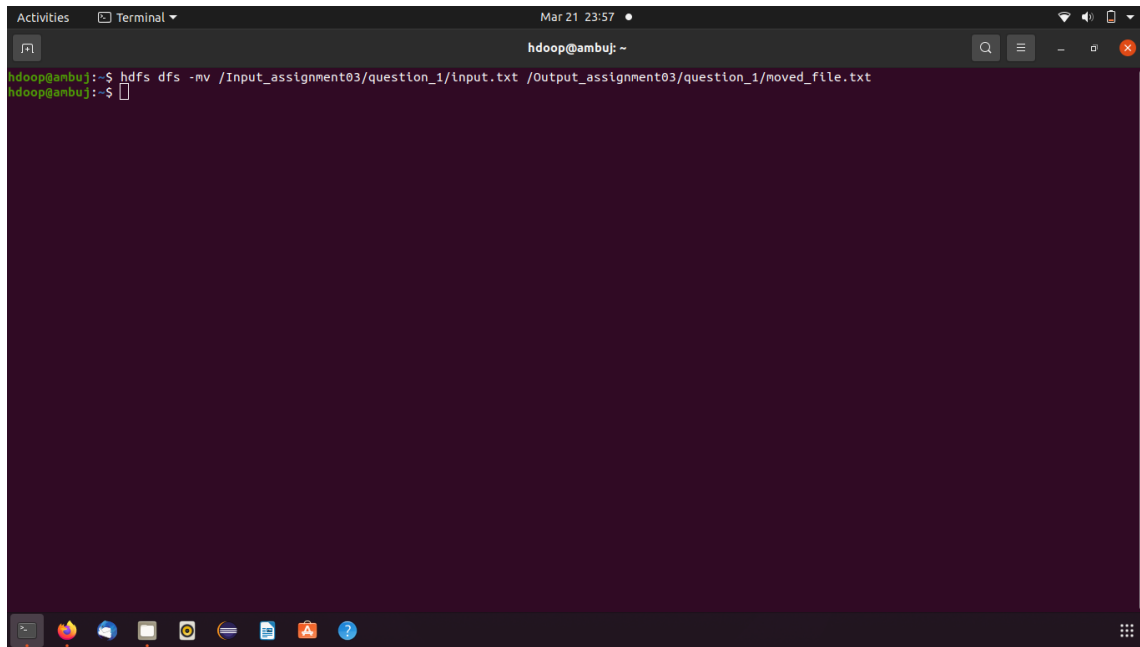
Similar to 'copyToLocal' command, 'get' command is also used to transfer the file from HDFS to local.

```
hadoop@ambuj:~$ cd hadoop-3.3.1/
hadoop@ambuj:~/hadoop-3.3.1$ cd/sbin/
hadoop@ambuj:~/hadoop-3.3.1/sbin$ ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ambuj]
Starting resourcemanager
Starting nodemanagers
hadoop@ambuj:~/hadoop-3.3.1/sbin$ hdfs dfs -get /output_assignment03/question_1/moved_file.txt /home/hadoop/pig/getfile.txt
hadoop@ambuj:~/hadoop-3.3.1/sbin$
```

Fig: using 'get' command to transfer files from HDFS to local

e.) 'mv' -

'mv' command is used to move any given file from source path to destination path. In this case, the file gets deleted from the source path as well.

A screenshot of a Linux terminal window. The window has a title bar with 'Activities', 'Terminal', and a dropdown arrow. The system clock shows 'Mar 21 23:57'. The terminal prompt is 'hadoop@ambuj: ~'. The command 'hdfs dfs -mv /Input_assignment03/question_1/input.txt /Output_assignment03/question_1/moved_file.txt' has been entered and executed. The prompt is now 'hadoop@ambuj:~\$' followed by a cursor. The terminal background is dark purple. The bottom of the window shows a taskbar with various application icons.

```
hadoop@ambuj:~$ hdfs dfs -mv /Input_assignment03/question_1/input.txt /Output_assignment03/question_1/moved_file.txt
hadoop@ambuj:~$
```

Fig: using 'mv' command to move file from one folder to another

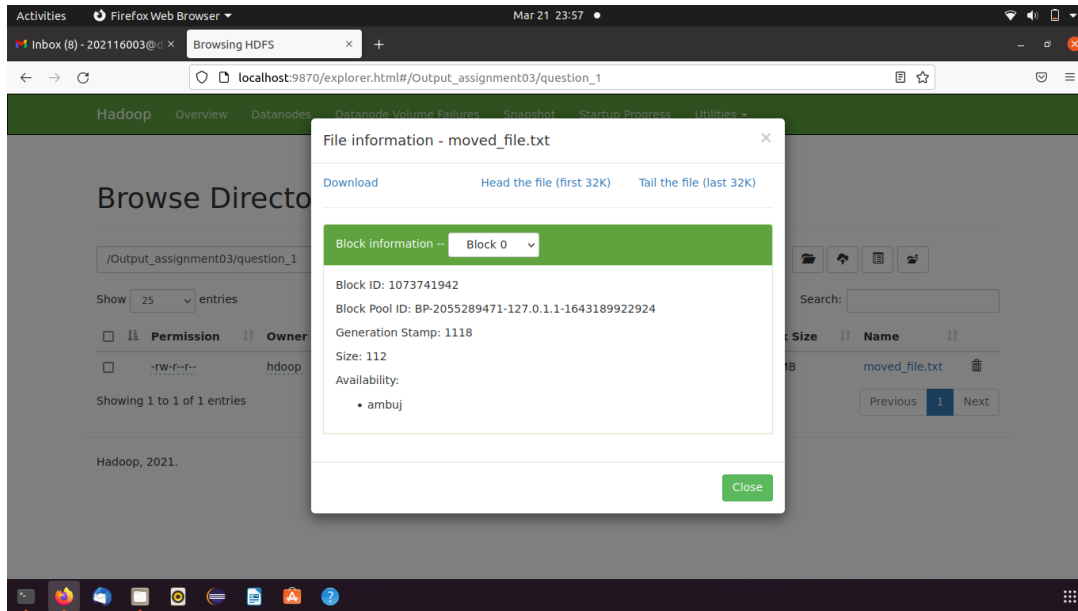


Fig: Cross-checking the newly created file at destination path

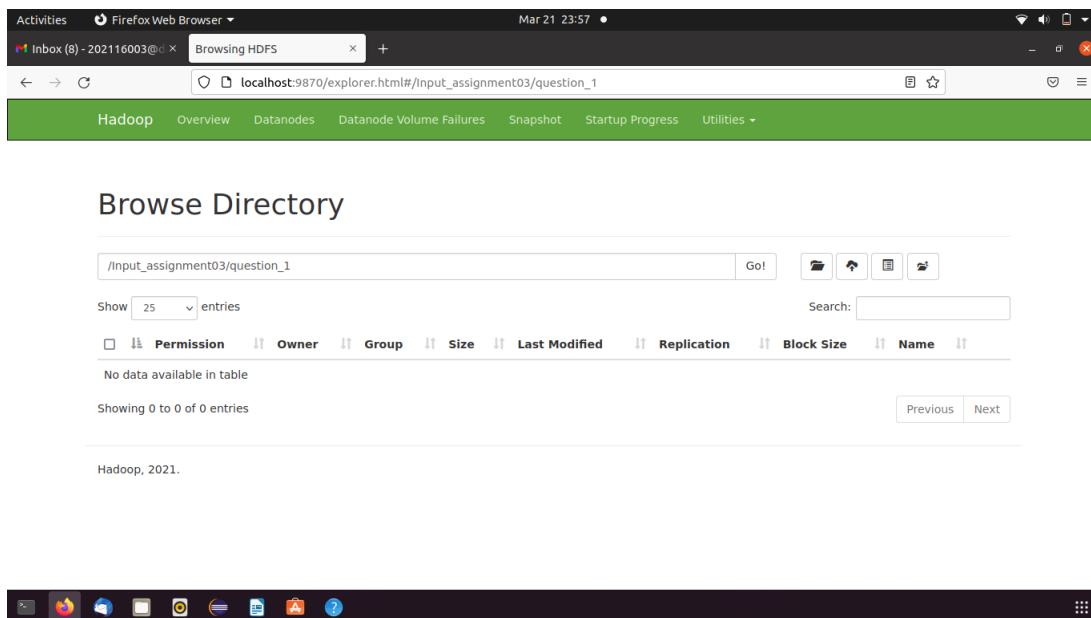
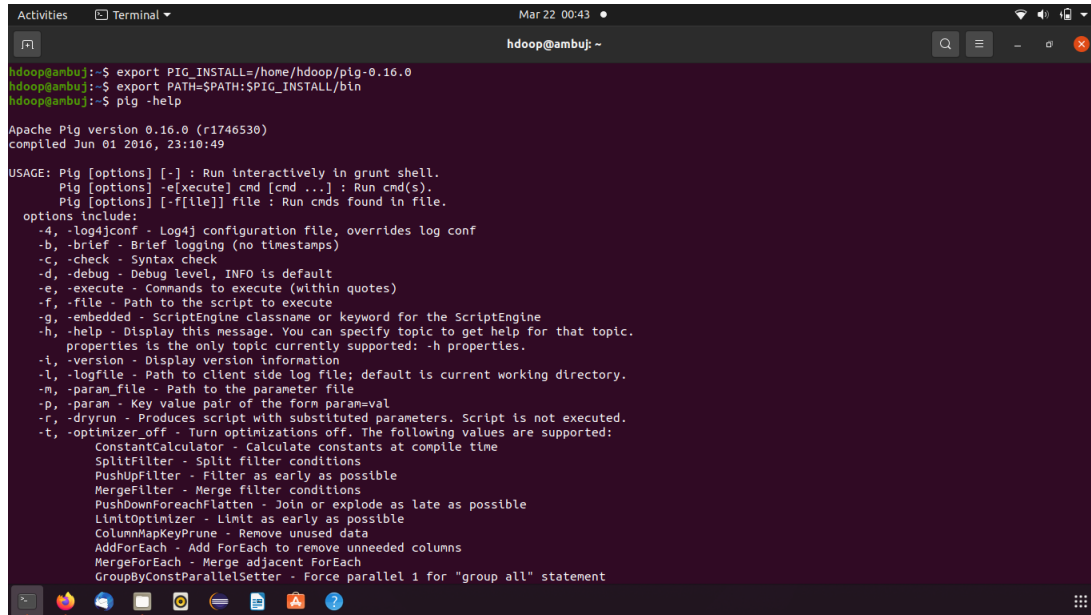


Fig: Cross-checking file deletion at source path

2nd Question:

Pig setup:

After setting up the pig, we need to pass the paths to the setup-



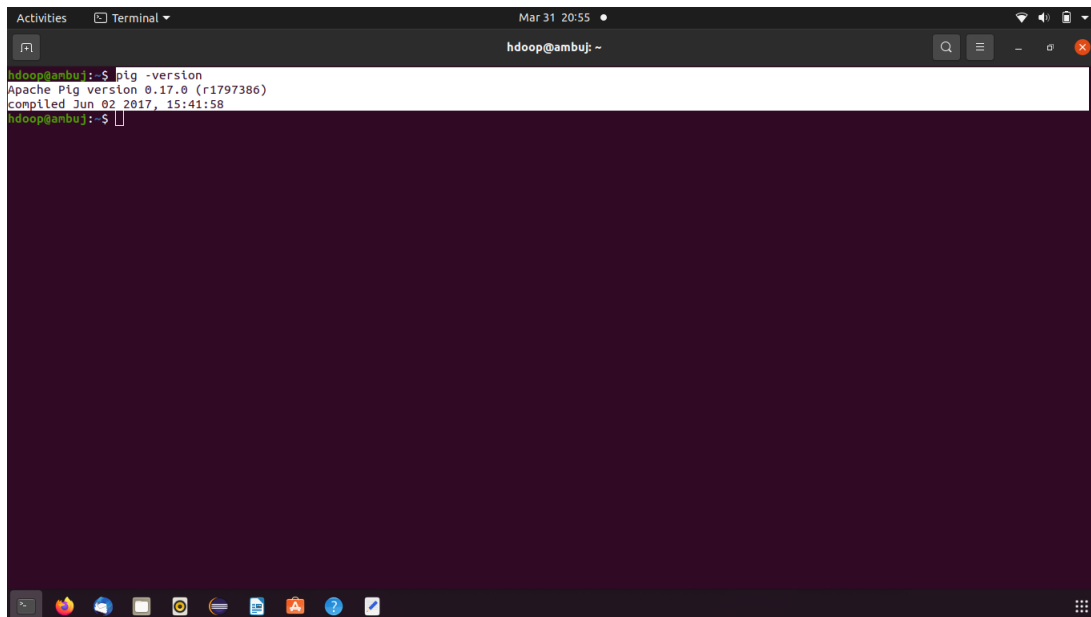
```
hadoop@ambuj:~$ export PIG_INSTALL=/home/hadoop/pig-0.16.0
hadoop@ambuj:~$ export PATH=$PATH:$PIG_INSTALL/bin
hadoop@ambuj:~$ pig -help

Apache Pig version 0.16.0 (r1746530)
compiled Jun 01 2016, 23:10:49

USAGE: Pig [options] [-] : Run interactively in grunt shell.
      Pig [options] -e[execute] cmd [cmd ...] : Run cmd(s).
      Pig [options] [-f[file]] file : Run cmds found in file.

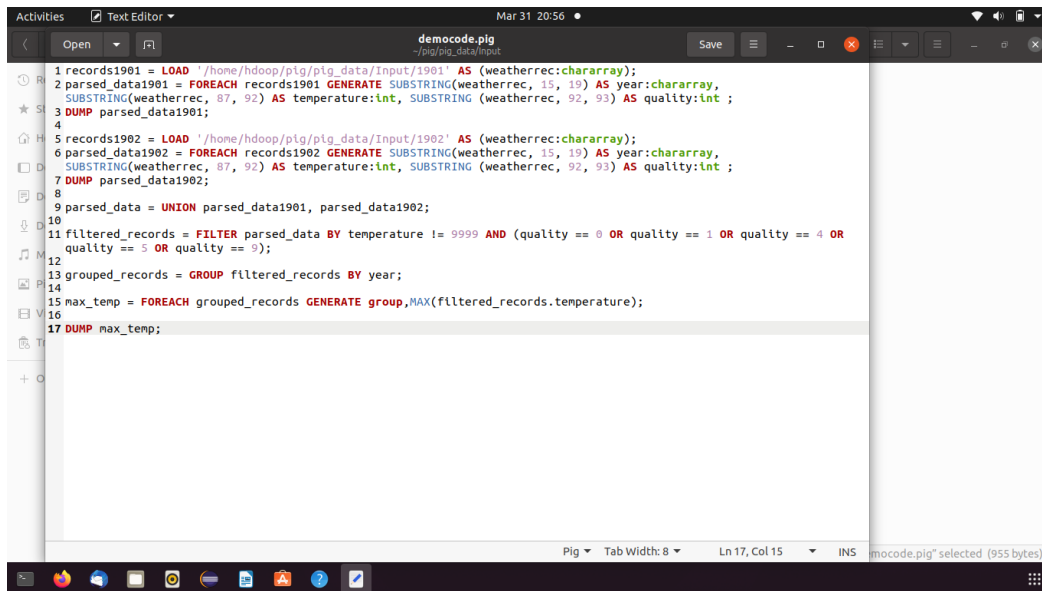
options include:
  -4, -log4jconf - Log4j configuration file, overrides log conf
  -b, -brief - Brief logging (no timestamps)
  -c, -check - Syntax check
  -d, -debug - Debug level, INFO is default
  -e, -execute - Commands to execute (within quotes)
  -f, -file - Path to the script to execute
  -g, -embedded - ScriptEngine classname or keyword for the ScriptEngine
  -h, -help - Display this message. You can specify topic to get help for that topic.
        properties is the only topic currently supported: -h properties.
  -i, -version - Display version information
  -l, -logfile - Path to client side log file; default is current working directory.
  -m, -param_file - Path to the parameter file
  -p, -param - Key value pair of the form param=val
  -r, -dryrun - Produces script with substituted parameters. Script is not executed.
  -t, -optimizer_off - Turn optimizations off. The following values are supported:
        ConstantCalculator - Calculate constants at compile time
        SplitFilter - Split filter conditions
        PushUpFilter - Filter as early as possible
        MergeFilter - Merge filter conditions
        PushDownForEachFlatten - Join or explode as late as possible
        LimitOptimizer - Limit as early as possible
        ColumnMapKeyPrune - Remove unused data
        AddForEach - Add ForEach to remove unneeded columns
        MergeForEach - Merge adjacent ForEach
        GroupByConstParallelSetter - Force parallel 1 for "group all" statement
```

Fig: setting up the paths for the pig



```
hadoop@ambuj:~$ pig -version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
hadoop@ambuj:~$
```

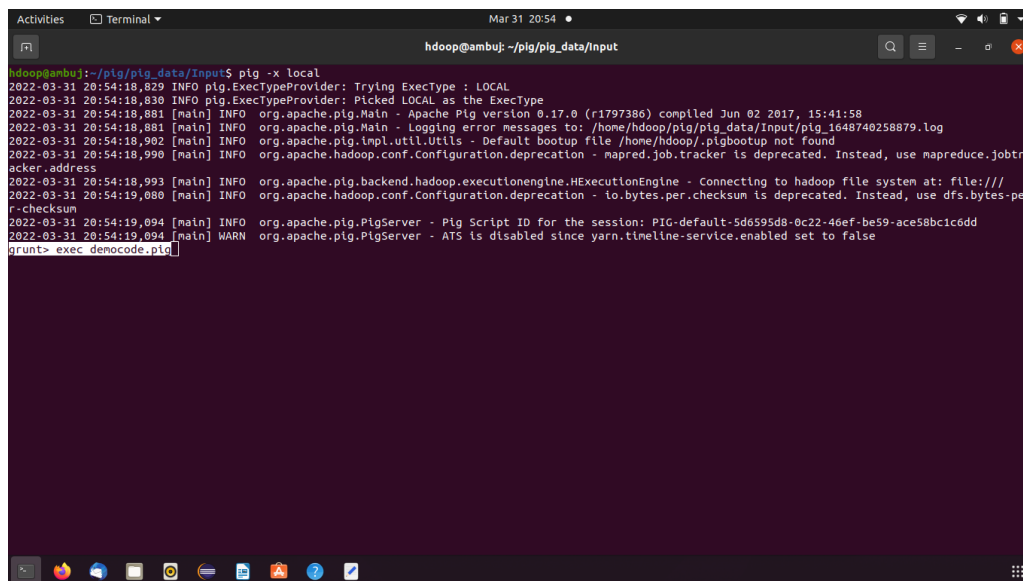
Fig: Checking the pig setup



The screenshot shows a text editor window titled 'democode.pig' with the following Pig Latin code:

```
1 records1901 = LOAD '/home/hadoop/pig/pig_data/Input/1901' AS (weatherrec:chararray);
2 parsed_data1901 = FOREACH records1901 GENERATE SUBSTRING(weatherrec, 15, 19) AS year:chararray,
3 SUBSTRING(weatherrec, 87, 92) AS temperature:int, SUBSTRING(weatherrec, 92, 93) AS quality:int;
4 DUMP parsed_data1901;
5 records1902 = LOAD '/home/hadoop/pig/pig_data/Input/1902' AS (weatherrec:chararray);
6 parsed_data1902 = FOREACH records1902 GENERATE SUBSTRING(weatherrec, 15, 19) AS year:chararray,
7 SUBSTRING(weatherrec, 87, 92) AS temperature:int, SUBSTRING(weatherrec, 92, 93) AS quality:int;
8 DUMP parsed_data1902;
9 parsed_data = UNION parsed_data1901, parsed_data1902;
10
11 filtered_records = FILTER parsed_data BY temperature != 9999 AND (quality == 0 OR quality == 1 OR quality == 4 OR
12 quality == 5 OR quality == 9);
13 grouped_records = GROUP filtered_records BY year;
14
15 max_temp = FOREACH grouped_records GENERATE group, MAX(filtered_records.temperature);
16
17 DUMP max_temp;
```

Fig: Code to work on the 1901 and 1902 files



The screenshot shows a terminal window with the following output:

```
hadoop@ambuj: ~/pig/pig_data/Input
hadoop@ambuj:~/pig/pig_data/Input$ pig -x local
2022-03-31 20:54:18,829 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2022-03-31 20:54:18,830 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2022-03-31 20:54:19,081 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2022-03-31 20:54:18,881 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig/pig_data/Input/pig_1648740258879.log
2022-03-31 20:54:18,902 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2022-03-31 20:54:18,990 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtr
acker.address
2022-03-31 20:54:18,993 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2022-03-31 20:54:19,080 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-pe
r-checksum
2022-03-31 20:54:19,094 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-5d6595d8-0c22-46ef-be59-ace58bc1c6dd
2022-03-31 20:54:19,094 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> exec democode.pig
```

Fig: Starting the pig locally and executing the code file

```

Job Stats (time in seconds):
JobId  Maps  Reduces MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime
job_local371775809_0003 2  1  n/a  n/a  n/a  n/a  n/a  n/a  filtered_records,grouped_records,max_temp,pars
ed_data,parsed_data1901,parsed_data1902,records1901,records1902  GROUP_BY,COMBINER  file:/tmp/temp-1771164005/tmp-1518473027,

Input(s):
Successfully read 6565 records from: "/home/hadoop/pig/pig_data/Input/1902"
Successfully read 6565 records from: "/home/hadoop/pig/pig_data/Input/1901"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp-1771164005/tmp-1518473027"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local371775809_0003

2022-03-31 20:54:50,094 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-03-31 20:54:50,096 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-03-31 20:54:50,097 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-03-31 20:54:50,103 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-31 20:54:50,104 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-pe
r-checksum
2022-03-31 20:54:50,104 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-31 20:54:50,106 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-03-31 20:54:50,106 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1901,317)
(1902,244)
grunt>

```

Fig: Final results

Short Note on pig -

Apache pig was performed to analyze larger sets of data representing them as data flows. Map-reduce algorithm takes a lot of time to perform map and reduce operations. Thus, we can say that apache Pig is an abstraction over MapReduce. Pig has 2 main components – pig latin and execution environment. Pig is built on top of hadoop. It has below advantages over map-reduce algorithm -

- Less development time
- Easy to learn
- Procedural language
- Dataflow
- Easy to control execution

Similarly, Apache Pig has the following disadvantages as well:

- Errors of Pig
 - Not mature
 - Support
 - Minor one
 - Implicit data schema
-