Problem Statement 01:

Working with HDFS Commands

1. Create a folder in HDFS by name “dir01” and move input1.txt , input2.txt and input3.txt into /dir01.

```
127 login: mavricbdhnov017
mavricbdhnov017@127.0.0.1's password:
Last login: Mon Nov 27 15:35:33 2023 from localhost
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -ls /user/$USER/
Found 5 items
drwx------   - mavricbdhnov017 mavricbdhnov017          0 2023-11-25 07:00 /user/mavricbdhnov017/.Trash
drwxr-xr-x   - mavricbdhnov017 mavricbdhnov017          0 2023-11-27 08:11 /user/mavricbdhnov017/.sparkStaging
drwx------   - mavricbdhnov017 mavricbdhnov017          0 2023-11-27 04:01 /user/mavricbdhnov017/.staging
drwxr-xr-x   - mavricbdhnov017 mavricbdhnov017          0 2023-11-22 07:56 /user/mavricbdhnov017/output_22
drwxr-xr-x   - mavricbdhnov017 mavricbdhnov017          0 2023-11-22 07:12 /user/mavricbdhnov017/workspace
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/$USER/dir01
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -ls /user/$USER/dir01
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -copyFromLocal input1.txt input2.txt input3.txt /user/$USER/dir01
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -ls /user/$USER/dir01
Found 3 items
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017       9510 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input1.txt
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017      19718 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input2.txt
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017      50537 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input3.txt
[mavricbdhnov017@ip-10-1-1-204 ~]$
```

2. List only the file names present in “/dir01”

```
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -ls /user/$USER/dir01
Found 3 items
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017       9510 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input1.txt
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017      19718 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input2.txt
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017      50537 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input3.txt
[mavricbdhnov017@ip-10-1-1-204 ~]$
```

3. Change the replication factor for the content present in directory “/dir01” to 5 and display the replication factor for the files present in “/dir01”.

```
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -setrep 5 /user/$USER/dir01
Replication 5 set: /user/mavricbdhnov017/dir01/input1.txt
Replication 5 set: /user/mavricbdhnov017/dir01/input2.txt
Replication 5 set: /user/mavricbdhnov017/dir01/input3.txt
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -ls /user/$USER/dir01
Found 3 items
-rw-r--r--   5 mavricbdhnov017 mavricbdhnov017       9510 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input1.txt
-rw-r--r--   5 mavricbdhnov017 mavricbdhnov017      19718 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input2.txt
-rw-r--r--   5 mavricbdhnov017 mavricbdhnov017      50537 2023-11-27 15:43 /user/mavricbdhnov017/dir01/input3.txt
[mavricbdhnov017@ip-10-1-1-204 ~]$
```

4. Create a folder in HDFS by name “scenario01” and create directory “level01” inside “scenario01” directory and create another directory “level02” inside directory “level01”. Once the required directories are created copy input1.txt to scenario01, input2.txt to level01 and input3.txt to level02 and finally recursively print only the file names present in directory scenario01

```
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/$USER/scenario01
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/$USER/scenario01/level01
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/$USER/scenario01/level01/level02
```

```
[mavricbdhnov017@ip-10-1-1-204 ~]$ hdfs dfs -ls -R /user/mavricbdhnov017/scenario01 | grep -e '/user/mavricbdhnov017/scenario01/*$*'
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017       9510 2023-11-27 16:27 /user/mavricbdhnov017/scenario01/input1.txt
drwxr-xr-x   - mavricbdhnov017 mavricbdhnov017          0 2023-11-27 16:28 /user/mavricbdhnov017/scenario01/level01
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017      19718 2023-11-27 16:28 /user/mavricbdhnov017/scenario01/level01/input2.txt
drwxr-xr-x   - mavricbdhnov017 mavricbdhnov017          0 2023-11-27 16:28 /user/mavricbdhnov017/scenario01/level01/level02
-rw-r--r--   3 mavricbdhnov017 mavricbdhnov017      50537 2023-11-27 16:28 /user/mavricbdhnov017/scenario01/level01/level02/input3.txt
[mavricbdhnov017@ip-10-1-1-204 ~]$
```

Problem Statement 02 :

Working with YARN Commands Commands

1. Run MapReduce Program and capture the application Id of the job.

```
[mavricbdhnov017@ip-10-1-1-204 jars]$ yarn jar hadoop-mapreduce-examples-3.0.0-cdh6.2.1.jar wordcount  /user/$USER/dir01/ /user/$USER/output_02/
WARNING: YARN_OPTS has been replaced by HADOOP_OPTS. Using value of YARN_OPTS.
23/11/27 17:08:37 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/11/27 17:08:38 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/mavricbdhnov017/.staging/job_1700812329856_1575
23/11/27 17:08:38 INFO input.FileInputFormat: Total input files to process : 3
23/11/27 17:08:38 INFO mapreduce.JobSubmitter: number of splits:3
23/11/27 17:08:38 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics
-publisher.enabled
23/11/27 17:08:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1700812329856_1575
23/11/27 17:08:39 INFO mapreduce.JobSubmitter: Executing with tokens: []
23/11/27 17:08:39 INFO conf.Configuration: resource-types.xml not found
23/11/27 17:08:39 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/11/27 17:08:39 INFO impl.YarnClientImpl: Submitted application application_1700812329856_1575
23/11/27 17:08:39 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1700812329856_1
575/
23/11/27 17:08:39 INFO mapreduce.Job: Running job: job_1700812329856_1575
23/11/27 17:08:49 INFO mapreduce.Job: Job job_1700812329856_1575 running in uber mode : false
23/11/27 17:08:49 INFO mapreduce.Job:  map 0% reduce 0%
23/11/27 17:08:54 INFO mapreduce.Job:  map 100% reduce 0%
```

2. Re-run the MapReduce program and kill the application using the yarn command.

```
-publisher.enabled
23/11/27 17:15:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1700812329856_1577
23/11/27 17:15:15 INFO mapreduce.JobSubmitter: Executing with tokens: []
23/11/27 17:15:16 INFO conf.Configuration: resource-types.xml not found
23/11/27 17:15:16 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/11/27 17:15:16 INFO impl.YarnClientImpl: Submitted application application_1700812329856_1577
23/11/27 17:15:16 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1700812329856_1
577/
23/11/27 17:15:16 INFO mapreduce.Job: Running job: job_1700812329856_1577

^C[mavricbdhnov017@ip-10-1-1-204 jars]$ yarn application -kill 1700812329856
WARNING: YARN_OPTS has been replaced by HADOOP_OPTS. Using value of YARN_OPTS.
23/11/27 17:15:37 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Exception in thread "main" java.lang.IllegalArgumentException: Invalid ApplicationId prefix: 1700812329856. The valid ApplicationId should start with p
refix application
        at org.apache.hadoop.yarn.api.records.ApplicationId.fromString(ApplicationId.java:112)
        at org.apache.hadoop.yarn.client.cli.ApplicationCLI.killApplication(ApplicationCLI.java:605)
        at org.apache.hadoop.yarn.client.cli.ApplicationCLI.killApplication(ApplicationCLI.java:585)
        at org.apache.hadoop.yarn.client.cli.ApplicationCLI.run(ApplicationCLI.java:293)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:76)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:90)
        at org.apache.hadoop.yarn.client.cli.ApplicationCLI.main(ApplicationCLI.java:102)
```

3. List all the applications which are RUNNING state

```
[mavricbdhnov017@ip-10-1-1-204 jars]$ yarn application -list -appStates RUNNING
WARNING: YARN_OPTS has been replaced by HADOOP_OPTS. Using value of YARN_OPTS.
23/11/27 17:16:17 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Total number of applications (application-types: [], states: [RUNNING] and tags: []):2
                Application-Id      Application-Name        Application-Type          User           Queue                   State             Final-St
ate            Progress                        Tracking-URL
application_1700812329856_0630        kafkaStreaming                   SPARK    nuvelabs321010  root.default                RUNNING               UNDEFI
NED                10% http://ip-10-1-1-204.ap-south-1.compute.internal:18088/history/application_1700812329856_0630/1
application_1700812329856_0515          Spark shell                    SPARK    nuvelabs321010  root.default                RUNNING               UNDEFI
NED                10% http://ip-10-1-1-204.ap-south-1.compute.internal:18088/history/application_1700812329856_0515/1
[mavricbdhnov017@ip-10-1-1-204 jars]$
```

4. View the logs of any of the jobs which are already completed.

```
2023-11-27 17:15:37,918 INFO [fetcher#4] org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput: Read 1264 bytes from map-output for attempt_1700812
329856_1577_m_000000_0
2023-11-27 17:15:37,918 INFO [fetcher#4] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1264, inMem
oryMapOutputs.size() -> 3, commitMemory -> 1987, usedMemory ->3251
2023-11-27 17:15:37,919 INFO [fetcher#4] org.apache.hadoop.mapreduce.task.reduce.ShuffleSchedulerImpl: ip-10-1-2-103.ap-south-1.compute.internal:13562
freed by fetcher#4 in 547ms
2023-11-27 17:15:37,919 INFO [EventFetcher for fetching Map Completion Events] org.apache.hadoop.mapreduce.task.reduce.EventFetcher: EventFetcher is in
terrupted.. Returning
2023-11-27 17:15:37,935 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: finalMerge called with 3 in-memory map-outputs and 0 on-d
isk map-outputs
2023-11-27 17:15:37,952 INFO [main] org.apache.hadoop.mapred.Merger: Merging 3 sorted segments
2023-11-27 17:15:37,952 INFO [main] org.apache.hadoop.mapred.Merger: Down to the last merge-pass, with 3 segments left of total size: 3232 bytes
2023-11-27 17:15:37,997 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merged 3 segments, 3251 bytes to disk to satisfy reduce m
emory limit
2023-11-27 17:15:37,998 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merging 1 files, 1657 bytes from disk
2023-11-27 17:15:37,999 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2023-11-27 17:15:37,999 INFO [main] org.apache.hadoop.mapred.Merger: Merging 1 sorted segments
2023-11-27 17:15:38,029 INFO [main] org.apache.hadoop.mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3240 bytes

End of LogType:syslog.shuffle
**********************************************************************

[mavricbdhnov017@ip-10-1-1-204 jars]$ ^C
[mavricbdhnov017@ip-10-1-1-204 jars]$ `
Display all 1760 possibilities? (y or n)
[mavricbdhnov017@ip-10-1-1-204 jars]$ yarn logs -applicationId application_1700812329856_1577
```