Exercise 1:

1. Data is structured as data blocks and is spread across Hadoop cluster over different Data Nodes. Each data block is of 128 MB.

Directory tree :-

Using command :-

hadoop dfs -ls -R /data/ghcnd | awk '{print $8}' | \

> sed -e 's/[^-][^\/]\*\//--/g' -e 's/^/ /' -e 's/-/|/'

|---countries

|---daily

|-----1763.csv.gz

|-----1764.csv.gz

|-----1765.csv.gz

…

…

…

|-----2020.csv.gz

|---inventory

|---states

|---stations

1. There are total 258 years in daily.

Command : - hdfs dfs -count -q hdfs:///data/ghcnd/daily

1. The size of all the data is 15.5 GB. This is obtained by following command : -

hdfs dfs -du -s -h /data/ghcnd

The size of daily is also 15.5 GB as rest of the files are relatively very small.

Using the following command we can see size of all sub-directories in ghcnd :-

hdfs dfs -du -h /data/ghcnd

This will generate following output :-

3.6 K 28.6 K /data/ghcnd/countries

15.5 G 124.0 G /data/ghcnd/daily

30.1 M 241.2 M /data/ghcnd/inventory

1.1 K 8.5 K /data/ghcnd/states

9.4 M 75.5 M /data/ghcnd/stations