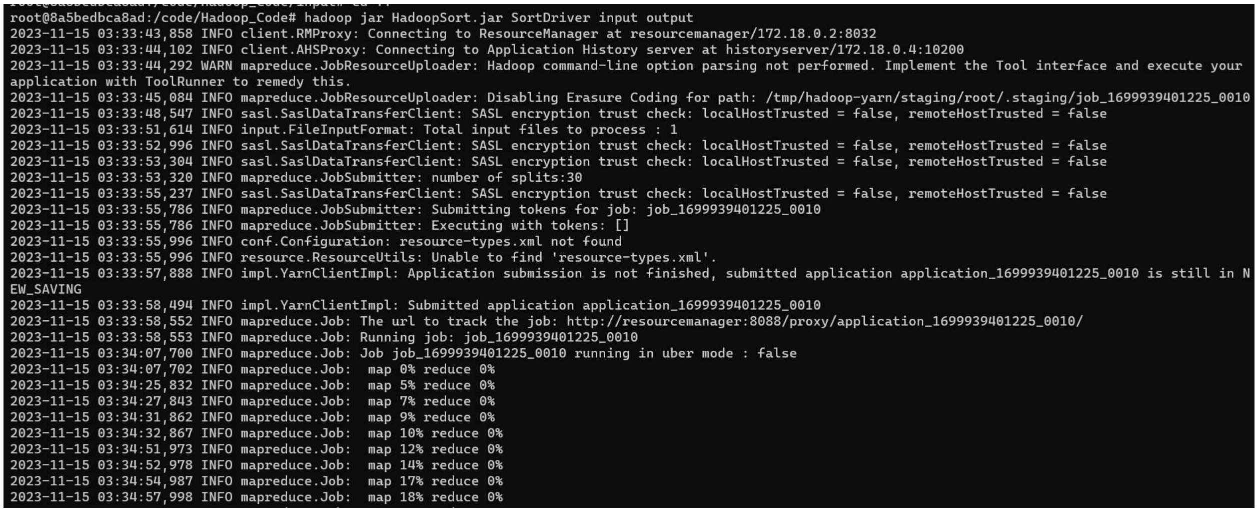
ReadME file

Problems faced: I completed the setup and all Hadoop and spark configurations but after 3 days i was mot able to access my node for 2 days i tried all at last I hard rebooted the instance and started my work again

Setup

Hadoop Environment settings:

- Ensure you r root
- Used Java 8 for compilation of libraries and hadoop3.2.1 for Hadoop Sort
- I created Jar file and SortDriver and just needed to run by using this cmd in the hdfs environment
- **hadoop jar HadoopSort.jar SortDriver ***/input/path*** ****/output/path******
- Please specify your gen sort input and output path for the output
- It will map and reduce jobs and will give output like this.
- Please delete the output path after every time using the cmd because in essence there will be a directory or else use a different directory for different output
- Please find the screenshots of my work for any doubt



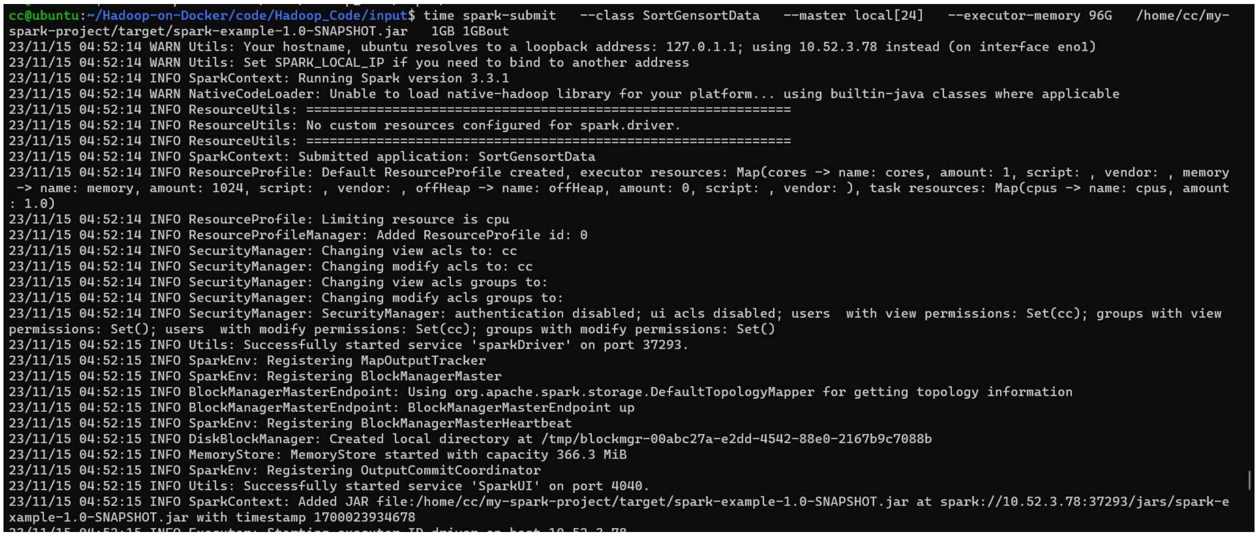
```
root@8a5bedbca8ad:/code/Hadoop_Code# hadoop jar HadoopSort.jar SortDriver input output
2023-11-15 03:33:43,858 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.2:8032
2023-11-15 03:33:44,102 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.4:10200
2023-11-15 03:33:44,292 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
2023-11-15 03:33:45,084 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1699939401225_0010
2023-11-15 03:33:48,547 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-11-15 03:33:51,614 INFO input.FileInputFormat: Total input files to process : 1
2023-11-15 03:33:52,996 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-11-15 03:33:53,304 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-11-15 03:33:53,320 INFO mapreduce.JobSubmitter: number of splits:30
2023-11-15 03:33:55,237 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-11-15 03:33:55,786 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1699939401225_0010
2023-11-15 03:33:55,786 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-11-15 03:33:55,996 INFO conf.Configuration: resource-types.xml not found
2023-11-15 03:33:55,996 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-11-15 03:33:57,888 INFO impl.YarnClientImpl: Application submission is not finished, submitted application application_1699939401225_0010 is still in N
EW_SAVING
2023-11-15 03:33:58,494 INFO impl.YarnClientImpl: Submitted application application_1699939401225_0010
2023-11-15 03:33:58,552 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1699939401225_0010/
2023-11-15 03:33:58,553 INFO mapreduce.Job: Running job: job_1699939401225_0010
2023-11-15 03:34:07,700 INFO mapreduce.Job: Job job_1699939401225_0010 running in uber mode : false
2023-11-15 03:34:07,702 INFO mapreduce.Job:  map 0% reduce 0%
2023-11-15 03:34:25,832 INFO mapreduce.Job:  map 5% reduce 0%
2023-11-15 03:34:27,843 INFO mapreduce.Job:  map 7% reduce 0%
2023-11-15 03:34:31,862 INFO mapreduce.Job:  map 9% reduce 0%
2023-11-15 03:34:32,867 INFO mapreduce.Job:  map 10% reduce 0%
2023-11-15 03:34:51,973 INFO mapreduce.Job:  map 12% reduce 0%
2023-11-15 03:34:52,978 INFO mapreduce.Job:  map 14% reduce 0%
2023-11-15 03:34:54,987 INFO mapreduce.Job:  map 17% reduce 0%
2023-11-15 03:34:57,998 INFO mapreduce.Job:  map 18% reduce 0%
```

Proof of execution the environment

Spark Environment setup

- Ensure you r root
- sudo apt install maven
- Spark-3.3.1-bin-hadoop3 this is my spark version ensure to use it

- In the project do mvn package
- This program sorts by first 10bytes taken as key and used 24 cores and 96GB memory
- Use the below command to generate sorted data
- spark-submit   --class SortGensortData   --master local[24]   --executor-memory 96G my-spark-project/target/spark-example-1.0-SNAPSHOT.jar   input output
- Ensure to specify to the jar generated by the mvn package
- Where in the input field pls mention the file to be sorted and leave output as it is
- Please find screenshots of my work in the report for any doubt
- After running the program you find parts of spark use this command for final output
- cat part-* > outputfile.txt



```
cc@ubuntu:~/Hadoop-on-Docker/code/Hadoop_Code/input$ time spark-submit   --class SortGensortData   --master local[24]   --executor-memory 96G   /home/cc/my-
spark-project/target/spark-example-1.0-SNAPSHOT.jar   1GB 1GBout
23/11/15 04:52:14 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 10.52.3.78 instead (on interface eno1)
23/11/15 04:52:14 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
23/11/15 04:52:14 INFO SparkContext: Running Spark version 3.3.1
23/11/15 04:52:14 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/11/15 04:52:14 INFO ResourceUtils: ==============================================================
23/11/15 04:52:14 INFO ResourceUtils: No custom resources configured for spark.driver.
23/11/15 04:52:14 INFO ResourceUtils: ==============================================================
23/11/15 04:52:14 INFO SparkContext: Submitted application: SortGensortData
23/11/15 04:52:14 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory
 -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount
: 1.0)
23/11/15 04:52:14 INFO ResourceProfile: Limiting resource is cpu
23/11/15 04:52:14 INFO ResourceProfileManager: Added ResourceProfile id: 0
23/11/15 04:52:14 INFO SecurityManager: Changing view acls to: cc
23/11/15 04:52:14 INFO SecurityManager: Changing modify acls to: cc
23/11/15 04:52:14 INFO SecurityManager: Changing view acls groups to:
23/11/15 04:52:14 INFO SecurityManager: Changing modify acls groups to:
23/11/15 04:52:14 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(cc); groups with view
permissions: Set(); users  with modify permissions: Set(cc); groups with modify permissions: Set()
23/11/15 04:52:15 INFO Utils: Successfully started service 'sparkDriver' on port 37293.
23/11/15 04:52:15 INFO SparkEnv: Registering MapOutputTracker
23/11/15 04:52:15 INFO SparkEnv: Registering BlockManagerMaster
23/11/15 04:52:15 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
23/11/15 04:52:15 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
23/11/15 04:52:15 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/11/15 04:52:15 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-00abc27a-e2dd-4542-88e0-2167b9c7088b
23/11/15 04:52:15 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
23/11/15 04:52:15 INFO SparkEnv: Registering OutputCommitCoordinator
23/11/15 04:52:15 INFO Utils: Successfully started service 'SparkUI' on port 4040.
23/11/15 04:52:15 INFO SparkContext: Added JAR file:/home/cc/my-spark-project/target/spark-example-1.0-SNAPSHOT.jar at spark://10.52.3.78:37293/jars/spark-e
xample-1.0-SNAPSHOT.jar with timestamp 1700023934678
23/11/15 04:52:15 INFO Executor: Starting executor ID driver on host 10.52.3.78
```

Proof of execution

Note :: valsort output cant able to verify inside hdfs environment so just pasted screenshot in the report head and tail as successful sort executions and logs generate after program execution