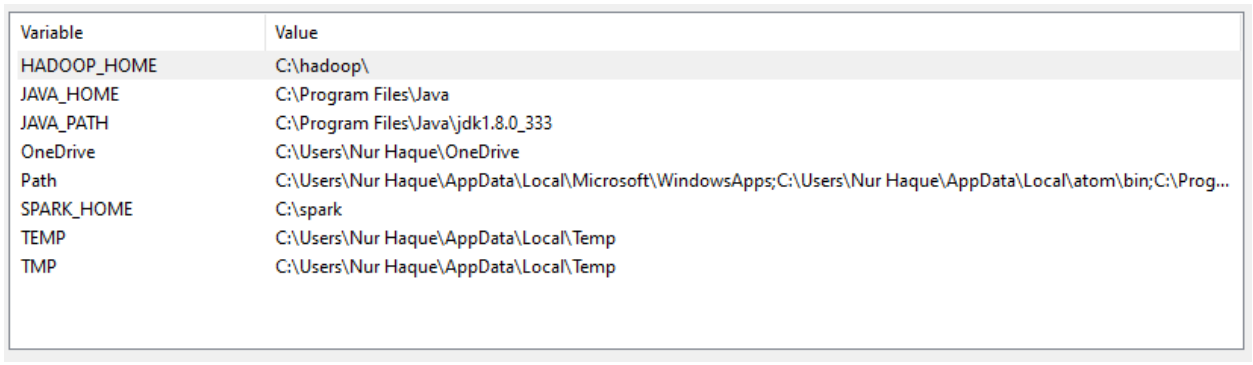
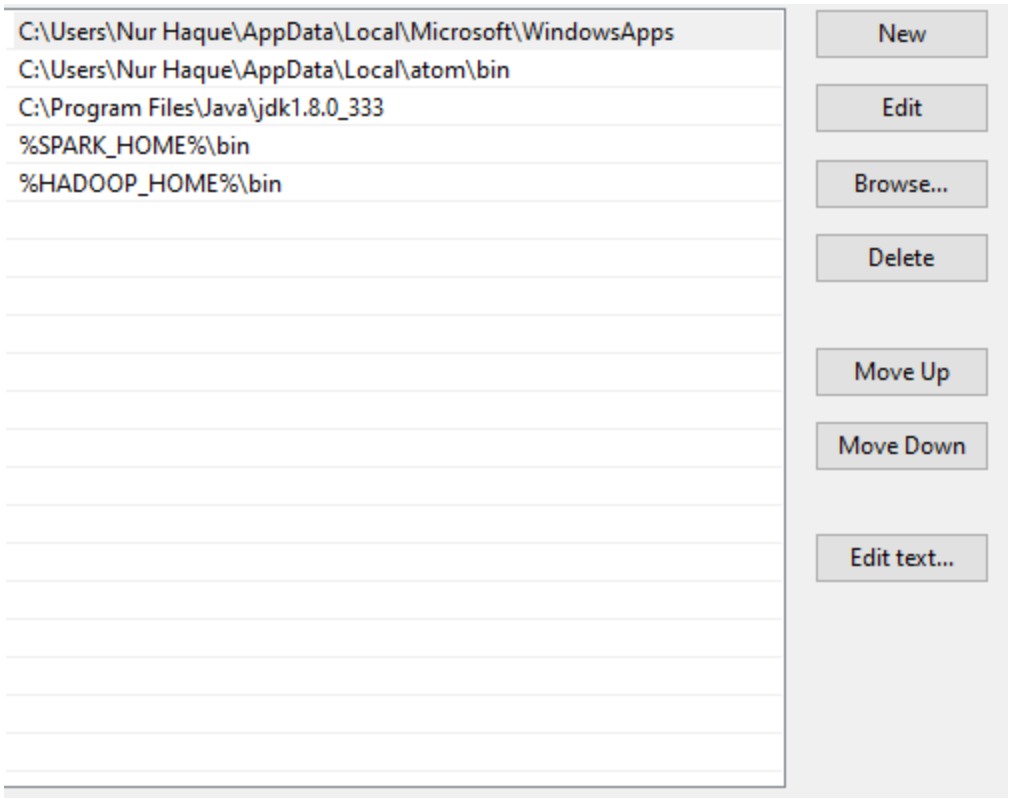
Assignment3_Haque

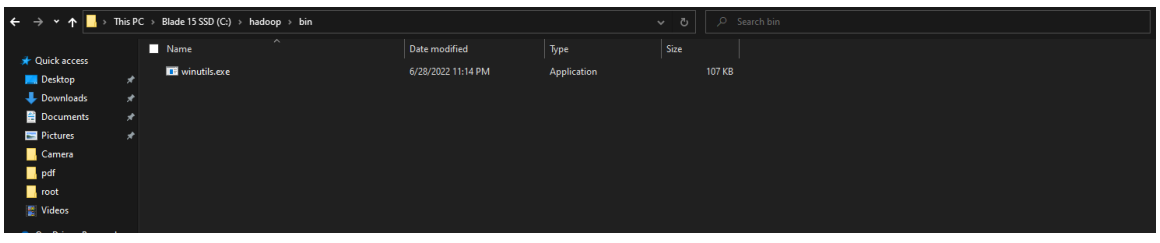
Spark Download in windows

1. Set the environment variables in windows

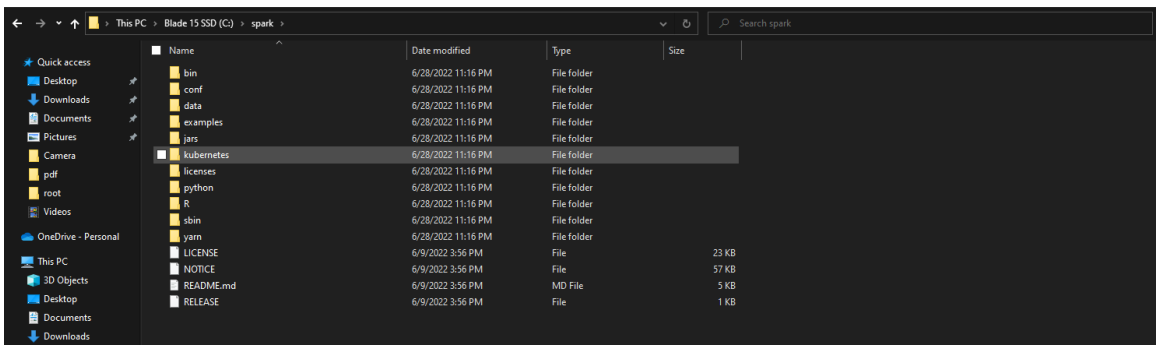




Download Hadoop executable



Download Spark



Run Spark using the following commands

```
spark-shell
```

```
spark-shell                                                                    ( 11:21:58 PM
22/06/28 23:44:34 WARN Utils: Your hostname, LAPTOP-IUI6PNV0 resolves to a loopback address: 127.0.1.1; using 172.30.230
.189 instead (on interface eth0)
22/06/28 23:44:34 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/06/28 23:44:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
22/06/28 23:44:49 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/06/28 23:44:49 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
Spark context Web UI available at http://172.30.230.189:4042
Spark context available as 'sc' (master = local[*], app id = local-1656474290294).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.0
      /_/

Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 11.0.15)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

```
pyspark
```

```
pyspark                                                                        ( 11:46:52 PM
Python 3.8.10 (default, Mar 15 2022, 12:22:08)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
22/06/28 23:47:03 WARN Utils: Your hostname, LAPTOP-IUI6PNV0 resolves to a loopback address: 127.0.1.1; using 172.30.230
.189 instead (on interface eth0)
22/06/28 23:47:03 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/06/28 23:47:06 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
22/06/28 23:47:09 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/06/28 23:47:09 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.3.0
      /_/

Using Python version 3.8.10 (default, Mar 15 2022 12:22:08)
Spark context Web UI available at http://172.30.230.189:4042
Spark context available as 'sc' (master = local[*], app id = local-1656474429522).
SparkSession available as 'spark'.
>>>
```

Execute the code using the following command

```
spark-submit wordCount.py /home/nur_haque/spark/Assignment1/test.txt 5
```

```
22/06/28 23:38:33 WARN Utils: Your hostname, LAPTOP-IUI6PNV0 resolves to a loopback address: 127.0.1.1; using 172.30.230.189 instead (on interface eth0)
22/06/28 23:38:33 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
/home/nur_haque/spark/Assignment1/test.txt
5
22/06/28 23:38:36 INFO SparkContext: Running Spark version 3.3.0
22/06/28 23:38:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/06/28 23:38:36 INFO ResourceUtils: ==============================================================
22/06/28 23:38:36 INFO ResourceUtils: No custom resources configured for spark.driver.
22/06/28 23:38:36 INFO ResourceUtils: ==============================================================
22/06/28 23:38:36 INFO SparkContext: Submitted application: Word Count Spark
22/06/28 23:38:36 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memo
ry, amount: 4096, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
22/06/28 23:38:36 INFO ResourceProfile: Limiting resource is cpu
22/06/28 23:38:36 INFO ResourceProfileManager: Added ResourceProfile id: 0
22/06/28 23:38:36 INFO SecurityManager: Changing view acls to: nur_haque
22/06/28 23:38:36 INFO SecurityManager: Changing modify acls to: nur_haque
22/06/28 23:38:36 INFO SecurityManager: Changing view acls groups to:
22/06/28 23:38:36 INFO SecurityManager: Changing modify acls groups to:
22/06/28 23:38:36 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(nur_haque); groups with view permiss
ions: Set(); users  with modify permissions: Set(nur_haque); groups with modify permissions: Set()
22/06/28 23:38:37 INFO Utils: Successfully started service 'sparkDriver' on port 36473.
22/06/28 23:38:37 INFO SparkEnv: Registering MapOutputTracker
22/06/28 23:38:37 INFO SparkEnv: Registering BlockManagerMaster
22/06/28 23:38:37 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/06/28 23:38:37 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/06/28 23:38:37 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
22/06/28 23:38:37 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-9052e59e-60fc-42c1-a3d3-cd70070c6331
22/06/28 23:38:37 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
22/06/28 23:38:38 INFO SparkEnv: Registering OutputCommitCoordinator
22/06/28 23:38:38 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/06/28 23:38:38 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
22/06/28 23:38:38 INFO Utils: Successfully started service 'SparkUI' on port 4042.
22/06/28 23:38:38 INFO Executor: Starting executor ID driver on host 172.30.230.189
22/06/28 23:38:38 INFO Executor: Starting executor with user classpath (userClassPathFirst = false): ''
22/06/28 23:38:39 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 44675.
22/06/28 23:38:39 INFO NettyBlockTransferService: Server created on 172.30.230.189:44675
22/06/28 23:38:39 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/06/28 23:38:39 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:39 INFO BlockManagerMasterEndpoint: Registering block manager 172.30.230.189:44675 with 434.4 MiB RAM, BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:39 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:39 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:40 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 127.3 KiB, free 434.3 MiB)
22/06/28 23:38:40 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 23.6 KiB, free 434.3 MiB)
22/06/28 23:38:40 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 172.30.230.189:44675 (size: 23.6 KiB, free: 434.4 MiB)
22/06/28 23:38:40 INFO SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:0
22/06/28 23:38:41 INFO FileInputFormat: Total input paths to process : 1
22/06/28 23:38:41 INFO SparkContext: Starting job: collect at /home/nur_haque/wordCount.py:27
22/06/28 23:38:41 INFO DAGScheduler: Registering RDD 3 (reduceByKey at /home/nur_haque/wordCount.py:24) as input to shuffle 0
22/06/28 23:38:41 INFO DAGScheduler: Got job 0 (collect at /home/nur_haque/wordCount.py:27) with 1 output partitions
22/06/28 23:38:41 INFO DAGScheduler: Final stage: ResultStage 1 (collect at /home/nur_haque/wordCount.py:27)
```

```
22/06/28 23:38:38 INFO Executor: Starting executor ID driver on host 172.30.230.189
22/06/28 23:38:38 INFO Executor: Starting executor with user classpath (userClassPathFirst = false): ''
22/06/28 23:38:39 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 44675.
22/06/28 23:38:39 INFO NettyBlockTransferService: Server created on 172.30.230.189:44675
22/06/28 23:38:39 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/06/28 23:38:39 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:39 INFO BlockManagerMasterEndpoint: Registering block manager 172.30.230.189:44675 with 434.4 MiB RAM, BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:39 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:39 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 172.30.230.189, 44675, None)
22/06/28 23:38:40 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 127.3 KiB, free 434.3 MiB)
22/06/28 23:38:40 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 23.6 KiB, free 434.3 MiB)
22/06/28 23:38:40 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 172.30.230.189:44675 (size: 23.6 KiB, free: 434.4 MiB)
22/06/28 23:38:40 INFO SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:0
22/06/28 23:38:41 INFO FileInputFormat: Total input paths to process : 1
22/06/28 23:38:41 INFO SparkContext: Starting job: collect at /home/nur_haque/wordCount.py:27
22/06/28 23:38:41 INFO DAGScheduler: Registering RDD 3 (reduceByKey at /home/nur_haque/wordCount.py:24) as input to shuffle 0
22/06/28 23:38:41 INFO DAGScheduler: Got job 0 (collect at /home/nur_haque/wordCount.py:27) with 1 output partitions
22/06/28 23:38:41 INFO DAGScheduler: Final stage: ResultStage 1 (collect at /home/nur_haque/wordCount.py:27)
22/06/28 23:38:41 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
22/06/28 23:38:41 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 0)
22/06/28 23:38:41 INFO DAGScheduler: Submitting ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/nur_haque/wordCount.py:24), which has no missing parents
22/06/28 23:38:41 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 12.5 KiB, free 434.2 MiB)
22/06/28 23:38:41 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 7.5 KiB, free 434.2 MiB)
22/06/28 23:38:41 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 172.30.230.189:44675 (size: 7.5 KiB, free: 434.4 MiB)
22/06/28 23:38:41 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1513
22/06/28 23:38:41 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/nur_haque/wordCount.py:24) (first 15 tasks are for partitions Vector(0))
22/06/28 23:38:41 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks resource profile 0
22/06/28 23:38:41 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0) (172.30.230.189, executor driver, partition 0, PROCESS_LOCAL, 4505 bytes) taskResourceAssignments Map()
22/06/28 23:38:41 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
22/06/28 23:38:42 INFO HadoopRDD: Input split: file:/home/nur_haque/spark/Assignment1/test.txt:0+64
22/06/28 23:38:43 INFO PythonRunner: Times: total = 801, boot = 772, init = 28, finish = 1
22/06/28 23:38:43 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1624 bytes result sent to driver
22/06/28 23:38:43 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1926 ms on 172.30.230.189 (executor driver) (1/1)
22/06/28 23:38:43 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
22/06/28 23:38:43 INFO PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 60909
22/06/28 23:38:43 INFO DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/nur_haque/wordCount.py:24) finished in 2.163 s
22/06/28 23:38:43 INFO DAGScheduler: looking for newly runnable stages
22/06/28 23:38:43 INFO DAGScheduler: running: Set()
22/06/28 23:38:43 INFO DAGScheduler: waiting: Set(ResultStage 1)
22/06/28 23:38:43 INFO DAGScheduler: failed: Set()
22/06/28 23:38:43 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at collect at /home/nur_haque/wordCount.py:27), which has no missing parents
22/06/28 23:38:43 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 9.5 KiB, free 434.2 MiB)
22/06/28 23:38:43 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 5.7 KiB, free 434.2 MiB)
22/06/28 23:38:43 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 172.30.230.189:44675 (size: 5.7 KiB, free: 434.4 MiB)
22/06/28 23:38:43 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1513
22/06/28 23:38:43 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[6] at collect at /home/nur_haque/wordCount.py:27) (first 15 tasks are for partitions Vector(0))
22/06/28 23:38:43 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks resource profile 0
22/06/28 23:38:43 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1) (172.30.230.189, executor driver, partition 0, NODE_LOCAL, 4271 bytes) taskResourceAssignments Map()
22/06/28 23:38:43 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
22/06/28 23:38:43 INFO ShuffleBlockFetcherIterator: Getting 1 (129.0 B) non-empty blocks including 1 (129.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/06/28 23:38:43 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 40 ms
22/06/28 23:38:44 INFO PythonRunner: Times: total = 33, boot = -1027, init = 1060, finish = 0
22/06/28 23:38:44 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1771 bytes result sent to driver
22/06/28 23:38:44 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 172 ms on 172.30.230.189 (executor driver) (1/1)
22/06/28 23:38:44 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
22/06/28 23:38:44 INFO DAGScheduler: ResultStage 1 (collect at /home/nur_haque/wordCount.py:27) finished in 0.196 s
22/06/28 23:38:44 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
22/06/28 23:38:44 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
22/06/28 23:38:44 INFO DAGScheduler: Job 0 finished: collect at /home/nur_haque/wordCount.py:27, took 2.559365 s
[('hadoop', 4), ('spark', 3), ('pig', 2), ('hive', 1), ('hbase', 1)]
22/06/28 23:38:44 INFO SparkContext: Invoking stop() from shutdown hook
22/06/28 23:38:44 INFO SparkUI: Stopped Spark web UI at http://172.30.230.189:4042
22/06/28 23:38:44 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/06/28 23:38:44 INFO MemoryStore: MemoryStore cleared
22/06/28 23:38:44 INFO BlockManager: BlockManager stopped
22/06/28 23:38:44 INFO BlockManagerMaster: BlockManagerMaster stopped
22/06/28 23:38:44 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/06/28 23:38:44 INFO SparkContext: Successfully stopped SparkContext
22/06/28 23:38:44 INFO ShutdownHookManager: Shutdown hook called
22/06/28 23:38:44 INFO ShutdownHookManager: Deleting directory /tmp/spark-4a8422d6-0634-47fc-aaf6-3c881376bf78
```