

Big Data and Large Scale Computing

Lab Report -03

July 21, 2025

Name : Rohan Baghel
Student ID: 202116011

About

Hadoop

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Apache Pig Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all the data manipulation operations in Hadoop using Pig.

Features of Apache Pig:

- For performing several operations Apache Pig provides rich sets of operators like the filters, join, sort, etc.
- Easy to learn, read and write. Especially for SQL-programmer, Apache Pig is a boon.
- Join operation is easy in Apache Pig.
- Fewer lines of code.

Question 1

Configure a Hadoop cluster to run in pseudo-distributed mode and do the following tasks:

1. Run the commands (1) `mkdir`, (2) `rm`, (3) `copyFromLocal`, (4) `copyToLocal` and observe the output.
2. Read “Managing Files with the Hadoop File System Commands” in chapter 5 from [3]. Try out any 5 commands from Table 5-2 other than those asked for in (1-a).

Answer :

(a)

To run the Command in HDFS, firstly we need to start the Hadoop cluster.

- start the NameNode and DataNode
- start the YARN resource and nodemanagers

Use the command to all the nodes at once

```
$ hadoop start-all.sh
```

To check if all the daemons are active and running as Java processes use:

```
$ jps
```

Now run the commands

1. **mkdir**

Creates directories on one or more specified paths. Its behavior is similar to the Unix `mkdir -p` command, which creates all directories that lead up to the specified directory if they don't exist already.

```
$ hadoop fs -mkdir /new_file_name
```

2. **rm**

Deletes one or more specified files. This command doesn't delete empty directories or files. To bypass the trash (if it's enabled) and delete the specified files immediately, specify the `-skipTrash` option.

```
$ hadoop fs -rm /folder/sample.txt
```

It will remove `sample.txt` file

3. **copyFromLocal**

Works similarly to the put command, except that the source is restricted to a local file reference.

```
$ hadoop fs -copyFromLocal sample.txt /folder1/folder2
```

It will copy sample.txt from local machine to folder2

4. **copyToLocal**

Works similarly to the get command, except that the destination is restricted to a local file reference.

```
$ hadoop fs -copyToLocal /folder1/folder2/sample.txt sample2.txt
```

It will copy sample.txt to local machine as sample2.txt

(b)

Some extra commands

1. **put**

Copies files from the local file system to the destination file system. This command can also read input from stdin and write to the destination file system.

```
$ hadoop fs -put sample.txt /folder1/folder2
```

It will copy sample.txt from local machine to folder2

2. **get**

Works similarly to the get command, except that the destination is restricted to a local file reference.

```
$ hadoop fs -get /folder1/folder2/sample.txt sample2.txt
```

It will copy sample.txt to local machine as sample2.txt

3. **cp**

Copies one or more files from a specified source to a specified destination. If you specify multiple sources, the specified destination must be a directory.

```
$ hadoop fs -cp /folder1/folder2 /folder1/folder3
```

It will copy files of folder2 to another folder name folder3

4. **rmdir**

The `rmdir` command removes the directory, specified by the `Directory` parameter, from the system. The directory must be empty before you can remove it, and you must have write permission in its parent directory. Use the `ls -al` command to check whether the directory is empty. The directory must not contain any folder or file.

```
$ hadoop fs -rmdir /folder/removing_directory
```

It will delete/remove the directory permanently

5. **du**

Displays the size of the specified file, or the sizes of files and directories that are contained in the specified directory. If you specify the `-s` option, displays an aggregate summary of file sizes rather than individual file sizes. If you specify the `-h` option, formats the file sizes in a "human-readable" way.

```
$ hadoop fs -du -h /folder/folder2/sample.txt
```

It returns the size of file, `-h` is used for human readable.

Question 2

Refer to chapter 11 (Pig) from [2], and read up to “Generating Examples”. Try out the installation and other commands given in there.

Under the topic “An Example”, with regard to working in “grunt” shell on the weather data set example, you have two options: (1) download the data files for years 1901 and 1902 from <https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all>(directedfrom<http://www.hadoopbook.com/code.html>) and run all the commands for these years only, or (2) You go the website of NCDC and look out for free datasets. If available, you download one and work with it. For help in downloading the data set from NCDC website, you can look into https://serc.carleton.edu/hydromodules/steps/downloading_ncd.html. If none of the options work out, you may try out the commands on some random dataset/files of your choice.

Our purpose in this exercise is to learn the basic working in Pig.

Answer

Firstly download the Apache Pig for installation from the link

<https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz>

above link is for apache pig version 0.17.0, you can download another versions form the official website <https://pig.apache.org/releases.html>

After Downloading move it to a specific folder where you want to install it.

And follow the following process.

- Open the terminal within the same folder where the file is been moved. and run the command

```
$ tar -xvf pig-0.17.0.tar.gz
```

- Set the Environment Variables

- open .bashrc file which is hidden in home directory or can use

- ```
$ nano .bashrc
```

- at the last of the .bashrc attach

```
#PIG VARIABLES
```

```
export PIG_HOME=/home/rohan/pig.0.17.0
```

```
export PATH= $ PATH:$PIG_HOME/bin
```

```
export PIG_CLASSPATH= $ PIG_HOME$/conf:$HADOOP_INSTALL/etx/hadoop/bin
```

```
export PIG_CONF_DIR=$PIG_HOME/conf
```

```
export PIG_CLASSPATH=$PIG_CONF_DIR
```

After that save the .bashrc file

Now run the command in terminal

```
$ source .bashrc
```

- Check the pig version

```
$ pig -version
```

If it runs fine your installation is completed

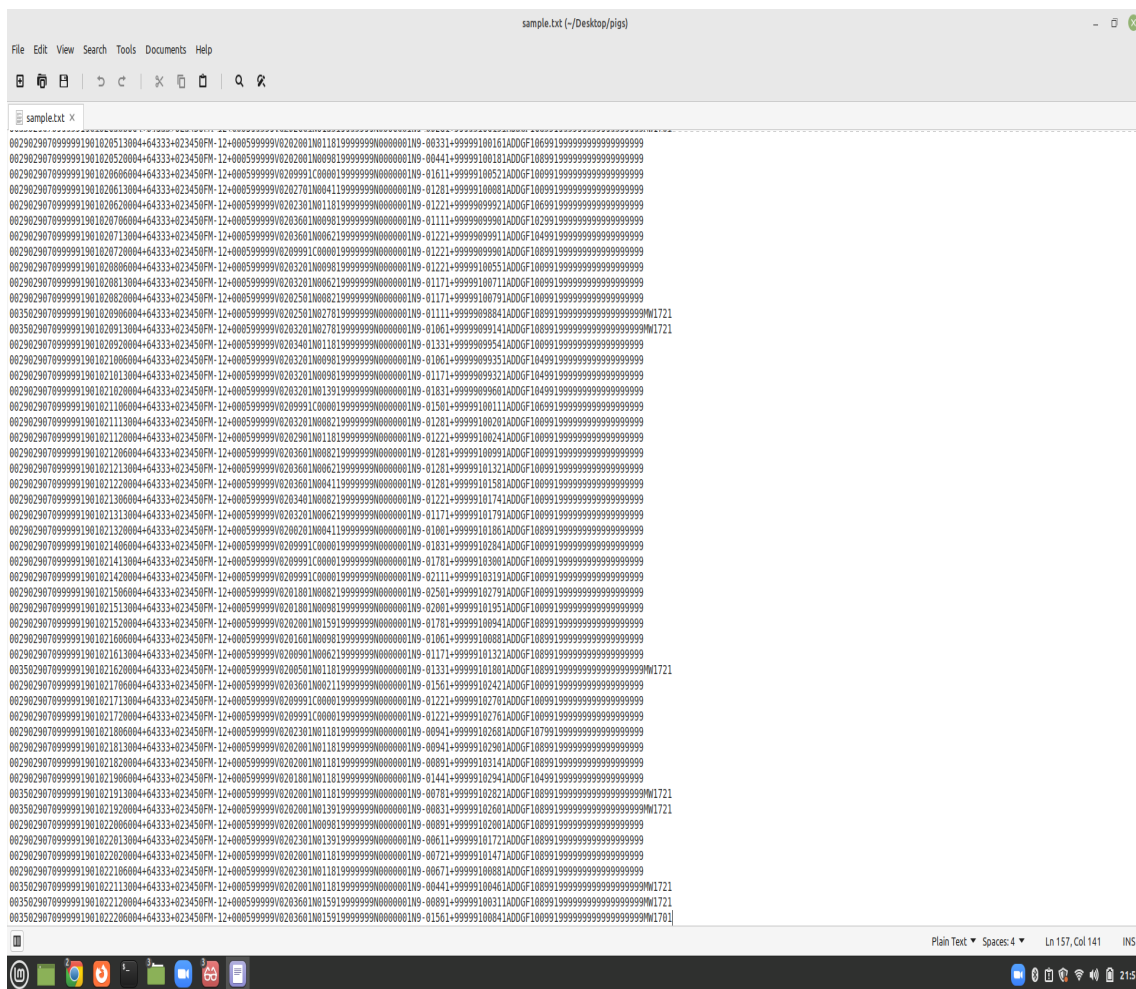
## Execute dataset

Now Download the dataset form the site

<https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all>

Keep the files in a file so that the file can be used for input purpose.

## Input file



```
sample.txt (-/Desktop/pigs)
File Edit View Search Tools Documents Help
sample.txt x
0029029070999991901820513004+64333+023450FM-12+000599999V0202001M01181999999N000000IN0-00331+99999100161A00GF1069919999999999999999
0029029070999991901820520004+64333+023450FM-12+000599999V0202001M00901999999N000000IN0-00441+99999100181A00GF1089919999999999999999
0029029070999991901820606004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01011+99999100521A00GF1009919999999999999999
0029029070999991901820613004+64333+023450FM-12+000599999V0202301M00411999999N000000IN0-01201+99999100081A00GF1009919999999999999999
0029029070999991901820620004+64333+023450FM-12+000599999V0202301M01181999999N000000IN0-01221+99999999921A00GF1069919999999999999999
0029029070999991901820706004+64333+023450FM-12+000599999V0203001M00901999999N000000IN0-01111+99999999901A00GF1029919999999999999999
0029029070999991901820713004+64333+023450FM-12+000599999V0203001M00621999999N000000IN0-01221+99999999911A00GF1049919999999999999999
0029029070999991901820720004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01221+9999999991A00GF1009919999999999999999
0029029070999991901820806004+64333+023450FM-12+000599999V0203201M00901999999N000000IN0-01221+99999100531A00GF1009919999999999999999
0029029070999991901820813004+64333+023450FM-12+000599999V0203201M00621999999N000000IN0-01171+99999100711A00GF1009919999999999999999
0029029070999991901820820004+64333+023450FM-12+000599999V0202501M00821999999N000000IN0-01171+99999100791A00GF1009919999999999999999
0035029070999991901820906004+64333+023450FM-12+000599999V0202501M02781999999N000000IN0-01111+9999999841A00GF1009919999999999999999M1721
0035029070999991901820913004+64333+023450FM-12+000599999V0203201M02781999999N000000IN0-01061+99999999914A00GF1009919999999999999999M1721
0029029070999991901820920004+64333+023450FM-12+000599999V0203401M01181999999N000000IN0-01331+9999999954A00GF1009919999999999999999
0029029070999991901820930004+64333+023450FM-12+000599999V0203201M00901999999N000000IN0-01061+99999999351A00GF1049919999999999999999
0029029070999991901821013004+64333+023450FM-12+000599999V0203201M01391999999N000000IN0-01831+99999999601A00GF1049919999999999999999
0029029070999991901821020004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01501+99999100111A00GF1069919999999999999999
002902907099999190182106004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01201+99999100201A00GF1009919999999999999999
002902907099999190182110004+64333+023450FM-12+000599999V0203201M01181999999N000000IN0-01221+99999100241A00GF1009919999999999999999
0029029070999991901821120004+64333+023450FM-12+000599999V0203601M00821999999N000000IN0-01281+99999100991A00GF1009919999999999999999
002902907099999190182120004+64333+023450FM-12+000599999V0203601M00621999999N000000IN0-01281+99999101321A00GF1009919999999999999999
0029029070999991901821213004+64333+023450FM-12+000599999V0203601M00411999999N000000IN0-01281+99999101581A00GF1009919999999999999999
002902907099999190182126004+64333+023450FM-12+000599999V0203401M00621999999N000000IN0-01221+99999101791A00GF1009919999999999999999
002902907099999190182130004+64333+023450FM-12+000599999V0203201M00411999999N000000IN0-01061+99999101861A00GF1009919999999999999999
0029029070999991901821320004+64333+023450FM-12+000599999V0206201M00411999999N000000IN0-01061+99999101861A00GF1009919999999999999999
0029029070999991901821406004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01831+99999102041A00GF1009919999999999999999
0029029070999991901821413004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01781+99999103001A00GF1009919999999999999999
0029029070999991901821506004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-02111+99999103191A00GF1009919999999999999999
0029029070999991901821513004+64333+023450FM-12+000599999V0201801M00901999999N000000IN0-02001+99999101951A00GF1009919999999999999999
0029029070999991901821520004+64333+023450FM-12+000599999V0201801M00901999999N000000IN0-01781+99999100941A00GF1009919999999999999999
0029029070999991901821606004+64333+023450FM-12+000599999V0201601M00901999999N000000IN0-01061+99999100881A00GF1009919999999999999999
0029029070999991901821620004+64333+023450FM-12+000599999V0200901M00621999999N000000IN0-01171+99999101321A00GF1009919999999999999999
003502907099999190182163004+64333+023450FM-12+000599999V0200501M01181999999N000000IN0-01331+99999101001A00GF1009919999999999999999M1721
0029029070999991901821706004+64333+023450FM-12+000599999V0203601M00211999999N000000IN0-01501+99999102421A00GF1009919999999999999999
0029029070999991901821713004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01221+99999102701A00GF1009919999999999999999
0029029070999991901821720004+64333+023450FM-12+000599999V0209991C00001999999N000000IN0-01221+99999102701A00GF1009919999999999999999
0029029070999991901821806004+64333+023450FM-12+000599999V0202301M01181999999N000000IN0-00941+99999102681A00GF1079919999999999999999
0029029070999991901821813004+64333+023450FM-12+000599999V0202001M01181999999N000000IN0-00941+99999102901A00GF1009919999999999999999
0029029070999991901821820004+64333+023450FM-12+000599999V0202001M01181999999N000000IN0-00891+9999910314A00GF1009919999999999999999
002902907099999190182186004+64333+023450FM-12+000599999V0201801M01181999999N000000IN0-01441+99999102941A00GF1049919999999999999999
0035029070999991901821913004+64333+023450FM-12+000599999V0202001M01181999999N000000IN0-00781+99999102821A00GF1009919999999999999999M1721
0035029070999991901821920004+64333+023450FM-12+000599999V0202001M01391999999N000000IN0-00831+99999102601A00GF1009919999999999999999M1721
0029029070999991901822006004+64333+023450FM-12+000599999V0202801M00901999999N000000IN0-00891+99999102001A00GF1009919999999999999999
0029029070999991901822020004+64333+023450FM-12+000599999V0202301M01391999999N000000IN0-00611+99999101721A00GF1009919999999999999999
0029029070999991901822026004+64333+023450FM-12+000599999V0202001M01181999999N000000IN0-00721+99999101471A00GF1009919999999999999999
002902907099999190182212004+64333+023450FM-12+000599999V0202301M01181999999N000000IN0-00671+99999100881A00GF1009919999999999999999
003502907099999190182213004+64333+023450FM-12+000599999V0202001M01181999999N000000IN0-00441+99999100461A00GF1009919999999999999999M1721
0035029070999991901822126004+64333+023450FM-12+000599999V0203601M01591999999N000000IN0-00891+99999100311A00GF1009919999999999999999M1721
0035029070999991901822206004+64333+023450FM-12+000599999V0203601M01591999999N000000IN0-01561+99999100041A00GF1009919999999999999999M1701
```

Figure 1: Output

- Open terminal and the run the command

```
$ pig -x local
```

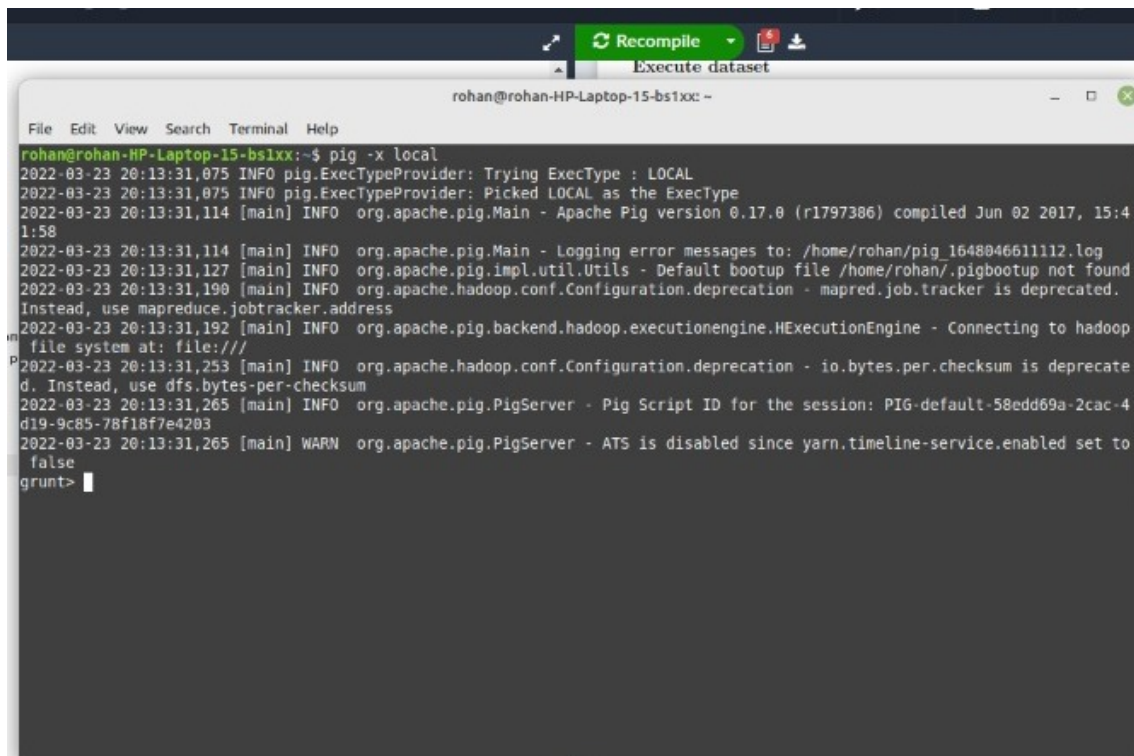


Figure 2: command shell

## Command shell

- now run in grunt

```
grunt> records = LOAD '/home/rohan/Desktop/pig_exc/sample.txt'
 AS (weatherrec:chararray);
```

don't forget to change the location of folder where the file is located Now run

- ```
grunt> parsed_data=FOREACH records GENERATE SUBSTRING(weatherrec,15,19)
      AS year:chararray, SUBSTRING(weatherrec, 87, 92)
      AS temperature:int, SUBSTRING(weatherrec,92,93) AS quality:int;
```

And than DUMP the passed_data as

```
grunt> DUMP parsed_data
```

handy to add commonly used file paths (especially because Pig does not perform file-name completion) or the names of any user-defined functions you have created.

```
rohan@rohan-HP-Laptop-15-bstxx: -
File Edit View Search Terminal Help
2022-03-23 20:51:42,962 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-23 20:51:42,963 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-03-23 20:51:42,963 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-23 20:51:42,965 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-03-23 20:51:42,965 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1901,-78,1)
(1901,-72,1)
(1901,-94,1)
(1901,-61,1)
(1901,-56,1)
(1901,-28,1)
(1901,-67,1)
(1901,-33,1)
(1901,-28,1)
(1901,-33,1)
(1901,-44,1)
(1901,-39,1)
(1901,0,1)
(1901,6,1)
(1901,0,1)
(1901,6,1)
(1901,6,1)
(1901,6,1)
(1901,-11,1)
(1901,-33,1)
(1901,-50,1)
(1901,-44,1)
(1901,-28,1)
(1901,-33,1)
(1901,-33,1)
(1901,-50,1)
(1901,-33,1)
(1901,-28,1)
(1901,-44,1)
(1901,-44,1)
(1901,-44,1)
(1901,-39,1)
(1901,-50,1)
(1901,-44,1)
(1901,-39,1)
```

Figure 3: parsed_data output

- Now Filter the record using

```
grunt> filtered_records=FILTER parsed_data BY temperature!=9999 AND
(quality==0 OR quality==1 OR quality==4 OR quality==5 OR quality==9);
```

you can check the output by DUMP the filtered_records

- Now Group the records by years

```
grunt> grouped_records = GROUP filtered_records BY year;
grout> DUMP grouped_records
```

- to get information about the grouped_records use

```
grunt> DESCRIBE grouped_records;
```


- ```
grunt> max_temp = FOREACH grouped_records GENERATE group,
 MAX(filtered_records.temperature);
```

```
grunt> DUMP max_temp
```

- ```
grunt> ILLUSTRATE max_temp;
```