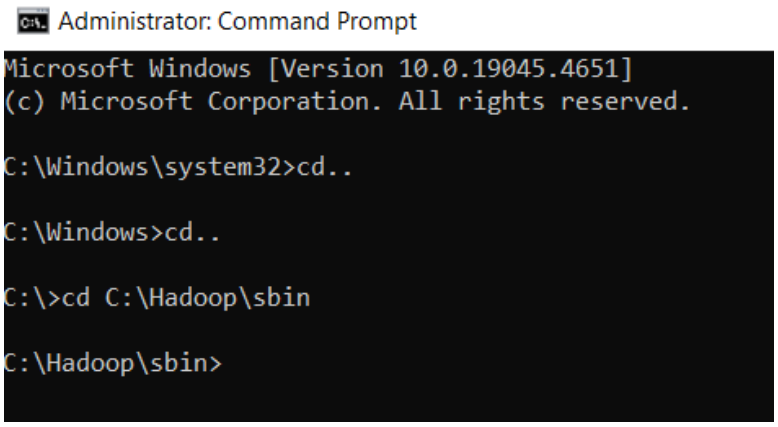
Hadoop Streaming – Wordcount Using Mapreducer in Hadoop

Steps:

1. Open command prompt and run as administrator

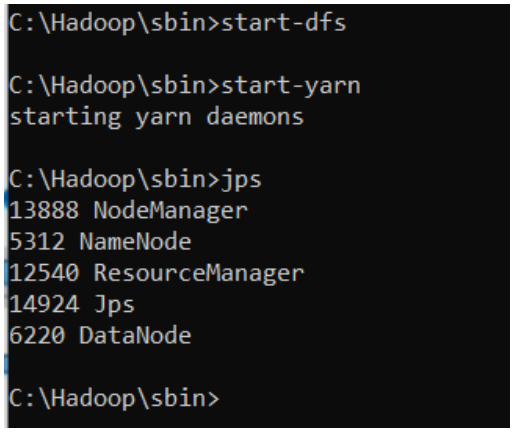
   Go to hadoop sbin directory



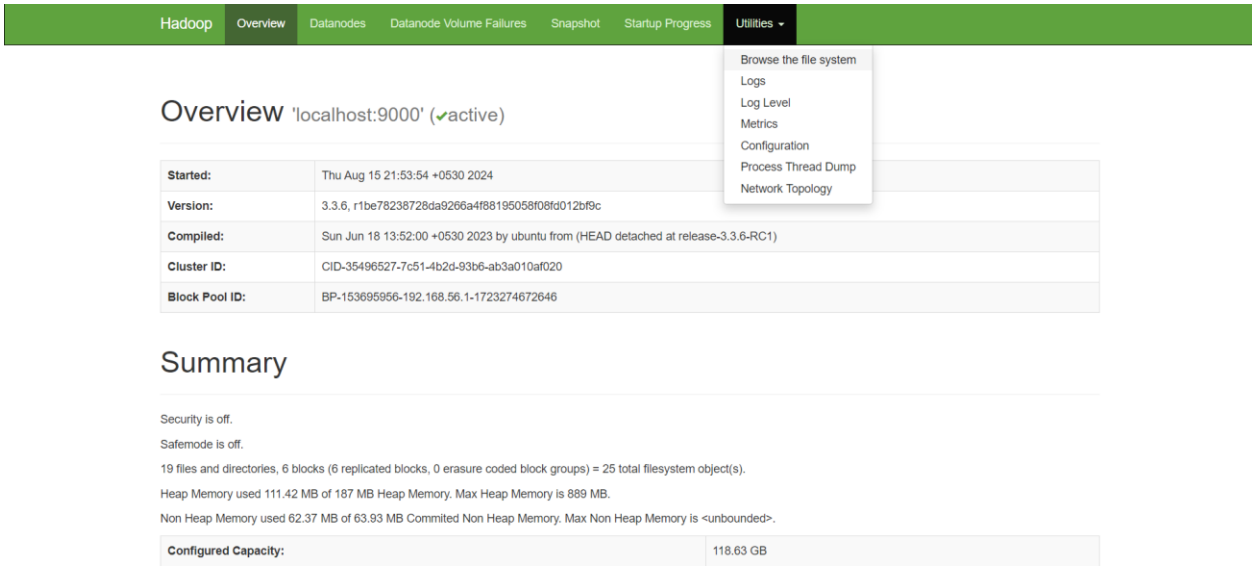
2. Start dfs and yarn services

   start-dfs

   start-yarn



3. Open the browser and go to the URL localhost:9870



4. Create input directory in Hadoop

```
C:\Hadoop\sbin>hadoop fs -mkdir /in
```

5. Upload the input.txt in Hadoop file system

```
C:\Hadoop\sbin>hadoop fs -put C:/input.txt /in

C:\Hadoop\sbin>hadoop fs -ls /in/
Found 1 items
-rw-r--r--   1 Admin supergroup         42 2024-08-15 22:28 /in/input.txt

C:\Hadoop\sbin>hadoop fs -cat /in/input.txt
hello world
hi all
hello all
all the best

C:\Hadoop\sbin>
```

6. Then perform the map-reduce operation using the command

```
C:\Hadoop\sbin>hadoop jar C:/Hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.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\sbin>
```

```
C:\Hadoop\sbin>hadoop jar C:/Hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar wordcount /in /out
2024-08-15 23:04:15,703 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-15 23:04:17,070 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
Admin/.staging/job_1723740547060_0005
2024-08-15 23:04:17,659 INFO input.FileInputFormat: Total input files to process : 2
2024-08-15 23:04:18,181 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-15 23:04:18,462 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1723740547060_0005
```

7. To view the output, type the following command

hadoop fs -cat /user/output/part-00000

```
C:\Hadoop\sbin>hadoop fs -cat /out/part-r-00000
all     3
best    1
hello   2
hi      1
the     1
world   1
```

8. Once the map reduce operations are performed successfully, the output will be present in the specified directory.

“/out/part-r-00000”