Hadoop Map Reduce Commands

Command 1: Create a Directory in HDFS

```bash
hdfs dfs -mkdir /user
```

Description: This command creates a directory named `user` in the Hadoop Distributed File System (HDFS). This is typically the top-level directory under which user-specific directories are organized.

Command 2: Create a Subdirectory in HDFS

```bash
hdfs dfs -mkdir /user/rawat
```

Description: This command creates a subdirectory named `rawat` (your username) inside the `/user` directory in HDFS. This is where you will store your input files and where the output of your MapReduce job will be written.

Command 3: Run a MapReduce Job

```bash
hadoop jar "C:\Hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar" -input /user/rawat/test_input/test.txt -output /user/rawat/test_output -mapper "python3 mapper.py" -reducer "python3 reducer.py" -file C:\Users\rawat\Desktop\mapper.py -file C:\Users\rawat\Desktop\reducer.py
```

Description: This command executes a MapReduce job using Hadoop Streaming. It specifies the input and output paths in HDFS, the mapper and reducer scripts to use, and the locations of these scripts on the local filesystem. The job processes the file `test.txt` located in the `/user/rawat/test_input` directory and writes the output to the `/user/rawat/test_output` directory.

Command 4: List the Output Directory Contents

```bash
hadoop fs -ls /user/rawat/test_output
```

Description: This command lists the files in the `/user/rawat/test_output` directory. The output shows two files:

- `_SUCCESS`: Indicates that the MapReduce job completed successfully.
- `part-00000`: Contains the results of the MapReduce job.

Output:
```
Found 2 items

-rw-r--r--    3 rawat supergroup          0 2024-09-03 22:20
/user/rawat/test_output/_SUCCESS

-rw-r--r--    3 rawat supergroup         54 2024-09-03 22:20
/user/rawat/test_output/part-00000
```

Explanation:

- `_SUCCESS` file is an indicator that the job completed successfully without errors.
- `part-00000` is the output file containing the result of the MapReduce job.

Command 5: View the Output File Content

```bash
hadoop fs -cat /user/rawat/test_output/part-00000
```

Description: This command displays the contents of the `part-00000` file, which contains the results of your MapReduce job.

Output:
```
Hello   1
World   1
```

file   1

hadoop 1

is   1

testing 1

this   1

```
```

Explanation:

Each line shows a word from the input file and its count (number of occurrences). This is the result of the MapReduce job where `mapper.py` emitted words and counts, and `reducer.py` aggregated these counts.

Summary

- **Command 1 & 2:** Create necessary directories in HDFS.
- **Command 3:** Run a MapReduce job with specified mapper and reducer scripts.
- **Command 4:** Verify the output directory contents.
- **Command 5:** View the final results of the MapReduce job.